University of ULSTER

INTELLIGENT SYSTEMS RESEARCH CENTRE

PROCEEDINGS

IMVIP 2014

2014 IRISH MACHINE VISION AND IMAGE PROCESSING

27-29 AUGUST 2014
UNIVERSITY OF ULSTER, MAGEE,
DERRY-LONDONDERRY,
NORTHERN IRELAND

Edited By

Sonya Coleman, Bryan Gardiner and Dermot Kerr

CSRI
>>COMPUTER SCIENCE RESEARCH INSTITUTE

syntouch

IPRCS
Irish Pattern Recognition and Classification Society

IMVIP 2014

TABLE OF CONTENTS

KEYNOTE SPEAKERS

3 DIMENSIONAL DATA PROCESSING

APPLICATIONS

MEDICAL IMAGING

BIOINSPIRED IMAGING

VIDEO PROCESSING

IMAGE PROCESSING

POSTERS

IMVIP 2014

PREFACE

The Organising Committee extends a warm welcome to all speakers and delegates of the 2014 Irish Machine Vision and Image Processing Conference (IMVIP 2014). This year it is hosted at the University of Ulster under the organisation of the School of Computing and Intelligent Systems and the Intelligent Systems Research Centre.

The IMVIP Conference is Ireland's primary meeting for those researching in the fields of machine vision and image processing. The conference has been running since 1997 and provides a forum for the exchange of ideas and the presentation of research conducted both in Ireland and worldwide.

IMVIP is a single track conference consisting of high quality previously unpublished contributed papers focussing on both theoretical research and practical experiences in all areas. After a rigorous review process, 22 papers were selected for oral presentation and a further 12 for poster presentation; we wish to sincerely thank the members of the Programme Committee for generously giving their time, effort and expertise in reviewing the submissions.

Continuing the tradition of inviting high-profile speakers to IMVIP, we are delighted to have three high-profile speakers give keynote talks: Professor Hideo Sato from Keio University, Tokyo with a talk entitled "Vision-based 3D sensing and visualization for real world applications", Professor Stephen Marshall from University of Strathclyde with a talk entitled "Hyperspectral Image Processing and its applications" and Professor Ingmar Posner from University of Oxford with a talk entitled "Driven Learning for Driving: Why Autonomous Cars Need Introspection".

IMVIP 2014 is run in association with the Irish Pattern Recognition and Classification Society (IPRCS), a member organisation of the International Association for Pattern Recognition (IAPR)

Sonya Coleman
School of Computing and Intelligent Systems
University of Ulster
August 2014

CHAIRS AND COMMITTEES

CONFERENCE CHAIRS:

- General Chair: Sonya Coleman, University of Ulster, Northern Ireland
- Co Chair: Dermot Kerr, University of Ulster, Northern Ireland
- Co Chair: Bryan Gardiner, University of Ulster, Northern Ireland

ORGANISING COMMITTEE:

- Sandra Moffett, University of Ulster, Northern Ireland
- Hubert Cecotti, University of Ulster, Northern Ireland
- John Wade, University of Ulster, Northern Ireland
- Phillip Vance, University of Ulster, Northern Ireland
- Emmett Kerr, University of Ulster, Northern Ireland
- Paula Sheerin, University of Ulster, Northern Ireland

PROGRAMME COMMITTEE:

- Andrew Donnellan, Tallaght Institute of Technology
- Andy Shearer, National University of Ireland Galway
- Antonio Fernández, University of Vigo, Spain
- Artzai Picón, TECNALIA-Infotech, Spain
- Bryan W. Scotney, University of Ulster
- Cem Direkoglu, Dublin City University
- Danny Crookes, The Queen’s University of Belfast
- David Monaghan, Dublin City University
- David Vernon, University of Skövde, Sweden
- Donald Bailey, Massey University, New Zealand
- Fionn Murtagh, Science Foundation Ireland/University of London, UK
- Francesco Bianconi, University of Perugia, Italy
- George Moore, University of Ulster
- Hiroshi Sako, Hosei University, Japan
- Jane Courtney, Dublin Institute of Technology
- John Barron, The University of Western Ontario, Canada
- John McDonald, National University of Ireland Maynooth
- John Winder, University of Ulster
- Kathleen Curran, University College Dublin
- Kenneth Dawson-Howe, Trinity College Dublin
- Kevin McGuinness, Dublin City University
- Madonna Herron, University of Ulster
- Martin McGinnity, University of Ulster
- Noel O’Connor, Dublin City University
- Paul McKevitt, University of Ulster
- Paul Miller, Queen’s University of Ulster
- Philip Morrow, University of Ulster

- Rami Albatal, Dublin City University
- Reyer Zwiggelaar, Aberystwyth University, UK
- Robert Sadleir, Dublin City University, Ireland
- Rozenn Dahyot, Trinity College Dublin
- Sally McClean, University of Ulster
- Shan Suganthan, Smart Sensors Ltd, Bath, UK
- Sudeep Sarkar, University of South Florida, USA
- Tom Naughton, National University of Ireland Maynooth

SPONSORS:

- Intelligent Systems Research Centre
- Computer Science Research Institute
- Syntouch
- Irish Pattern Recognition and Classification Society

IMVIP 2014

KEYNOTE SPEAKERS

KEYNOTE SPEAKERS

Hideo Saito - Keio University

Title: Vision-based 3D sensing and visualization for real world applications

In computer vision area, 3D sensing technologies have extensively been studied. For making such technologies be used in practical applications, there are still a lot of difficulties to adapt real world problems. In this talk, I introduce my recent challenges on real world applications that can be solved by vision-based 3D sensing. One application is on-site information visualization using mixed and augmented reality, in which vision-based object/camera pose estimation plays a significant role. Other application is 3D-video/Free viewpoint video using multi-view sensing/capturing based on 3D modeling of target scenes.

Hideo Saito received his Ph.D. degree in Electrical Engineering from Keio University, Japan, in 1992. Since then, he has been on the Faculty of Science and Technology, Keio University. In 1997 to 1999, he had joined into Virtualized Reality Project in the Robotics Institute, Carnegie Mellon University as a visiting researcher. Since 2006, he has been a full Professor of Department of Information and Computer Science, Keio University. He served as program co-chair of ISMAR (International Symposium on Mixed and Augmented Reality) 2008 and 2009. He is now serving as an Program Co-Chair of ACCV (Asian Conference on Computer Vision) 2014. He is a president of MVA Organization, which currently organizes the 14th IAPR International Conference on Machine Vision Applications (MVA2015). His research interests include computer vision, mixed reality, virtual reality, and 3D video analysis and synthesis.

Stephen Marshall - University of Strathclyde

Title: Hyperspectral Image Processing and its applications

Prof Stephen Marshall received a first class honours degree in Electrical and Electronic Engineering from the University of Nottingham in 1979 and a PhD in Image Processing from University of Strathclyde in 1989. His research activities have been focussed in the area of Non Linear Image Processing. In this time, he has pioneered new design techniques for morphological filters based on a class of iterative search techniques known as genetic algorithms. The resulting filters have been applied as four-dimensional operators to successfully restore old film archive material.

In recent years he has established the Hyperspectral Imaging Centre at the University of Strathclyde. The aims to provide solutions to industrial problems through applied research and Knowledge Exchange. He has published over 200 conference and journal papers on these topics including IET, IEEE, SPIE, SIAM, ICASSP, VIE and EUSIPCO. He has also been a reviewer for these and other journals and conferences. He is a Fellow of the Institution of Engineering and Technology (IET). He has also been successful in obtaining research funding from National, International and Industrial sources. These sources include EPSRC, EU, Rolls Royce, BT, DERA, the BBC and Scottish Enterprise, TSB, NERC and EDF Energy.

Ingmar Posner - University of Oxford

Title: Driven Learning for Driving: Why Autonomous Cars Need Introspection

Classification precision and recall have been widely adopted by roboticists as canonical metrics to quantify the performance of learning algorithms. This talk advocates that for robotics applications, which often involve mission critical decision making, good performance according to these standard metrics is desirable but insufficient to appropriately characterise system performance. Against the backdrop of an autonomous driving application - the Oxford RobotCar Project (http://mrg.robots.ox.ac.uk/robotcar/) - we will introduce and motivate the importance of a classifier's introspective capacity: the ability to mitigate potentially overconfident classifications by an appropriate assessment of how qualified the system is to make a judgement on the current test datum. The talk will provide an intuition as to how this introspective capacity can be achieved and systematically investigates it in a selection of classification frameworks commonly used in robotics.

Ingmar Posner is an Associate Professor in Engineering Science at the University of Oxford and one of the two PIs leading the Mobile Robotics Group (MRG). His expertise lies in the design and implementation of information engineering techniques that enable an autonomous agent to interpret complex, dynamic environments in a way which permits robust decision-making, planning and exploration online and in real-time. His research tackles questions such as what semantic information can be inferred about the environment the robot has traversed (e.g. what type of structures, what objects can be found? What type of terrain is it travelling on?) and how this knowledge can feed into the decision-making process of an autonomous agent such as a self-driving car? His research track record includes award winning work on semantic mapping, active perception and 3D reconstruction. Building on his successes to date, Ingmar's current research focus lies on closing the action-perception loop in semantic mapping to enable robust robot decision-making and online planning and exploration in the context of, amongst others, autonomous transport.

IMVIP 2014

3 DIMENSIONAL DATA PROCESSING

Specular 3D Object Tracking by View Generative Learning

Yukiko Shinozuka, Francois de Sorbier and Hideo Saito
Keio University
3-14-1 Hiyoshi, Kohoku-ku
223-8522 Yokohama, Japan
shinozuka@hvrl.ics.keio.ac.jp

Abstract

This paper proposes a novel specular 3D object tracking method. Our method works with texture-less specular objects and objects with background reflections on the surface. It is a keypoint-based tracking using a view generative learning. Conventional local features are robust to scale and rotation, but keypoint matching fails when the viewpoint significantly changes. We apply a view generative learning to improve the robustness to viewpoint changes. To be robust to large appearance changes, our method does view-dependent rendering for generating views and stores all the descriptors of the keypoints on the generated images and its 3D-position in the reference database called ”feature table”. We conducted quantitative evaluation on the object pose and showed our method outperforms compared with the other view generative learning methods in terms of tracking accuracy and learning process.

Keywords: View Generative Learning, 3D Object Tracking, Local Feature, Feature Table, Specular Object

1 Introduction

Object pose estimation is necessary to augment a virtual object on a real environment for augmented reality. To estimate the object pose, it is required to find correspondences between a reference dataset and an input image for vision-based methods. There are two types of methods based on the target objects; planar model-based methods [Lepetit and Fua, 2006] and 3D model-based methods [Drummond et al., 2002]. Our proposed algorithm takes the latter solution to track a 3D object.

Keypoint matching is one of the solutions to find correspondences. Local features such as scale-invariant feature transform (SIFT) [Lowe, 2004] is well-known as a keypoint extractor and descriptor. It is robust to rotation, translation and illumination changes, but there is a limit for affine transformation. There are plenty of studies of local features to improve this limitation. Harris-affine [Baumberg, 2000] and Maximally-stable extremal region detector (MSER) [Matas et al., 2002] are known for the invariance to affine transformation. However there is no descriptor for each of them. That means even if the keypoints are extracted, the feature description will be different from the ones extracted on the image before transformation. Affine-SIFT (ASIFT) [Morel and Yu, 2009] is also known as affine invariant. This method applies several possible transformation before matching.

A generative learning method is proposed as another solution for keypoint matching. It is a learning method which uses local features which is not invariant to affine transformation. It virtually generates the possible views and extracts keypoints from them. If the same keypoint is extracted from different views, this point is considered as ”stable keypoint”. Stable keypoints are robust under strong perspective view changes. For creating a reference database, machine learning process is often conducted. Lepetit *et al.*’s method [Lepetit and Fua, 2006] uses randomized trees with huge amount of training dataset, whereas Thachasongtham *et al.*’s method [Thachasongtham et al., 2013] uses k-means clustering with much less dataset. However, in

both methods, they consider a target object is covered with Lambartian surface, so they do not take highlight and specular areas into consideration.
Our motivation is to estimate 3D object pose by vision-based 3D object tracking using view generative learning. Our contributions are to propose a novel view generative learning method ”feature table” to track a specular object. In the experiments, we compare with other generative learning methods (randomized trees [Lepetit and Fua, 2006] and k-means [Thachasongtham et al., 2013]). We evaluate the rotation matrix and translation vector of the target object and computational time. Our experimental results show that our proposed method outperforms in tracking specular 3D objects with less training datasets.

2 Related Works

This section refers to other generative learning tracking methods and the recent trials on specular object tracking.

As already mentioned in section 1, local features such as SIFT [Lowe, 2004] is not invariant to perspective transformation. A generative learning is proposed to improve this weakness for keypoint matching. It is the learning method which generates the possible views by affine or perspective transformation and selects robust stable keypoint.

Randomized trees method [Lepetit and Fua, 2006] is a generative learning method which considers a keypoint matching problem as a patch classification problem. It applies affine transformations to the image patches around the extracted keypoints and trains with them by randomized trees. It requires large amount of training dataset for learning. Learning process computational costs time due to the size of the dataset and the recursive algorithm of randomized trees, but tracking runs quite fast.

Thachasongtham *et al.* propose a generative learning method with k-means clustering for 3D object tracking [Thachasongtham et al., 2013]. His method is similar to randomized trees, but there are three main differences. In randomized trees method, keypoints are extracted once before affine transformation whereas Thachasongtham *et al.*’s method extracta the keypoints from every generated patterns. For learning, they apply k-means to determine the centroid of the stable keypoint. For the datasize of the learning data, k-means method requires much less than randomized trees method.

However, both methods assume that a surface of a target object is covered with Lambartian surface. Therefore when the specular reflection occurs, it is hard to extract the keypoints from the same area from different views.

Torki *et al.* propose that a regression was a key to estimate a 3D object pose with specular highlight [Torki and Elgammal, 2011]. The regression is calculated from the 2D-position of each keypoints and its descriptors from the video sequences. They succeed in estimating rotation of cars. Netz *et al.* consider high light is one of the features of the image, then use the specular as features [Netz and Osadchy, 2011]. Our method uses the same concept that the highlight area can be characteristics in the images.

3 View Generative Learning – Feature Table

This section refers to the algorithm of a generative learning method. A generative learning is proposed to be robust to viewpoint changes. It is a keypoint-based method which virtually generates the possible images of different viewpoints and extracts keypoints from them for the creation of the reference database. If the keypoint is extracted at the same position in 3D world coordinate from different views, this point is considered as a ”stable keypoint”. Dataset consists of the descriptors of these stable keypoints and their position in 3D world coordinate.

The algorithm is shown in Figure 1. The method can be divided into learning and tracking phase. The learning phase has to be done off-line phase before tracking.

There are main two differences between our method and other view generative learning methods [Lepetit and Fua, 2006] [Thachasongtham et al., 2013]. Both points contribute to im-

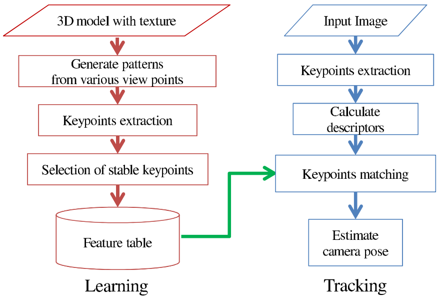
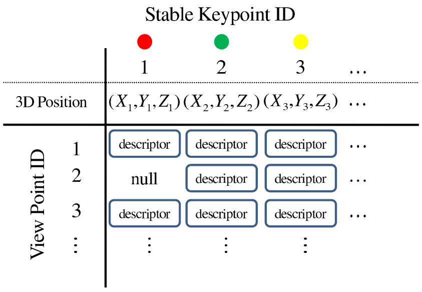


Figure 1: Overview



Stable Keypoint ID

		● 1	● 2	● 3	…
3D Position		(X_1,Y_1,Z_1)	(X_2,Y_2,Z_2)	(X_3,Y_3,Z_3)	…
View Point ID	1	descriptor	descriptor	descriptor	…
	2	null	descriptor	descriptor	…
	3	descriptor	descriptor	descriptor	…
	⋮	⋮	⋮	⋮	

Figure 2: Feature Table

prove the robustness to appearance changes such as highlight. First of all, we conduct view-dependent rendering to create the possible views whereas the conventional methods only do affine-transformation. This process enables to include the highlight area in the database. Second point is to create ”feature table”. Feature table is a table with descriptors of the stable keypoints. Its vertical axis is for viewpoint ID, and the horizontal for stable keypoint ID as shown in Figure 2. We store all the descriptors extracted from the possible views in the table. This process enables to absorb the difference of the descriptors on the same stable keypoint.

4 Learning

4.1 Generate Views

We require a 3D model as an input data for learning. The images from various viewpoints are generated virtually from it. There are two important concerns in this phase. First point is the background of the learning phase. If the object is static in a scene and the background texture is available, the 3D model of the background is reconstructed and we also learn the background. If the texture of the background is not available, we generate the virtual views with a random colored background to be robust to the noisy background. Second point is lighting condition. We generate the views by view-dependent rendering to extract the keypoints around the highlight areas.

The camera pose setting is important to generate the virtual views. Since local feature such as SIFT is scale invariant, there is no need to take the distance from the object into consideration. It means the distance between the virtual viewpoint and the object scene does not have to be changed for learning different views. The different rotation angles of the camera also do not have to be learned, because SIFT is rotation invariant. Thus, we change only two angles, the longitude ϕ and the latitude θ for generating different viewpoint images for generative viewpoint learning.

4.2 Keypoints Extraction and Stable Keypoints

We extract the keypoints from the generated patterns by local feature. Each keypoint p on the image is reprojected to p' in the 3D world coordinate by perspective matrix P. The perspective matrix is already given in section 4.1. The equation of the reprojection is shown in equation (1).

$$p'_i \sim Pp_i \tag{1}$$

We compare the Euclidean distance of the reprojected points from different views. If their Euclidean distance is under the threshold, these points are considered as the same point in 3D world coordinate. The keypoints with high repeatability are called ”stable keypoints” because they can be extracted from other viewpoint images. We store the stable keypoints with high repeatability. We sort the stable keypoints in order of repeatability and store the top N stable keypoints. If the target object has less-texture, the number of the stable kepoints can be lower than threshold N. If it happens, we store all the stable keypoints in the database.

4.3 Creating Feature Table

Feature table is a table with descriptors of the stable keypoints. Its vertical axis is for viewpoint ID, and the horizontal for stable keypoint ID as Figure 2 shows. The descriptor at the same keypoint can be described differently depending on highlights and viewpoint changes. Therefore, all the descriptors from the generated images are stored in our method, whereas the conventional methods did not consider the differences.

Each stable keypoint has multiple descriptors and one position in the 3D world coordinate. The position is calculated by getting the centorid of the keypoints in each stable keypoint group.

5 Tracking

To estimate the object pose for tracking, the projection matrix is calculated by referring to the feature table. After extracting local feature on an input image, we find the nearest descriptor by fast approximate nearest neighbor matching in the feature table. To decrease false matching, we apply nearest-neighbor distance ratio between the first (D_A) and second closest (D_B) as Mikolajczyk *et al.* mentioned in [Mikolajczyk et al., 2005]. If the keypoint fulfills equation (2), it is considered as a correct correspondence. We set $\tau = 0.6$ in our experiments. After finding the correspondences, we use a robust estimator RANSAC and calculate projection matrix with the 2D and 3D positions of the keypoints.

$$\frac{|D_A|}{|D_B|} < \tau \tag{2}$$

6 Experimental Results

6.1 Parameters and System Configuration

We conducted two experiments and compared our method with randomized trees [Lepetit and Fua, 2006] and k-means method [Thachasongtham et al., 2013].

In the first experiments, we set a texture-less specular object **Box** as a target object. We conducted the learning in a random color background. We rotated the object from -15 degrees to 85 degrees in longitude ϕ and 0 to 360 degrees in latitude θ for every 10 degrees.The distance between the camera and the object was set as 40 cm.In the second experiment, we tracked a object with background reflection **Teapot**. We used the texture and 3D structure of the background in the learning. We rotated the object every one degree in latitude where its longitude equals to zero degree. The distance between the camera and the object was set as 30 cm.

In both experiments, the video sequences are created by computer graphic to get the ground truth. We used the same video for learning and testing. We set the number of the stable keypoint N equals to 2000. All the experiments were implemented on Windows 7, 64 bits, Intel Core i7-3930K 3.20GHz CPU, 16.00GB RAM and GeForce 310 589MHz GPU. We chose SIFT-GPU [of North Carolina,] for local feature.

6.2 Texture-less Specular Object

This section shows the tracking result of object **Box**. Figure 3 shows our proposed ”feature table” worked the best of all. Randomized trees method did not track the object in any frame because randomized trees is designed for large amount of database, but the number of the stable keypoints was only a few due to less-texture for object **Box**. (The number of the stable keypoint was 807.) This result shows our method works with much less training data.

Figure 4 shows the L-2 norm error of rotation matrix and the error of translation vector in each axis. They show the huge translation error occurred often in k-means method compared with feature table.

Table 1: Computational Time

Learning method	**Box**	**Teapot**	Tracking method	**Box**	**Teapot**
Randomized Trees [sec]	11304	3115	Randomized Trees [msec]	541	108
K-means [sec]	340	106	K-means [msec]	415	762
Feature Table [sec]	379	114	Feature Table [msec]	26612	22050

6.3 Object with Background Reflection

This section shows the result of object **Teapot**. Figure 3 shows k-means and feature table method tracked 3D object whereas randomized trees did not did not track the object in any frame. It shows our method outperformed in tracking. It is because the descriptors on the same stable keypoints are too different from each other so that the other methods did not absorb these differences.

We evaluated the results on 3D object pose (rotation matrix and translation vector) in Figure 5. The translation errors are better in our proposed method and there is not much difference in rotation.

6.4 Computational Time

Table 1 shows the computational time for learning and tracking. The computational time of randomized trees method for learning cost more than that of the others. It is because the algorithm of randomized tees is recursive. The tracking time of feature table was the slowest of all because the database is not compressed and the size is the largest.

7 Conclusion

This paper proposed a novel specular 3D object tracking method ”feature table”. It worked with the texture-less specular objects and objects with background reflections. Our contributions are following two points. We used the idea that highlight or non-feature area should be included in database to be robust to large appearance changes. Second point is that our method applied a generative learning with less training dataset to improve the robustness to viewpoint changes.

Our experimental results showed our method outperformed in terms of tracking accuracy compared with other methods [Lepetit and Fua, 2006] [Thachasongtham et al., 2013]. Speed in tracking should be improved, but our method required less training computational time with less training data.

References

[Baumberg, 2000] Baumberg, A. (2000). Reliable feature matching across widely separated views. In *CVPR*, pages 1774–1781.

[Drummond et al., 2002] Drummond, T., Society, I. C., and Cipolla, R. (2002). Real-time visual tracking of complex structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:932–946.

[Lepetit and Fua, 2006] Lepetit, V. and Fua, P. (2006). Keypoint recognition using randomized trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1465–1479.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

[Matas et al., 2002] Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, pages 36.1–36.10. BMVA Press.

[Mikolajczyk et al., 2005] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72.

[Morel and Yu, 2009] Morel, J.-M. and Yu, G. (2009). Asift: A new framework for fully affine invariant image comparison. *SIAM J. Img. Sci.*, 2(2):438–469.

[Netz and Osadchy, 2011] Netz, A. and Osadchy, M. (2011). Using specular highlights as pose invariant features for 2d-3d pose estimation. In *CVPR*, pages 721–728. IEEE.

[of North Carolina,] of North Carolina, U. Siftgpu.

[Thachasongtham et al., 2013] Thachasongtham, D., Yoshida, T., Sorbier, F., and Saito, H. (2013). 3d object pose estimation using viewpoint generative learning. volume 7944, pages 512–521. Springer Berlin Heidelberg.

[Torki and Elgammal, 2011] Torki, M. and Elgammal, A. (2011). Regression from local features for viewpoint and pose estimation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2603–2610.

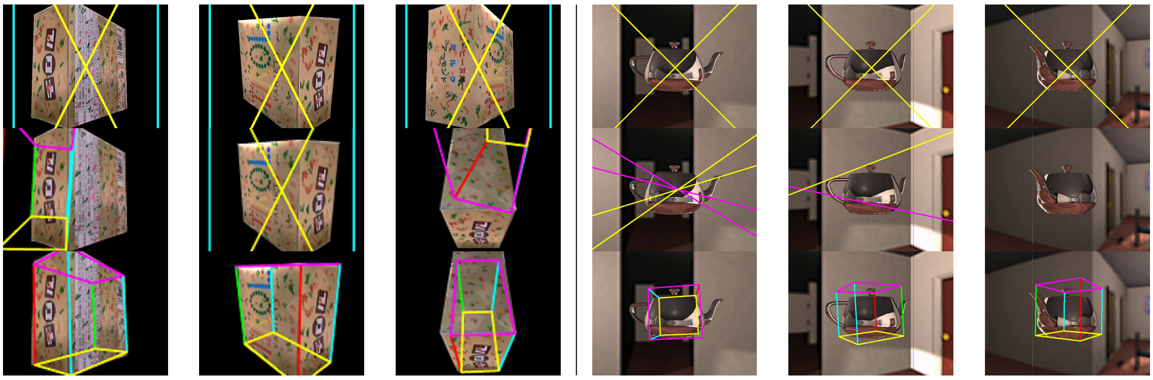
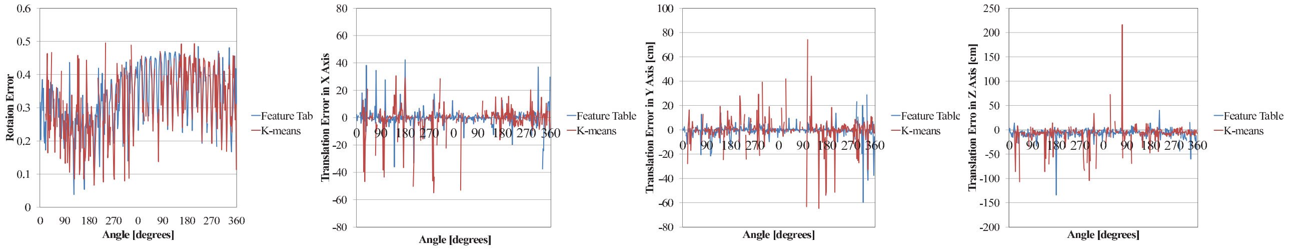
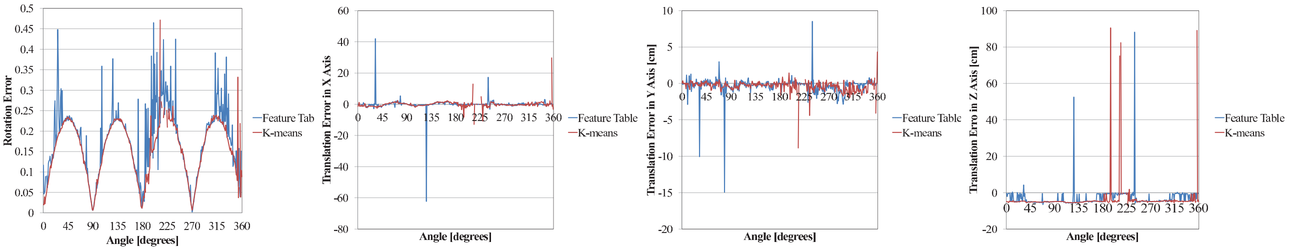


Figure 3: Tracking Results (Top : Randomized Trees, Middle : K-Means, Bottom : Feature Table) (Left : **Box** where $(\theta, \phi) = (-5, 140), (15, 220), (55, 180)$ in degrees, Right : **Teapot** 0,28,239 degrees



(a) Rotation (b) Translation in X Axis (c) Translation in Y Axis (d) Translation in Z Axis

Figure 4: Average Error of Object Pose (Texture-less Specular Object)



(a) Rotation (b) Translation in X Axis (c) Translation in Y Axis (d) Translation in Z Axis

Figure 5: Average Error of Object Pose (Background Reflection)

Mesh from Depth Images Using GR^{2}T

Mairead Grogan & Rozenn Dahyot
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
mgrogan@tcd.ie, Rozenn.Dahyot@tcd.ie
www.scss.tcd.ie/~mgrogan/IMVIP.html

Abstract

This paper proposes an algorithm for inferring a 3D mesh using the robust cost function proposed by Ruttle et al. [12]. Our contribution is in proposing a new algorithm for inference that is very suitable for parallel architecture. The cost function also provides a goodness of fit for each element of the mesh which is correlated to the distance to the ground truth, hence providing informative feedback to users.

Keywords: 3D reconstruction, Depth images, Generalised Relaxed Radon Transform.

1 Introduction

To capture the 3D shape of an object using low cost hardware opens interesting perspectives for non-specialist users for archiving, reproducing or displaying objects. Several solutions have already been proposed. For instance Autodesk [1] reconstructs a 3D object from multiple colour images with offline processing on the cloud. Using depth images recorded by the Kinect sensor, Microsoft has proposed a real-time algorithm called Kinect Fusion for computing a 3D mesh of an environment [9] on desktop computers. Project Tango [7] pushes this further by reconstructing a 3D scene in real-time on a mobile device using an integrated depth sensor. However the success of all algorithms depends on the recorded data available for inferring the 3D scene. It is therefore important to give feedback about the estimated mesh to help the user improve the results by, for instance, recording more data of under-exposed areas of the scene. In this paper, we propose to use the cost function recently proposed by Ruttle et al. [12] for 3D reconstruction from depth images. This cost function corresponds to a probability density function over the 3D space allowing us to directly give a measure of confidence for each vertex and edge on the inferred mesh, giving feedback about the local quality of the mesh. This cost function is also designed to be robust [5] to noise, and any new recorded data point contributes to the overall cost function in an additive and localised fashion. After a brief review (section 2), we propose a new algorithm in section 3 that is highly parallelizable. Section 4 presents quantitative and qualitative 3D reconstructions obtained using our approach with comparison to ground truth and Ruttle et al.'s reconstructions. Conclusions and future work are discussed in section 5.

2 State of the Art

Reconstructing a 3D mesh from RGB-D cameras is an area of intense research in computer vision. In the past, algorithms have mostly consisted of three steps - denoising the depth images from several camera views, converting them to 3D point clouds and aligning the point clouds to recreate the surface [3, 9]. Izadi et al. proposed a real time 3D reconstruction algorithm (KinectFusion) using depth information [9]. An iterative pipeline is implemented which processes each depth image consecutively and uses a volumetric surface representation to generate a mesh of the scene. As a preprocessing step, a bilateral filter is applied to the raw depth data in order to reduce noise. This step ignores the uncertainty associated with the depth information and pixel positions as well as resulting in a loss of important information. A vertex and normal map of the scene from the first depth map are then computed using the connectivity of the pixels in the depth image.

For each consecutive depth map the pose of the camera is estimated and the depth information is fused with a volumetric truncated signed distance function (TSDF) [4] representing the scene. This representation gives a signed value to each voxel in the scene, depending on how far it is from the surface of the object. There is no measure of confidence associated with the vertices on the object's surface and the distances given to each voxel are not calculated using a robust objective function, but using a weighted distance measure. In order to render the surface in the scene a per pixel ray cast is performed. Each pixel's corresponding ray is calculated and marched starting from it's minimum depth value until the surface interface is found. This fully parallel mapping algorithm takes full advantage of GPU processing hardware and scales naturally with processing and memory resources.

Recently, methods have been proposed which generate a density function from 3D depth images or point clouds [12, 13]. To find points on the surface these density functions are then explored using either gradient ascent algorithms or marching cubes. Ruttle et al. [12] proposed to accurately infer the 3D shape of an object captured by a depth camera from multiple view points. The Generalised Relaxed Radon Transform (GR^{2}T) [5] is used here to merge all depth images in a robust kernel density estimate that models the surface of an object in the 3D space. The kernel is tailored to capture the uncertainty associated with each pixel in the depth images and the resulting cost function is defined for a 3D location Θ as:

$$\overline{\text{lik}}(\Theta) = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c} \sum_{i=1}^{N_c} p_\epsilon(F(x_c^{(i)}, \Theta, \Psi_c)), \tag{1}$$

where C is the number of recorded camera views (with known camera parameters $\Psi_c, \forall c = 1, \cdots, C$) and N_c is the number of pixels in the image generated by camera c, $x_c^{(i)}$ is a triplet of values recorded by camera c corresponding to the 2D location and the depth value of the pixel i. This function accounts for uncertainties in the observations via the probability density function p_ϵ that is chosen Gaussian. F is a link function associated with the pin-hole camera model connecting a 3D position to its projection in an image plane [8, 12]. Suitable values must also be chosen for the parameters h_1, h_2 and h_3, which account for noise in the pixel and depth values.

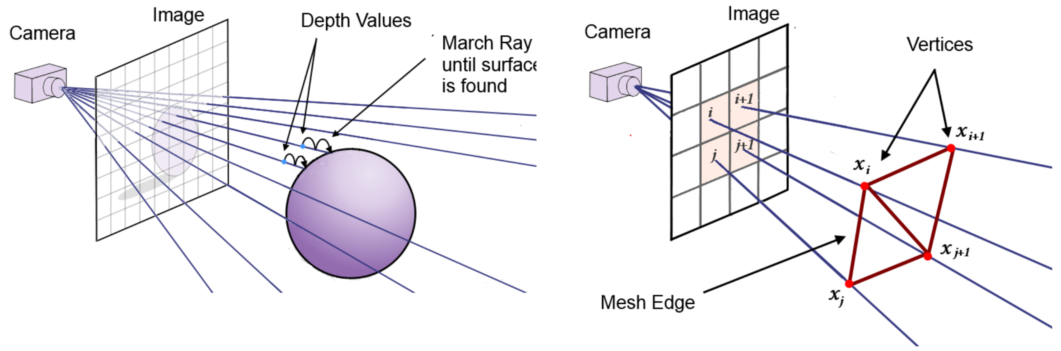
To extract a surface mesh using the cost function $\overline{\text{lik}}(\Theta)$, Ruttle et al. proposed a two stage process [12]. First maxima of the cost function are extracted. These are then connected in a second step by finding vertices and edges that connect them by following the ridge created by the object's surface in the cost function $\overline{\text{lik}}(\Theta)$. Both algorithms correspond to gradient ascent algorithms, the first is initialised by several positions in the 3D space to converge to several local maxima of $\overline{\text{lik}}(\Theta)$, while the second algorithm is initialised with the output of the first algorithm. This approach is simplified further by considering 2D slices in the 3D space, and performing the optimisation in parallel in these 2D manifolds.

This surface exploration technique is time consuming and inefficient. As an alternative to this two stage process, we propose an algorithm constraining the solution to a 1D manifold (a ray in the 3D space) to estimate a vertex on the surface of the object. This process is performed for each pixel in the depth images independently and the resulting approach is highly parallelizable. Connectivity between vertices is inferred automatically using pixel neighborhood information from the depth images.

3 Mesh Inference from $\overline{\text{lik}}(\Theta)$

A mesh is a discrete representation of a continuous coloured 3D surface and is made up of a number of different elements. In our approach, we used the .ply format, which can be used to store a variety of mesh properties including vertices, edges, faces, vertex colour, edge colour and vertex normals. We focus on efficiently inferring the vertices, faces and edges of the mesh as well as creating informative colour information which indicates the likelihood value of a particular vertex and edge.

In order to explore the density $\overline{\text{lik}}(\Theta)$ and determine which 3D points are most likely to be on the object's surface we propose generating a separate mesh for each camera view by casting a ray from the camera centre through each pixel in the image into 3D space, as shown in Figure 1a. This ray is then marched, starting from the calculated depth value, until the maximum likelihood value $\hat{\Theta}$ is found.



(a) Ray based optimisation of $\overline{\text{lik}}(\Theta)$.

(b) Generating the edges of the mesh.

Figure 1: Ray based strategy for Mesh from Depth images.

In order to calculate $\hat{\Theta}$, the point of maximum likelihood on the ray passing through the camera centre $C_{\Psi_c} \in \mathbb{R}^3$ and pixel (x_1, x_2) , we maximize the following:

$$\hat{\Theta} = \arg\max_{\beta} \ \overline{\text{lik}}(\Theta) \tag{2}$$

subject to the constraint that

$$\Theta = C_{\Psi_c} + \beta\vec{n}, \ \beta \in \mathbb{R}, \tag{3}$$

where $\vec{n}$ is the direction of the ray. As we can express Θ in terms of known parameters C_{Ψ_c} and $\vec{n}$, and a one dimensional latent variable β, we have reduced our latent space from three dimensions to one dimension. This greatly reduces the computational cost of our optimisation problem. W and H represent the width and height of the image in pixels and f_x represent the focal length in the horizontal axis. We define the horizontal field of view to be

$$\phi = \arctan\left(\frac{W/2}{f_x}\right). \tag{4}$$

The distance between the camera and the projection plane d is given by the formula $d = \frac{1}{tan(\phi)}$. We let $a = \frac{W}{H}$ and define $\vec{n} = (n_1, n_2, n_3)$ to be:

$$\begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} = \left(\begin{array}{c|c} R(\Psi) & \begin{matrix} \psi_4 \\ \psi_5 \\ \psi_6 \end{matrix} \end{array} \right)^{-1} \begin{pmatrix} \frac{2a}{W} & 0 & -a & 0 \\ 0 & \frac{-2}{H} & 1 & 0 \\ 0 & 0 & d & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ 1 \\ 1 \end{pmatrix}, \tag{5}$$

where $R(\Psi)$ is the camera rotation matrix and (ψ_4, ψ_5, ψ_6) is the camera translation vector [14, 6]. Given an observed depth value of x_3 at the pixel (x_1, x_2), the point on the ray which is a distance x_3 from the camera centre is given by $\beta^{(0)} = \frac{x_3}{\gamma}$ (initial guess) where

$$\gamma = \frac{\overrightarrow{n} \cdot C_{\Psi_c}}{\parallel \overrightarrow{n}, C_{\Psi_c} \parallel}. \tag{6}$$

A Newton Raphson gradient ascent algorithm is implemented to iteratively update the position on the ray until the point with maximum likelihood is found. This process is repeated for each pixel in the image, and generates a point cloud of the portion of the surface visible from camera c. This is repeated for all C cameras, generating C point clouds. Our approach is highly parallelizable as each ray can be computed and marched independent of the other rays.

Using this method, points are found at regular intervals along the surface and each point has a corresponding pixel in the image. We use the connectivity of the pixels to create the edges of the mesh. We consider four pixels i, $i+1$, j and $j+1$, as seen in Figure 1b, which make up a 2×2 square in the image. We also consider the 3D points x_i, x_{i+1}, x_j and x_{j+1} that were found by tracing the ray through each of these pixels. We create four edges $[x_i, x_{i+1}]$, $[x_j, x_{j+1}]$, $[x_i, x_j]$, $[x_{i+1}, x_{j+1}]$ between these points since their corresponding pixels are connected in a horizontal or vertical direction. Then, in order to ensure that the faces of the mesh are triangular in shape, we also create an edge between points x_i and x_{j+1}. This is a very simple meshing algorithm which is easy to implement and eliminates the need to cluster the data or calculate vertex neighbourhoods as in other meshing algorithms [11, 2, 10].

We set the colour value of each vertex Θ in the .ply mesh according to its likelihood value $\overline{\text{lik}}(\Theta)$. For each edge in the mesh, the barycentre B of the edge is calculated. The colour value of the edge is then set according to the value of $\overline{\text{lik}}(B)$. This can be seen in Figure 2. This allows the user to see which vertices and edges have a high or low likelihood, and which regions of the object may have been poorly scanned.

4 Experimental Results

Our approach was first applied to the ground truth Stanford Bunny mesh (size $10 \times 13 \times 13$ in cm). Autodesk 3DS max was used to generate 12 depth images of the bunny, with no noise added to the depth values (apart for the digitisation process in creating the projected depth images). The camera parameters were assumed to be known. We set the pixel bandwidths to $h_1 = h_2 = 2$ and depth bandwidth to $h_3 = .001$. For each camera view and corresponding depth image, a mesh was generated using our method. Figure 2 presents four meshes: the colour of each vertex in a mesh represents the likelihood that the vertex is on the bunny surface. Blue vertices have a low likelihood value and many appear at the edge of the mesh as their rays do not intersect with the bunny. Vertices with a low probability can be easily removed from the mesh by thresholding the likelihood values (second row of Figure 2). These colour values also illustrate which regions of the object have been poorly scanned, allowing the user to scan them in order to ensure that a more reliable mesh is generated.

The average time taken to converge to the point with highest likelihood on a given ray is .5162 seconds, with a standard deviation of .7080 seconds (non optimised Matlab code on a single core). The average number of iterations needed per ray is 71.2106. Our algorithm is highly parallelizable (as each ray can be marched independently) and optimising it to perform on the GPU would result in considerable speedup.

In Figure 3 we compare our reconstructed results to those obtained by Ruttle et al. in [12]. The colours of each vertex in these meshes represent the distance to the closest point on the ground truth Stanford Bunny mesh. The algorithm proposed by Ruttle et al. performs well apart from concave regions such as the neck and between the ears. Their meshing algorithm creates edges between points on different ears, and between points on the head and back. The red vertices on the bunny in Figure 3 (f) and (h) represent these meshing errors. Our results (top row of Figure 3) show that our meshing algorithm has eliminated these errors as it only considers vertices and edges with a high likelihood value.

We computed the average distance between the reconstructed meshes and the ground truth bunny mesh. The average distance is 0.000527m for our algorithm. This can be compared to 0.000711m obtained with Ruttle et al.'s algorithm [12]. We also investigated the correlation between the likelihood value $\overline{\text{lik}}(\Theta)$ of a vertex Θ, and the distance between Θ and the ground truth Stanford Bunny mesh. We have found that a threshold on the likelihood values can easily

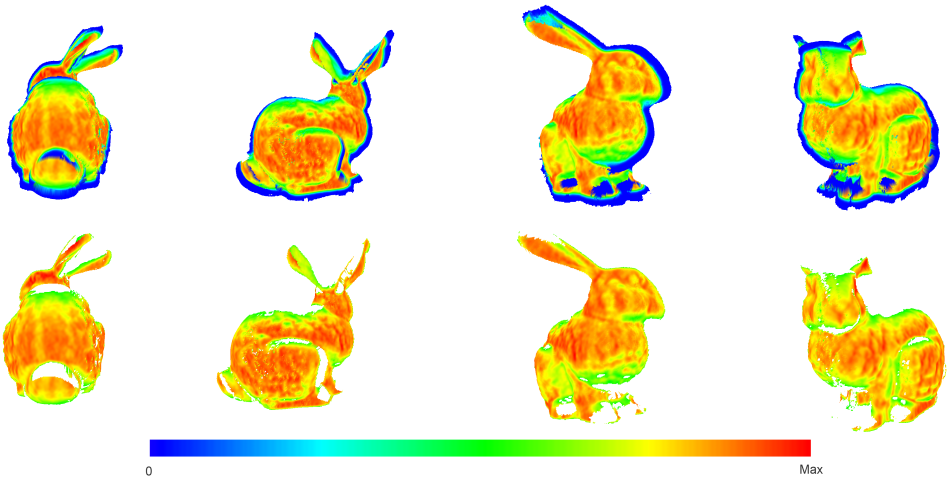


Figure 2: Meshes from 4 camera views (visualisation with Meshlab meshlab.sourceforge.net): with all vertices and edges shown with their probability (top) and when low probability vertices are deleted (bottom). The Max value on the colour scale refers to the maximum value of $\overline{\text{lik}}(\Theta)$ found on each mesh.

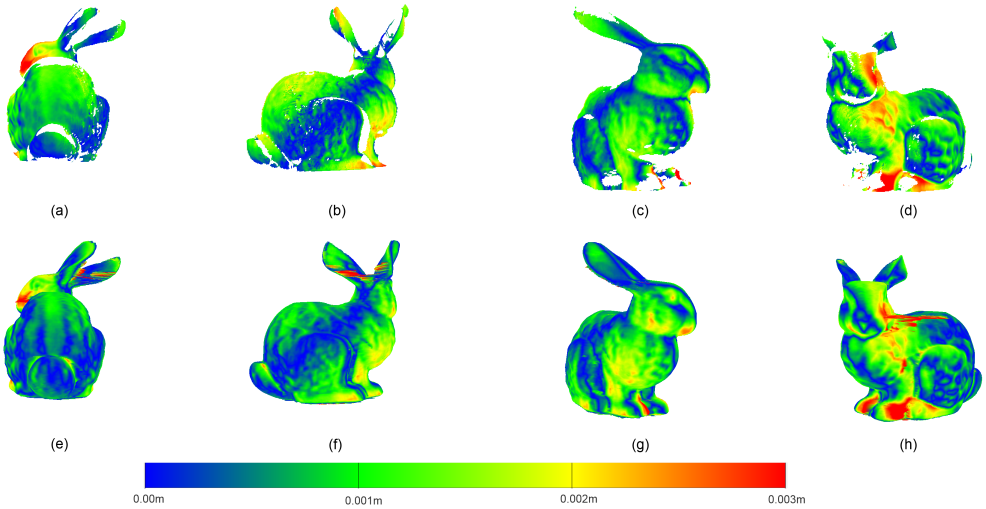


Figure 3: Visual comparison with ground truth (using Cloud Compare software www.cloudcompare.org): our method (top row), and Ruttle et al. [12] (bottom row).

be set to keep 73.65% of vertices, amongst which only 3% are vertices far from the ground truth (or 97% of vertices with high likelihood are close to the ground truth). This indicates that $\overline{\text{lik}}(\Theta)$ is a good measure of confidence for each element of the mesh. It can provide users with feedback to improve areas of the surface that have low likelihood and are therefore very likely to be far from the ground truth. Because this ground truth mesh is not available in general, $\overline{\text{lik}}(\Theta)$ can be used as a good substitute.

We have run several experiments with noisy depth images. Similarly to Ruttle et al., the cost function is robust to noise and our algorithm is not affected by this noise on the depth values.

5 Conclusion

We have proposed another algorithm for optimisation of the robust cost function proposed by Ruttle et al. [12]. This new approach allows us to infer a mesh of vertices and edges for each camera view including a measure of uncertainty for each element in the mesh. Future work will focus on stitching together the meshes created for each camera view so that a single mesh is generated. Considering the confidence value associated with each vertex will ensure those vertices with a higher likelihood will be given preference.

Acknowledgements

This work has been supported by a Ussher scholarship from Trinity College Dublin (Ireland), and partially supported by EU FP7-PEOPLE-2013-IAPP GRAISearch grant (612334).

References

[1] Autodesk. 123d catch. `http://www.123dapp.com/catch`, retrieved 04/2014.

[2] D. Bradley, T. Boubekeur, and W. Heidrich. Accurate multiview reconstruction using robust binocular stereo and surface meshing. In *Proceeding of CVPR*, 2008.

[3] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1180, 2010.

[4] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 303–312, New York, NY, USA, 1996. ACM.

[5] R. Dahyot and J. Ruttle. Generalised relaxed radon transform (GR^{2}T) for robust inference. *Pattern Recognition*, 46(3):788 – 794, 2013.

[6] J.D. Foley. *Computer Graphics: Principles and Practice*. Addison-Wesley systems programming series. Addison-Wesley, 1996.

[7] Google. Project tango. `https://www.google.com/atap/projecttango/`, retrieved 04/2014.

[8] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[9] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *ACM Symposium on User Interface Software and Technology*, 2011.

[10] S. J. Owen. A survey of unstructured mesh generation technology. In *International Meshing Roundtable*, pages 239–267, 1998.

[11] F. Remondino. From point cloud to surface: the modeling and visualization problem. *International Workshop on Visualization and Animation of Reality-based 3D Models*, 34:5, 2003.

[12] J. Ruttle, C. Arellano, and R. Dahyot. Robust shape from depth images with GR^{2}T. *Pattern Recognition Letters*, 2014. in press.

[13] J. Süßmuth and G. Greiner. Ridge based curve and surface reconstruction. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, SGP '07, pages 243–251, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association.

[14] A. Watt. *3D computer graphics (3. ed.)*. Addison-Wesley-Longman, 2000.

Comparison of geometric and texture-based features for facial landmark localization in 3D

Luke Gahan[1], **Federico M. Sukno**[2,1] **and Paul F. Whelan**[1]
[1]Centre for Image Processing & Analysis
Dublin City University, Dublin, Ireland
[2]Department of Information and Communication Technologies
Pompeu Fabra University, Barcelona, Spain

Abstract

The localisation of facial landmarks is an important problem in computer vision, with applications to biometric identification and medicine. The increasing availability of three-dimensional data allows for a complete representation of the facial geometry, overcoming traditional limitations inherent to 2D, such as viewpoint and lighting conditions. However, these benefits can only be fully exploited when the processing concentrates purely on the geometric information, disregarding texture. This fact is particularly interesting when addressing the localisation of anatomical landmarks, as it is not clear to date whether geometric information can be used to fully replace texture (e.g. the localisation of the eye corners and the lips is believed to be strongly linked to texture clues).

In this paper we present a quantitative study of 3D landmark localization based on geometry, texture or a combination of both, integrated in a common framework based on Gabor filters that has reported state of the art results. We target 10 facial landmarks and find that, while the algorithm performs poorly for the nose tip with a mean 3D error of 6.15mm, the remaining landmarks are all localised with an error under 3.35mm, with the outer eye corners and mouth corners performing particularly well. Interestingly, geometry and texture achieved comparable results for the inner eye corners and mouth corners, while texture clearly outperformed geometry for the outer eye corners.

1 Introduction

Facial landmark localisation is the primary step in a number of computer vision systems including facial recognition, facial pose estimation, medical diagnostics and multimedia applications. Historically most landmark localisation algorithms have used standard 2D images. Such systems, no matter how accurate, are always going to be limited by the fact that they are operating on dimensionally reduced representations of 3D objects. A significant amount of extra information about the human face is contained in the 3D spatial dimension.

A number of different approaches have been taken with regard to localising facial landmarks in 3D images. Geometry based techniques have received a good deal of attention. Segundo et al. present an effective system which uses surface classification techniques in order to localise landmarks [Segundo et al., 2010]. The authors record a 3D localisation error of under 10mm for 90% of images in their test set. Creusot et al. combine machine learning and a large number of geometric techniques in their system [Creusot et al., 2013]. The authors note that while this system does not outperform others in terms of accuracy, it does perform quite well in terms of robustness. Since the algorithm used is not sequential in nature, a failure to detect certain landmarks does not influence the localisation of subsequent landmarks. This system provides a framework for landmark localisation and leaves potential for future improvement.

Zhao et al. present a statistical model based approach in [Zhao et al., 2011]. This system works well in challenging situations where there is facial occlusion and/or very expressive faces. This system learns the spatial relationships between different landmarks and uses this in conjunction with local texture and range information. The authors use Principal Component

Analysis (PCA) to create a statistical facial feature map. This is essentially a combination of individual geometry (landmark coordinates), shape (range images) and texture (texture images) models. The authors report a mean 3D error rate of below 5.07mm for all 15 facial landmarks.

Perakis et al. use local shape descriptors to localise facial landmarks [Perakis et al., 2013, Passalis et al., 2011]. These local shape descriptors characterise the shape profile at a given landmark. By evaluating the shape index at a landmark in a number of training images a model can be constructed. These descriptors are generated by examining the principal curvature and spin image at a landmark. A facial landmark model is then created. This is used to constrain the relative locations of detected landmarks. Models are also created for the left and right hand side of the face. These are used to deal with profile or semi-profile faces. The systems achieves relatively good results with a mean 3D error of below 5.58mm for all 8 targeted landmarks.

One particular approach which has received increased attention in recent years is the use of Gabor filters for facial landmark localisation [Movellan, 2002]. Jahanbin et al. use Gabor filter banks for landmark localisation in [Jahanbin et al., 2008]. This technique implements the same landmark localisation procedure as Wiscott et al. used in their Elastic Bunch Graph Match system (without the elastic constraint) [Wiskott et al., 1997]. While the authors do not present in depth results in this particular paper, it does serve as a basis for later work carried out by the same research group [Gupta et al., 2010b]. This particular system combines curvature detection, Gabor filters and expert knowledge of the human face to localise landmarks using anthropometric information based on the work carried out by Farkas et al. in the medical field [Farkas and Munro, 1987]. This information plays a vital role in establishing a sensible search region which is then examined to further improve the accuracy of localisation.

An interesting element of the work by Gupta et al. [Gupta et al., 2010b] is that Gabor filters are applied to both range and texture and their framework allows for a direct integration of both sources of information. However, the authors did not provide a detailed analysis of this aspect and results were limited to 2D standard deviation errors, which hampers a thorough comparison to other approaches. In this work we present a quantitative analysis of landmark localization errors when using texture, range or both sources of information at the same time. We use the framework developed by Gupta et al. and reproduce the results reported originally, which allows to also calculate the mean 3D error to make results comparable to related work. We find that the inclusion of both texture and range information always yields the best results, although the benefit of range was negligible in some cases. Interestingly, for the inner eye corners and mouth corners the error results were similar for all three tested alternatives.

2 Automatic Landmark Localisation Using Anthropometric Information

The landmark localisation procedure carried out remains as faithful as possible to the method developed by Gupta et al. [Gupta et al., 2010b]. Generally speaking the algorithm first uses curvature information to detect an approximate location for a particular landmark. Using anthropometric information a search region is defined around this approximation and the position is then refined using as described below. The 10 landmarks localised are the nose tip, with points and root center, inner and outer eye corners and mouth corners.

Nose Tip (prn): The Iterative Closest Point (ICP) algorithm is used to register each face in the database to a frontal template face. These aligned images are used in all subsequent steps. Once all images have been aligned the manually localised tip of the template face is taken as an approximate location for tip of the nose in all images. A window of 96 mm x 96mm is then defined around this approximated nose tip. Since all faces have been frontally aligned, the actual nose tip is present in this large window for all cases. This means that the method is not fully automated since it relies on the manually localised tip of the template face.

It has been observed that the Gaussian surface curvature of the tip of the nose is distinctly elliptical (K $>$0,) [Moreno et al., 2003, Segundo et al., 2010, Creusot et al., 2013]. For this reason the Gaussian surface curvature (σ = 15 pixels) is evaluated within the search region about the nose tip approximation. The maximum Gaussian curvature within the region is taken as final location of the nose tip (prn).

Nose Width Points (al-al): These points are localised by first defining a search region around the detected nose tip. The size of this window (42 mm x 50 mm) is defined based on the mean and standard deviation values published by Farkas [Farkas and Munro, 1987]. A Laplacian of Gaussian edge detector (σ = 7 pixels) is then used within this region. Moving in a horizontal direction from the nose tip, the first edge encountered is considered to be the nose contour and is retained. Then, points of negative curvature are detected by generating an unwrapped chain code for the nose contour and using a derivative of Gaussian filter on this one dimensional signal to detect points of critical curvature [Rodriguez and Aggarwal, 1990]. Nose width points are finally selected from the critical points immediately above and below the vertical coordinate of the nose tip. The widest of these are selected as nose width points.

Inner Eye Corner (en-en) & Center of Nose Root (m'): A search region for the left and right inner eye corners is defined using the location of the detected nose tip and nose width points. The vertical limit defined based on the fact that for the average adult, the distance between inner eye corners and the tip of the nose in the vertical direction is 0.3803 times the distance between the tip of the nose and the top point of the head [Farkas and Munro, 1987, Gupta et al., 2010b]. Gupta et al. allow for variations in the measure by setting the upper vertical limit at $(prn_y + 0.3803 \times 1.5|prn_y - V_y|)$, where V_y is the Y coordinate of the highest vertical point in the 3D model. The horizontal limit is obtained by using the locations of the nose width points and the nose tip. Specifically, horizontal limits are defined from the nose tip to $al_{x,left/right} \pm 0.5|al_{x,left} - al_{x,right}|$ for the left and right inner eye corners.

The Gaussian curvature within this region is evaluated and the location of maximum curvature is used as an approximation for the location of the inner eye corner ($\sigma = 15$ pixels). Finally a region of 20mm x 20mm is defined around this peak of Gaussian curvature.

The location of inner eye corners are then refined with a modified version of the EBGM technique [Jahanbin et al., 2008, Wiskott et al., 1997]. In brief, this technique involves comparing the Gabor coefficients generated for each pixel in the search region with the coefficients for the landmarks of 89 training images. These 89 images consist of neutral and expressive faces. The images are selected in an attempt to cover as much feature variance as possible (i.e. closed/open mouth and eyes). 80 Gabor coefficients (known as a Gabor jet) are generated at each landmark for each of the example images. A filter bank of 40 Gabor filters is used (5 scales x 8 orientations). 40 coefficients are generated for both range (3D) and texture (2D) images. While the specific parameters of these filters are not provided in [Gupta et al., 2010b], we used the filter bank outlined in by Wiscott et al. [Wiskott et al., 1997]. Note that, for the database used, all images should be scaled by $\frac{1}{3}$ when Gabor filtering is applied.The final location of the inner eye corner is obtained by finding the pixel which has a Gabor jet most similar to that of any training landmark. The similarity score is given in equation (1):

$$S(\overrightarrow{J}, \overrightarrow{J'}) = \frac{\sum_{i=1}^{40/80} a_i a'_i \cos(\Phi_i - \Phi'_i)}{\sqrt{\sum_{i=1}^{40/80} a_i^2 \sum_{i=1}^{40/80} a_i'^2}} \qquad (1)$$

where J and J' are the jets to be compared, defined as $J_j = a_j \, e^{i\phi_j}$. Where a is the magnitude and ϕ is the phase of the Gabor coefficient at a given pixel.The jets contain either 40 or 80 coefficients depending on which form of EBGM is to be used. Gupta et al. chose to use 2D and 3D Gabor coefficients. In this work 2D, 3D and 2D+3D results are compared. The center of the nose root is determined by finding the mid-point between the two inner eye corners.

Outer Eye Corners (ex-ex): A search region for the outer eye corners is defined based on the location of the detected inner eye corners as per [Gupta et al., 2010b]. This 20 x 34 mm region is evaluated using the same search procedure as used for the inner eye corners. Gupta et al.

chose to use 2D EBGM search as the outer eye corner region does not have distinct enough curvature characteristics. In this work all three EBGM techniques are evaluated.

Mouth Corners (ch-ch): The lip curvature is examined in order to determine a search region for the mouth corners. The Gaussian curvature of both the upper and lower lips is elliptical in nature. The regions immediately above the upper lip and below the lower lip are hyperbolic ($K < 0$). These properties can be used to define upper and lower search limits for the mouth corners. The horizontal limits are defined by $[(al_{x,left} - 0.7|al_{x,left} - al_{x,right}|), (al_{x,left})]$ for ch_{left} and analogously for ch_{right}. In order to remove noise a certain amount of smoothing must be carried out when calculating Gaussian curvature. In some cases the Gaussian curvature of the upper or lower lip is too weak and cannot be localised. In such cases the troughs in Gaussian curvature immediately above and below the lip region are used as limits. While these are usually stronger features than the lips, errors can arise when searching for peak mean curvature in the next stage of the algorithm as there is a high mean curvature along the jaw line.

The mean curvature ($\sigma = 2$ pixels) is then calculated for the defined search region. Since the mouth corners are regions of high mean curvature the peak curvature value in this region is taken as an estimate for of the mouth corner. A 30mm x 11mm search region is defined around these mouth corner estimates. The same EBGM procedure used to localise the eye corners is also used to precisely localise the mouth corners. Gupta et al. chose to use 2D+3D EBGM. In this work 2D, 3D and 2D+3D EBGM results are compared.

3 Experimental Results & Discussion

3.1 Test Data

The performance of the landmark localisation algorithm is evaluated using the Texas 3DFR database [Gupta et al., 2010a]. It contains high resolution (751 x 501 pixels, 0.32 mm per pixel) pairs of portrait and range images from 118 healthy adult subjects. 25 facial landmarks have been manually located. Both range and portrait images were acquired simultaneously using a regularly calibrated stereo vision system and the data was filtered, interpolated and smoothed to remove impulse noise and large holes [Gupta et al., 2010a]. From the 1149 portrait-range pairs of the database, 89 were used in the EBGM search and the remaining 1060 were used as test data.

3.2 Landmark Localisation Results

The landmark localisation results obtained for the Texas 3DFR database are given in Table 2. All results are given in millimetres. As mentioned previously Gupta et al. do not provide 3D error results [Gupta et al., 2010b]. Thus, we compared our results to the ones originally provided, in terms of 2D standard deviation and confirmed that our implementation a faithfully reproduced the original method (Table 1).

The mean error result of the nose tip is noticeably larger than the localisation of the other landmarks. On closer examination it appears that in all cases the detected nose tip is above the manually localised nose tip (in the Y direction). This can clearly be seen in the boxplot in Figure 1. This figure shows clearly that the median value for the X error is 0mm as expected in a normal error distribution. The Y distribution is extremely skewed to one side of the manually localised nose tip (a negative Y error is above the manual location for an upright face). Since the standard deviation of the Y error is relatively small it seems that the issue is that the peak of Gaussian curvature does not correspond to the same location the manual annotators have identified as the nose tip.

The mean error results obtained for the nose width points are reasonable while the standard deviations are impressive, especially when using the modified EBGM technique. A 3D mean error of under 2mm is recorded for both inner eye corners. The outer eye corners which are slightly more difficult to localise are detected with a mean error of under 2.6mm. A mean

	X std. dev (mm)		Y std. dev (mm)		2D std. dev (mm)	
Landmark	**Gupta**	**This Method**	**Gupta**	**This Method**	**Gupta**	**This Method**
PRN	1.045	0.766	1.680	1.714	1.978	1.705
AL Left	0.721	0.647	1.655	0.710	1.805	0.739
Al Right	0.798	0.546	1.646	0.814	1.829	0.818
EN Left	1.488	1.249	1.245	0.908	1.940	1.363
EN Right	1.354	1.378	1.344	0.792	1.908	1.417
M'	1.355	1.415	1.811	1.010	2.261	1.417
EX Left	1.795	1.727	1.285	1.047	2.208	1.850
EX Right	2.126	1.940	1.384	1.248	2.537	2.149
CH Left	1.948	1.749	0.933	1.692	2.160	2.321
CH Right	1.976	1.429	1.045	0.844	2.235	1.460

Table 1: Error standard deviation results comparison with Gupta et al. [Gupta et al., 2010b]

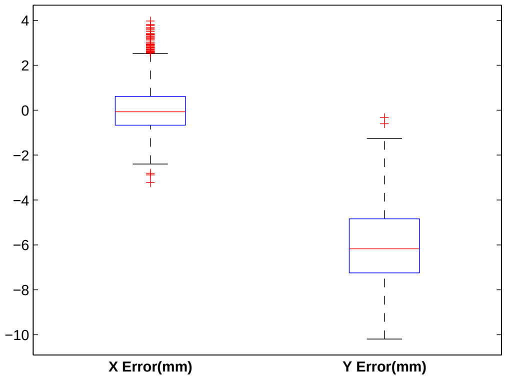


Figure 1: Vertical Prn Error Bias

error of below 2.16mm is achieved for both mouth corners. The algorithm does have particular difficultly with faces where facial hair is present. This is as expected when using Gabor filters as there is a significantly different response to a Gabor filter when facial hair is present.

One interesting point to note is that the three worst results obtained are for the three landmarks localised using techniques which do not involve training. The training stage of EBGM uses manual landmark locations. This means that when EBGM is used, the algorithm searches for a location on an unknown image which is most similar to the training data, which is based on manual locations. For the nose tip and width points the algorithm searches for a particular image feature (e.g. maximum Gaussian curvature) which is said to be present at that landmark. Perhaps using EBGM for all landmarks might yield better performance. Another possible issue could be marker bias. No details are provided about how many annotators are used but using separate annotators for test and training data could be a possible solution.

Landmark	Prn	Al_L	Al_R	En_L	En_R	M'	Ex_L	Ex_R	Ch_L	Ch_R
3D mean	6.15	3.35	3.31	1.82	1.75	2.76	2.48	2.59	2.16	2.02
3D stdev	1.75	1.65	1.88	1.50	1.52	1.59	2.58	2.99	3.04	2.15

Table 2: Landmark localisation error

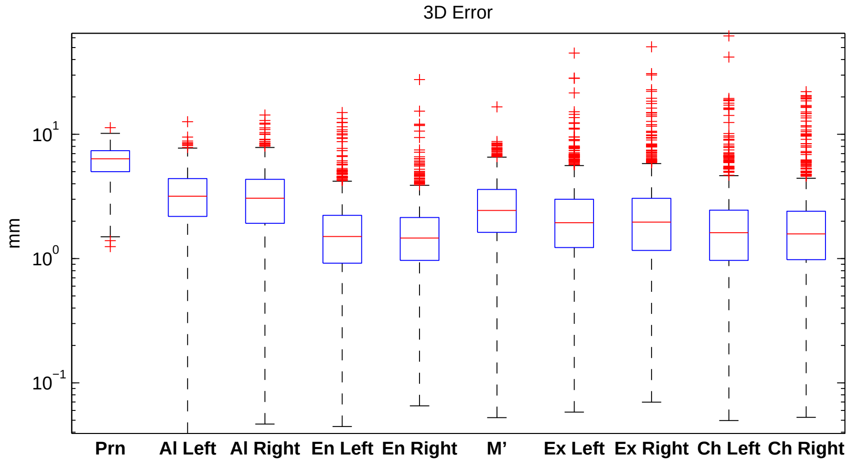


Figure 2: 3D error boxplot

3.3 Texture & Range Comparison

The inner eyes and outer mouth corners are detected using 2D + 3D EBGM while 2D EBGM is used for the outer eye corners. The same similarity metric is used in each case (1) with the only difference being the coefficients examined.

Landmark	2D EBGM	3D EBGM	2D+3D EBGM
En Left	1.83 ± 1.53	2.18 ± 1.70	1.82 ± 1.50
En Right	1.75 ± 1.55	1.99 ± 1.58	1.75 ± 1.52
Ex Left	2.48 ± 2.58	5.10 ± 5.28	2.39 ± 2.12
Ex Right	2.59 ± 2.99	8.91 ± 7.22	2.49 ± 2.27
Ch Left	2.20 ± 2.83	2.54 ± 2.89	2.16 ± 3.04
Ch Right	2.15 ± 2.44	2.20 ± 1.61	2.02 ± 2.15

Table 3: 2D, 3D & 2D+3D EBGM comparison, in terms of 3D error (mean ± std. dev.)

Interestingly, Table 3 shows that for the inner and outer eye corners the inclusion of range coefficients improves localisation results. Gupta et al. use 2D + 3D for the inner eye corner while they choose to use just 2D for the outer eye corners. The results obtained here suggest that a similar improvement in localisation could be achieved with the inclusion of range information. While it is clear that just using 3D information results in poor localisation performance it should be noted that the 3D information only influences the result of localisation when a 3D coefficient is more similar to one of the training image coefficients than any of the 2D coefficients. This means that in some individual cases the inclusion of 3D information may adversely affect localisation but for the entire database the average error is reduced.

With regard to the mouth corners the use of texture and range information results in the best mean error performance. This is the same as the behaviour for the other landmarks. Once again the worst mean error is recorded when just range information is used.

It is clear that in all cases examined the inclusion of more information (texture & range) in the EBGM stage results in better overall localisation. This suggests that the similarity score and the procedure Gupta et al. use for choosing the landmark location works quite well. It suggests that in the majority of cases the inclusion of extra information leads to enhanced localisation performance. Obviously there is a computational overhead to be considered when including this extra information but in cases where speed isn't an issue it seems that the inclusion of 2D and 3D information leads to the best localisation performance.

Since the 2D and 3D EBGM techniques are directly comparable, Table 3 shows that for all landmarks examined texture information yields better results. Though for the inner eye corners and mouth corners this difference is quite small.

4 Conclusion

We have shown that the method developed by Gupta et al. achieves state of the art landmark localisation results. The one weak point is the localisation of the nose tip which is quite poor. Even though the localisation of the tip is poor it does not appear to adversely affect the localisation of subsequent landmarks where the location of the nose tip is used to define a search region. Another better performing method, such as that used by Segundo et al., could perhaps be used for the localisation of the nose tip [Segundo et al., 2010].

It was determined that for the EBGM stage, the inclusion of both texture and range information yields the best results. Interestingly, for the inner eye corners and mouth corners the error results recorded are similar for each of the EBGM methods. For the outer eye corner 3D EBGM performed quite poorly, with 2D and 2D+3D obtaining similar results. This suggests that for outer eye corner detection, 2D EBGM could be used without a significant ($\sim$ 0.3mm) decrease in mean error.

Acknowledgments

The authors would like to thank their colleagues in the Face3D Consortium (www.face3d.ac.uk), and financial support from the Wellcome Trust (WT-086901 MA) and the Marie Curie IEF programme (grant 299605, SP-MORPH)..

References

[Creusot et al., 2013] Creusot, C., Pears, N., and Austin, J. (2013). A machine-learning approach to keypoint detection and landmarking on 3D meshes. *Int. J. Comput. Vis*, 102(1-3):146–179.

[Farkas and Munro, 1987] Farkas, L. G. and Munro, I. R. (1987). *Anthropometric facial proportions in medicine*. Charles C. Thomas Publisher.

[Gupta et al., 2010a] Gupta, S., et al. (2010a). Texas 3D face recognition database. In *Proc. SSIAI 2010*, pages 97–100.

[Gupta et al., 2010b] Gupta, S., Markey, M. K., and Boxxxvik, A. C. (2010b). Anthropometric 3D face recognition. *Int. J. Comput. Vis*, 90(3):331–349.

[Jahanbin et al., 2008] Jahanbin, S., Bovik, A. C., and Choi, H. (2008). Automated facial feature detection from portrait and range images. In *Proc. SSIAI 2008*, pages 25–28.

[Moreno et al., 2003] Moreno, A. B., et al. (2003). Face recognition using 3D surface-extracted descriptors. In *Proc. IMVIP 2003*.

[Movellan, 2002] Movellan, J. R. (2002). Tutorial on gabor filters. *Open Source Document*.

[Passalis et al., 2011] Passalis, G., et al. (2011). Using facial symmetry to handle pose variations in real-world 3D face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1938–1951.

[Perakis et al., 2013] Perakis, P., et al. (2013). 3d facial landmark detection under large yaw and expression variations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1552–1564.

[Rodriguez and Aggarwal, 1990] Rodriguez, J. J. and Aggarwal, J. (1990). Matching aerial images to 3D terrain maps. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(12):1138–1149.

[Segundo et al., 2010] Segundo, M. P., et al. (2010). Automatic face segmentation and facial landmark detection in range images. *IEEE Trans. Syst., Man, Cybern. B*, 40(5):1319–1330.

[Wiskott et al., 1997] Wiskott, L., et al. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779.

[Zhao et al., 2011] Zhao, X., et al. (2011). Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-dimensional statistical facial feature model. *IEEE Trans. Syst., Man, Cybern. B*, 41(5):1417–1428.

Indoor rigid sphere recognition based on 3D point cloud data

Jifang Duan, Kishan Lachhani, Hadi Baghsiahi, Eero Willman, David R. Selviah
Department of Electronic and Electrical Engineering
University College London (UCL)
Torrington Place
WC1E 7JE London
Jifang.duan.13@ucl.ac.uk, d.selviah@ucl.ac.uk

Abstract

This paper presents a method for recognising spherical shapes in 3D point cloud XYZ coordinate data obtained by scanning an indoor environment using a LIDAR scanner. Firstly, bilateral smoothing is performed to smooth the surfaces consisting of points. Then, the surface curvature and surface roughness of each point in the scan are extracted by analysing the point cloud data. Finally, a three layer multilayer perceptron neural network trained by the Levenberg-Marquardt algorithm is used to automatically distinguish points belonging to spheres from all the other points making use of extracted features. A novel feedback technique is applied in which the neural network is used several times on the recognised data.

Keywords: Point cloud data, Object recognition, Neural networks, 3D laser scanning, LIDAR.

1 Introduction

The development of 3D scanning and camera projection technology make it possible for people to have better access to large amounts of accurate 3D point cloud data. The point cloud data, recorded by 3D LIDAR scanners, is used in a variety of fields, including architecture, medical science, surveying and mapping. It is widely applied in generating 3D models, undertaking metrology inspection and performing medical imaging.

In general, point cloud data itself is not directly usable in most 3D applications as it occupies a lot of memory and storage, requiring further analysis and processing. For example, point clouds can be converted to mesh models or CAD models for further use. In the field of architecture, the Building Information Modelling (BIM) concept was introduced in recent years. BIM describes the whole life cycle of a project and gives very detailed information of everything related to the building, including cost, construction, project and facility management. Currently, BIM is mainly used at the start of construction projects where laser scanning may not be that useful since 3D models of buildings would normally already exist. However, with the expansion of the BIM industry, existing buildings will require BIM as well, and that will offer a market for laser scanning and modelling automation. At present, the conversion from point cloud data obtained from a scanned building to BIM is typically performed by manual means, which is time consuming and labour intensive. In order to facilitate the procedure, S. Oesau et al. [1] proposed a method using feature sensitive primitive extraction and graph-cut for automatic reconstruction of permanent structures, such as walls, floors and ceilings. X. Xiong et al. [2] succeeded in identifying and modelling the main visible structural components of an indoor environment. Apart from large planar areas, windows and doorways are also able to be identified by applying this method. The question is whether other shapes such as spheres can also be automatically identified.

This paper focuses on indoor sphere recognition. Spheres have the characteristic of rotational symmetry, which can be regarded as a distinctive feature for alignment between different scans. As a basic geometric shape, spheres of small or large radius appear everywhere in buildings. Being able to recognise spherical objects can be seen as a start for the recognition of more complex 3D objects. The Hough transform is a powerful tool in shape analysis. O. Ogundana et al. [3] extended the strategy for detecting circles in 2D images to detecting spheres in 3D point clouds. RANSAC is also a useful tool for 3D sphere extraction [4]. This paper proposes a new method for sphere recognition. Our algorithm can be divided into four main steps: 1) bilateral smoothing, in which the point clouds are smoothed; 2) calculation of the surface curvature and the surface roughness; 3) multilayer perceptron neural networks are trained using supervised learning by the Levenberg-Marquardt algorithm, and used to distinguish points belonging to spheres from other points; 4) low-density filtering, in which low-density points are removed from the point cloud.

2 Methodology

2.1 Bilateral smoothing

The 3D point cloud data is obtained by Faro Focus 3D LIDAR. Due to the tolerances of the scanner itself, the 3D data, inevitably, contains range noise. In order to improve the accuracy and reliability to point cloud computation, it is important to de-noise and smooth the point cloud. The bilateral filter, introduced by Tomasi and Manduchi [5], is a non-linear, edge-preserving and noise-reducing filter, which was first used to filter images. It has a simple and intuitive formulation and can be adapted to point cloud data easily and successfully [6-7].

In 3D point cloud data smoothing, let, $\mathbf{p}$, be the 3D coordinates of a point in the scan. After the application of the filter, updating, $\mathbf{p}$, as is given in Eq. (1):

$$\mathbf{p}' = \mathbf{p} + b \cdot \mathbf{n} \tag{1}$$

Where, $\mathbf{n}$, is the surface normal of the point, b , is the bilateral smoothing factor defined as follows:

$$b = \frac{\sum_{\mathbf{p_i} \in R} W_c\left(\left\|\mathbf{p} - \mathbf{p_i}\right\|\right) \cdot W_s\left(\left|\left\langle \mathbf{n}, \mathbf{n_i} \right\rangle\right|\right) \cdot \left\langle \mathbf{n}, \mathbf{p} - \mathbf{p_i} \right\rangle}{\sum_{\mathbf{p_i} \in R} W_c\left(\left\|\mathbf{p} - \mathbf{p_i}\right\|\right) \cdot W_s\left(\left|\left\langle \mathbf{n}, \mathbf{n_i} \right\rangle\right|\right)} \tag{2}$$

Where, R, is the spherical neighbourhood of $\mathbf{p}$, $\mathbf{p_i} \in R$ apart from the point $\mathbf{p}$, $\left\|\mathbf{p} - \mathbf{p_i}\right\|$ is the distance between point $\mathbf{p}$ and $\mathbf{p_i}$, $\left\langle \mathbf{n}, \mathbf{n_i} \right\rangle$ is the angle between vector $\mathbf{n}$ and $\mathbf{n_i}$. The closeness smoothing filter is a standard Gaussian filter with parameter σ_c: $W_c(x) = e^{-x^2/2\sigma_c{}^2}$. A feature-preserving weight function with parameter σ_s is defined as: $W_s(x) = e^{-x^2/2\sigma_s{}^2}$.

2.2 Surface curvature and surface roughness

The surface curvature (SC) of each point of the scan is computed from eigenvalues of a local 3 by 3 covariance matrix [8] of a certain region of interest around the point. This region is usually taken to be a sphere. The radius of the sphere is required to be chosen wisely. It should be much smaller than the size of the scanned spheres while relatively larger than the surface thickness of the scanned objects. The surface roughness can be obtained by comparing surface normals of neighbouring points. Again this is calculated over a spherical region of interest whose radius is set to be the same as that for the SC calculation. The surface normal of each point is calculated using principle component analysis and covariance analysis [9] with an Octree-based 3D-grid method [10] for efficient neighbouring point searching.

Theoretically, the average SC of a sphere should be invariant despite any changes in the distance between the parts of the sphere and the scanner and the variable density of points, and the SC of each point on one sphere should be the same. However, in practice this is not the case. The point cloud data in the experiments is from a single scan, so objects are partially scanned and will have point cloud boundary edges. The discrepancy in surface curvature calculation is mainly due to the effect of edges and the spatially varying range noise.

2.3 Neural networks trained by the Levenberg-Marquardt algorithm

Artificial neural networks (ANN) can detect complex nonlinear relationships between dependent variables and separate and distinguish different classes of pattern. Several algorithms can be used for the training procedure of an ANN. The Levenberg-Marquardt (LM) algorithm [11] is selected for this research. The LM algorithm is a combination of the steepest descent method and the Gauss-Newton algorithm. It inherits the stability of the steepest descent method and the speed advantage of the Gauss-Newton algorithm.

In this research, two features, surface curvature and surface roughness, are extracted from point cloud data for each point. They are used for the inputs of a 3 layer multilayer perceptron ANN for each point. Four spheres and two non-sphere backgrounds selected from the scan are used for training this ANN. After being well trained but not overtrained, the ANN model is applied to each point in a sample area in order to identify points belonging to spheres. Repeated application of the ANN helps in obtaining better discrimination results. Each time after the ANN model is applied, the points identified by the ANN as not belonging to a spherical surface are removed from the scene. This reduction of number of points results in the modification of the curvature and the roughness values for points remaining in the scene when they are recalculated before putting back into the same ANN. As a result each time the ANN acts as a filter removing non-sphere points and the SC and SR are recalculated and the ANN is applied again. After applying the ANN for several times, there still remain some points that do not belong to spheres, so a low-density filter is applied, which is described in section 2.3, to remove these points.

2.3 Low-density filtering

For each point, the number of its neighbouring points within a certain spherical volume of radius, r, around that point is calculated. The distance, d, between each point and the LIDAR is also computed. Points, whose neighbouring points count is smaller than a threshold number, th, are removed from the point cloud. The threshold, th, is calculated as follows:

$$th = \frac{\sigma r^2}{2d^2} \tag{3}$$

where, σ, is the number of points per unit area at 1 metre away from the scanner.

2.4 Flowchart of methodology

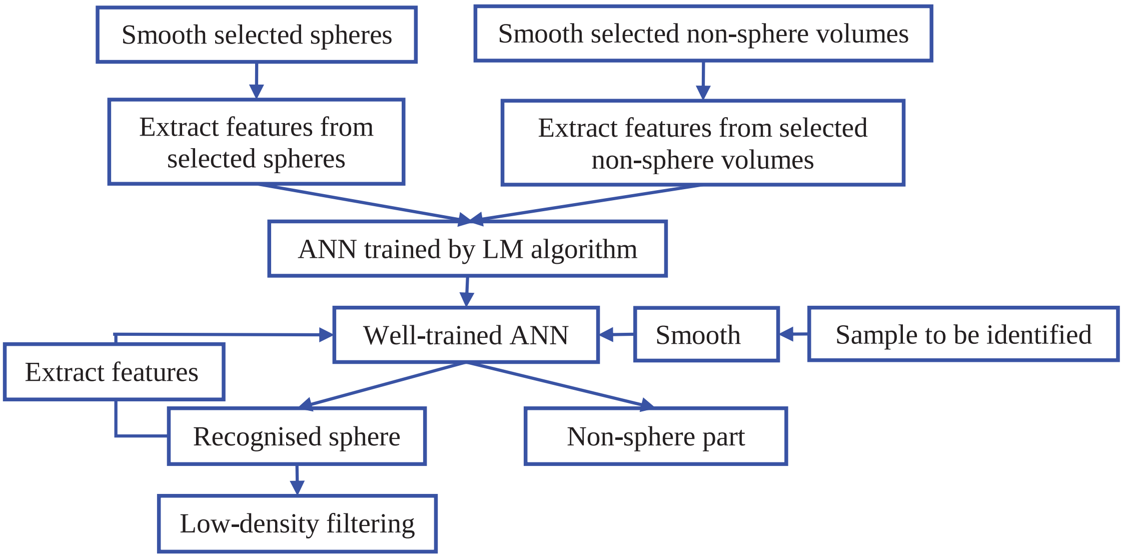
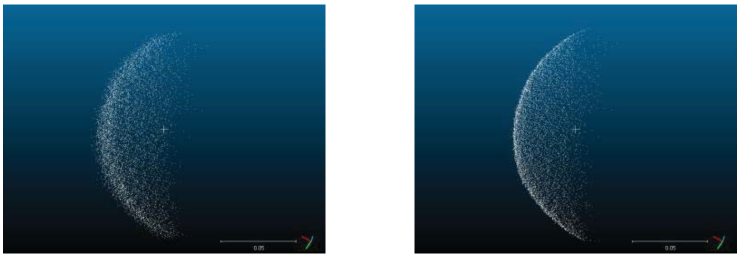

Smooth selected spheres
Smooth selected non-sphere volumes
Extract features from selected spheres
Extract features from selected non-sphere volumes
ANN trained by LM algorithm
Well-trained ANN
Smooth
Sample to be identified
Extract features
Recognised sphere
Non-sphere part
Low-density filtering

Figure 1: Flowchart of methodology

3 Experimental results

In the experiment, eight spheres were randomly placed in an indoor environment. The scene was then scanned with Faro Focus 3D LIDAR. Figure 2 shows a side view of a partially captured sphere. After applying bilateral smoothing, the surface of this sphere appears smoother and better defined. A quantitative comparison of the calculated surface thicknesses for spherical objects and the scene floor before and after smoothing is presented in Table 1. Filtering improves the accuracy in computing the curvature and the roughness measures for each point, which are illustrated in Figure 3 and Figure 5. Figure 4 shows the calculated curvature value distribution on the surface of a partially captured sphere. Near the edges of the partial sphere, the curvature is relatively low, with a value close to that of a flat surface. Conversely, points on the sphere, located further away from the edges possess relatively higher curvature. Figure 6 plots both the curvature and the roughness of a sphere after smoothing shows that there exists a relationship between the two. These two different properties of the sphere surface have a similar tendency.

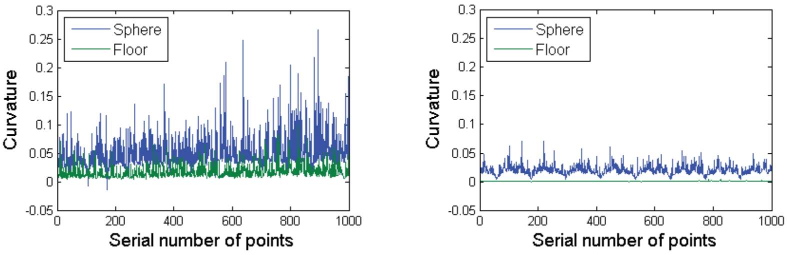


a. Before smoothing *b. After smoothing*

Figure 2: Comparison of a scanned sphere before/after smoothing

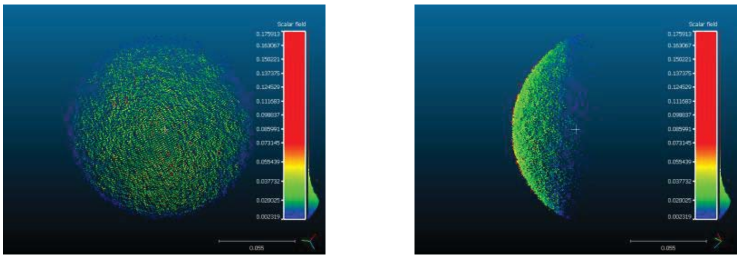
Surface thickness	Sphere (2 m away from the scanner)	Sphere (4.5 m away from the scanner)	Floor (2 m away from the scanner)	Floor (4.5 m away from the scanner)
Before smoothing (mm)	20	15	12	3
After smoothing (mm)	8	6	2	1

Table 1: Comparison of the surface thicknesses of scanned objects before and after smoothing



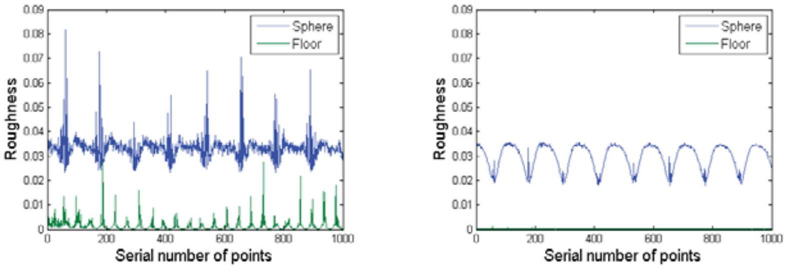
a. Before smoothing *b. After smoothing*

Figure 3: Comparison of surface curvature (a) before and (b) after smoothing. The curvature is calculated for each point on a selected sphere and for each point on a selected area of floor



a. Front view *b. Side view*

Figure 4: Calculated surface curvature distribution over a sphere



a. Before smoothing *b. After smoothing*

Figure 5: Comparison of surface roughness before and after smoothing. The roughness is calculated for each point on the selected area of the sphere and for each point on the selected area of the floor

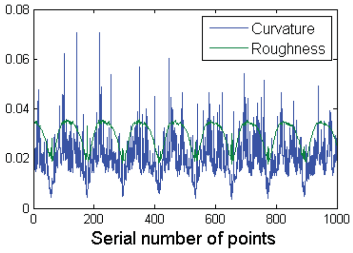


Figure 6: Surface curvature and roughness for points of a selected area of the sphere after smoothing

The distance between each of these eight scanned spheres and the scanner was in the range from 2 m to 6 m. As is shown in Figure 7 and Figure 8, the implementation of bilateral smoothing reduces the variability of average curvature and average roughness for spheres placed at different distances with respect to the LIDAR scanner. For detecting objects, it is desirable that the calculated curvature and roughness are independent of their distances to the scanner.

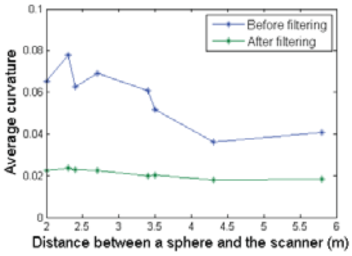


Figure 7: Comparison of the average curvature before/after smoothing between spheres

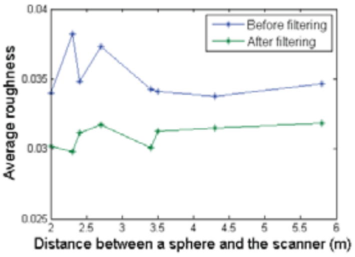
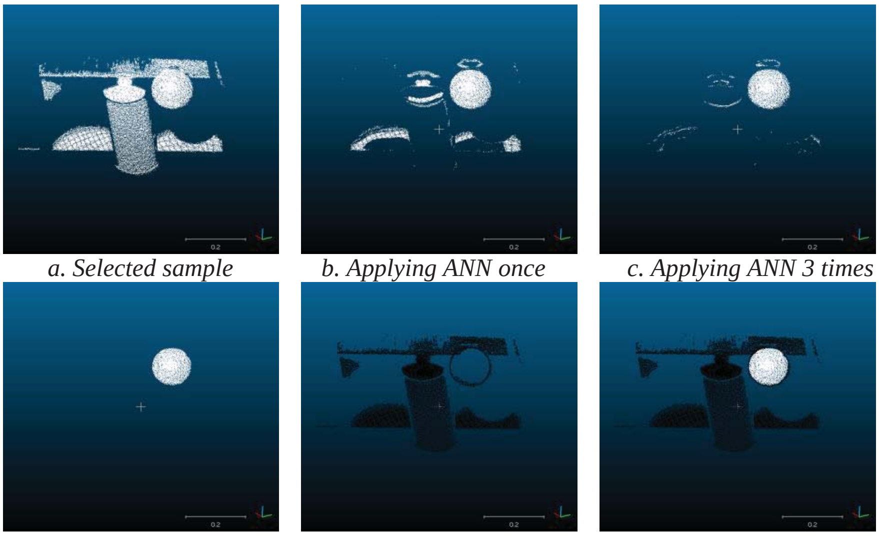


Figure 8: Comparison of the average roughness before/after smoothing between spheres

Four of these scanned spheres and two non-sphere volumes were selected to train the ANN model. Parameters for the ANN chosen for this experiment are presented in Table 2. Figure 9 (a) shows the sample area selected from a scan, which includes a sphere, a lamp and a flat surface. This data was not used in the training of the ANN model. Figure 9 (a) - (c) show the effect of repeatedly applying the ANN to the data. After this, low-density filtering described in section 2.3 is applied to the remaining points. The result of this is shown in Figure 9 (d).

Number of neurons for input/hidden/output layer	2; 10; 2
Transfer function	Tan-sigmoid, Log-sigmoid
Percentage of data for training/validation/testing	70%; 15%; 15%

Table 2: Parameters for ANN



a. Selected sample *b. Applying ANN once* *c. Applying ANN 3 times*

d. Recognised sphere *e. Points other than the recognised sphere* *f. Overview*

Figure 9: Recognition process. (a) Input data. (b)-(c), result of applying ANN repeatedly to input data. (d) remaining points after ANN and density filtering of input data. (e) and (f) Points identified to lie on non-spherical surfaces (black) and points classified to lie on surface of sphere (white).

4 Conclusions

In this paper, we present an artificial neural network pattern recognition approach to detect points in a point cloud that define the surface of a spherical object. Our method is able to correctly distinguish points belonging to spheres from other points in the environment which may also include other curved surfaces present. A novel feedback technique is applied in which the neural network is used several times on the input data.

References

[1] O. Sven, L. Florent and A. Pierre (2014). Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut. *ISPRS Journal of Photogrammetry and Remote Sensing*, 90: 68-82.

[2] X. Xuehan, A. Antonio and A. Burcu (2013). Automatic creation of semantically rich 3D building models from laser scanner data. *Automation in Construction*, 31: 325-337.

[3] O. Olatokunbo, C. Russell and L. Richard (2007). Fast Hough transform for automated detection of spheres in three-dimensional point clouds. *Optical Engineering*, 46 (5): 051002

[4] R. Schnabel, R. Wahl and R. Klein (2007). Efficient RANSAC for point-cloud shape detection. *Computer Graphics Forum*, 26 (2): 214-226

[5] C. Tomasi and R. Manduchi (1998). Bilateral filtering for gray and colour images. *Sixth International Conference on Computer Vision*, 839-846.

[6] S. Fleishman, I. Drori, and D. Cohen-Or (2003). Bilateral mesh denoising. *ACM Transactions on graphics*, 22: 950-953.

[7] D. Xiaoyuan, J. Xiaofeng, H. Chuangang and W. Yumei (2010). Bilateral filtering denoising algorithm for point-cloud model. *Computer Applications and Software*, 20 (7): 245-264.

[8] M. Pauly, M. Gross and LP. Kobbelt (2002). Efficient simplification of point-sampled surfaces. *IEEE Visualization 2002*, 163-170.

[9] M. Pauly (2003).Point primitives for interactive modeling and processing of 3D geometry. For the degree of Doctor for Sciences. Federal Institute of Technology of Zurich.

[10] H. Woo, E. Kang and SY. Wang (2002). A new segmentation method for point cloud data. *International Journal of Machine Tools & Manufacture*, 42: 167-178.

[11] J. Shawash and D. Selviah (2013). Real-Time nonlinear parameter estimation using the Levenberg-Marquardt algorithm on field programmable gate arrays. *IEEE Transactions on Industrial Electronics*, 60 (1): 170-176.

Error Metric for Indoor 3D Point Cloud Registration

Kishan Lachhani, Jifang Duan, Hadi Baghsiahi, Eero Willman, David R. Selviah
Department of Electronic and Electrical Engineering
University College London (UCL)
Torrington Place
London, WC1E 7JE
kishan.lachhani.13@ucl.ac.uk, d.selviah@ucl.ac.uk

Abstract

An increase in commercial availability of 3D scanning technology has led to an increase of 3D perception for a variety of applications. High quality scanners require to be stationary and so multiple scans are required and subsequently need to be registered. A new error metric for registration based on the deviation of registered planar surfaces is introduced here and compared with a commonly used metric: mean square point-to-point distance. Four different sets of features are used to register six scans, the point-to-point errors are compared to the new error metric, planar surface deviation, and a disparity is observed for certain sets of features. The two metrics agree as to which sets of features gave the best registration but disagree as to which set produced the worst registration. It is concluded that further analysis and evaluation is required to determine which metric is more meaningful as a representative measure of registration accuracy and to also investigate other error metrics.

Keywords: Point Cloud Registration, 3D Laser Scanning, LIDAR, Feature Recognition, Principal Component Analysis

1 Introduction

LIDAR is a remote sensing technology that measures direction and range, similar to RADAR, except that it uses a laser. The laser beam from the LIDAR illuminates a surface; the surface may scatter some of the light back to the LIDAR from which the distance is determined using either phase-shift or time information. Stationary LIDARs perform a sweeping scan of their environment creating a point in 3D space for each range measurement; such a collection of points is referred to as a point cloud. For a dense, high resolution and accurate point cloud at relatively long ranges, LIDAR is type of technology typically used though there are other options available. The increase in commercial availability of such technologies means that 3D perception continually gains importance in applications such as 3D mapping and navigation, architecture, augmented reality, robotics and gaming.

A LIDAR scanner is used here to collect point cloud data due to its cost, accuracy and availability. Like many other similar technologies it is a stationary unit and so multiple scans are required from multiple vantage points to attempt to reduce of impact of occlusions and capture a complete or near-complete point cloud. The multiple scans lead to the common problem in computer vision of registration. The task of registration is to place the individual point clouds in the same spatial reference frame by estimating rigid body transformations between the datasets. The problem is difficult because the correspondences of the datasets and the precise location of the scanner are unknown a priori; the difficulty in obtaining this information accurately means that it is problematic to evaluate meaningful and truly representative registration errors.

The purpose of finding such transformations is to acquire a more complete dataset which increases its usability and reliability which is important for many applications. Due to the relatively high accuracy of 3D laser scanning technology (typically ±2 mm at 25 m), the importance of accurate registration becomes of greater significance since the range accuracy is a key limiting factor of registration accuracy. The required registration accuracy is ultimately specified by the client and the application. The point cloud data may be used to provide an approximate representation of the scanned environment for visualisation applications or it may be used for some other application where metrology is of greater importance such as BIM (Building Information Modelling). BIM is one of the most significant drivers for 3D scanning technology and 3D imaging [1], and in the UK, government mandate states that by 2016, public sector centrally procured construction projects will be delivered using level 2 BIM [2].

Scanning processes in BIM have a required minimum accuracy, however, the 'Client Guide to Scanning and Data Capture' published by the BIM Task Group [3] advocate a functional performance approach rather than a prescriptive approach to establishing this number. This essentially means that considering the technology used for data capture, construction tolerances, budgetary restrictions and other constraints the best achievable accuracy should be sought.

In the next section, we discuss point cloud registration error metrics associated with the most popular registration methods such as ICP (Iterative Closest Point) [4]. ICP often requires good initial alignment typically achieved by feature-based registration. We also introduce our new error metric which measures the deviation of planar or near-planar surfaces in registered scans. We address the problem of registration error metrics of point clouds in scenarios where the true value is not observable. In real-world applications, an accurately and precisely measured true value is too difficult to obtain so we resort to measuring quantities about objects from the scene from which we can infer the degree of registration accuracy. Typically this may be corresponding points or nearby points and planes, however, since we work with indoor 3D scans in which there is typically an abundance of planar features, we use these to assess the degree of registration accuracy.

Real point clouds with planar regions have some surface deviation arising from range noise (ranging accuracy), surface profile and poor registration. In the ideal case, truly planar surfaces produce point clouds restricted to two dimensions such that there is no surface deviation; furthermore, perfectly registered ideal point clouds would also return no surface deviation. If we can account for range noise and surface profile in our error metric then planar surface as an error metric should provide representative registration errors.

2 Registration Error Metrics

A method which is often used for registration of two point clouds is ICP (Iterative Closest Point); in this algorithm, one point cloud, the reference or the target is kept constant, while the other one is transformed to minimise the distance of the closest points between the reference and target. The rigid body transformation is iteratively evaluated for the revised closest points. ICP is very popular due to its simplicity, however, it only works very well in ideal cases, subsequently there are a very large number of ICP variants (around 400 papers in the past 20 years with ICP in the title or abstract) [5] which enable it to be more robust or faster but the basic principal remains the same which is that the distance between iteratively revised closest points are minimised.

A paper by Rusinkiewicz and Levoy [6] reviews some of the efficient variants of ICP and classifies these as affecting one of the 6 stages of the algorithm: selection, matching, weighting, rejecting, assigning an error metric and minimising the error metric. Most of the variants aim to add speed and robustness to the algorithm, but here we are concerned with the accuracy of registration which is ultimately determined by the error metric. The metric specified in the original ICP paper [4] is the sum of squared point-to-point distances, other metrics include a combination of point-to-point and difference in colour [7], point-to-plane [8] and point-to-line [9] distances. Certain metrics may behave better than others in certain cases in terms of converging to the ground truth but their limitations are intrinsic to ICP. Whether the ICP variant uses points, planes, lines or anything else, the limitation lies in the fact that corresponding references are chosen by proximity. After numerous iterations the error may converge, but it may or may not converge to the true value; in either case, this cannot be known from summing squares distances between points which are not truly corresponding.

Another limitation of ICP is that it converges monotonically to local minima and the final result is very dependent on the initial conditions, for this reason, most ICP variants require a good initial estimate to increase the likelihood of converging to a global minimum. The initial estimate is typically evaluated using methods which are more robust to range noise such as feature-based registration. Feature-based registration algorithms attempt to identify truly corresponding points and to minimise the sum of these squared point-to-point distances. They are limited in accuracy due to range noise and on the premise that corresponding points are only truly corresponding within a certain tolerance. Nonetheless, they are a popular tool in determining coarse registration which is typically followed by fine registration performed by ICP.

Even with a good initial estimate, ICP is very susceptible to range noise which is something which is typical of real data; Low and Lastra [10] have shown that rate of convergence and likelihood of convergence to a global minimum can be improved by supressing noise through smoothing of smoothly varying surfaces. In our collected data, as mentioned previously, the data is relatively low in noise though there are many points which are perturbed by noise (44 million points per scan), the accuracy of range measurements is accurate within a standard deviation of 2 mm at 25 m. This level of noise and number of points may prove to inhibit ICPs likelihood of converging to the global minimum [11].

Recently there has been a re-emergence and increased interest in registration of point clouds represented as Gaussian mixture models [12]–[14], these models do not require pair-wise correspondences in the same way as ICP or feature based registration algorithms but instead use a probabilistic approach to reduce correspondence mismatch errors. Due to the simplicity and popularity of ICP, here we compare our metric with point-to-point/plane only and analysis on the mixture model registration errors is reserved for future work.

3 Planar Surface Deviation (PSD)

There are a number of advantages to using planar surfaces for an error metric. Particularly in our application of indoor 3D scanning, planar surfaces such as walls, ceilings and furnishings are typically found in abundance. Also, planar regions are identified by many points, at least hundreds if not thousands within a 25 cm radius depending on the distance of the scanner to the surface. Such a large number of points can be utilised to supress noise (by averaging or plane-fitting, for example) without deforming the structure by smoothing or other pre-processing. Lastly, unlike feature points, planes are localised in one dimension; relatively small variations in the other two dimensions do not alter the distance normal to corresponding registered planes and subsequently do not significantly affect the error metric. As a result, errors on planar surfaces should be calculated on multiple orthogonally orientated samples.

To evaluate the surface deviation of planar regions in combined registered scans, knowledge of the surface normal is required first; the normal is the direction in which we determine the surface deviation. First we identify the query points, which are the points which lie at the centre of the regions of interest; such points can be identified using shape detection algorithms such as RANSAC [15] and Hough transforms [16], however, for simplicity and proof of concept we identify such points manually here.

Next we identify points in the neighbourhood of our query point and use these to calculate the normal. Surface normal estimation can be achieved in many different ways (see [17]), the simplest is based on first order 3D plane fitting outlined in [18], which is essentially a least-square plane fitting estimation problem. The surface estimation problem is reduced to an eigenvector and eigenvalue analysis (or principal component analysis) of a 3D covariance matrix created from the neighbourhood of points around the query point. The surface normal is then estimated by the eigenvector corresponding to the smallest eigenvalue which corresponds to the direction of smallest variance [19]; additionally, the square root of the eigenvalue determines the standard deviation along the corresponding eigenvector. By comparing the standard deviation of individual planar regions to the registered and combined region, we obtain our planar surface deviation metric.

PSD is well suited to our application of indoor 3D scanning due to the typical abundance of planar surfaces in many buildings, however, scenes that lack plentiful planar surfaces would deem this metric far less useful. Since most new scanning applications are concerned with buildings and large structures, it is fair to say that PSD would be suited for many applications. Additionally, it should be noted that this metric, currently, does not identify corresponding planes but only evaluates errors for nearby planes which are assumed to be corresponding; this means that the usefulness of the metric is determined by reasonably well registered scans which will depend on the search radius for neighbourhood points and the size of the plane itself. For example, if the registration returns planes which are not very close together then they may be excluded from the neighbourhood around your query point and provide an overly optimistic registration error. Another limitation of PSD depends upon the surfaces themselves, the surface may have a certain profile which cannot easily be known from the scan data, such a profile would manifest as surface deviation. Additionally the surface profile will be measured differently depending on the position of the scanner relative to the surface normal; this would cause misinterpretation of the true position of the surface.

4 Method

We scanned the nanotechnology laboratory in the Department of Electronic and Electrical Engineering at UCL from 6 vantage points using a LIDAR (see Figure 1). The main types of features used for registration here are checkerboard targets which are strategically placed on walls and planar surfaces which are also used for registration. The first step of registration is to identify the features, 3 sets of features are extracted from the scans (automatically identified checkerboards (AI CHB), manually identified checkerboards (MI CHB) and automatically identified planes (AI Planes)) and a 4th set is acquired by using total station surveying instruments (total station identified checkerboards (TSI CHB)). To register two scans, the correspondences are identified between features from sets A and B. These correspondences are then used to determine the transformation required to minimise the distance between the corresponding features. Following this, the correspondences are then used to perform fine registration to minimise the distances further. We then compared the mean correspondence distance (also referred to as CD or point-to-point distance) after fine registration with the planar surface deviation (PSD) of a number of planar regions. Since the laboratory is rectangular, we take the mean PSD of five regions (typically containing many thousands of points) from the long walls, short walls and the ceiling, respectively labelled X, Y and Z. The errors, in both cases, are evaluated for the final registration of the 6 scans.

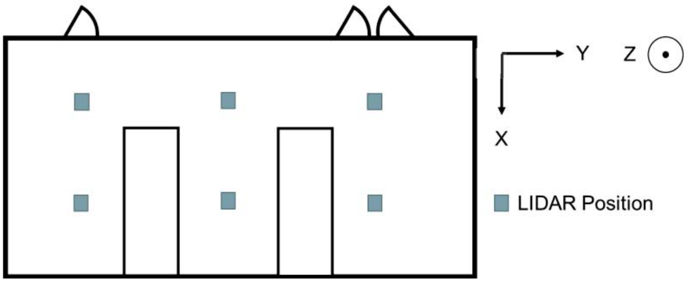


Figure 1 – Laboratory layout indicating workbenches and LIDAR position.

Method	Feature Set A	Feature Set B
Manual CHB	MI CHBs	MI CHBs
Automatic CHB	AI CHBs	AI CHBs
Total Station CHB	AI CHBs	TSI CHBs
Automatic Planes	AI Planes	AI Planes

Table 1 – Table identifying the features used in each registration method. MI – Manually Identified, AI – Automatically Identified, TSI – Total Station Identified and CHB – Checkerboard.

5 3D Scans and Results

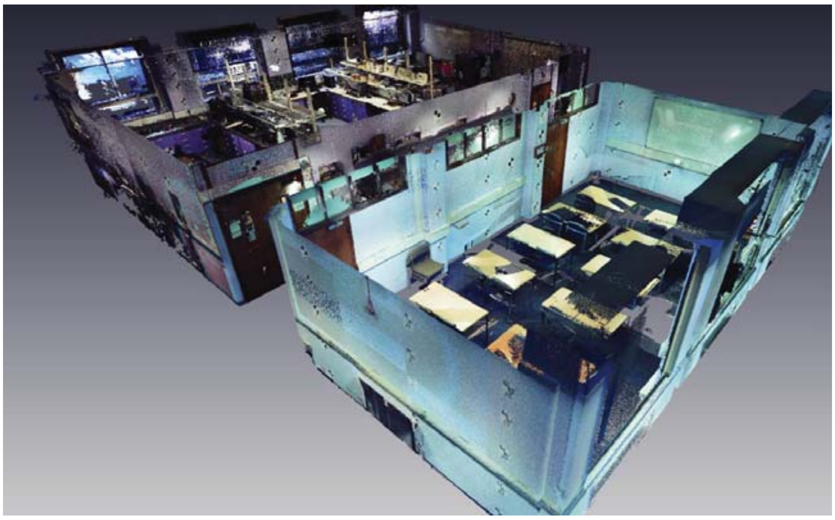


Figure 2 - Screenshot of the registered scans of the laboratory and classroom.

		Manual CHB	Automatic CHB	TSI CHB	Automatic Planes
	Correspondences	38	218	33	100
CD	**Mean (mm)**	0.8	0.6	2.7	6.7
	Deviation (mm)	0.4	0.7	1.8	6.9
PSD	**X (mm)**	0.8	0.8	16.3	1.3
	Y (mm)	0.9	1.2	10.0	3.2
	Z (mm)	10.2	10.1	4.0	6.2

Table 2 - Table of alignment errors measured by CD (correspondence distance or point-to-point distances of corresponding points) and PSD (planar surface deviation) for the 4 methods. Mean and deviation of all correspondences are determined for CD. PSD is evaluated for a set of 3 orthogonal surfaces.

6 Conclusion

It can be seen from Table 2 that CD and PSD agree that Manual and Automatic CHBs provide good alignment. CD states that Automatic planes produce the worst alignment while PSD states that Total Station CHBs produce the worst alignment. Though both alignment error measures agree as to which sets of features produce the best alignment, they disagree as to which produce the worst alignment. Further testing and evaluation is required to determine which method is the more meaningful measure of error. The CD error method hides variations in error in X, Y and Z so multiple checkerboards on the walls reduce the misalignment error in X and Y while hiding the much larger misalignment error in Z.

This work attempts to assess the accuracy of registration in a novel way by using the spatial deviation in the direction of the surface normal of overlapping planar regions from different scans. Planar features are found in abundance indoors, and using additional information from these planar regions gives for a more detailed analysis of the final registration. In the future, we intend to include more information in PSD, to also extend the method to better account for the type of surface in question and to analyse a wider range of metrics.

7 References

[1] BIM Task Group, “A report for the Government Construction Client Group,” 2011. [Online]. Available: http://www.bimtaskgroup.org/wp-content/uploads/2012/03/BIS-BIM-strategy-Report.pdf. [Accessed: 13-May-2014].

[2] H. Government, “Industrial strategy: government and industry in partnership - Building Information Modelling,” 2012. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/34710/12-1327-building-information-modelling.pdf. [Accessed: 13-May-2014].

[3] BIM Task Group, “Client Guide to 3D Scanning and Data Capture,” 2013. [Online]. Available: http://www.bimtaskgroup.org/wp-content/uploads/2013/07/Client-Guide-to-3D-Scanning-and-Data-Capture.pdf. [Accessed: 13-May-2014].

[4] P. J. Besl and N. D. McKay, “A Method for Registration of 3-D Shapes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 586–606, Apr. 1992.

[5] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat, “Comparing ICP variants on real-world data sets,” *Auton. Robots*, vol. 34, no. 3, pp. 133–148, Feb. 2013.

[6] S. Rusinkiewicz and M. Levoy, “Efficient variants of the ICP algorithm,” *Proc. Third Int. Conf. 3-D Digit. Imaging Model.*, pp. 145–152, 2001.

[7] A. Johnson and M. Hebert, “Surface registration by matching oriented points,” in *Proceedings. International Conference on Recent Advances in 3-D Digital Imaging and Modeling (Cat. No.97TB100134)*, 1997, pp. 121–128.

[8] Y. Chen and G. Medioni, “Object modeling by registration of multiple range images,” in *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, 1992, pp. 2724–2729.

[9] A. Censi, “An ICP variant using a point-to-line metric,” *2008 IEEE Int. Conf. Robot. Autom.*, pp. 19–25, May 2008.

[10] A. Lastra, “Reliable and rapidly-converging ICP algorithm using multiresolution smoothing,” in *Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings.*, 2003, no. 3dim, pp. 171–178.

[11] D. Simon, “Fast and accurate shape-based registration,” PhD Dissertation, Carnegie Mellon University, 1996.

[12] B. Jian and B. Vemuri, “Robust point set registration using gaussian mixture models,” *Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1633–1645, 2011.

[13] C. Arellano and R. Dahyot, “Mean shift algorithm for robust rigid registration between Gaussian Mixture Models,” *Signal Process. Conf. (EUSIPCO), 2012 Proc. 20th Eur.*, pp. 1154–1158, 2012.

[14] J. Ruttle, C. Arellano, and R. Dahyot, “Robust shape from depth images with GR2T,” *Pattern Recognit. Lett.*, pp. 1–12, Jan. 2014.

[15] R. Bolles and M. Fischler, “A RANSAC-Based Approach to Model Fitting and Its Application to Finding Cylinders in Range Data.,” *IJCAI*, pp. 637–643, 1981.

[16] P. Hough, “Method and means for recognizing complex patterns,” *US Pat. 3,069,654*, 1962.

[17] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, “Comparison of surface normal estimation methods for range sensing applications,” *2009 IEEE Int. Conf. Robot. Autom.*, pp. 3206–3211, May 2009.

[18] R. B. Rusu, “Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments,” *KI - Künstliche Intelligenz*, vol. 24, no. 4, pp. 345–348, Aug. 2010.

[19] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, “Surface reconstruction from unorganized points,” *ACM SIGGRAPH Comput. Graph.*, vol. 26, no. 2, pp. 71–78, Jul. 1992.

IMVIP 2014

APPLICATIONS

Effects of Different Mixtures of Features, Colours and SVM Kernels on Wheat Disease Classification

Punnarai Siricharoen, Bryan Scotney, Philip Morrow and Gerard Parr
School of Computing and Information Engineering
University of Ulster, Coleraine, BT52 1SA
siricharoen-p@email.ulster.ac.uk

David Gibson, Nishan Canagarajah
Department of Computer Science
University of Bristol, Bristol, BS8 5UB
gibson@cs.bris.ac.uk

Abstract

This paper assesses how the combination of different features, colour models and SVM kernels affect the classification performance of wheat disease identification. The basic approach consists of pre-processing, feature extraction, and classification. Five colour models (greyscale, RGB, HSV, YCbCr and L*a*b*), four different feature sets (Haralick, Tamura, First-order statistics and HOG features) and three kernels for a support vector machine (linear, RBF and polynomial) are assessed in terms of overall performance accuracy. Image datasets including non-diseased, Yellow Rust diseased and Septoria diseased leaves have been acquired under controlled conditions. The results show that homogeneity and contrast or energy features combined with basic statistical information such as mean, skewness and kurtosis, and visually perceptual features consisting of directionality, contrast and coarseness, which are extracted from YCbCr images using a classification model based on a linear kernel, produce the highest classification accuracy with low computational complexity

Keywords: Plant Disease Classification, Feature Extraction, Combination of Features, SVM Kernel

1 Introduction

Crop diseases can lead to a substantial decrease in both quantity and quality of agricultural products worldwide. To reduce such losses, early notification or continuous monitoring of crops is required. However, it is expensive, time-consuming and labour-intensive for experts to accurately diagnose the symptoms appearing on a plant, especially in remote areas.

When plants are infected, they can exhibit a range of symptoms, for example, colour spots or colour bands on the leaves, fruit, stems, or seeds. Recently, various image processing techniques have been developed for automated disease detection. Automated systems for plant disease identification have played a role in agriculture not only for rapid detection but also reducing human error. However, to apply an automated classification system to real plant diseases, robust imaging methods are required. The aim of this work is to develop an automated plant disease identification system using image processing. We briefly discuss previous literature on plant disease classification in Section 2, give the details of our methodology and proposed classification system in Section 3, and describe experimentation and evaluation in Section 4. Finally, conclusions are given in Section 5.

2 Literature Review

General automated systems have four main components: pre-processing, segmentation, feature extraction, and classification. Firstly, pre-processing techniques are applied to handle data differences arising from different lighting conditions, or capture devices. The methods used include, but are not limited to, Colour Transformation, Colour Correction using a colour chart, and image enhancement. The choice of colour model is crucial in representing an image for statistical processing. The standard colour model in computer displays is the RGB model. The channels in RGB are highly correlated, so RGB is unlikely to be the best model for describing information [1]. The CIELAB (L*a*b*) model provides a normalised and more visually uniform chromaticity and more perceptually uniform luminance [1]. The HSV colour space is an intuitive colour model for describing data as well as the HSI colour model, which is designed for image processing [2]. Moreover, YCbCr is a model applied in digital video. The Cb and Cr components are employed as independent two-dimensional distributions which are unaffected by brightness [3]. Secondly, the pre-processed data are clustered into several groups such as background and particularly foreground, as regions of interest will be segmented out and used in the next process. Widely used methods include K-means clustering as a fast and simple technique [4]–[8], Fuzzy c-means [9] as it is more flexible than K-means, and Otsu's thresholding [4], [10], [11] as a robust binary classification. Thirdly, feature extraction is applied to calculate disease pattern representation in segmented areas. Various features have been used including features derived from the greyscale spatial independence matrix [4], [6], [12]–[15], shape and properties of regions of interest [9], [10], [16] and first-order statistical features [9]. Other potential features are widely applied in different applications such as HOG features [17] and visual perception features [18]. However, there is little evidence to demonstrate the effectiveness of features especially for plant disease patterns. In addition, most research studies apply individual sets of features [4], [6], [12], [13], [15], [19] or combine the whole set of different features in the system [10], [16]. The more features that are calculated, the more computational time is required, and so some studies have applied principal component analysis to remove correlated features [16], [20]. Fourthly, the classification models are constructed mainly using Neural Networks (NNs) and Support Vector Machine (SVM). Neural Networks are widely used for many applications in intelligent systems because of their ability to perform non-linear modelling. However, NNs have drawbacks in high computation complexity, a tendency for over-fitting, and lack of explainable relationships amongst inputs, outputs and variables [21]. The Support Vector Machine is a well-known classification method that is generalisable and able to cope with non-separable data [22].

Our classification system comprises pre-processing using different colour transformations, feature extraction using many types of features, and classification using SVM. The system is proposed to investigate the potential of different colour components, various features and SVM kernels affecting plant disease classification performance. Image segmentation is omitted in the automated system as we assume that the leaves are already segmented out from the background (though we recognise that this is far from a trivial task).

3 Methodology and Proposed Classification System

An automated classification system is proposed and is shown in Figure 1. The system consists of three main steps: pre-processing, feature extraction and classification. We assume that leaves have been segmented out from the background and that only one main leaf in each image is considered.

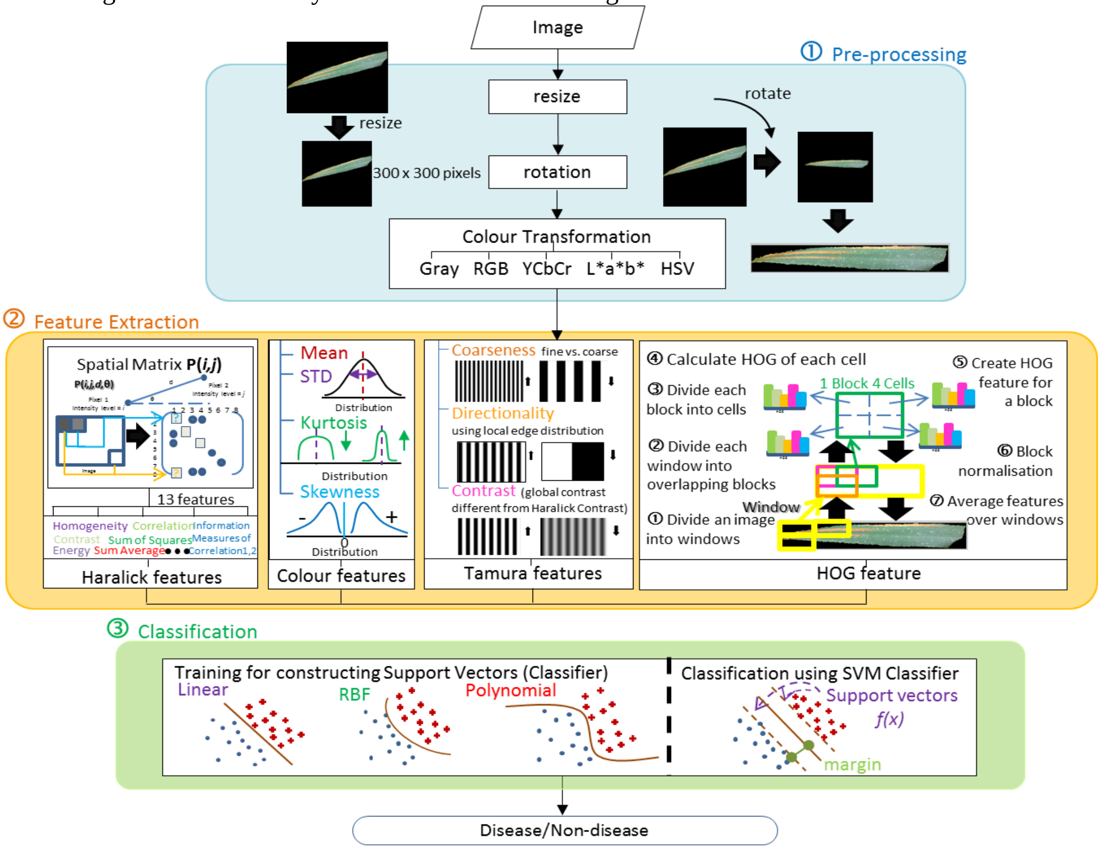


Image
① Pre-processing
resize
300 x 300 pixels
resize
rotation
rotate
Colour Transformation
Gray RGB YCbCr L*a*b* HSV
② Feature Extraction
Spatial Matrix P(i,j)
P(i,j,d,θ)
d
θ
Pixel 1 Intensity level = i
Pixel 2 Intensity level = j
Image
1 2 3 4 5 6 7 8
13 features
Homogeneity Correlation Information Measures of Correlation1,2
Contrast Sum of Squares
Energy Sum Average
Haralick features
Mean
STD
Distribution
Kurtosis
Distribution
Skewness
-
+
0
Distribution
Colour features
Coarseness fine vs. coarse
Directionality using local edge distribution
Contrast (global contrast different from Haralick Contrast)
Tamura features
④ Calculate HOG of each cell
③ Divide each block into cells
1 Block 4 Cells
⑤ Create HOG feature for a block
② Divide each window into overlapping blocks
⑥ Block normalisation
Window
① Divide an image into windows
⑦ Average features over windows
HOG feature
③ Classification
Training for constructing Support Vectors (Classifier)
Linear
RBF
Polynomial
Classification using SVM Classifier
Support vectors
f(x)
margin
Disease/Non-disease

Figure 1 Proposed Classification System

3.1 Pre-processing

Images were obtained from the UK Food and Environment Research Agency [23] by the University of Bristol and the leaves in these images are previously segmented manually from the clutter in order to constrain the background conditions. Because of the diversity of captured images in terms of distance and orientation, pre-processing (see block ① in Figure 1) is then applied to standardise the images.. Firstly, a resize operation is performed to each image using nearest-neighbour interpolation. To mitigate for the differences in direction of angle of the captured leaves, the orientation angles of the leaves are calculated in terms of the directions of the major axes of the leaf ellipses, and then rotation is automatically applied to align the main leaf horizontally. Finally, the rotated leaf is cropped to remove empty space (background in black pixels). Image resizing and cropping is used to reduce processing time of subsequent feature extraction and the rotation process increases the reliability of the extracted features. For this system, five colour spaces, greyscale, RGB, YCbCr, L*a*b* and HSV, are used to evaluate their effects on classification performance.

3.2 Feature Extraction

After the images have been pre-processed, the patterns in the images are extracted in terms of Haralick features, first-order statistical features, Tamura features and HOG features (see block ② in Figure 1).

3.2.1 Haralick Features

Haralick features are developed through the grey-level co-occurrence matrix which measures the spatial relationships in image intensity [24]. This relationship is in terms of a matrix of relative frequencies *P(i,j,d,θ)* between two neighbouring pixels, one with intensity level *i* and another with intensity level *j*, and separated by distance *d* at directional angle *θ*. Thirteen features based on a normalised matrix *p(i,j)* are calculated. Firstly, Angular Second Moment or Energy (EN) is a measure of uniformity of an image:

$$\mathrm{EN} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \{\mathrm{p(i,j)}\}^2 \tag{1}$$

Contrast (CON) measures the variations or spatial frequency of intensity levels for a greyscale image at the reference positions and their neighbours:

$$\mathrm{CON} = \sum_{\substack{\mathrm{n=0} \\ |\mathrm{i-j}|=\mathrm{n}}}^{\mathrm{N_g-1}} \mathrm{n}^2 \left\{ \sum_{\mathrm{i=1}}^{\mathrm{N_g}} \sum_{\mathrm{j=1}}^{\mathrm{N_g}} \mathrm{p(i,j)} \right\} \tag{2}$$

where $\mathrm{N_g}$ is the number of quantised intensity levels.

Homogeneity (HOM) is a measure of local homogeneity of an image. HOM will be high when an image is homogeneous because it takes the inhomogeneous area into account less than the homogeneous area by use of a weighting factor:

$$\mathrm{HOM} = \sum_{\mathrm{i}} \sum_{\mathrm{j}} \frac{1}{1+(\mathrm{i-j})^2} \mathrm{p(i,j)} \tag{3}$$

Other features include Correlation, Sum of Squares, Sum Average, Sum Variance, Entropy, Difference Variance, Difference Entropy, and First and Second Information Measures of Correlation.

3.2.2 Tamura Features

As some features do not correspond to well-explained image patterns, Tamura features, which are based on visual perception, are also introduced [18]. These features include coarseness, directionality, contrast, line-likeness, regularity, and roughness. The first three features are selected in our classification system and the latter three features are omitted as they are combinations of the first three features. Contrast measures the polarization of distribution of black and white in an intensity image. Contrast by Tamura is a global property based on the skewness value of the image which differs from local contrast by Haralick:

$$\mathrm{CON\_T} = \sigma/(\alpha_4)^{1/4}\text{: } \textit{where } \alpha_4 = \frac{\mu_4}{\sigma^4} \tag{4}$$

where $\alpha_4 = \frac{\mu_4}{\sigma^4}$, μ_4 is Kurtosis value.

Coarseness (COAR) measures the block size in an image that is frequently repeated. The element size is large when that image is coarse, whereas the image has fine textures when the element size is small. Directionality (DIR) measures the frequency of transition from one colour value to another. An image with high pattern frequency tends to have a higher degree of directionality. Images with the same pattern but different direction have the same degree of directionality.

3.2.3 First-order Statistical Features (Colour Feature)

These features measure basic statistical information in an image. Mean (MEAN, μ) and standard deviation (STD, σ) determine a global average and variation of an intensity image, respectively. Skewness (SKEW) measures asymmetry of the intensity probability distribution (*p(i)*). Skewness can have negative or positive values:

$$\mathrm{SKEW} = \sigma^{-3} \sum_{i=1}^{N} (i-\mu)^3 p(i) \tag{5}$$

Kurtosis (KUR) measures the shape of the probability distribution relative to the standard normal distribution, and the value is based on the fourth moment of the data. A higher kurtosis value implies a broader and taller distribution shape, whereas a lower kurtosis indicates a narrower and shorter distribution shape:

$$\mathrm{KUR} = \sigma^{-4} \sum_{i=1}^{N} (i-\mu)^4 p(i) \tag{6}$$

3.2.4 Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients measures the local edge distribution of an image [17]. HOG features are calculated by the following processes. Firstly, an image is divided into a number of windows, and each window contains overlapping blocks. Each block comprises non-overlapping cells for which histograms of gradient orientation are computed. To cope with a variety of local contrast, normalization is applied to the HOG vector for each block separately. All vector components from each block are combined to create the HOG descriptors. The final HOG descriptor is an average of all vector components from sliding windows in the image.

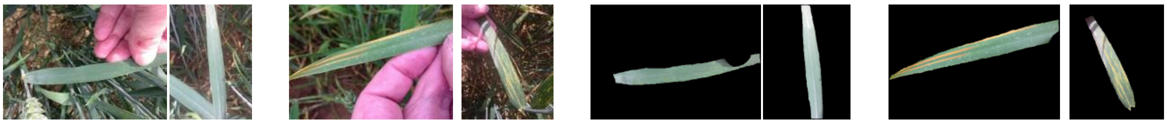
3.4 Classification

Support Vector Machine (SVM) has been shown to obtain high classification performance and good generalization from various studies. It is a supervised learning method which is based on the basic concept of searching for an optimum hyperplane to separate training data into two classes with maximum margin (see block ③ in Figure 1) [22].

In some classification problems, a simple linear hyperplane cannot be applied to divide groups of data efficiently, especially when handling data that have a number of dimensions. Thus, a non-linear function is represented in the discrimination function instead of a linear function. Assuming the weight vector is a linear combination of training data, the function is expressed in terms of a dot product of kernels. This kernel function, which avoids the explicit mapping of data into the high-dimensional feature space, plays a vital role in improving performance. The widely used types of kernel functions include Radial Basis Function (RBF), Polynomial, and Multilayer perceptron [22].,To determine the hyperplane function which divides particular datasets properly is crucial. The last step of the process aims to discover the best fitting model from three kernel characteristics, namely linear, polynomial, or RBF.

4 Experimental Results

The system proposed in Figure 1 is implemented in MATLAB 2012b. Image data were collected by the UK Food and Environment Research Agency [23] (Figure 2 (a)-(b)). The leaves in each image were initially segmented out from the background manually (Figure 2(c)-(d)). Firstly, the system was tested for binary classification (i.e., yellow rust disease or non-disease) using 5-fold cross-validation on 50 non-diseased and 50 Yellow Rust diseased leaves. For pre-processing, the images are resized into 300x300 pixels, then rotated to arrange the leaves horizontally, and finally cropped to remove major areas of background (now represented as black pixels). RGB images are converted to greyscale, HSV, YCbCr, or L*a*b* models, from which each model is used individually or a combination of different colour models is used.



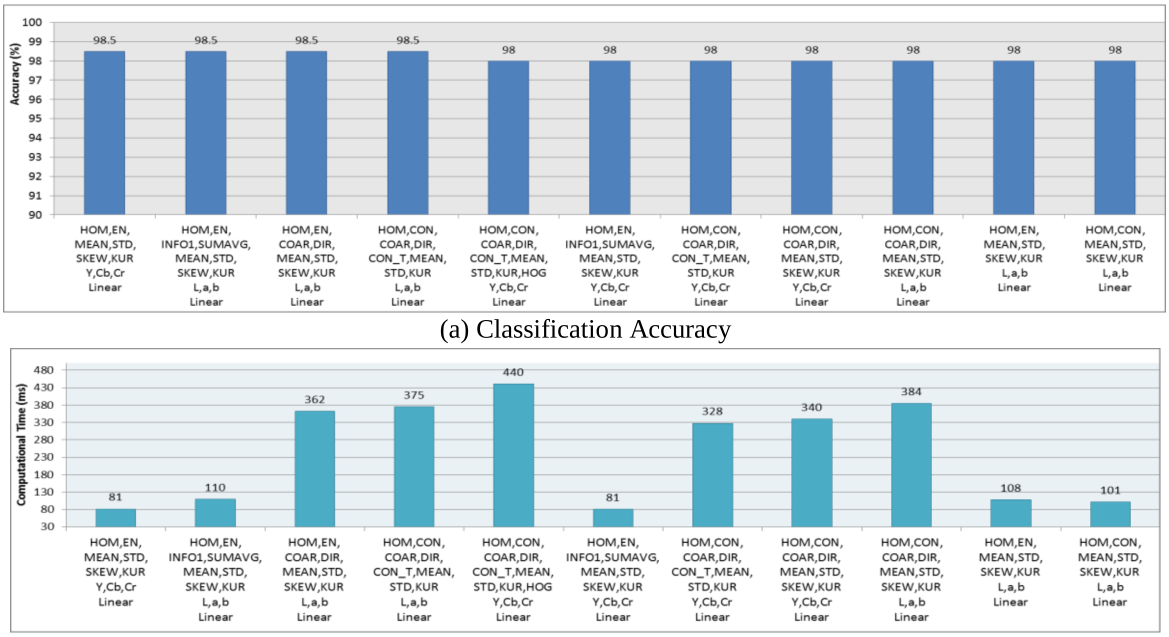
(a) Non-disease (b) Yellow Rust disease (c) Segmented non-disease (d) Segmented Yellow Rust

Figure 2 Wheat diseased and non-diseased images in the experiment

Thirteen Haralick features are created using relative information for pixels in an image with their consecutive neighbouring pixels at a zero directional angle. Also, the intensity levels of an image are quantised into 8 levels, creating an 8x8 spatial matrix for each colour component. Histogram of oriented gradients is calculated using average HOGs from all sliding windows in an image. Optimal window size is empirically set to 32x64 pixels with 50% overlap between blocks. Each block contains 2x2 cells and each cell size is 8x8 pixels. For SVM classifier, the scaling factor of the RBF kernel and the polynomial degree is set empirically to 1 and 2, respectively.

Initially, the combinations of four popular main Haralick features (homogeneity, energy, contrast and correlation) were investigated. It was found that the combinations of homogeneity and energy/contrast produced the highest binary classification accuracy, so these combinations were carried into the next phase of experiments. For the other nine Haralick features, different entropy, sum variance and first information measures of correlation have potential to improve classification accuracy; whereas combination of all four colour features and all three Tamura features have dominant on the classification accuracy. Hence, all combinations of four colour features and three Tamura features were carried into the next phase of combinations among feature sets for bi-class classification and multi-class classification.

Performance of the system for binary classification using a mixture of different features is shown in Figure 3. From the results, YCbCr and L*a*b* colour spaces have a significant influence on classification performance. Similarly, the SVM linear kernel is the best classifier to describe feature distribution for this dataset compared to RBF or polynomial kernels. Homogeneity of Haralick features and colour features (mean, skewness and kurtosis) are effective features to classify disease or non-disease leaf images as it is seen that these features always present in the best eleven accuracy results up to 98.5% in Figure 3(a).

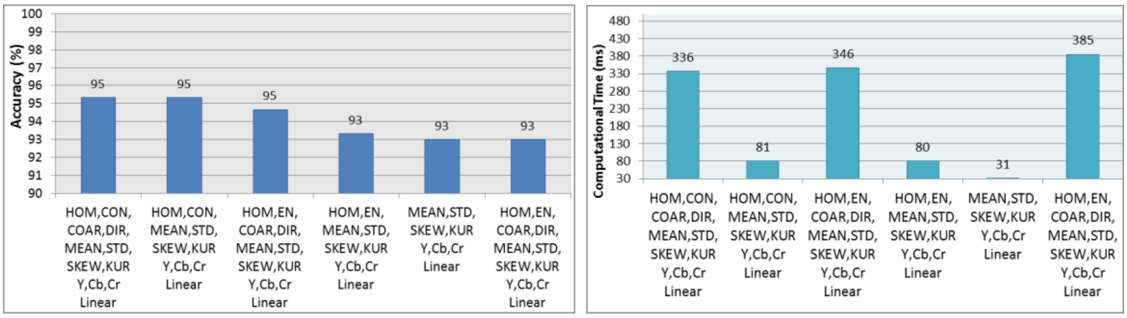


(a) Classification Accuracy

(b) Computational time

Figure 3 Top eleven accuracies for wheat binary classification (disease/non-disease)

Computational complexity for each classification is shown in Figure 3(b). It is simple to calculate Haralick and colour features, so the combinations containing only those features show low computational times as shown in the first bar in the chart in Figure 3(b); the classification time is less than 100 milliseconds. The computational times of Tamura and HOG features are relatively higher than the other two feature sets as they require many steps to calculate each feature; the fifth bar in the chart shows that with these additional two feature sets the computational time increases by about 300 milliseconds (from 81 msec to 440 msec).



(a) Classification Accuracy

(b) Computational time

Figure 4 Top six accuracies for wheat non-disease and disease

These feature sets were also investigated for multi-class classification (non-disease, Yellow Rust and Septoria diseases) as shown in Figure 4. Similar to binary classification, the best six mixtures of features, including two Haralick features (homogeneity and contrast or energy), two Tamura features (coarseness and directionality) and colour features, give the highest accuracy of up to 95%. When we combine HOG with other features we found that there is little improvement in accuracies compared to the use of HOG alone, and classification accuracies are less than for the feature combinations without HOG. The processing time for computing only the colour feature set is the least at approximately 30 milliseconds, and the classification accuracy shows a high value at 93% (fifth bar in the chart in Figure 4(a)-(b)).

5 Conclusion

Since using too many features may cause over-fitting and also require more computational time, the classification system discussed in this paper is proposed to assess effectiveness of features from different feature sets, such as Haralick features, first-order statistical features, Tamura features and HOG features. The results show that the most efficient features include, but are not limited to, two Haralick features (mixture of homogeneity and contrast or energy), three Tamura features and colour features. Use of only the HOG features is also a potential approach for classification, but the combination of features mentioned above performs better in both binary and multi-class classification. Also, processing times in calculating Haralick and colour features are less than for Tamura and HOG features. Following the experiments, it is planned that the effective sets of features be implemented in lightweight systems, such as a mobile application for real plant disease classification.

Acknowledgement:

This work was supported by a University of Ulster Vice Chancellor's Research Scholarship. We are grateful to EPSRC-DST funded India-UK Advanced Technology Centre (IU-ATC) project for providing some crop disease databases.

References:

[1] A. McAndrew, *Introduction to Digital Image Processing with MATLAB*. Boston: Thomson Course Technology, 2004.

[2] R. C. Gonzalez, R. E. Wood, and Steven L. Eddins, *Digital Image Processing using MATLAB*. New Jersey: Pearson Education, Inc., 2004.

[3] S. Kai, L. Zhikun, S. Hang, and G. Chunhong, "A Research of maize disease image recognition of Corn Based on BP Networks," in *2011 Third Int. Conf. Measuing Technol. Mechatronics Autom.*, 2011, pp. 246–249.

[4] H. Al Hiary, S. Bani Ahmad, M. Reyalat, M. Braik, and Z. ALRahamneh, "Fast and Accurate Detection and Classification of Plant Diseases," *Int. J. Comput. Appl.*, vol. 17, no. 1, pp. 31–38, 2011.

[5] S. Bashir and N. Sharma, "Remote Area Plant Disease Detection Using Image Processing," *IOSR J. Electron. Commun. Eng.*, vol. 2, no. 6, pp. 31–34, 2012.

[6] D. Al Bashish, M. Braik, and S. Bani-Ahmad, "A framework for detection and classification of plant leaf and stem diseases," in *Signal Image Process. ICSIP 2010 Int. Conf. on*, 2010, pp. 113–118.

[7] S. R. Dubey and A. S. Jalal, "Detection and Classification of Apple Fruit Diseases Using Complete Local Binary Patterns," in *2012 Third Int. Conf. Comput. Commun. Technol.*, 2012, pp. 346–351.

[8] H. Wang, G. Li, Z. Ma, and X. Li, "Image recognition of plant diseases based on backpropagation networks," in *Image Signal Process. (CISP), 2012 5th Int. Congr.*, 2012, pp. 894–900.

[9] M. El-Helly, R. Ahmed, and S. El-Gammal, "An Integrated Image Processing System for Leaf Disease Detection and Diagnosis," in *Proc. 1st Indian Int. Conf. Artif. Intell. IICAI 2003*, 2003, pp. 1182–1195.

[10] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu, and B. Yang, "Application of Support Vector Machine for Detecting Rice Diseases Using Shape and Color Texture Features," *2009 Int. Conf. Eng. Comput.*, pp. 79–83, 2009.

[11] J. Pang, Z. Bai, J. Lai, and S. Li, "Automatic segmentation of crop leaf spot disease images by integrating local threshold and seeded region growing," *2011 Int. Conf. Image Anal. Signal Process.*, pp. 590–594, Oct. 2011.

[12] S. Ananthi and S. V. Varthini, "Detection and Classification of Plant Leaf Diseases," *Int. J. Res. Eng. Appl. Sci.*, vol. 2, no. 2, pp. 763–773, 2012.

[13] S. B. Dhaygude and N. P. Kumbhar, "Agricultural plant Leaf Disease Detection Using Image Processing," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 2, no. 1, pp. 599–602, 2013.

[14] D. G. Kim, T. F. Burks, J. Qin, and D. M. Bulanon, "Classification of grapefruit peel diseases using color texture feature analysis," *Int. J. Agric. Biol. Eng.*, vol. 2, no. 3, pp. 41–50, 2009.

[15] R. Pydipati, T. F. Burks, and W. S. Lee, "Identification of citrus disease using color texture features and discriminant analysis," *Comput. Electron. Agric.*, vol. 52, no. 1–2, pp. 49–59, Jun. 2006.

[16] H. Wang, G. Li, Z. Ma, and X. Li, "Image recognition of plant diseases based on principal component analysis and neural networks," *2012 8th Int. Conf. Nat. Comput.*, pp. 246–251, May 2012.

[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 886–893, 2005.

[18] H. Tamura, S. Mori, and T. Yamawaki, "Textural Features Corresponding to Visual Perception," *Syst. Man Cybern. IEEE Trans.*, vol. 8, no. 6, pp. 460–473, 1978.

[19] T. A. Pham, "Optimization of Texture Feature Extraction Algorithm," 2010.

[20] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–45, Sep. 2010.

[21] A. K. Jain, J. Mao, H. Road, and S. Jose, "Artificial Neural Networks : A Tutorial," in *IEEE Comput. Spec. Issue Neural Comput.*, 1996, pp. 1–49.

[22] N. Christianini and J. Shawe-Taylor, *An Introduction to Support Vector Machine*. Cambridge University Press, 2000.

[23] "The Food & Environment Research Agency." [Online]. Available: http://fera.co.uk/.

[24] R. M. Haralick, K. Shanmugam, and DinsteinIts'shak, "Textural Features for Image Classification," *Syst. Man Cybern. IEEE Trans.*, vol. SMC-3, no. 6, pp. 613–621, 1973.

Optical flow for background subtraction in moving transport

Sriram Varadarajan, Paul Miller and Huiyu Zhou
The Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
svaradarajan01@qub.ac.uk, p.miller@qub.ac.uk, h.zhou@ecit.qub.ac.uk

Abstract

In recent years, safety of passengers on moving transport has been of paramount importance and this has led to an increase in the number of CCTV cameras fitted on buses and trains for monitoring purposes. Automated Computer Vision systems are expected to perform background subtraction on the footage from these cameras for various reasons like person detection, tracking etc. However, the scene is quite unique especially because of the dynamic background present in the bus window region. Normal background subtraction techniques cannot differentiate between this dynamic background and the foreground inside the bus. As the scene outside the bus is dependent on the motion of the bus, we investigate the use of optical flow to counteract this problem. We provide the results based on our initial experiments and show that this is a promising approach for this type of scenario while suggesting improvements to our initial approach.

Keywords: Optical flow, Mixture of Gaussians, Background Subtraction, Video Surveillance

1 Introduction

Classifying a scene into foreground or background is an essential component in many Computer Vision systems. Foreground is the part of the scene that one may be interested in segmenting for further analysis and background forms the rest of the scene. In simple scenarios, the background is static and does not change while the foreground usually changes with time. But real time scenes have a more complex background because of various factors like motion in the background, shadows or reflections in the scene, change in illumination etc. All these factors make the segmentation problem quite difficult. In this paper, we look into analysing CCTV footage on moving transport like buses and trains. This scenario is different from general background subtraction problems because the background is a combination of fixed and moving elements. The scene in the window region of the bus changes dynamically based on the environment outside the bus. This is hard to predict especially when the bus or train is in motion. Most of the above mentioned complexities like reflections, changes in illumination also occur in this region.

This paper extracts the motion information of the bus by using optical flow and combines it with the general background subtraction output to differentiate between the dynamic background and the foreground region. This differentiation is the key problem in this type of scenario and normal background subtraction methods that do not use any motion information fails to do this. The unique difference between the foreground and the dynamic background is that the dynamic background region is highly influenced by the motion of the bus whereas the foreground is not. Apart from this single factor, both the regions vary in a similar manner in terms of intensity, colour etc. Thus, the motion information is used to separate the foreground and the dynamic background from the background subtraction output. A brief overview of the background subtraction technique and the optical flow method are provided in the next two sections followed by the combination of the two with some examples from the experimental results. Finally, the possible hurdles with this approach are discussed with suitable suggestions and ideas provided for these problems.

2 Background subtraction

Classic background subtraction approaches work by modelling the scene based on the pixel distributions. There has been a lot of research done in this field to model backgrounds ranging from simple static ones to complex dynamic cases. Probably the most widely used method in this area is the Mixture

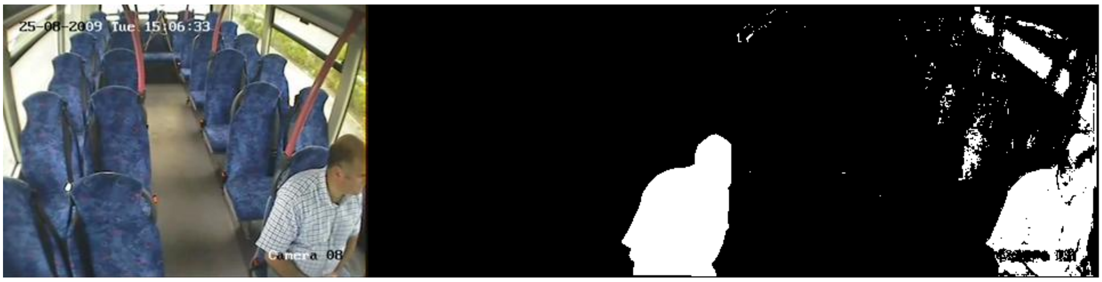


Figure 1: Typical CCTV image, the corresponding ground truth and a typical background subtraction output

of Gaussians (MoG) approach by Stauffer and Grimson [8] where the background modelling is done as a multi modal combination of Gaussian distributions. Non-parametric modelling has also been used in the past [3] to model pixels without the assumption of an inherent probability distribution. Wallflower [9] is a background maintenance algorithm that combines three different components (pixel-level, region-level and frame-level) to subtract even dynamic backgrounds from the scene. Their system works well for most of the dynamic background cases like illumination changes, shadows etc. There have been various other MoG algorithms [1] developed that are derivatives of the original Stauffer and Grimson algorithm. The problem with all these background subtraction algorithms is that in this particular scenario of a moving transport, they are not very effective as the dynamic background has very similar properties as the foreground. Pixel modelling alone will not be able to differentiate between the two regions.

Mixture modelling is done based on the assumption that the background pixels are constant over a large period of time and therefore can be modelled to have larger weights and low variances whereas the foreground pixels are constantly changing thus modelling them with smaller weights and high variances. Any pixel location can have a combination of these background and foreground pixels and classification is done by building a background model and checking whether a pixel falls within this model or not. The probability of observing a pixel value is given by

$$p(x^t) = \sum_{k=1}^{K} \omega_k^t * \mathcal{N}(x^t|\mu_k^t, \Sigma_k^t) \tag{1}$$

where x^t is the pixel value at time instant t, K is the number of mixtures, $\omega_k, \mu_k, \Sigma_k$ are the weights, means and variances of the different mixtures and $\mathcal{N}$ denotes a Gaussian distribution.

The model is updated over time by updating the parameters of the closest mixture $\{\omega_k^t, \mu_k^t, \Sigma_k^t\}$ that matches the current pixel x^t with the update equations (2)-(4). A match is said to be found only if the mixture is within a certain distance from the pixel. If not, a new mixture is created replacing the mixture with the lowest weight at that particular time instant.

$$\omega_k^t = (1-\alpha) * \omega_k^{t-1} + \alpha \tag{2}$$

$$\mu_k^t = (1-\rho) * \mu_k^{t-1} + \rho * x^t \tag{3}$$

$$\Sigma_k^t = (1-\rho) * \Sigma_k^{t-1} + \rho * (x^t - \mu_k^t)^2 \tag{4}$$

Figure 1 shows the major problem when using background subtraction algorithms in this environment. The person is the foreground and the scene outside the window is part of the background in this scene. However, as the bus is in motion, the background part changes unpredictably and this results in the output of the background subtraction algorithm including this part of the scene as well.

Numerous other pixel based background subtraction algorithms are available in literature, some more sophisticated than others, but the general theme here is to model the pixel distribution based on their intensity or colour. The problem with all these algorithms is that they will view the window region as another foreground as it has characteristics very similar to any typical foreground dealt with by these algorithms.

3 Flow Information

The idea of using optical flow to aid in the segmentation is based on the knowledge that as the vehicle moves, the scene in the window region appears to move in the opposite direction of the motion of the

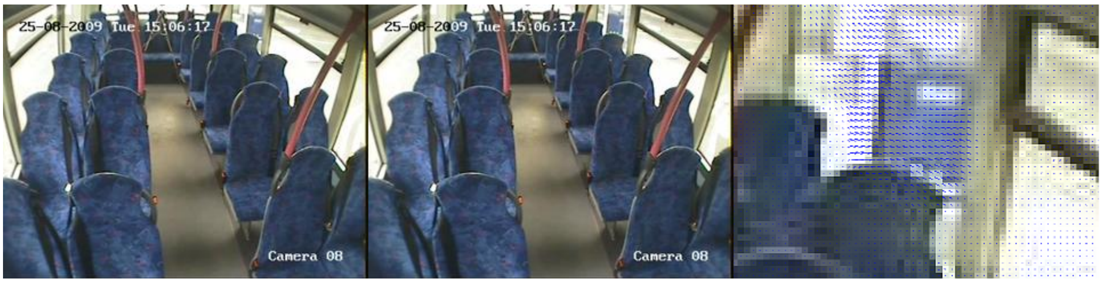


Figure 2: Two consecutive frames from an empty bus scene and the corresponding flow field for the bus window region

vehicle. The problem with traditional background subtraction methods is that they fail to differentiate between the actual foreground (people or objects inside the bus) and the dynamic background (scene outside the bus). However, with the motion information available, it can be used to separate the window region from the region of interest inside the bus.

A well-known global method is the Horn-Schunck approach [4] that assumes that the optical flow is smooth over the entire image. Lucas-Kanade method [7] is a widely used local approach that minimises the square difference between two images over a region. Global methods provide a dense field of flow vectors whereas local methods are less sensitive to noise in the image. Some researchers have tried to get the best of both worlds by combining the local and the global methods [2]. In this work, we use a multiscale coarse to fine approach based on [2] to compute a dense flow field [6].

Previously, optical flow has been combined with background subtraction in [10] to extract moving objects in noisy environment and in [5] for detection motion in large crowds. In this paper, the background subtraction output obtained from the online Mixture of Gaussians is combined with the optical flow method in the following manner. The flow vectors are used to separate the scene into various bins based on the angle at each pixel. From these bins, the ones corresponding to the window region are chosen to generate a magnitude map. This map is then used to subtract the window region from the output of the background subtraction method with the help of a logical operator and produce the final output containing only the foreground.

4 Experiments and Results

Our experiments were performed on our dataset created by using a CCTV camera inside a bus. In the video, the bus is initially empty and it starts moving after a while following which a person enters the scene and takes a seat in front of the bus. Fig. 2 shows two consecutive frames from a typical scene from the CCTV camera inside an empty bus that is in motion and the corresponding flow vectors after applying optical flow on these two frames. Taking a closer look in Fig. 2, it shows the direction is predominantly in a direction opposite to that of the direction of the bus. It should also be noted that the motion vectors are only well-defined in areas that are not uniform throughout. This could be due to the presence of an object or a structure in the dynamic background region. This is a common issue while using optical flow as it is known to not work properly in general in a uniform region devoid of any texture. But on the other hand, if the region is uniform, it can be compared to a static background without much change in the colour information thus not posing much of a problem in subtracting it.

In Fig. 3, a frame is shown from the video with a person present in the scene along with its magnitude heat map. This heat map is generated by normalising the magnitude of motion vectors and using a grayscale colour map to display the different levels of magnitude, with black corresponding to the lowest values and white corresponding to the highest values. This figure shows that the dynamic background region has higher magnitudes of motion when compared to the foreground region which in turn has a higher magnitude compared to the static background region inside the bus. The magnitude in the foreground region arises from the slight swaying of the person while sitting inside the moving bus. This type of map is the most commonly encountered scenario in the moving transport however, there can be times when the foreground region can have higher magnitudes comparable to the dynamic background region, for example, when the person walks into or out of the scene.

Now, once the motion vectors are obtained, the pixels are clustered into bins based on their angles. This is done in order to separate the image into similar regions based on the direction of motion. We used 12 bins to divide the angles (360°). This can be looked upon as an angle based histogram. However, this might result in some stray pixels from the inside of the bus to be clustered along with the window

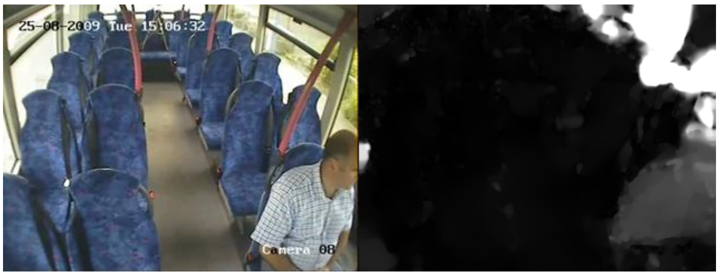


Figure 3: Typical scene with person and the corresponding magnitude heat map

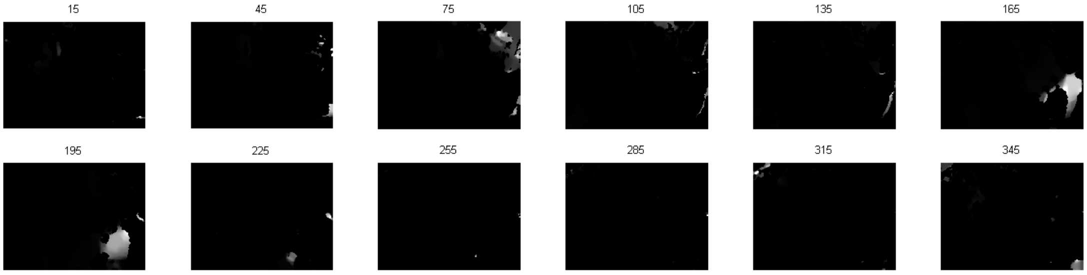


Figure 4: The magnitude heat maps clustered into different bins based on angles. Notice how the window region falls predominantly under different bins than the ones that have the person in them

region. In order to separate these regions, the magnitude of motion is used in unison with the angles. The magnitude of motion is much larger in the window region when compared to the other regions of the bus. Therefore, when the appropriate angles are chosen, they can be multiplied based on their magnitude to differentiate between regions.

Fig. 4 shows an example frame where the entire frame is split into 12 subplots based on the 12 bins mentioned above. This means the image under bin 45 shows all pixels having vectors with angles between 30° and 59°. The magnitude of these vectors have been normalised over grayscale as above. Note that this number 12 is an arbitrary choice and was chosen so that the number of bins are not too many and not too few. In this figure, the window region falls entirely under the bins with central angles 45, 75, 105. The foreground is mostly included under the bins with central angles 165 and 195. It was noted from our experiments that the dynamic background mostly fell under the angles 60 and 150 while the foreground did not stay under specific bins continuously from frame to frame. This is because the person usually sways in his seat thus causing the dominant angles of the motion vectors to change randomly whereas the motion of the scene outside the bus mostly moved at an angle opposite to the motion of the bus. This reiterates our idea at the beginning of the paper. In Fig. 5, the window region is isolated by selecting the appropriate bins based on the central angles mentioned above. It can be seen that these angles are opposite to that of the direction of motion of the bus.

Now, once this region has been isolated, it can be combined with the output of the background subtraction method by a simple logical operation that retains the foreground but removes the dynamic background from the output. Figure 6 shows the three results side by side for various frames from the

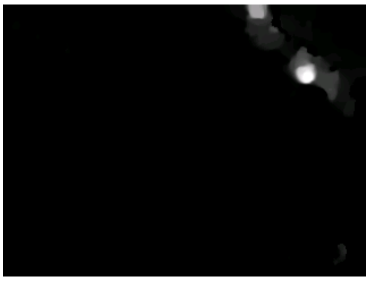


Figure 5: Clustering of the bins containing the angles between 60 and 150. This image clearly shows that the window region has been isolated from the foreground

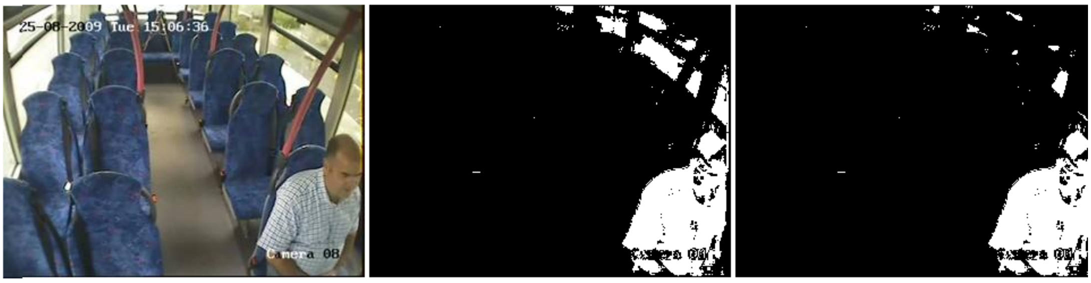


Figure 6: Segmentation result for an example frame from the video. Left: Original Frame; Middle: Output of the Background Subtraction method; Right: Final Output combining the Optical flow and Background Subtraction outputs

Table 1: Quantitative Analysis

Frame no.	True Positive Rate		False Positive Rate	
	BGSub Output	Final Output	BGSub Output	Final Output
91	NO FG	NO FG	3.84	1.93
104	76.29	76.26	3.99	0.83
119	84.20	82.83	4.05	2.17
134	86.64	84.85	4.08	2.06
147	86.63	84.60	4.20	1.94
163	83.22	83.00	3.59	1.01

video sequence. It can be clearly seen that the dynamic background region is mostly removed from the final output.

Table 1 shows the true positive rates and the false positive rates for a few different frames obtained by comparing the results with our manually labelled groundtruth images. It can be seen that the false positive rate reduces in the final output which corresponds to the subtraction of the dynamic background region. Note that frame number 91 does not have any foreground in it hence the true positive rate cannot be calculated for it. Fig. 6 shows the results for a typical frame (Frame no. 163 from Table 1). The middle image of Fig. 6 is the output of the mixture of Gaussians approach while the right column shows the output after combining with the flow information. It is evident from Fig. 6 that the decrease in false positives is because the pixels in the window region are subtracted quite well with the combined method.

The experiments show that there is a correlation between the motion of the bus and the dynamic background region. This motion information helps remove the background from the segmentation output. However, we noticed few drawbacks in using optical flow for generating motion vectors. The first one is quite obvious in that optical flow is known to be quite sensitive to noise and the quality of the CCTV camera footage is usually not very good and it tends to affect the output of the flow vectors. This in turn affects the final segmentation output as it is dependent on the optical flow output.

It was mentioned previously that the person swaying in his seat results in flow vectors in random directions. There is a possibility that the direction of motion of the person could mirror the direction of motion of the scene in the window region. This results in bad clustering of the foreground and the dynamic background into same bins. One such example is shown in Fig. 7 where part of the foreground region is subtracted along with the background. This happened in a few frames in our experiments and the consequence is that the foreground also gets subtracted along with the background while performing the logical operation. The method of choosing the bins to isolate the dynamic background region is done in a heuristic manner and although it works fine in most cases, it is not optimal but additional information like the speed of the bus could be used to aid in the clustering process. This is an area we will be looking into, in the future. We will also be looking at the combination of the optical flow based motion features and the pixel colour features in the background modelling process to produce a better statistical model of the scene.

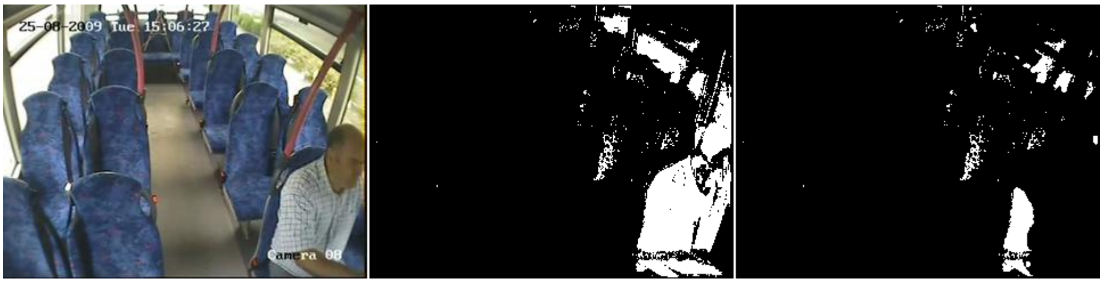


Figure 7: Segmentation output for an example frame when the angle of the foreground region matches the dominant angles of the bus window region. Left: Original Frame; Middle: Output of the Background Subtraction method; Right: Final Output combining the Optical flow and Background Subtraction outputs

5 Conclusion

This paper is an initial investigation into this unique topic of background subtraction in moving transport. We showed why probability based background subtraction methods fail to perform well in this scenario. We looked into the use of optical flow to obtain motion information which would aid in better subtraction of the dynamic background region. The initial results look promising and future research directions include more complex cases like multiple people in the foreground and combining the colour and flow features in the background modelling step.

References

[1] T. Bouwmans, F. El Baf, and B. Vachon. Background modeling using mixture of gaussians for foreground detection: A survey. *Recent Patents on Computer Science*, 1(3):219–237, November 2008.

[2] A. Bruhn, J. Weickert, and C. Schnrr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61:211–231, 2005.

[3] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. In *PROCEEDINGS OF THE IEEE*, pages 1151–1163, 2002.

[4] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[5] W. Li, X. Wu, K. Matsumoto, and H.-A. Zhao. Foreground detection based on optical flow and background subtract. In *Communications, Circuits and Systems (ICCCAS), 2010 International Conference on*, pages 359–362, 2010.

[6] C. Liu, W. Freeman, E. Adelson, and Y. Weiss. Human-assisted motion annotation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[7] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Seventh International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[8] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2246–2252, 1999.

[9] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Seventh International Conference on Computer Vision*, pages 255–261, 1999.

[10] D. Zhou and H. Zhang. Modified gmm background modeling and optical flow for detection of moving objects. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 2224–2229, 2005.

Vignetting Modeling and Correction for a Microlens-based Light Field Camera

Shan Xu
School of Physics
National University of Ireland, Galway
s.xu1@nuigalway.ie

Nicholas Devaney
School of Physics
National University of Ireland, Galway
nicholas,devaney@nuigalway.ie

Abstract

Microlens-based light field cameras, which are capable of recording both angular and spatial information of light, are already commercially available as consumer commodities. Intrinsically, the large f-number and the use of a microlens array introduce a more severe vignetting effect than a conventional camera. Proper devignetting is required to reconstruct a high quality light field from the captured 2D raw image. In this paper, a 2D Gaussian kernel is proposed to model the microlens image vignetting and local parameters are estimated by solving a nonlinear optimization problem. We assume this kernel is smoothly varying across the whole image. The global parameters are estimated by polynomial fitting. Our results show that it accurately predicts the vignetting effect. We also demonstrate successful vignetting correction based on our modeling prediction.

Keywords: Light Field Imaging, Vignetting, Gaussian Kernel, Non-linear Optimization

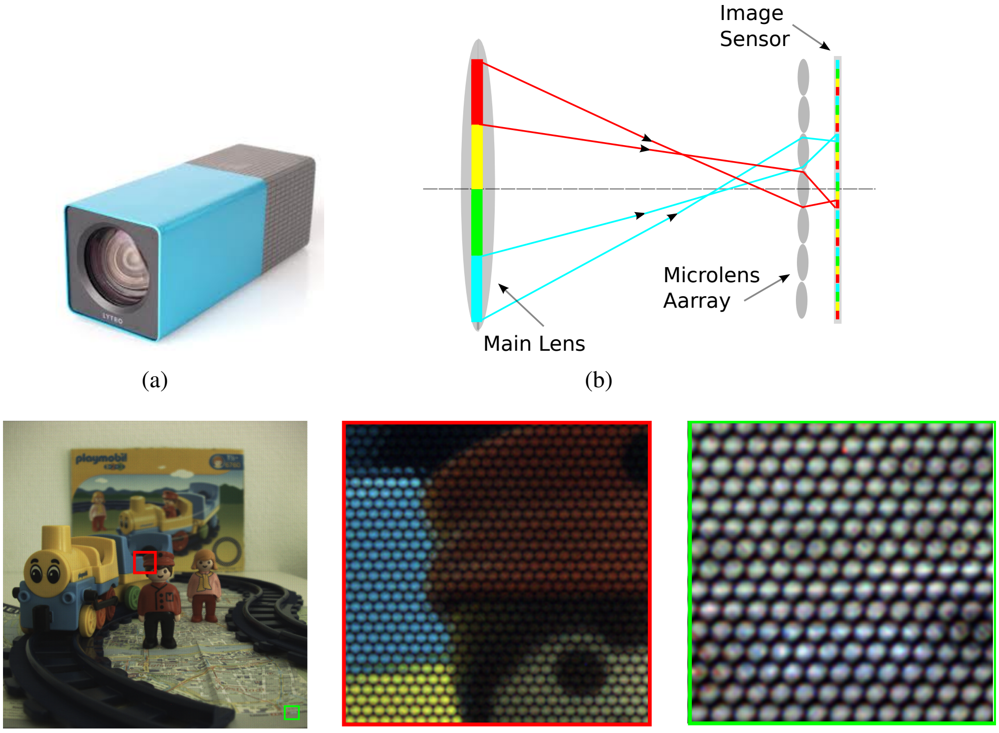
1 Introduction

In contrast to conventional cameras which capture light intensity, light field cameras capture the radiance of the light. A single pixel of a 2D image sensor collects photons coming from all directions and converts them to electrons. To avoid the loss of the directional information, an array of microlenses is inserted between the main lens and the image sensor [Ng et al., 2005]. With this optical setup, the light rays inside the camera body can be parameterized by 4D coordinates [Levoy and Hanrahan, 1996]. Knowing the travel path of light rays from objects to the sensor enables novel applications such as refocusing, changing perspective and depth estimation. However, in such a spatially multiplexing device, the price to pay is the significant loss of spatial resolution.

Vignetting refers to the gradual fall-off of light intensity from the center of the image captured by an imaging system [Smith, 2007]. It is a common imaging phenomenon and is usually corrected in the camera processing pipeline along with other processing steps such as demosaicing, distortion correction and noise reduction. The sources of vignetting can be classified into natural vignetting, optical vignetting, mechanical vignetting and pixel vignetting. Short focal lengths, large apertures and large format sensors introduce more severe vignetting effects in imaging systems.

A microlens-based light field camera exhibits severe vignetting effects due to two factors. First, the pixels underneath each microlens can be seen as a large format sensor. As a result of natural vignetting, also known as the cosine-fourth law, the edge pixels receive much less light than the central ones. Second, a microlens-based light field camera has a constant, large

f-number[1], usually as large as $f/2$. For such an optical configuration in a conventional camera, vignetting effects are always significant.



(a)

(b)

(c)

Figure 1: (a) The Lytro camera. (b) The internal structure of a microlens based light field camera. The pixels underneath the microlens record the directional information of the light rays. (c) Detail of a raw image captured by a Lytro camera. The closeup view (green window) shows the distorted microlens images as evidenced by the non-circular shapes caused by optical vignetting.

2 Related Work

Vignetting correction is a well-explored topic in optics and computer vision. It is also known as flat-field correction (FFC) in astronomy. The most simple and efficient approach to FFC can be written [Lindfors, 1998],

$$I'(x,y) = \frac{I(x,y) - I_B(x,y)}{I_U(x,y) - I_B(x,y)} M \tag{1}$$

where I_B is a dark frame, I_U is the uniform illumination image, I is the uncalibrated image, I' is the flat field corrected image and M is a normalizing constant.

Evidently, the above approach is simple and robust for any optical imaging system. However, it requires a large amount of memory to store the reference images with different optical configurations, i.e. focal distance, exposure time, etc. Accurate prediction of light intensity fall-off and modeling the vignetting effect can significantly reduces the memory requirement . A quadratic polynomial function [Sawchuk, 1977] and a hyperbolic function [Yu et al., 2004] have been proposed to approximate the light fall-off profile across the image. In [K. Sooknanan, 2012], a multi-frame approach was proposed for removing vignetting of underwater video sequences.

According to our knowledge, our work is the first attempt to model the vignetting effect for a light field camera. In [Dansereau et al., 2013], a light field camera decoding and calibration

[1] The f number or focal ratio is defined as the focal length divided by the aperture diameter

pipeline is proposed, but the vignetting is corrected by a traditional FFC approach. In contrast, our work demonstrates modeling and removing vignetting with global and local parametric fitting.

3 Microlens-based Light Field Camera Image Formation

As illustrated in Fig.1(b), in order to avoid the loss of angular information, a microlens array is inserted at the image plane of the main lens [Ng et al., 2005].[2] With this optical setup, each microlens image is an image of the exit pupil, viewing at different angles on the image plane. Ideally, the microlens image has a circular shape. In practice, due to light rays being blocked by the apertures of the main lens elements, we observed severe distortions for the off-axis microlens images as shown in Fig.1(c). This effect is also discussed in [Aggarwal et al., 2001].

4 Vignetting Modeling

4.1 Local Parametric Kernel Estimation

In [Kee et al., 2011], a spatially smooth 2D kernel is proposed for estimating the image blur kernel to restore images. Inspired by their work, we choose a two-dimensional Gaussian function $G(\mathbf{\Sigma}, a)$ to fit the light intensity profile of each individual microlens image. The orientation is controlled by single one parameter ρ. All the parameters can be obtained by minimizing the energy function,

$$\{\hat{\mathbf{\Sigma}}, \hat{a}\} = \arg\min_{\mathbf{\Sigma},a} \sum_{m=0}^{M} \sum_{n=0}^{N} ||I_{m,n} - G(\mathbf{\Sigma}, a)||^2 \tag{2}$$

where the 2D Gaussian function is expressed as,

$$G(\mathbf{\Sigma}, a) = \frac{exp(-\boldsymbol{x}^T \mathbf{\Sigma}^{-1} \boldsymbol{x})}{a|\mathbf{\Sigma}|^{1/2}} \tag{3}$$

The covariance matrix $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{4}$$

$\boldsymbol{x}$ is the spatial coordinate vector in the microlens image. The Levenberg-Marquardt algorithm [Levenber, 1944] is used to solve this non-linear optimization problem. Fig.2 shows the fitting performance for different microlens images at different locations. Our model ensures the errors are consistently at an acceptable level.

4.2 Global Parametric Fitting

The orientation of each microlens image varies smoothly in spatial coordinates. Polynomial functions are employed to model the global variations of the local parameters a, $\sigma 1$, $\sigma 2$ and ρ. To balance the approximation accuracy and computational complexity, we use a fourth order polynomial. Table 1 shows the impact of the polynomial order on the final approximation result. Fig.3 shows a real example of the global fitting result for the local parameters. We discard the 10% estimated local parameters from the poorest fit microlens images to obtain global parameters a, $\sigma 1$, $\sigma 2$ and ρ.

[2] In an alternative optical setup, the so-call focused plenoptic camera [Lumsdaine and Georgiev, 2008], a microlens array is positioned at the image plane of the main lens.

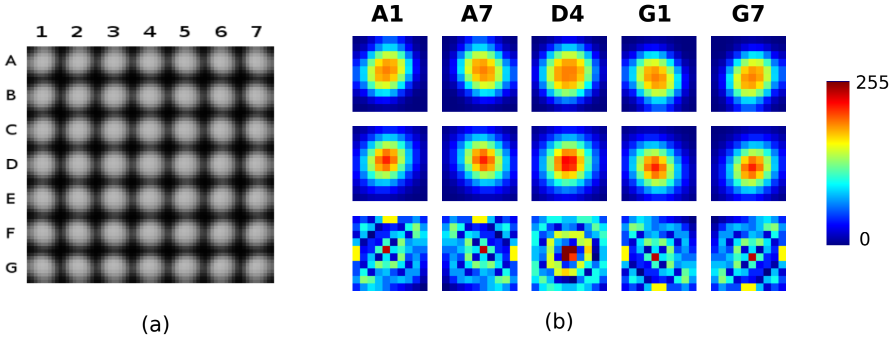


Figure 2: (a) The ground truth microlens array image generated by simulation. (b) Microlens image fitting results at different locations. The first row is the ground truth microlens image, second row is the parametric fitting result, the third row is residual errors. For the purpose of visualization, the errors are magnified by a factor of 10.

	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
a	10.0558	0.0200	0.0189	0.0171	0.0168	0.0160
σ_1	0.1391	0.0684	0.0676	0.0549	0.0543	0.0526
σ_2	0.1583	0.0707	0.0686	0.0520	0.0512	0.0493
ρ	15.9008	21.9774	1.9110	1.4530	1.4195	1.3246

Table 1: The parameter estimation errors from 1st to 6th order polynomial fitting. Based on the performance and complexity, 4th order polynomial fitting is our choice.

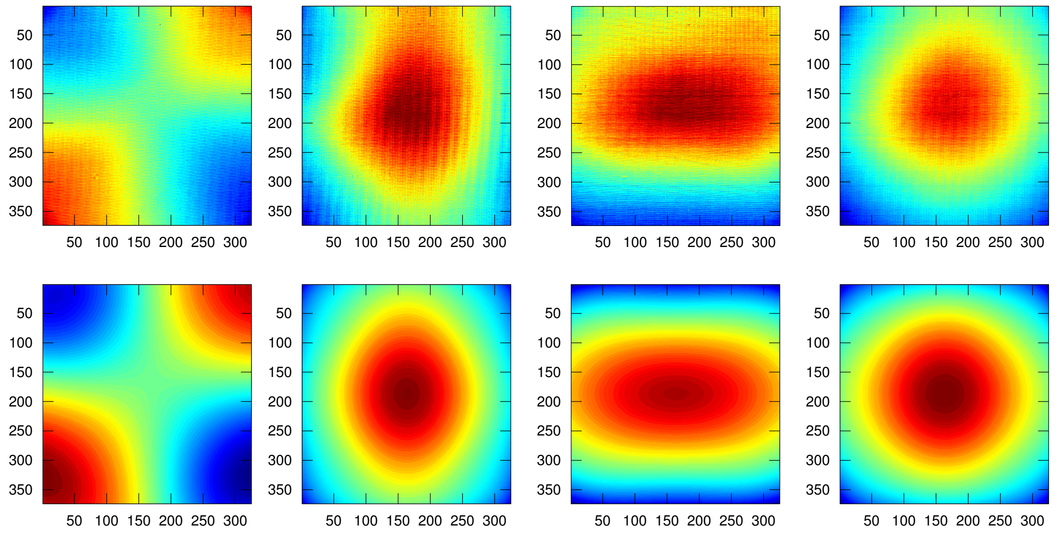


Figure 3: Global fitting result for the four local parameters. Top row are the measured results. Bottom row are the polynomial fitting results. From left to right, the figure corresponds to parameter ρ, σ_1. σ_2 and a respectively.

5 Experimental Result

Our experiment is based on the first commercially available consumer light field camera, the Lytro[3]. It has approximately 360 by 380 microlenses. There are around 10 by 10 pixels under each individual microlens. The resolution of the image sensor is 3,280 x 3,280 pixels. To maintain the uniform illumination, we place a diffuser in front of the Lytro camera. A sequence of uniform illumination images and a dark frame with the same zoom and exposure settings are

[3] www.lytro.com

captured. The average image of the sequence is used for our modeling. In order to verify our vignetting correction result, we reconstruct the multiview array image from the light field raw data. Detailed implementation can be found in [Dansereau et al., 2013] and [Cho et al., 2013]. In Fig.4(b), (c), we show that our approach significantly reduces the intensity variation between views and provides smooth vignetting correction in single view images.

6 Conclusion

In this paper, we present the modeling and correction of vignetting for a microlens-based light field camera. Using our approach, the vignetting effect can be successfully compensated. However, there are some over-compensated regions around the edge of the corrected multiview image array. The vignetting compensation factor is much larger in those regions because light falls dramatically at the edge of the microlens. The errors of the microlens center position estimation leads to the prediction errors. In our future work, we will investigate how to accurately identify the center of each microlens image and how to compose an effective boundary constraint to suppress the over compensation.

References

[Aggarwal et al., 2001] Aggarwal, M., Hua, H., and Ahuja, N. (2001). On cosine-fourth and vignetting effects in real lenses. In *ICCV*, pages 472–479.

[Cho et al., 2013] Cho, D., Lee, M., Kim, S., and Tai, Y.-W. (2013). Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In *ICCV*. IEEE.

[Dansereau et al., 2013] Dansereau, D. G., Pizarro, O., and Williams, S. B. (2013). Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *CVPR*. IEEE.

[K. Sooknanan, 2012] K. Sooknanan, A. Kokaram, D. C. G. B. J. W. N. H. (2012). Improving underwater visibility using vignetting correction. In *Visual Information Processing and Communication,SPIE*.

[Kee et al., 2011] Kee, E., Paris, S., Chen, S., and Wang, J. (2011). Modeling and removing spatially-varying optical blur. In *Computational Photography (ICCP)*, pages 1–8. IEEE.

[Levenber, 1944] Levenber, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quart. J. Appl. Maths*.

[Levoy and Hanrahan, 1996] Levoy, M. and Hanrahan, P. (1996). Light field rendering. In *Computer Graphics Proceedings, Annual Conference Series, 1996 (ACM SIGGRAPH '96 Proceedings)*, pages 31–42.

[Lindfors, 1998] Lindfors, J. A. S. . J. M. B. . K. K. (1998). Flat-field correction technique for digital detectors.

[Lumsdaine and Georgiev, 2008] Lumsdaine, A. and Georgiev, T. (2008). Full resolution lightfield rendering. Technical report, Adobe.

[Ng et al., 2005] Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., and Hanrahan, P. (2005). Light Field Photography with a Hand-Held Plenoptic Camera. Technical report.

[Sawchuk, 1977] Sawchuk, A. A. (1977). Real-time correction of intensity nonlinearities in imaging systems. *IEEE Trans. Computers*, 26(1):34–39.

[Smith, 2007] Smith, W. J. (2007). *Modern Optical Engineering*. McGraw-Hill.

[Yu et al., 2004] Yu, W., Chung, Y., and Soh, J. (2004). Vignetting distortion correction method for high quality digital imaging. In *ICPR*, pages 666–669.

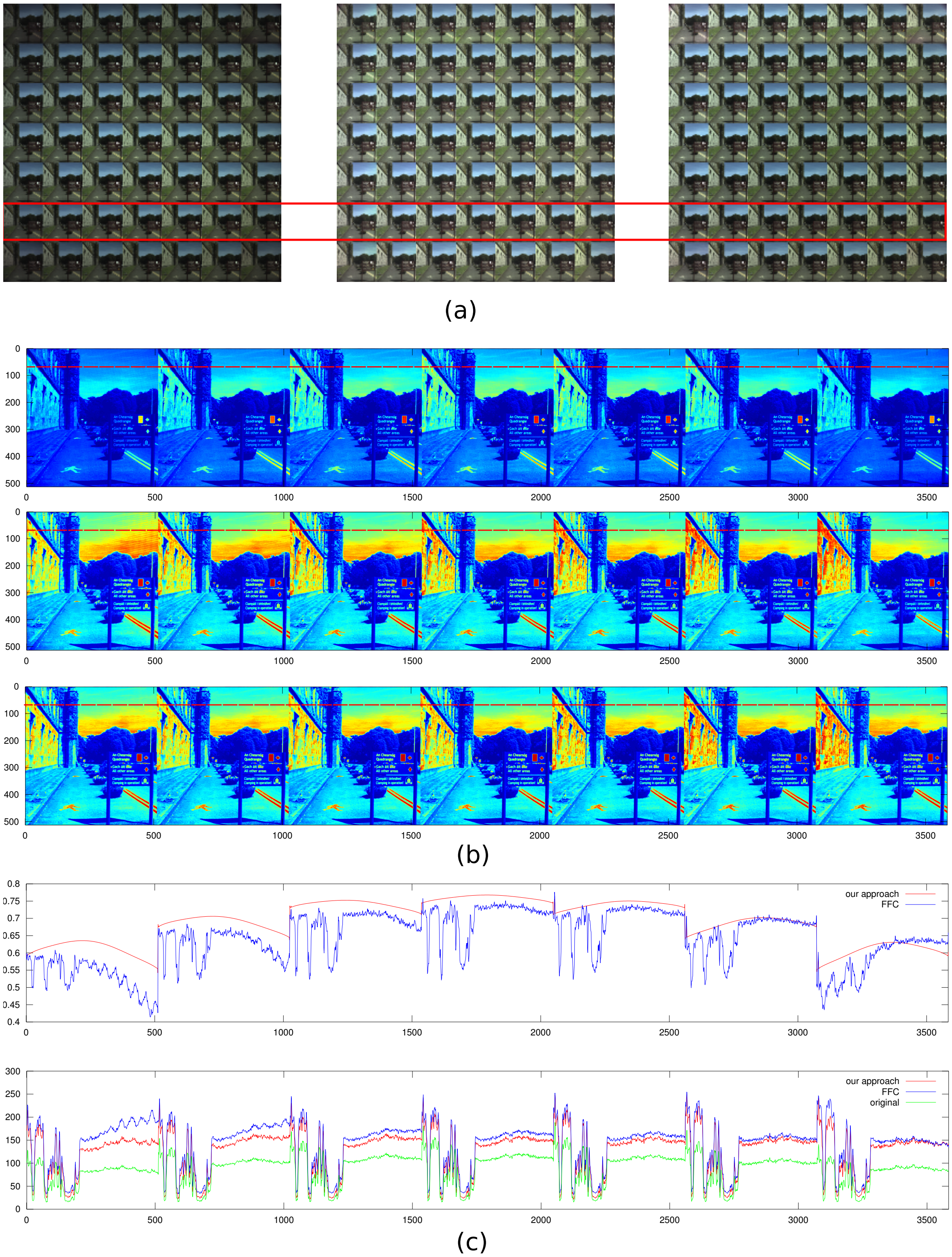


Figure 4: (a) The 7 by 7 multiview image array of a natural scene reconstructed from raw light field data. Each view's spatial resolution is 512 by 512. From left to right: without vignetting correction, with FFC approach vignetting correction, our approach. (b) The intensity profile of a horizontal 7 views. From Top to bottom: without vignetting correction, with FFC approach vignetting correction, our approach. (c) Top is the correction coefficients plot. Bottom is the intensity profile plot.

Classification of Seabed Type from Underwater Video

Steven Tyner[1], **James Wilson**[2] **and David Corrigan**[1]
[1] Department of Electronic and Electrical Engineering
[2] Department of Zoology
Trinity College Dublin
Dublin, Ireland.
{tyners, jwilson, corrigad}@tcd.ie

July 8, 2014

Abstract

This paper describes a method for the classification of seabed type from a video captured by a camera mounted on a towed vehicle that is dragged along the sea floor. Classification of seabed type is important for the mapping of marine habitats. Unlike other methods that are based on various sonar technologies, the proposed method is based purely on video frames. The aim is to tell from a single frame, what seabed type is present. A supervised learning approach is adopted, with a total of 5 different seabed types being represented. We developed a set of 6 image features to characterise the visual appearances of these seabed types. Both k-Nearest Neighbours (kNN) and Support Vector Machine (SVM) classifiers are implemented based on this feature set. Our analysis shows that is possible to achieve a cross-validation error of 10% for the 5-class problem.

Keywords: Marine Surveillance, Seabed Classification, Habitat Mapping, Supervised Learning

1 Introduction

Knowledge of seabed habitats is important for both environmental and commercial purposes [Robinson et al., 2009]. Environmental Impact Assessments (EIAs) are often required for the licensing of marine development or fishing operations. Furthermore, EU legislation, namely the Habitats Directive [European Community, 2007], requires member states to contribute to a European ecological network specifying special areas of conservation within their territories. To this end, each state must conduct seabed surveys to establish if areas of conservation exist within their territories [Sotheran et al., 1997].

At present the majority of seabed surveys are conducted using various sonar technologies such as backscatter and AGDS [Collier and Brown, 2005]. These acoustic methods, however, provide results which are hard to interpret and require ground-truthing (often by the acquisition of sample grabs or underwater photography) to establish the true seabed type [Blondel and Murton, 1997]. Video is less commonly used to acquire survey data as it requires the extra expense of mounting cameras on vehicles that are either towed or remotely operated [Robinson et al., 2009, Davie et al., 2008]. However, it is easier to visually identify seabed type from video and also allows the identification of marine flora and fauna that are indicative of habitat present.

The aim of our work is to develop a system than can automatically classify seabed type from an underwater video captured by a camera mounted on a towed vehicle dragged along the sea floor. This would reduce the amount of time spent on manual analysis of these surveys and hence accelerate the acquisition and analysis of new data. We treat the classification as an image texture analysis problem and we use 6 features to characterise the texture of each video frame. Unlike related work [Pican et al., 1998, Davie et al., 2008], our method adopts a

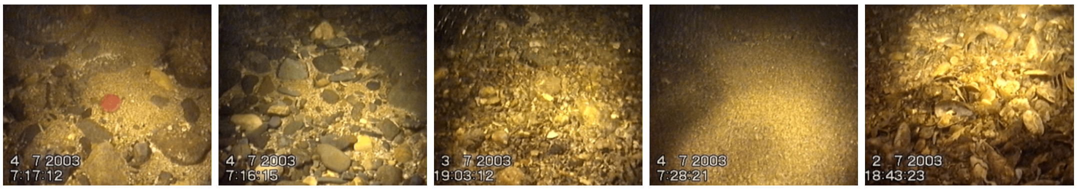


Figure 1: Examples of the 5 types of seabeds classified by our method. From Left to Right: Boulders, Cobbles, Pebbles, Sand and Shells.

Table 1: The number of examples of each class in the training set.

Class	Boulders	Cobbles	Pebbles	Sand	Shells
# Images	11	4	11	66	19

supervised learning approach. This allows the classifier to be explicitly trained to determine type according to accepted definitions (eg. the EUNIS (`eunis.eea.europa.eu`) or JNCC (`jncc.defra.gov.uk/page-1584`) classification systems) and, given sufficient training data, would allow the classifier to work under different lighting conditions, geographical locations and with different vehicles. A second key difference is that classification is performed at frame resolution rather than identifying multiple seabed types within a frame. This is a reasonable approach as the field of view is typically much less than the recommended 5×5 m^2 resolution specified by the EUNIS mapping instructions.

The remainder of the paper is organised as follows. The next section describes the test data and how it is used to generate the ground truth for both the training and testing of the classifiers. This is followed by a description of the feature set. Section 3 provides details of the classifiers implemented and presents the results of our experiments. The paper concludes with a discussion of our results and outlines avenues of future work.

2 Methodology

A total of 111 images were extracted from an underwater video from the HABMAP dataset [Robinson et al., 2009] and were manually classified into five seabed types of Boulders, Cobbles, Pebbles, Sand and Shells. Figure 1 shows an example of a video frame containing each seabed type and the number of examples of each type is given in Table 1. Only a small number of the video frames were suitable due to the large amount of motion blur present and because the camera was periodically stationary. Since the salient image content is unchanging while the camera is stationary, features are only estimated for one frame of each stationary period. This frame was extracted manually. All of the images were first pre-processed to mitigate against the uneven lighting present. This was performed by filtering the image with a low-pass filter with a narrow bandwidth and subtracting it from the original image.

From each of these pre-processed images six features were extracted for use in the seabed classifier. The features were designed to exploit the various image texture properties of each of the seabed classes. Each feature, bar one, was scaled so as to have zero mean and unit variance. This step was performed as the scale of each feature differed greatly and so added an error to the classification results.

Number of Edges

The first feature calculated is the number of edge pixels in the example image. This is an obvious choice since, for example, the sand class would contain fewer edges than other classes. An edge map was estimated from the *edge()* function in Matlab using the Prewitt edge detector [Prewitt, 1970]. The threshold for all of the frames was chosen to be the mean of the threshold values for all of the training examples obtained by setting the automatic threshold flag in the edge function. The number of edges feature is given by the total number of non-zero values in the edge map. An obvious drawback of this method is the falsely inflated number of edges introduced by the time and date stamp in the bottom left corner. For training and testing of the classifier with the same dataset this offset in the number of edges would remain relatively constant and so can be ignored. However, to make the feature useful in general, this could be accounted for by ignoring the parts of the frame where the timestamp is located.

Mean Colour

As colour is a prominent feature across the seabed types, with sand exhibiting a greater yellow hue compared to the remaining classes, a mean colour feature was calculated by taking the mean of each of the RGB channels and returning the average of these three values.

Discrete Wavelet Transform Coefficient Energies

The Discrete Wavelet Transform (DWT) is a useful tool for texture analysis [Smith and Chang, 1994] and breaks down the texture into various bands according to texture orientation and frequency content. We developed three separate features based on the energy of the horizontal (C_H), vertical (C_V) and diagonal (C_D) bands at the third level of the wavelet transform [Arivazhagan et al., 2005, Kociolek et al., 2001]. For this paper the 'bior2.2' wavelet, a symmetric biorthogonal wavelet, was selected for the DWT. Each of the features, E_x, are calculated from the DWT coefficients, C_x, by

$$E_x = \frac{1}{N} \sum C_x^2 \tag{1}$$

where x = H, V or D and N is the number of coefficients.

Co-occurrence Matrix Correlation

A Grey-Level Co-occurrence Matrix (GLCM) is a statistical method used to examine the texture of an image through the spatial relationship between the pixels [Nanthagopal and Sukanesh, 2012]. The GLCM is created by calculating the number of times pixels with intensity i and j, separated by a distance d in a specified direction, occur in an input image [Haralick et al., 1973, Pican et al., 1998]. The GLCM is constructed with the (i,j) location of the matrix representing the calculated number of occurrences of the i and j pixels.

The GLCM is, in fact, calculated from a quantised version of the input image. The default quantisation is used for this feature which scales all of the grey level intensities to integers between 1 and 8. This decreases the size of the GLCM to an 8x8 matrix and so lowers memory overhead and computation time and allows the matrix to be more densely populated. However, the distance and direction of the pixel pairs, known as the offset, was altered from the default of one pixel distance in the horizontal plane. For this feature the distances were chosen as 1, 2, 3, 8, 16 and 32 pixels so as to give a range of GLCMs based on increasing distance [Ghazali et al., 2007]. In addition, GLCMs were estimated for both the horizontal and vertical orientations, giving a total of 12 GLCMs.

Haralick proposed a number of features suitable for texture analysis that can be extracted from a GLCM [Haralick et al., 1973]. In this work the GLCM correlation was chosen as a feature and is calculated on each of the GLCMs after having been normalised to sum to one. The correlation of a GLCM is calculated as

$$Correlation = \frac{\sum_{i}^{N} \sum_{j}^{N} (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{2}$$

where $p(i,j)$ is the normalised GLCM and where μ_x, μ_y, σ_x and σ_y are the means and standard deviations respectively of the sums of the rows and columns.

Table 2: Cross-validation errors for the *one-v-the-rest* SVM Classifiers.

Kernel Type	**Boulders**	**Cobbles**	**Pebbles**	**Sand**	**Shells**
Linear	29%	19%	8%	5%	5%
RBF ($\sigma = 1$)	10%	18%	6%	5%	3%
Polynomial (3^{rd} order)	5%	15%	9%	4%	5%

This correlation provides an indication of how correlated a pixel is to its neighbour over the whole image. This value ranges between -1 and +1, corresponding to a perfectly positively or negatively correlated image. As the correlation of constant textures is non-determined, feature scaling was not performed on this feature. The correlation feature is the mean of the coefficients for the 12 GLCMs.

3 Results

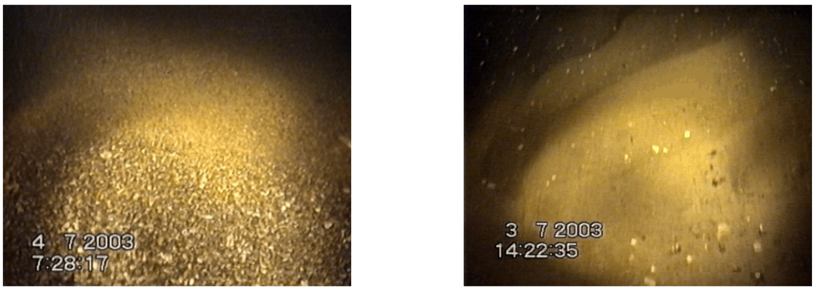
The k-Nearest Neighbours (kNN) and Support Vector Machine (SVM) classifiers were both developed to classify the seabed images. The kNN classifier is referred to as a lazy classifier and implementation of a multi-class classifier is straightforward. However, it is computationally inefficient at classifying unknown examples when the amount of training data is large. The key parameter in the kNN classifier is the value of k, which implicitly defines the complexity of the decision boundary between the classes. On the other hand, SVMs are much more efficient at classifying unknown examples once training has been completed. However, optimising an SVM for more than 2 classes is not straightforward. The notion of a kernel is central to SVMs as it allows complex decision boundaries between the classes. In our experiments, we tested both the value of k in the kNN and the Polynomial and Radial Basis Function (RBF) kernels of the SVM to minimise the classification error on the training data.

As SVMs are designed for 2-class problems, we trained a one-v-the rest SVM classifier for each of the 5 classes. By performing a K-fold cross-validation with 10 folds, the average cross-validation error can be estimated on the training set and this value is used to optimise either the polynomial degree or sigma parameter for the respective kernels. From our experiments it was determined that the optimal degree of the polynomial kernel was 3 and achieved an average cross-validation error of 7.6% for the 5 one-v-the rest classifiers. The optimal value of the sigma parameter for the RBF was 1, resulting in a slightly higher than average cross-validation error of 8.4%. For comparison we estimated the average cross validation error for the default linear kernel as 13.2%. A summary of the classification errors for each classifier is given in Table 2.

We used a kNN to implement a full 5-class classifier and use K-fold cross-validation to optimise the value of k. The optimum result was achieved with a value of k of 1 (ie. A nearest Neighbour Classifier) and the cross-validation error achieved was 10%.

As expected with these results the linear SVM kernel causes the greatest classification error due to the lack of linearly separable data. The more complex, non-linear polynomial and Gaussian kernels had a lower classification error with the polynomial kernel outperforming the Gaussian with the optimal order of three. It is also expected that this classification error is less than the kNN classifier due to the the fact that it is classifying between 5 rather than 2 classes.

Throughout the testing of both classifiers it was found that there was a constant misclassification of several of the sand images. Analysis of these specific images revealed they depicted coarse sand, as shown in Figure 2, not the smooth, silty sand that would typically be associated with a sandy seabed. As such the features detected this increased level of coarseness and likely viewed these images as resembling the features of the pebble images. The Joint Nature Conservation Committee (JNCC) classification framework defines the various classes based predominantly on the physical size of the individual stones or grains of sand. Hence, there



(a) Coarse sand

(b) Sand

Figure 2: An example of a coarse sand image which is miss-classified as pebbles compared to a 'standard' sand image.

is scope for miss-classifications where the sizes are near the specified boundaries between the classes.

4 Conclusion

This paper has proposed a method for the automatic detection of seabed type from a video taken from a towed or remotely operated vehicle. The key idea is to use a supervised learning approach that allows a classifier to be trained to accepted definitions of seabed type and given sufficient training data should be robust to variations in lighting conditions and geographical location. Our experiments show that on a small dataset it is possible to get a classification error of 10% for the 5 class problem.

Acquiring more labelled videos of seabeds is key to improving the performance of the classifications. It will allow both improvements in the design of the classifiers and on the design of features as well as to obtain a more reliable estimate of the accuracy of the classifier. For example, including data from more than one survey will establish the accuracy of the classifier on newly acquired marine surveys and will ensure that the features used are not overfit to the training data. We would also like to explore multi-class implementations of the SVM as well as Neural Networks as frameworks for the multi-class classification problem.

Acknowledgments

This work was funded in part by a grant from Science Foundation Ireland (Project Number 08/IN.1/I2112). The authors would like to thank Dr. Bryn Jones from Natural Resources Wales for providing some of the training data for this study.

References

[Arivazhagan et al., 2005] Arivazhagan, S., Ganesan, L., and Angayarkanni, V. (2005). Color texture classification using wavelet transform. In *Sixth International Conference on Computational Intelligence and Multimedia Applications*, pages 5–10.

[Blondel and Murton, 1997] Blondel, P. and Murton, B. J. (1997). *Handbook of seafloor sonar imagery*. Wiley Chichester, UK.

[Collier and Brown, 2005] Collier, J. and Brown, C. (2005). Correlation of sidescan backscatter with grain size distribution of surficial seabed sediments. *Marine Geology*, 214(4):431–449.

[Davie et al., 2008] Davie, A., Hartmann, K., Timms, G., de Groot, M., and McCulloch, J. (2008). Benthic habitat mapping with autonomous underwater vehicles. In *OCEANS 2008*, pages 1–9.

[European Community, 2007] European Community (2007). Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora.

[Ghazali et al., 2007] Ghazali, K. H., Mansor, M. F., Mustafa, M. M., and Hussain, A. (2007). Feature extraction technique using discrete wavelet transform for image classification. In *5th Student Conference on Research and Development - SCOReD*, number December, pages 5–8.

[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(610-621).

[Kociolek et al., 2001] Kociolek, M., Materka, A., Strzelecki, M., and Szczypinski, P. (2001). Discrete wavelet transform - derived features for digital image texture analysis. In *International Conference on Signals and Electronic Systems*, number September, pages 163–168.

[Nanthagopal and Sukanesh, 2012] Nanthagopal, A. P. and Sukanesh, R. (2012). Wavelet statistical texture features-based segmentation and classification of brain computed tomography images. *IET Image Processing*, 7(1):25–32.

[Pican et al., 1998] Pican, N., Trucco, E., Ross, M., Lane, D., Petillot, Y., and Tena Ruiz, I. (1998). Texture analysis for seabed classification: co-occurrence matrices vs. self-organizing maps. In *IEEE Oceanic Engineering Society. OCEANS 1998*, volume 1, pages 424–428. Ieee.

[Prewitt, 1970] Prewitt, J. M. S. (1970). Object enhancement and extraction. *Picture processing and Psychopictorics*.

[Robinson et al., 2009] Robinson, K. A., Darbyshire, T., Van Landeghem, K., Lindenbaum, C., McBreen, F., Creaven, S., Ramsay, K., Mackie, A. S. Y., Mitchell, N. C., Wheeler, A., Wilson, J. G., and OBeirn, F. (2009). Habitat mapping for conservation and management of the southern irish sea (habmap). i: Seabed surveys. studies in marine biodiversity and systematics from the national museum of wales.

[Smith and Chang, 1994] Smith, J. R. and Chang, S.-F. (1994). Transform features for texture classification and discrimination in large image databases. In *IEEE International Conference on Image Processing*, volume 3, pages 407–411. IEEE.

[Sotheran et al., 1997] Sotheran, I., Foster-Smith, R., and Davies, J. (1997). Mapping of marine benthic habitats using image processing techniques within a raster-based geographic information system. *Estuarine, Coastal and Shelf Science*, 44:25–31.

Application of Background Modelling to Acoustic Monitoring of Underwater Environment

Albert Akhriev, Mark Purcell, John Sheehan, Michael Barry
IBM Research Ireland, Dublin
albert_akhriev@ie.ibm.com

Abstract

We are performing two conceptual transformations on captured underwater acoustic data. Through the use of spectrograms we transform this data into a visual format, upon which we can apply standard image processing techniques. Then, we map events detected back into the acoustic domain, associating visual artefacts with acoustic events.

Keywords: Acoustic monitoring, spectrogram image, visual analysis.

1 Introduction

One common strategy for studying acoustic signals is to display a recorded signal as a spectrogram image. Spectrograms are used to visualize the frequency content of sounds in various applications. This includes acoustic monitoring of the underwater environment, our primary interest. Live spectrograms, obtained by sliding an *analysis window* over a continuous acoustic data stream, are well suited to acoustic monitoring and event detection.

A live spectrogram can be considered as a sequence of $N \times 1$ images, where N is the number of harmonics output by the Discrete Fourier Transform (DFT) after processing a subseries of the acoustic signal. This useful interpretation implies applicability of certain machine vision methods to time-series analysis. In fact, many acoustic events are of a very complicated nature, only becoming apparent via a 2-D spectrogram representation.

In this study, we have applied a background modelling approach to event detection in a sequence of spectrum “images”; just as streaming video is used for video-surveillance. Our ultimate goal is to build a real-time, underwater monitoring system for detection of “events of interest” without or limited knowledge of their underlying nature. This system ingests a real-time data feed, which is often too large to store for off-line processing (100’s of gigabytes per day). So the system must identify the presence of some perceptible events and store only a subset of time-series where the event has occurred.

2 Background Modelling

In video-surveillance applications background subtraction methods are used to detect scene changes. A comprehensive survey of modern background modelling and subtraction methods [Bouwmans, 2014] clearly demonstrates a large variety of ideas in this field, fuelled primarily by increasing demand on security systems.

For the task of acoustic monitoring, a high level of sophistication is not necessary. There are no sudden illumination changes, dynamic textures, “ghost” objects, shadows or many other video-surveillance specific problems. The following properties are typical for audio streams collected in underwater environment. (1) The background noise is relatively strong. Its amplitude slowly evolves over time depending on weather conditions, time of day, season, etc. (2) The normal ambient (background) noise can be influenced by equipment interference at some frequencies, resulting in a mixture of extraneous signals registered by a sensor. (3) In the absence of acoustic events, the distributions of the majority of spectral components are (nearly) independent, except for a few harmonics contaminated by equipment noise.

We consider background subtraction in the broader context of change detection. Given an input audio stream, the Fourier transform is applied to successive (overlapping) data segments. Then for each data segment the amplitudes of Fourier harmonics are computed. Obtained $N\times 1$ "image" of amplitudes is appended to the (infinite) live spectrogram image as the right-most column, where N is the number of harmonics with non-negative frequencies. At each entry ("pixel") of successive $N\times 1$ "images", the normal (background) process can be interrupted by an acoustic event. In what follows, the terms "foreground object", "pixel" and "pixel value", familiar to the computer vision community, will be often used instead of "acoustic event", "entry of Fourier spectrum" and "amplitude of Fourier harmonic" respectively.

Interestingly, the Fourier transform has been already employed for background modelling, for example [Porikli and Wren, 2005]. However, our use case is different. Here, we monitor the amplitudes of Fourier harmonics directly without relation to the objects in the image domain.

In this research we have chosen the popular Mixture of Gaussians Model (GMM) [Stauffer and Grimson, 1999], recently extended in [Zivkovic and van der Heijden, 2006]. We shall refer the latter as GMMA, which stands for *adaptive* GMM. GMMA provides a statistically sound and high performance mechanism to handle the mixture of components and decide their number. Other important extensions, including robust non-Gaussian mixtures [Bouwmans, 2014], are worthy of mention, but they were not examined here.

To handle the case of non-stationary background, it is common to store and gradually update the recent history of background observations without assuming any model [Bouwmans, 2014]. An interesting example of non-parametric method has been presented in [Barnich and Van Droogenbroeck, 2011], which, in addition to history tracking, also spreads information across neighbouring pixels to boost inter-pixel consistency. Our version of *background history tracking* algorithm, hereafter referred as BHT, shares these common ideas.

In the subsequent sections we shall present some real data results for extended GMMA and BHT, along with brief description of algorithms and our most important extensions to them.

3 BHT algorithm

BHT updates background history at each pixel replacing the oldest observation. This replacement is done not every frame but every M-th one, $M = 5\ldots30$, in order to cover larger time span. The history size (the number of past observations) H is chosen sufficiently large, $H \geq 50$. A counter variable at each pixel helps to ensure that only N/M pixels are updated upon arrival of a new frame of size N. The latter saves on computational time. In contrast to the original paper, we gradually replace outdated values in non-random way.

One interesting feature of [Barnich and Van Droogenbroeck, 2011] is that the history of neighbouring pixels is periodically updated by the samples of a pixel in question. Spreading information across the neighbourhood prevents the background model from sticking in a wrong state forever. We use a similar approach, introducing yet another time sub-sampling factor $V = 5\ldots30$. Each time N/V pixels copy their values into the history of randomly chosen neighbours in the small vicinity $\{-2, -1, +1, +2\}$. Recall, "images" are $N\times 1$ dimensional.

In the meantime we use the same measure of proximity between a new observation and pixel history as in [Barnich and Van Droogenbroeck, 2011]. The value belongs to background if at least $K = 3$ previously observed samples fell within distance threshold R. However, in BHT, R is not a hard-coded parameter but the one estimated from the data as a multiplicity of noise amplitude. Robust estimation of noise deviation is a computationally demanding procedure. A simpler approach uses the moving average in the following form:

$$\xi_k^{t+1} = (1-\alpha)\xi_k^t + \alpha\sqrt{\left|p_k^t - 2p_k^{t-1} + p_k^{t-2}\right|}, \qquad \sigma_k^t = const \cdot \left(\xi_k^t\right)^2, \tag{1}$$

where t is a discrete time, α is an update rate, $\alpha = 0.01$, p_k is the value of k-th pixel, the expression under square root is the 2nd order time derivative of a time-series at k-th pixel, ξ_k is its moving average estimation, σ_k is an estimation of noise deviation at the k-th pixel. The 2nd order time derivative in (1) eliminates any linear trend in data. The square root, on the other hand, softly diminishes the influence of outliers introducing some robustness in σ_n estimation. In all conducted experiments, parameter R was defined as $R = 3 \cdot \left(\xi_k^t\right)^2$ and evolves over time.

The algorithm in [Barnich and Van Droogenbroeck, 2011] relies on a so called *conservative update policy*, which "never includes a sample belonging to a foreground region in the background model". It works very well for slowly changing backgrounds, while most innovations fall within a fraction of proximity threshold R. To speed up recovery when a data stream starts with foreground dominance, we introduced a user-defined time threshold $T \gg H$, where H is the history size. If foreground values constitute the majority of the history, the normal background would be detected as an "event". If an event has been detected in 90% of cases during time interval T, the history is discarded and reinitialized. Obviously, very long-lived events do eventually become part of background, but this is the usual trade-off.

4 Extended GMMA

One major question in the practical realization of any (online) clustering method: "how many clusters?" GMMA [Zivkovic and van der Heijden, 2006] was particularly designed to answer this question by choosing an appropriate prior (Figueiredo & Jain 2002, Brand 1999) and automatically adjusting the number of components in a mixture.

The OpenCV implementation comes with reasonable default values of several crucial parameters of GMMA, and works very well on many standard video-sequences. However, when GMMA was applied to spectrogram analysis it showed much less impressive results than BHT. First of all, mixture models are sensitive to the selection of initial variance value (when a new component is added). Also, the minimum and maximum admissible variances matter, because otherwise a component can shrink or expand unpredictably.

To make GMMA less dependent on *ad-hoc* variance parameters, we have developed a relatively computationally expensive pre-processor, actually a clustering algorithm, which we refer to as *variance estimator*. To achieve real-time performance, a time-series, formed by successive values at a spectrogram entry, is divided into a 256-sample chunks. Selecting the proper time offsets, only $1/256$ of the total number of time-series' have to be processed upon arrival of a new spectrogram. Thus, while GMMA works all the time, the clusterizer gradually updates estimated mean variance of mixture components from time to time.

Development of clustering algorithm was motivated by unsatisfactory results obtained by more traditional approach based on information criterion (AIC, BIC, etc.), which often leads to under- or over-clustering given a sample of relatively small volume.

Our clustering algorithm uses a variant of global k-means [Xie and Jiang, 2010], for its elegant procedure for selection of candidate cluster centre, combined with fuzzy c-means, which adjusts partition once a new cluster has been added. One cluster is added in a time starting from a single cluster configuration. On each iteration the distribution of Mahalanobis distances between a point and the nearest cluster centre is formed. Distributions obtained for configurations with $C-1$ and C clusters are compared by Kolmogorov-Smirnov test. If the null hypothesis (distributions are similar) was rejected given significance level (0.1 by default), then we proceed adding more clusters, otherwise the former configuration with $C-1$ clusters is accepted and algorithm stops. For practical computational purposes, the maximum number of clusters is limited to 5. Once the clustering is finished, the minimum, average and maximum variances over all clusters are estimated, then smoothed with the previously obtained values using a simple moving average and fed into GMMA to impose restrictions on component variance.

With this variance estimator, GMMA demonstrates much better results but is still behind BHT. The reason is that often the mixture components are not well modelled by Gaussian distribution. Better results could be obtained using a non-Gaussian mixture (e.g. t-distribution) having a “heavy” tail, which seems more suitable for our use case. The advantage of GMMA, on the other hand, is analytically simpler and practically “faster” updating formulae. The tradeoff is that an expensive pre-processing could undermine the overall GMMA performance.

We have made yet another extension to GMMA, considering a multi-variate metric that comprises 2 to 4 recent observations, for example, $\mathbf{m} = (m^t, m^{t-1}, m^{t-2})$. When $\mathbf{m}$ is compared against component's center $\mathbf{c} = (c_1, c_2, c_3)$ all cyclic permutations are tried $\mathbf{c}^{(1)} = (c_1, c_2, c_3)$, $\mathbf{c}^{(2)} = (c_3, c_1, c_2)$, $\mathbf{c}^{(3)} = (c_2, c_3, c_1)$. The smallest value $\|\mathbf{m} - \mathbf{c}^{(k)}\|$ gives the distance to component. This simple scheme accounts for co-occurrence of successive pixels at the expense of the larger number of components and provides a significant improvement.

5 Experiments

In this section, we present the results for the GMMA and BHT algorithms with all the extensions as described above. To our knowledge, there is no well established framework for comparison of different methods of live spectrogram evaluation like changedetection.net. The critical issue is the process of extracting “ground-truth”, typically requiring manual data annotation. In the meantime, we assess detectability visually, inspecting the spectrogram of a data record. We are also investigating other options for semi-automatic labeling.

Our live data feed comes from hydrophones connected to an off-shore buoy. The buoy is equipped with an analog to digital conversion unit and has appropriate network connectivity. Unavoidably, the on-board systems produce artificial noise that can be seen in the first two rows on Figure 1 (thick horizontal lines). Thus, the system should reject persistent disturbance and register only new patterns.

In our experimental set-up, the hydrophone data stream is sampled at 500 KHz, then chopped into successive windows (50% overlapping) and fed into DFT engine to compute spectrum amplitudes. The size of the analysis window is selected to be smaller than one second; typically, for performance reasons, a power of two: $2^{15 \cdots 17}$. This reduces the best achievable frequency resolution, which is usually redundant for acoustic monitoring.

The first two rows on Figure 1 show the result of detection of acoustic events. The left-hand side present the results of the extended GMMA method, the middle show the fragments of live spectrogram without any highlighting, and on the right-hand side are the results of the BHT method. When harmonic amplitudes deviate significantly from the normal background, it is detected as a part of an event and highlighted in yellow. BHT additionally distinguishes strong $n < K-1$ evidence (yellow) and weak $n = K-1$ evidence (green), where the parameter K (minimum number of close background samples) was discussed in Section 3. For better visibility, only part of the spectrum is shown with exact numbers given in the caption.

In the first row, a boat is passing and stopping nearby the buoy. Its engine produces a typical spectral picture — a bunch of frequencies $k{\cdot}f_0$, $k = 1, 2, \ldots$, where f_0 relates to the engine rotation speed varying over time. The second row shows an unknown event, which looks like a train of vertical lines around 18 KHz. Note that persistent disturbances (electrical noise) — the horizontal lines crossing the spectrogram — were not detected as events.

We also detect other types of acoustic events. For example, Lloyd's mirror — an interference pattern between direct and reflected noise produced by a passing vessel. Detection of the presence of sea mammals is another point of practical interest. Mammal vocalization results in a number of very different acoustic events. We selected file examples of crabeater seal from the collection of the Alfred Wegener Institute for Polar and Marine Research (Am Handelshafen 12, 27570 Bremerhaven, Germany), and these are shown in the last two rows of Figure 1. The very short time-series does not provide sufficent event-free data for background adaptation. However, both methods do well on repeated play-back of the files.

Our multi-threaded software implementation of both extended GMMA and BHT is written in C++ and runs on an Intel I7 CPU, 1.7GHz, using 7 threads out of 8 available. The same real-time performance can be achieved with a single thread, if so called AVX vector instructions of modern CPUs are employed. The latter, however, requires further programming effort.

In general, our version of GMMA tends to produce slightly cleaner, but less accurate change mask. The bottleneck of this GMMA is an expensive step for rough variance estimation, which takes about 0.2-0.25 seconds for analysis window with 2^{17} entries, although we envisage several options for significant performance improvement.

6 Conclusion and Future Work

This paper summarizes our experience with the application of modern background modelling and subtraction methods, developed by computer vision community, to the problem of acoustic monitoring. This approach aims to facilitate collection of atypical acoustics events in large data sets and live (intense) data streams in an unattended manner. In many cases the model of event is not known in advance, so the traditional signal processing methods, like filtering, might not work. Another application of the proposed approach, useful for training various classifiers, is to instrument data annotation, which would be otherwise unmanageable given the large amount of data containing nothing but noise.

We have made certain extensions to existing state-of-the-art approaches, but we are still looking to other opportunities. For example, subspace methods like PCA and its numerous extensions [Monnet et al., 2003, Bouwmans, 2014], as well as sophisticated statistical models based on Support Vector Machines (SVM), e.g. [Tavakkoli et al., 2008], are the alternative algorithms worth considering. Cepstrum representation of input data rather than conventional spectrogram is yet another option.

Post-processing is a separate subject of research, intentionally omitted in this paper. The traditional morphological approach would work well on acoustic data, however, in many cases the detected event has an elongated structure(s). This property along with the strong correlation between successive spectrum images is worth exploiting when a change mask is constructed.

References

[Barnich and Van Droogenbroeck, 2011] Barnich, O. and Van Droogenbroeck, M. (2011). Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1723.

[Bouwmans, 2014] Bouwmans, T. (2014). Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*.

[Monnet et al., 2003] Monnet, A., Mittal, A., Paragios, N., and Ramesh, V. (2003). Background modeling and subtraction of dynamic scenes. In *ICCV*, pages 1305–1312.

[Porikli and Wren, 2005] Porikli, F. and Wren, C. (2005). Change detection by frequency decomposition: Wave-back. *Mitsubishi Electric Research Laboratories, TR2005-034*.

[Stauffer and Grimson, 1999] Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 246–252.

[Tavakkoli et al., 2008] Tavakkoli, A., Nicolescu, M., Bebis, G., and Nicolescu, M. N. (2008). A support vector data description approach for background modeling in videos with quasi-stationary backgrounds. *Int. Journal on Artificial Intelligence Tools*, 17(4):635–658.

[Xie and Jiang, 2010] Xie, J. and Jiang, S. (2010). A simple and fast algorithm for global k-means clustering. In *Second International Workshop on Education Technology and Computer Science (ETCS)*, volume 2, pages 36–40.

[Zivkovic and van der Heijden, 2006] Zivkovic, Z. and van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recogn. Lett.*, 27(7):773–780.

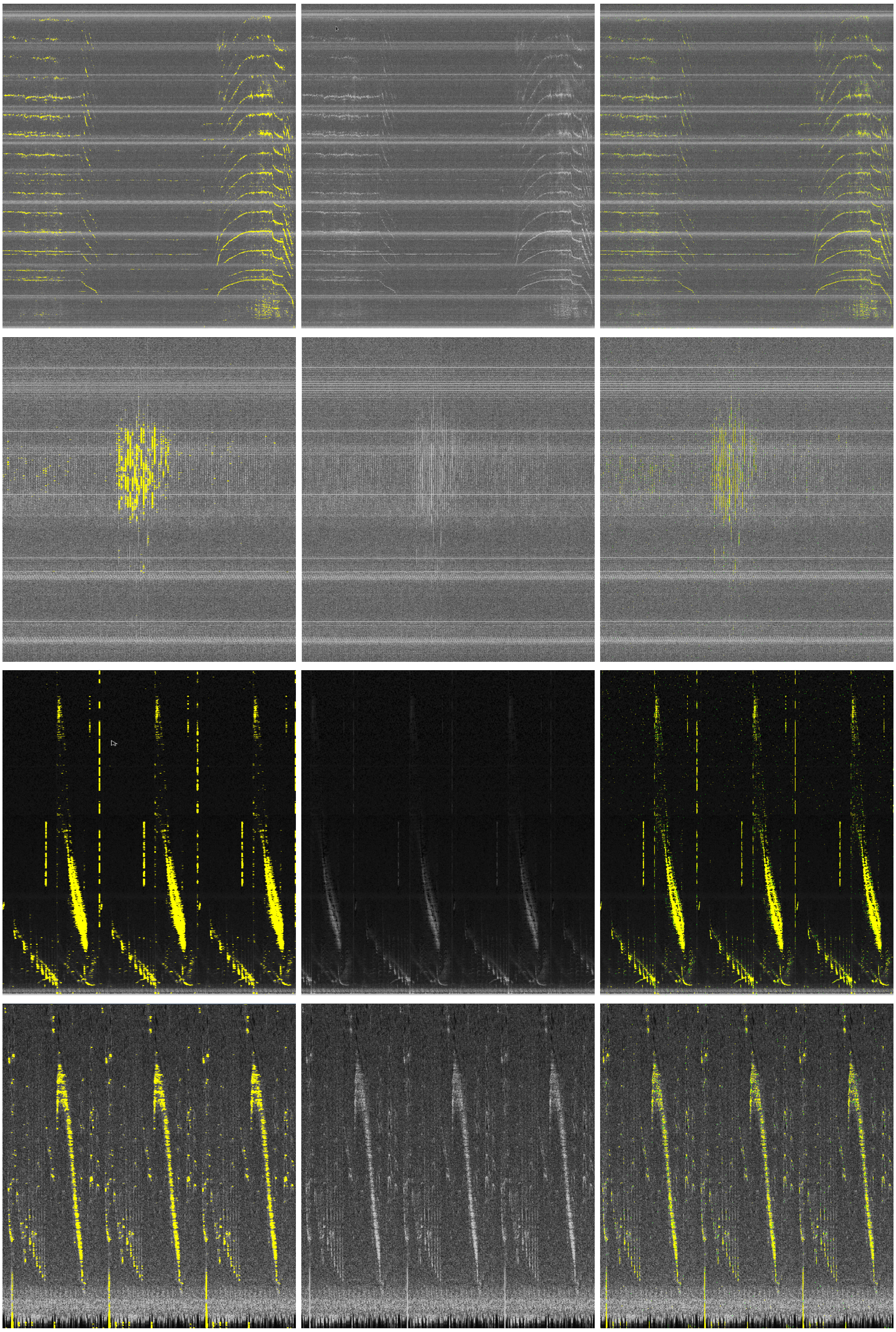


Figure 1: $\mathbf{X}$: time, $\mathbf{Y}$: frequency. *Left* column – results of GMMA, *middle* column – fragment of live spectrogram, *right* column – results of BHT. *First* row: a boat passing and stopping nearby sensor location, frequency range $[0, 10000]$ Hz, 300 seconds. *Second* row: unknown underwater event, frequency range $[15000, 20000]$ Hz, 164 seconds. *Third* row: crabeater seal voice, frequency range $[0, 7200]$ Hz, 60 seconds duration (repeated playback). *Fourth* row: crabeater seal voice, frequency range $[0, 1200]$ Hz, 60 seconds (repeated playback). Detected events are highlighted by *yellow* (strong evidence) and *green* (weak evidence) colours.

IMVIP 2014

MEDICAL IMAGING

Detection and Localisation of Prostate Cancer within the Peripheral Zone using Scoring Algorithm

Andrik Rampun, Reyer Zwiggelaar
Department of Computer Science, Aberystwyth University,Aberystwyth SY23 3DB, UK
yar, rrz@aber.ac.uk
Paul Malcolm
Department of Radiology, Norfolk Norwich University Hospital, Norwich NR4 7UY, UK
paul.malcolm@nnuh.nhs.uk

Abstract

The paper presents a novel computer aided diagnosis method for prostate cancer detection within the prostate's peripheral zone using a combination of different image features and grey level histogram analysis. The peripheral zone is subdivided into four regions and a scoring algorithm is employed to determine the most cancerous sub region based on specific metrics. The initial evaluation of this method is based on 200 MRI images from 40 patients and we achieved 89% accuracy with 0.89 and 0.88 sensitivity and specificity, respectively.

Keywords: Prostate Cancer Detection, MRI, Prostate Cancer Localisation

1 Introduction

More than 40,000 men are diagnosed with prostate cancer annually and it is expected to be the most common cancer by 2030 across the United Kingdom [1]. According to a recent article in [2] targeted biopsies would be better than the random ones which are currently used. Therefore, we propose a CAD method which can specifies a region(s) which has (have) the highest probability to be malignant, hence help radiologists to perform targeted biopsies and potentially improve the accuracy of prostate cancer diagnosis.

2 Modeling Prostate Peripheral Zone

Since 80% – 85% of the cancers arise in the peripheral zone (PZ) [3], we aim to detect prostate abnormality within that region. Note that, we did not perform prostate segmentation because all prostates were already delineated by an expert radiologist. Figure 1 shows our 2D prostate model where prostate's boundary and PZ's boundary are in black and magenta lines, respectively. The prostate's PZ is defined using the quadratic equation $y = ax^2 + bx + c$ based on three crucial coordinate points of the prostate which are v_1, v_2 and v_3. They are determined by the outmost x and y coordinates of the prostate boundary which are $x_{min}, x_{max}, y_{min}, y_{max}$ (see Figure 1). For example, x_{min} and y_{max} can be determined by taking the minimum x and maximum y coordinates along the prostate boundary. Moreover, the x coordinates of v_1 and v_3 are captured from x_{min} and x_{max} and their y coordinate is determined by taking the y coordinate between y_{min} and y_{max}. On the other hand, the x coordinate of v_2 is taken from the x coordinate

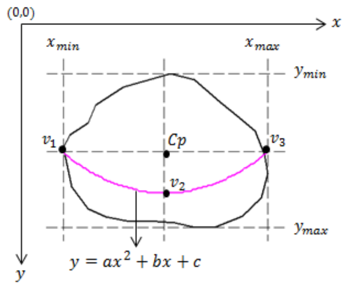


Figure 1: 2D Prostate Model

x_{min} and x_{max} and its y coordinate is determined by taking $\frac{3}{4}$ of the distance from y_{min} to y_{max}. Mathematically, these can be represented equations (1), (2), (3) and (4).

$$C_p = ((x_{min} + x_{max})/2, (y_{min} + y_{max})/2) \tag{1}$$

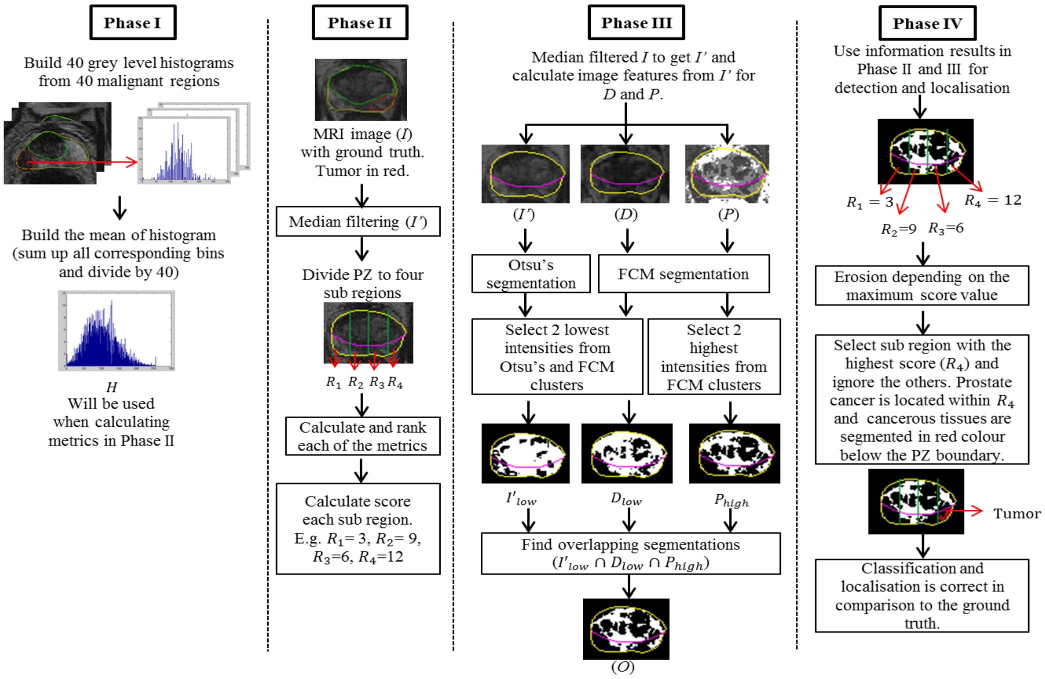
$$v_1 = (x_{min}, (y_{min} + y_{max})/2) \tag{2}$$

$$v_2 = ((x_{min} + x_{max})/2, y_{min} + ((y_{max} - y_{min}) \times \frac{3}{4})) \tag{3}$$

$$v_3 = (x_{max}, (y_{min} + y_{max})/2) \tag{4}$$

where C_p is the central point of the prostate. Once the coordinates of v_1, v_2 and v_3 are defined, we can determine the values of a, b and c and determine the PZ's boundary (magenta line).

3 Methodology



Phase I
Build 40 grey level histograms from 40 malignant regions
Build the mean of histogram (sum up all corresponding bins and divide by 40)
H
Will be used when calculating metrics in Phase II
Phase II
MRI image (I) with ground truth. Tumor in red.
Median filtering (I')
Divide PZ to four sub regions
R_1 R_2 R_3 R_4
Calculate and rank each of the metrics
Calculate score each sub region. E.g. R_1= 3, R_2= 9, R_3=6, R_4=12
Phase III
Median filtered I to get I' and calculate image features from I' for D and P.
(I')
(D)
(P)
Otsu's segmentation
FCM segmentation
Select 2 lowest intensities from Otsu's and FCM clusters
Select 2 highest intensities from FCM clusters
I'_{low}
D_{low}
P_{high}
Find overlapping segmentations ($I'_{low} \cap D_{low} \cap P_{high}$)
(O)
Phase IV
Use information results in Phase II and III for detection and localisation
$R_1 = 3$
$R_4 = 12$
R_2=9 R_3=6
Erosion depending on the maximum score value
Select sub region with the highest score (R_4) and ignore the others. Prostate cancer is located within R_4 and cancerous tissues are segmented in red colour below the PZ boundary.
Tumor
Classification and localisation is correct in comparison to the ground truth.

Figure 2: The proposed methodology starts from left to right (phase I to IV).

In phase I we construct a malignant histogram model. We use the malignant histogram model to calculate metrics and implements the scoring algorithm to determine the most cancerous sub region within the PZ (phase II). In phase III we compute image features and perform Fuzzy c-means (FCM) and Otsu's segmentation algorithms to segment cancerous tissues. Finally, we perform erosion to reduce false positives and use the results in phase II and III for detection and localisation purpose.

3.1 Preprocessing

Firstly, we perform median filtering on the original image (I) to reduce the amount of noise. Previous study has shown that median filter is effective at preserving sharp edges in MRI and in our case we want to preserve the information-bearing structures such as tumor's edge boundaries [4].

3.2 Construction of Histogram Model

We construct a histogram model (H) of 256 bins based on 40 malignant regions (40 patients) delineated by an expert radiologist. For every cancerous region we compute its normalised grey level histogram (by assigning each pixel to its corresponding grey level) and sum up all histograms (40) followed by diving each bin with the number of regions (in our case it is 40). All histograms are normalised so the sum of histogram's bins is equal to 1. This means, H represents the mean distribution of grey level based on 40 malignant regions.

3.3 Scoring Algorithm

We divide the PZ (under the magenta line) into four sub regions (R_1, R_2, R_3 and R_4) (Figure 2, phase II) which is similar to the prostate anatomy proposed in the European consensus guidelines division of prostate gland [5]. The main purpose of dividing the PZ into four sub regions is to enable the algorithm to determine the most cancerous sub region based on the highest score point thus reduce the number of false positives. The algorithm considers three metrics (subsection 3.3.1, 3.3.2 and 3.3.3) to determine score point value to be assigned to every sub region. For simplicity of the scoring algorithm design, we assign the least suspecious (based on the metrics values) sub region to be malignant with the lowest score followed by the most suspecious with the highest score (e.g. +1 is the least suspecious followed by +2,+3 and the most suspecious will be assigned +4). In principles, the scores could be any positive integer numbers as long as the sub region with a higher rank receives higher score than the lower ones.

3.3.1 Histogram Mean

For each R_n ($n \in \{1, 2, 3, 4\}$) we calculate the mean of the lower half of the histogram (H_{R_n}) because it represents the darker level of a region. Several studies have suggested that prostate cancer tissue tends to appear darker on a T2-weighted MRI image [6, 7]. In fact, contrast level in normal tissue is higher than cancerous tissues [8] and radiologists also tend to use darker regions to identify abnormality within the PZ [9]. Therefore higher probability to capture cancerous tissues by taking the lower half of the histogram. H_{R_n} indicates the frequencies of low grey levels distributed within a sub region. Higher H_{R_n} means higher number of low grey levels within the sub region. This indicates the sub region has darker appearance (higher probability to be malignant). We rank H_{R_n} from the lowest to the highest value and assign each R_n with an appropriate score. In this case R_1 receives the highest score point because it has the highest mean value (highest number of low grey levels, hence highest probability to be malignant). Equation (5) shows how this metric is calculated.

$$H_{Rn} = \frac{1}{N/2} \sum_{i=1}^{N/2} H_n(i) \tag{5}$$

where N is the number of bins and i represents the i^{th} bin in a histogram.

3.3.2 Histogram Intersection Distance

The second metric is histogram intersection distance (d_n) between H and each of H_{R_n}. The distance (e.g. the distance between H and H_{R_1}) indicates the histogram similarity between H (the histogram malignant model) and H_{R_1}. The smaller the distance the more similar H and H_{R_1} which means higher probability of R_1 being malignant, hence receives the higher score. However, since we are interested only in low grey levels, we only consider the lower half of the histograms to capture malignant tissues [6, 7, 9, 8]. The minimum and maximum values are 0 and 1, respectively. This metric can be calculated using Equation (6).

$$d_n = 1 - (\sum_{i=1}^{N/2} min\{H(i), H_{R_n}(i)\} / \sum_{i=1}^{N/2} H(i)) \tag{6}$$

Next, we rank d_n from the highest to the lowest using the same scores and assign each sub region with an appropriate score depending on its ranking.

3.3.3 Grey Level Co-Occurrence Matrix (GLCM) Feature

For the final metric, we calculate the GLCM contrast feature (C_{R_n}) for each of R_n. We chose GLCM contrast as it is found to be the most discriminate feature among GLCM features for the differentiation between cancerous and non-cancerous tissues [8]. R_n with the lowest contrast has the highest probability to be malignant [8], hence receives the highest score. To maximise the variations we took all orientations ($\theta = 0°, 45°, 90°, 135°$) and calculated the mean by summing up the contrast value of each orientation and divided it by the number of orientations. Subsequently, we ranked C_{R_n} from the highest to the lowest and assigned each sub region with an appropriate score.

3.4 Prostate Segmentation

We calculate the probability and local contrast image feature from the filtered original image (I') which can be calculated using equation (7) and (8), respectively. For an $I'(x, y)$ image, the probability value for the k^{th} grey level is:

$$P(x, y) = \frac{\#(I'(x, y) = k)}{R \times C} \tag{7}$$

where $\#(I'(x, y) = k)$ is the number of pixels at the k^{th} grey level in an $R \times C$ image, and as such each element in P is the probability value for a particular intensity level. Gharge and Kekre [10] used probability image for tumor demarcation in mammograms and MRI images while a study in [11] used probability image feature to segment cancerous regions within the peripheral zone. The local contrast image feature (D) is determined by

$$D(x, y) = W_{max} - W_{min} \tag{8}$$

where W_{max} and W_{min} are the maximum and minimum values, respectively within a 5×5 window. Litjens et al. [12] have shown that local contrast feature can improve the characterisation of tumour from normal tissues while Mukhopadhyay and Chande [13] showed qualitatively that capturing local contrast can enhance both regions' edges and textures. Subsequently, we segment image feature P and D individually into four different classes using a modified Fuzzy c-means algorithm proposed in [14] because it is robust in dealing with noises in medical images and used Otsu's segmentation method to segment I' for grey level segmentation. We chose four classes because there are four tissue categories in the prostate (two of them are associated with abnormal tissues) [15]. Since malignant regions within the PZ have dark appearance [3] we select segmented regions which correspond to the first two lowest intensities (indicated by the subscript 'low' in equation (9)) FCM clusters in D (the same in Otsu's clusters in I'). On the other hand, we selected two highest intensities (malignant region appears to be bright in P) FCM clusters in P. After selecting the regions of interest, we combine all binary segmentations by finding its overlapping region as showed in Figure 2, phase III. This process can be represented using the following equation

$$O = I'_{low} \cap D_{low} \cap P_{high} \tag{9}$$

To this point, we have segmented possible cancerous tissues (phase III) and have identified the most cancerous sub region according (phase II).

3.5 Abnormality Detection and Localisation

The propose method uses score points and O to identify whether cancer is truly present or not. The maximum and minimum total score points are 12 and 3, respectively. We use the combined binary segmentations (O) (Figure 2, phase III) and take the sub region which has the highest score. If a sub region has a maximum score point of 12 ($R_4 = 12$), we assume the level of confidence is strong and a malignant region is located within R_4. Therefore, we ignore all other sub regions (R_1, R_2 and R_3) and only consider the segmented area within R_4 (Figure 2, phase IV) for detection and localisation purpose. If there is no segmented area in R_4, we assume that it is normal (hence the whole slice is considered normal). Figure 3 shows the ground truth (malignant region is indicated in red) with O together with its score points in R_n. Without the scoring algorithm the segmentation within R_2 (false positive) in Figure 3, will be considered a malignant region which lead to incorrect localisation.

On the other hand, if the maximum score point is 11 ($R_3 = 11$), we assume the level of confidence is average and ignore all other sub regions but performed erosion with a 'disk' shape structuring element with size 1 (note that no erosion is performed if the maximum score point is 12 because the level of confidence is high). This ensures that the segmented areas are not noise which could lead to false positive result. If there are segmented areas after erosion, then we assume R_3 is malignant (otherwise normal). Finally, if the maximum score point is less than 11 ($R_3 \leq 10$), we assume the level of confidence is weak and all sub regions should be considered. We performed erosion with the same shape structuring element with size 2. The malignant region is the one with the biggest segmented area within R_n after erosion is performed.

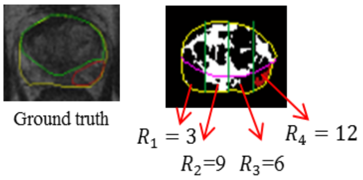


Figure 3: Tumor is located in R_4.

4 Experimental Results

We evaluated the proposed method based on 200 (105 malignant and 95 normal slices, excluded 40 slices used to construct H) prostate T2-Weighted MRI images (512×512) from 40 different patients aged 54 to , collected from Norwich. Each case has 4 to 6 slices through the central part of the prostate. The prostates, cancer and central zones were delineated by an expert radiologist on each of the MRI images. Data was analysed and classified as to whether the prostate contains cancer. The detection of cancer occurs when there are any retained segmented regions within R_n. Subsequently, we compared the result with the ground truth whether the prostate contains cancer regions or not. The proposed method achieved 89% accuracy (0.89 sensitivity and 0.88 specificity) with 7.5% false negatives and 7.5% false positives. In comparison to the state of the arts methods, it is extremely difficult to make quantitative comparisons due to the absence of public datasets. However, to have an overall qualitative estimate of the functioning of our method we compared with some of the recent studies. The method proposed by Artan and Yetik [4] achieved 82% accuracy (sensitivity =0.76 and specificity = 0.86) based on 15 patients. Niaf et al. [16] reported 0.89 sensitivity and 0.82 specificity for 30 patients. Futterer et al. [17] presented their results for six patients and their results show 0.83 for both sensitivity and specificity. Finally a method proposed in [11] reported 85% accuracy with 0.82 and 0.88 sensitivity and specificity, respectively. Although these comparisons are very subjective, the results achieved the state of the art in the literature qualitatively.

5 Conclusions

In conclusion, we have presented a novel method of prostate cancer detection and localisation within the PZ. Our idea is to subdivide the PZ into four regions and employed a scoring algorithm to determine the most cancerous sub region based on the cumulative score. In addition, we combined binary segmentations to segment the most cancerous tissues and performed different erosion sizes to reduce false positives and false negatives. Early evaluation have showed that the proposed method has the potential to help radiologists in corresponding to the current problem stated in [2].

References

[1] “Prostate cancer facts and figures,” 2013. http://prostatecanceruk.org/information/prostate-cancer-facts-and-figures/.

[2] “Prostate cancer tests miss severity in half of cases,” 2014. http://www.bbc.co.uk/news/health-26970132.

[3] Edge et al., S. B., *AJCC Cancer Staging Manual (7th Edition).* Springer, ISBN: 9780387884400, 2010.

[4] Artan, Y. and Yetik, I., “Prostate cancer localization using multiparametric mri based on semisupervised techniques with automated seed initialization,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1313–1323, 2012.

[5] Dickinson et al., L., “Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: recommendations from a european consensus meeting,” *Eur Urol*, vol. 59, no. 4, pp. 477–494, 2011.

[6] Garnick et al., M. B., *Harvard Medical School 2012: Annual Report on Prostate Diseases.* Harvard Medical School, 2012.

[7] Ginat et al., D. T., “Us elastography of breast and prostate lesions,” *Radiographics*, vol. 29, no. 7, pp. 2007–2016, 2009.

[8] Mohamed et al., S. S., “Region of interest identification in prostate trus images based on gabor filter,” in *IEEE 46th Midwest Symposium on Circuits and Systems*, vol. 1, pp. 415–419, Dec 2003.

[9] Taneja, S. S., “Imaging in the diagnosis and management of prostate cancer,” *Reviews in Urology*, vol. 6, no. 3, pp. 101–113, 2004.

[10] Gharge, S. and Kekre, H. B., *Segmentation of Medical Images. Ph.D. Thesis.* Mukesh Patel School of Technology Management and Engineering: India, 2013.

[11] Rampun et al., A., “Detection and localisation of prostate abnormalities.,” in *Proceeding in Computational and Mathematical Biomedical Engineering, Hong Kong*, pp. 204–208, 2013.

[12] Litjens et al., G. J. S., “Interpatient variation in normal peripheral zone apparent diffusion coefficient: effect on the prediction of prostate cancer aggressiveness,” *Radiology*, vol. 265, no. 1, pp. 260–266, 2012.

[13] Mukhopadhyay, S. and Chanda, B., “Local contrast enhancement of grayscale images using multiscale morphology.,” in *Proceeding in Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 17–24, 2000.

[14] Chen, Z. and Zwiggelaar, R., “A modified fuzzy c-means algorithm for breast tissue density segmentation in mammograms,” in *10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB)*, pp. 1–4, Nov 2010.

[15] Yin et al., M., “Diagnostic utility of p501s (prostein) in comparison to prostate specific antigen (psa) for the detection of metastatic prostatic adenocarcinoma,” *Diagnostic Pathology*, vol. 41, no. 2, 2007.

[16] Niaf et al., E., “Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric mri.,” *Phys Med Biol*, vol. 57, pp. 3833–3851, 2012.

[17] Futterer et al., J. J., “Prostate cancer localization with dynamic contrast-enhanced mr imaging and proton mr spectroscopic imaging,” *Radiology*, vol. 241, no. 2, pp. 449–458, 2006.

Automatic Analysis of Digital Retinal Images for Glaucoma Detection

Alexandre Guerre Jesús Martínez del Rincón Paul Miller
The Institute of Electronics, Communications and Information Technology
Queen's University Belfast, BT3 9DT
{a.guerre,j.martinez-del-rincon,p.miller}@qub.ac.uk

Augusto Azuara Blanco
Centre for Vision and Vascular Science
Queen's University Belfast, BT12 6BA
a.azuara-blanco@qub.ac.uk

Abstract

In this paper we propose a novel automated glaucoma detection framework for mass-screening that operates on inexpensive retinal cameras. The proposed methodology is based on the assumption that discriminative features for glaucoma diagnosis can be extracted from the optical nerve head structures, such as the cup-to-disc ratio or the neuro-retinal rim variation. After automatically segmenting the cup and optical disc, these features are feed into a machine learning classifier. Experiments were performed using two different datasets and from the obtained results the proposed technique provides better performance than approaches based on appearance. A main advantage of our approach is that it only requires a few training samples to provide high accuracy over several different glaucoma stages.

1 Introduction

Glaucoma is one of the most common causes of preventable blindness [1] and official population projections and epidemiological prevalence surveys predict that the number of glaucoma cases will increase by a third in the next twenty years [2]. Glaucoma, and those at risk of suffering from glaucoma, constitute a major part of the workload of secondary care eye services [3]. However, patient overload is not the only problem. Currently, referrals for suspected glaucoma are usually initiated by a community optometrist and then assessed at hospital by trained ophthalmologists. The reported diagnostic accuracy for detecting glaucoma by optometrists is suboptimal: only 20-30% of these referrals actually have glaucoma, and 45% of patients are discharged after their first visit [4]. This illustrates the inefficiency of current glaucoma detection methods and causes avoidable distress and worry to patients and carers. Interventions for optometrists, such as glaucoma training [5] or agreed guidelines [6], do not appear to affect the rates of false positive referrals. Even definitive glaucoma diagnosis, carried out by ophthalmologists, are not exempt from drawbacks: clinical optic nerve assessment is limited by subjectivity and reliance on examiner experience, while new diagnostic techniques for assessment of the structural changes at the optic nerve head (ONH) and retinal nerve fibre layer (RNFL) are expensive and therefore not widely available.

In this context, automatic detection methods are highly valuable for early glaucoma diagnosis [1], especially considering that glaucoma can be treated effectively if detected at an early stage. We propose a method based on the automatic analysis of the eye fundus, which brings together the expertise of human practitioners and the cost-effective advantages of computers. Given that digital fundus cameras are relatively inexpensive and are already widely available in optometrists' and hospital eye services, our system could potentially be deployed as a systematic screening programme for glaucoma.

Our method differs from other state-of-art systems in the usage of geometric parameters of the ONH structures that change in case of glaucoma disease: optic disc diameter, optic disc area, cup diameter, rim area, mean cup depth, etc. These features are extracted from the automatic segmentation of the structures and used for training a machine learning classifier which provides the final decision given a new fundus image. The usage of these features, traditionally employed in the manual analysis, has some competitive advantages regarding appearance based methods: they are less dependent on the camera model, they require a lower order of magnitude in the number of training images -dozens instead of

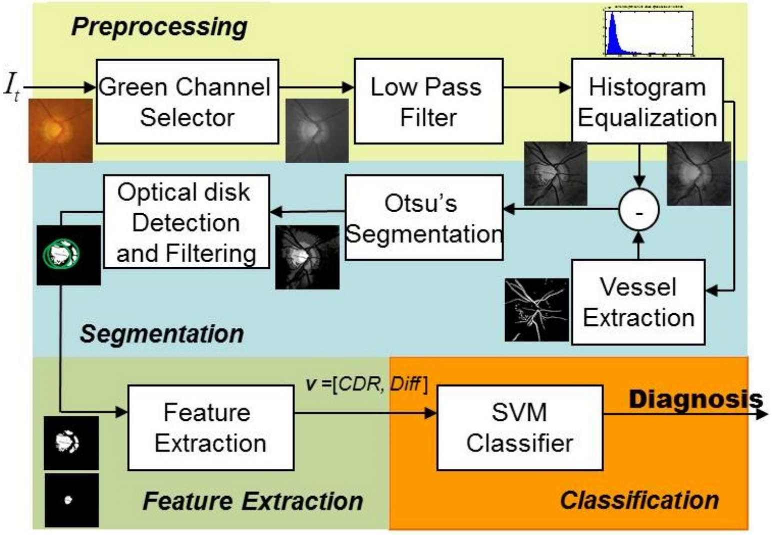


Figure 1: Block diagram of the proposed approach

hundred or thousands- and they allow detection of different levels of glaucoma even in the presence of other ocular pathologies.

1.1 State of the Art

Initial attempts to automatise the glaucoma diagnosis were based on quantitative parameters that can help to make the qualitative assessment more objective, reproducible and lead to a reduction of the observer variability. However, these methods were based on manually annotated ONH images [7, 20], or they lacked robustness and reliability. In response, researchers moved away from the extraction of geometrical features, which rely on good ONH segmentation, towards frameworks based on the pixel appearance of the whole retinal image [8, 16]. These methodologies were inspired by a pipeline previously used for face and object recognition [24] and have the advantage of not needing the segmentation of the ONH. Recently Bock et al. achieved 73% sensitivity and 85% specificity in the detection of glaucoma using a fully-automated analysis of monoscopic photographs [8] based on appearance. However, this is computationally expensive, their classification depends on the camera or machinery involved and they require hundreds of positive and negative glaucoma samples for retraining. The structure of the retina and how glaucoma affects its appearance is much more subtle than the differences between faces or objects for which these frameworks were originally designed. This makes it difficult for systems only based on appearance to detect glaucoma, specially in the early stages.

In the last few years there have been great advances, not only in medical image processing, but also in fundus cameras able to provide high resolution and low-noise retinal images. As a consequence, new studies have been performed showing high accuracy on the segmentation of the ONH characteristics such as disc area, disc diameter or the well established cup-to-disc ratio. Most successful approaches are based on ONH models able to automatically adjust to the image [14, 16, 17], although this implies additional training for model generation. In general, most of these approaches have only been validated on healthy eyes under assumptions that are not valid for glaucomatous eyes. Other approaches, tested on glaucomatous examples, have been validated against human segmentation, but they were not evaluated for diagnosis since they were neither input to a classifier, nor compared against pixel based approaches [12, 13, 15].

In this paper, we propose a computerised glaucoma diagnosis system which relies on the automatic extraction of high level geometrical features related to ONH structures. As a main advantage, it only requires a few training samples and provides high accuracy even in the presence of different stages of glaucoma. The approach is compared against a state-of-art methodology based on pixel appearance.

2 Methodology

Our image processing framework is structured in a standard 4-stage pipeline as depicted in Figure 1: (i) preprocessing, (ii) image-based segmentation, (iii) feature extraction and (iv) classification.

2.1 Preprocessing

Image normalisation is required to correct for variations caused by acquisition and illumination conditions. For this purpose, only the green channel is selected, as it has been shown as the most robust against variations [8]. After that, a low-pass filter [21] is applied to reduce the fine grain noise. Finally, histogram equalisation [21] is applied to ensure consistancy across images, Fig. 2(b).

2.2 Automatic segmentation

After the glaucoma specific preprocessing, the ONH structures needs to be segmented in order to extract the features. Our automatic segmentation methods aims to segment the disc and the cup. First, retinal vessels are located accurately using the Isotropic Undecimated Wavelet Transform (IUWT) and edge location refinement [9], Fig. 2(c). The resulting image is used as a mask to remove blood vessels and facilitate the segmentation of the different image regions, Fig. 2(d). After that, an iterative and multilevel variation of the Otsu's adaptative thresholding [22] is applied, which allows us to identify several different image regions [25]. Given the composition of retinal images, four thresholds are applied to segment the first and second brightest regions, corresponding to the cup and the rim respectively, Fig. 2(e). The morphological operator open is applied to filter noise without changing the feature size.

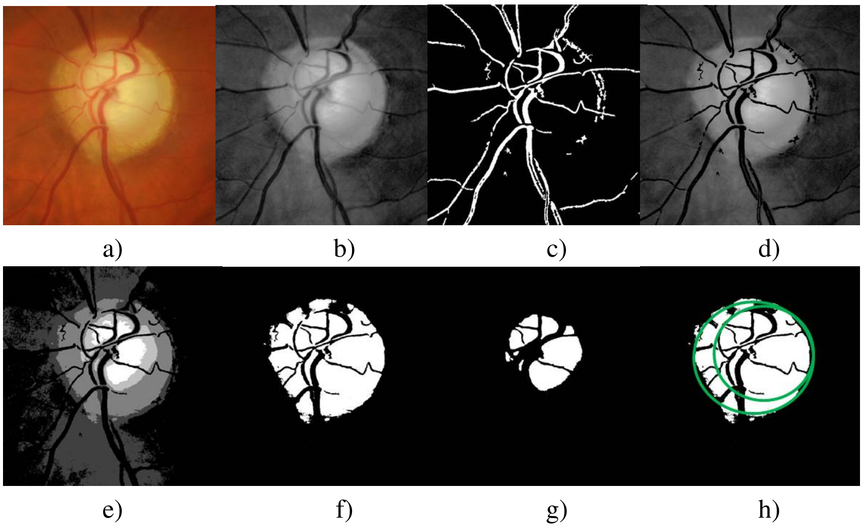


Figure 2: Segmentation process. From left to right, from top to bottom: a)original image, b) preprocessed image, c)Vessel mask, d) Vessel subtracted image, e) segmented image, f) segmented rim, g) segmented cup, h) optical discs candidates.

Although the algorithm gives an accurate segmentation, other image regions can be falsely detected as belonging to these ONH primary structures, especially in the presence of other retinal anomalies. In order to filter those false positives, an optical disc detection algorithm is employed. This detector applies a combination of the Circular Hough Transform with a scale invariant kernel operator, as described in [19], to detect circles within the retinal image. The primary goal is not to provide the geometrical parameters of the ONH, since its assumption about circularity may not be well-matched to the real shape of the ONH and may distort the feature values. Instead, the optical disc candidates are used for removing all those segmented areas outside the detected circles, thus filtering those false positives included by the region segmentation. More than one candidate is allowed, since the goal is not to uniquely identify the optical disc center, but to filter wrongly segmented pixels outside the ONH, Fig. 2(h).

2.3 Feature extraction

We hypothesise that geometric features measured from the segmentation of the disc and rim are of greater value than appearance features in detecting glaucoma. To this aim, two features are extracted and used in our framework:

Cup-to-disc ratio (CDR): The ratio of the vertical diameters of the inner cup and the outer disc rim is commonly used as an indicator of glaucoma likelihood or disease progression [10]. In our pipeline, the ratio is calculated by localising the highest and the lowest pixel in the vertical axis for both the rim and the cup segmented regions (see Fig. 3).

$$CDR = D_{cup}/D_{rim} \tag{1}$$

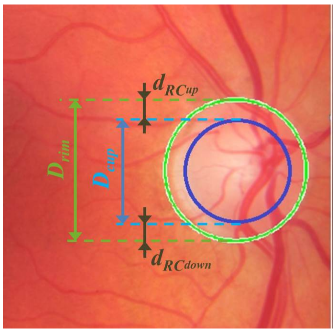


Figure 3: Feature extraction variables

Neuro-retinal rim width variation: The relative width of the neuro-retinal rim at different angular locations is known to differ between normal and glaucomatous discs. Normal subjects have a characteristic distribution, being widest at the inferior part of the disc, followed by decreasing width at the superior, nasal and temporal locations. Glaucomatous eyes typically do not follow such a pattern, which is commonly known as the ”ISN’T rule” [11]. In our system, the upper and lower widths of the rim are calculated as the distance between the highest point of the rim and the highest point of the cup and the distance between the lowest point of the rim and the lowest point of the cup respectively, see Fig. 3. Then, the feature is implemented as the difference between these vertical distances.

$$Diff = d_{RC_{up}} - d_{RC_{down}} \tag{2}$$

The above features are the most common features used by ophthalmologists and therefore likely to provide discriminative information for classification.

2.4 Classification

Although in the past these features have been used directly for diagnosis by applying a set of rules, mimicking what optometrists do manually, this is subjective, depending on the experience of the expert and the demography of the population. It also assumes that features are perfectly extracted, which is not always true due to segmentation difficulties. On the contrary, by feeding the above feature set into a robust classifier, more complex rules can be automatically inferred and deviations produced by segmentation failures accounted for. Therefore, in our implementation, a SVM classifier with linear kernel is used [23]. This linear classifier determines a maximum-margin and soft hyperplane that best separates the considered classes. Data is normalised and transformed via the linear radial basis kernel.

The decision to use a linear classifier is supported by the literature [16, 8], where linear classifiers have reported excellent results for glaucoma diagnosis. The choice of SVM over more traditional approaches such as nearest neighbours, regression, neural networks and discriminant analysis is supported by their reported advantages [23]: they do not require regularity in the data so it can be applied to data following an unknown distribution, it delivers a unique solution since the optimality problem is convex contrary to neural networks, it can be easily extended to non-linear nonparametric problems by replacing the linear kernel, it scales well to high dimensional data, and the trade-off between complexity and error can be controlled explicitly.

Although more complex classification pipelines could be applied in our framework, such as the double schema proposed in [8] or non linear kernels [16], it is not the scope of our paper to state the best classification technique but to prove the validity of geometrical features. Therefore, SVM is the classifier providing the best framework for comparison with other state-of-art methodologies without compromising future improvements of the system.

3 Experimental Results

Two different datasets have been used to validate the experiments and ensure that the conclusions are not dependant on the fundus camera. The first dataset was captured with a stereoscopic camera Kowa nonmyd WX. Only one of the two images provided was used since our goal is to evaluate monoscopic systems for screening, given their broader availability. The dataset is composed of 29 samples, 14 healthy eyes and 15 glaucomatous ones. The second dataset is a standard set, publicly available [18], which facilitates future comparison of our methodology with others. It contains 26 samples, 8 healthy and 18 glaucomatous discs. Both datasets contain different degrees of glaucoma, from very early stages

Table 1: Results over the 2 datasets. Four first rows show the state-of-art appearance features, while three middle rows show different variation of our framework and last two rows the combination of our pipeline with appearance features

	Dataset 1						Dataset 2					
Method	Acc	Sens	Spec	Prec	Recl	Fmes	Acc	Sens	Spec	Prec	Recl	Fmes
Intensity + PCA [8]	0.59	0.60	0.57	0.6	0.6	0.6	0.48	0.55	0.33	0.65	0.55	0.59
FFT + PCA [8]	0.45	0.40	0.50	0.46	0.40	0.43	0.38	0.45	0.22	0.56	0.45	0.50
Spline + PCA [8]	0.59	0.60	0.57	0.60	0.60	0.60	0.59	0.65	0.44	0.72	0.65	0.68
All appearance [8]	0.55	0.53	0.57	0.57	0.53	0.55	0.46	0.50	0.33	0.63	0.50	0.55
CDR (no circle detect.)	0.79	0.73	0.85	0.85	0.73	0.79	0.46	0.50	0.38	0.62	0.50	0.55
CDR	**0.89**	**0.93**	**0.85**	**0.88**	**0.93**	**0.90**	0.59	0.56	0.63	0.75	0.56	0.64
CDR + Diff	0.82	0.87	0.77	0.81	0.87	0.84	**0.71**	**0.69**	**0.75**	**0.85**	**0.69**	**0.76**
Intensity + PCA + CDR	0.69	0.73	0.64	0.69	0.73	0.71	0.50	0.56	0.38	0.64	0.56	0.60
All Appearance + CDR	0.62	0.60	0.64	0.64	0.60	0.62	0.46	0.56	0.25	0.60	0.56	0.58

to severe cases, as well as other disorders, such as hypermetrope, haemorrhages or peripapillary atrophy, that can make diagnosis difficult.

Different variations of our methodology were tested, using only the cup-to-disc ratio, the rim width variation or a combination of both. The circle detector and filter, used to reduce segmentation errors outside the ONH was also evaluated. All the parameters were setup experimentally and kept identical for all experiments and datasets in order to compare the methods in equal conditions and to avoid overfitting to specific cases. In order to validate our method and extract pertinent conclusions regarding the comparison between geometrical and appearance features, different appearance based methodologies -pixel values, fft coefficients B-spline coefficients and a combination of all- were implemented following the description, setup and conclusions by Block et al. [8]. All experiments were performed using leave-one-out cross validation.

Results are shown in Table 1 in terms of accuracy (Acc), sensitivity (Sens), specificity (Spec), precision (Prec), recall (Recl) and F measurement (Fmes). It can be seen how our framework provides accurate glaucoma diagnosis. The extraction and usage of geometrical features seems to provide superior diagnosis accuracy under realistic conditions: when the number of training images is small, they perform much better than appearance based. This explains the significant decrease in performance of appearance based feature compared to other results reported in the literature [16, 8], where hundred of examples were available. Since those features depend heavily on the camera setting, they require retraining for every camera model and therefore they are difficult to deploy in the real world.

Other conclusions can be drawn from these results. The optical disc detection and filtering gives a significant improvement in the final classification. The rim variation feature does not always provide a significant increase in accuracy, especially in the first dataset where the high resolution allows a perfect segmentation of disc and cup, but it plays a significant role for cheaper cameras. The second dataset complexity, with a much lower resolution, is reflected in the final performance of all the tested methods. Finally, by adding geometrical features to the appearance feature vector, results appear to invariably improve, which shows the potential of combining both methodologies with the potential of fully exploiting the advantages of both techniques.

4 Conclusions

In this paper, a method for glaucoma diagnosis, based on ONH segmentation of retinal images, is proposed. Our framework is able to accurately extract the cup and the rim of the optical disc to extract high level geometrical features. The obtained values are then used as input to a machine learning classifier, responsible for detecting glaucoma given a new retinal image.

Our approach has been designed for glaucoma screening in real world conditions. Experiments on varied datasets were performed to evaluate our schema with different cameras and resolutions, and both colour and black and white images. The proposed method achieved high accuracy rates overperforming state-of-art methodologies in real conditions, when small training sets are available. The experiments also validated the usage of geometrical features for glaucoma detection and as a complement to appearance based methods. As future work, a diagnosis study will be performed to ensure the validity of our conclusion in a larger scale and the potential of our framework for glaucoma screening and diagnosis in real life.

References

[1] Burr J, Mowatt G, Siddiqui MAR, Herandez R et al. (2007). The clinical and cost effectiveness of screening for open angle glaucoma: a systematic review and economic evaluation. *Health Technology Assessment*, 11(41).

[2] Tuck MW, Crick RP (2003). The projected increase in glaucoma due to an ageing population. *Ophthalmic and Physiological Optics*, 23(2):175-179.

[3] Harrison RJ, Wild JM, Hobley AJ (1988). Referral patterns to an ophthalmic outpatient clinic by general practitioners and ophthalmic opticians and the role of these professionals in screening for ocular disease. *Br Med J*, 297:1162-7.

[4] Bowling B, Chen SD, Salmon JF (2005). Outcomes of referrals by community optometrists to a hospital glaucoma service. *Br J Ophthalmol*, 89:1102-4.

[5] Patel UD, Murdoch IE, Theodossiades J (1920). Glaucoma detection in the community: does ongoing training of optometrists have a lasting effect?. *Eye*, 591-4.

[6] Vernon SA, Ghosh G (2001). Do locally agreed guidelines for optometrists concerning the referral of glaucoma suspects influence referral practice?. *Eye*, 15:458-63

[7] Schultz RO, Radius RL, Hartz AJ, Brown DB, Eytan ON, Ogawa GS, Kuhn E, Simons KB (1995). Screening for glaucoma with stereo disc photography. *J Glaucoma*, 4:177-82.

[8] Bock R, Meier J, Nyl LG, Hornegger J, Michelson G (2010). Glaucoma risk index: Automated glaucoma detection from color fundus images. *Med Image Anal*, 14:471-81.

[9] Bankhead P, Scholfield CN, McGeown JG, Curtis TM (2012). Fast retinal vessel detection and measurement using wavelets and edge location refinement. *PLoS ONE*, 7(3):1-12.

[10] Armaly MF, Sayegh RE (1969). The cup/disc ratio. The findings of tonometry and tonography in the normal eye. *Arch Opthalmol*, 82:191-6.

[11] Jonas JB, Gusek GC, Naumann GO (1988). Optic disc morphometry in chronic primary open-angle glaucoma. I. Morphometric intrapapillary characteristics. *Graefes Arch Clin Exp Ophthalmol*, 226(6):522-30.

[12] Hatanaka Y, Noudo A, Muramatsu C, Sawada A, Hara T, et al. (2010). Automatic Measurement of Vertical Cup-to-Disc Ratio on Retinal Fundus Images. *Lecture Notes in Computer Science*, 6165:64-72.

[13] Liu J, Wong DWK, Lim JH, Li H, Tan NM, Zhang Z, Wong TY, Lavanya R (2009). ARGALI: An Automatic Cup-to-Disc Ratio Measurement System for Glaucoma Analysis Using Level-set Image Processing. *International Conference on Biomedical Engineering IFMBE Proceedings*, 23:559-562

[14] Fondn I, Nez F, Tirado M, Jimnez S, Alemany P, Abbas Q, Serrano C, Acha B (2012). Automatic Cup-to-Disc Ratio Estimation Using Active Contours and Color Clustering in Fundus Images for Glaucoma Diagnosis, Image Analysis and Recognition. *Lecture Notes in Computer Science*, 7325:390-399

[15] Narasimhan K, Vijayarekha, K (2011). An Efficient Automated System For Glaucoma Detection Using Fundus Image. *Journal of Theoretical and Applied Information Technology*, 33:1

[16] Mookiah MRK, Acharya UR, Lim CM, Petznick A, Suri JS (2012). Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features, *Knowledge-Based Systems*, 33:73-82.

[17] Xu J, Chutatape O, Sung E, Zheng C, Chewteckuan P (2007). Optic disk feature extraction via modified deformable model technique for glaucoma analysis. *Pattern Recognition*, 2063-2076.

[18] Spaeth G (2014). An online resource for ophtalmologists, physicians, medical students and optometrists. Online: http://www.optic-disc.org/ [Last access: January 2014]

[19] Atherton TJ, Kerbyson DJ (1999). Size invariant circle detection. *Image and Vision Computing*, 17(11):795-803.

[20] Wollstein G, Garway-Heath DF, Hitchings RA (1998). Identification of early glaucoma cases with the scanning laser ophthalmoscope. *Ophthalmology*, 105(8):15571563.

[21] Acharya R (2005). *Image Processing: Principles and Applications*, Wiley-Interscience.

[22] Otsu N (1979). A threshold selection method from gray-level histograms. *IEEE SMCB*, 9(1):6266.

[23] Auria L, Moro R (2008). Support Vector Machines (SVM) as a technique for solvency analysis. *Discussion papers*, 811, DIW Berlin.

[24] Turk M, Pentland A (1991). Eigenfaces for recognition. *J. Cognit. Neurosci.*, 3(1):7186.

[25] Sezgin M, Sankur B (2004). Survey over image thresholding techniques and quantitative performance evaluation. *J. Electronic Imaging*, 13(1):146-165.

IMVIP 2014

BIOINSPIRED IMAGING

A Biologically Inspired Framework for Efficient Video Processing

Bryan Gardiner, Sonya Coleman
School of Computing and Intelligent Systems,
University of Ulster, Magee,
BT48 7JL, N. Ireland.
{b.gardiner, sa.coleman}@ulster.ac.uk

Bryan Scotney
School of Computing and Information Engineering,
University of Ulster, Coleraine,
BT52 1SA, N. Ireland.
b.scotney@ulster.ac.uk

Abstract

Improving the computational efficiency of video processing tasks has become a dominant issue with the ultimate goal of real-time processing. One methodology to achieve efficient processing on still images is the use of a hexagonal image representation, which permits efficient implementation of feature extraction. To enhance the efficiency of video processing tasks, this paper presents a hexagonal pixel-based framework for video processing, where a biologically inspired eye tremor approach is used for efficient application of processing algorithms. We demonstrate that this eye tremor approach is significantly faster than the use of conventional spiral convolution or the use of a neighbourhood address look-up table for hexagonal based video processing.

Keywords: Hexagonal imaging, eye tremor, video processing.

1. Introduction

Preliminary research into the human vision system perceived the human eye to be a sensor that closely relates to the operation of a pinhole camera [12]. More recently, it has been concluded that the human visual system is much more sophisticated, operating and functioning like a mini brain [9]. The use of hexagonal grid structures for image representation is inspired by the hexagonal structures of the fovea present in the human vision system. The fovea is responsible for sharp vision capture and is comprised of cones that are shaped and placed in a hexagonal arrangement [6], [13], [20]. Recent work using the hexagonal structure to imitate the human vision system includes biologically inspired fovea modelling [14] and the development of silicon retinas for robot vision [16], [25]. Some examples of work completed on processing images represented on a hexagonal grid include low-level feature extraction [4],[29] and efficient image rotation and translation [19].

Computer vision systems often possess similar characteristics and functionality to those of the human vision system. A camera is the basic sensing element, and its purpose as a visual input device has provided the opportunity for numerous digital processing procedures to assist humans in tasks such as video surveillance, object recognition and motion tracking. Visual tracking in video data can be described as the segmentation of an object from a sequence of video scenes, keeping track of its motion and orientation. Techniques such as video object detection and tracking are the initial steps for more complex processes, such as video context analysis and multimedia indexing [21]. Visual object tracking is closely related to object detection in still images but relies also on the motion characteristics of objects, i.e. the continuity of the object detection over time. The need for real-time object tracking for video analysis exists in many aspects of our daily lives, for instance: surveillance [28], assistive robotics [2] and traffic accident detection [26]. Moving object detection is considered to be the most important task in automated video systems, representing the low level image processing technique that is the basis of automated video analysis [1].

The state of the art in visual object detection has advanced considerably over the last 15 years [5], [1], [17], [22]. Recently, advances in other image processing areas have generated renewed interest in tracking: specifically, progress in the definition of features invariant to various imaging transformations [18], [3], online learning [27], [10], and object detection [8], [15]. In order to process video in an efficient manner, it is not only important to investigate reliable and robust detection algorithms but to investigate the new processes for applying such algorithms to video sequences as fast and efficiently as possible. Hence, this paper presents a hexagonal pixel-based framework for video tracking, demonstrating its capability for reducing computational overheads when processing frame sequences. This efficiency gain is possible for a number of reasons. Firstly the hexagonal lattice has a number of advantages over conventional grid structures [19], for example, equidistance between neighbouring pixels, greater angular resolution and a higher degree of freedom. Secondly, in the fovea of the human eye, where the photoreceptive fields of ganglion cells do not overlap [7], we may use non-overlapping neighbourhoods when processing image or video data. Typically detection algorithms are applied over each complete frame throughout the video sequence and hence use convolution neighbourhoods that overlap. In contrast, we have developed a framework within which the convolution neighbourhoods do not overlap. In addition, the human eye can be subjected to three types of movement: tremor, drift, and micro-saccades [23]. As a consequence of eye tremor - rhythmic oscillations of the eye - the human vision system does not process single static images, but a series of temporal images that are slightly off-set due to these involuntary eye movements. Therefore, we adopt this concept for video processing by off-setting neighbouring frames, each of which is partially processed using non-overlapping convolution neighbourhoods.

Although the framework presented can be used for many video processing algorithms, this paper uses edge detection as the application, as it provides a visually effective demonstration. In previous work [4] the finite element method was used to develop a systematic and efficient design procedure for operators for use with hexagonal images, and these operators will be used as a test bed to demonstrate the efficiency of the proposed framework.

2. Spiral Framework

In the spiral architecture [24] the addressing scheme for the spiral image, denoted by *S*, originates at the centre of the image (pixel index 0) and spirals out using one-dimensional indexing. Figure 1 shows the spiral addressing scheme for the central portion of an image. Pixel 0 may be considered as a layer 0 cluster. Pixel 0, together with its six immediate neighbours indexed in a clockwise direction (pixels 1,…,6) then form a layer 1 cluster centred at pixel 0.

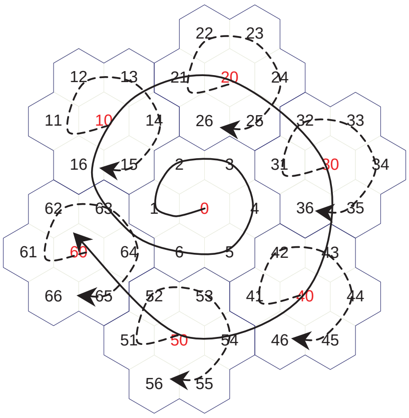


Figure 1: One-dimensional addressing scheme in the central region of the image

This layer 1 cluster may then be combined with its six immediately neighbouring layer 1 clusters, the centres of which are indexed as 10, 20, 30, 40, 50 and 60, to form a layer 2 cluster centred at pixel 0 (as shown in Figure 1); the remaining pixels in each of these layer 1 clusters are indexed in a clockwise direction in the same fashion as the layer 1 cluster centred at 0, (e.g., for the layer 1 cluster centred at 30, the pixel indices are 30, 31, 32, 33, 34, 35 and 36). The entire spiral addressing scheme is generated by recursive use of the clusters; for example, seven layer 2 clusters are combined to form a layer 3 cluster. Ultimately the entire hexagonal image may be considered to be a layer L cluster centred at 0 comprising 7^L pixels.

An important advantage of the spiral addressing scheme is that any location in the image can be represented by a single co-ordinate value, and hence the spiral image can be stored as a vector [5]. Spatially neighbouring pixels within any 7-pixel layer 1 cluster in the image remain neighbouring pixels in the one-dimensional image storage structure. This is a very useful characteristic when performing processing tasks on the stored image vector, and this contiguity property lies at the heart of our approach to achieve fast and efficient processing for visual object detection.

3. Biologically Inspired Framework

In this section a biologically inspired eye tremor approach is presented for efficient application of processing algorithms. This approach involves the development of an eye tremor framework that permits non-overlapping convolution of processing algorithms.

3.1 Simulating Eye Tremor

Following the approach presented in [8] we consider the spiral image I_0 to be the "base" image, corresponding to a particular frame in the video sequence. For the following six frames we denote further images, $I_j, j = 1,...,6$. The location of the origin of each of these frames is offset spatially from I_0 by a distance of one pixel in the image plane along one of the three natural hexagonal axis directions. This mechanism simulates the phenomenon of "eye tremor". In each image $I_j, j = 1,...,6$, the pixel with spiral address "*0*" represents the same spatial location in the scene as the pixel with spiral address "*j*" in I_0. (We are assuming that the camera is static.) The "centre" (i.e., the pixel with spiral address zero) of each image $I_j, j = 0,...,6$, is thus located at a pixel within the layer $\lambda = 1$ neighbourhood centred at the pixel with spiral address "0" in image I_0, as shown in Figure 2.

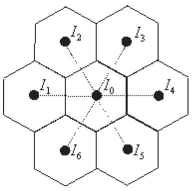


Figure 2. The 7 image centres in the eye tremor approach

Through use of the spiral architecture for pixel addressing, it is assumed that image I_0 is stored in a one-dimensional vector (with base-7 indexing). Using the spiral architecture the following frames $I_j, j = 1,...,6$, are stored similarly.

3.2 Non-Overlapping Convolution

For a given image I_0, convolution of the operator H_1 across the entire image plane is achieved by applying the operator sparsely to each of the seven frames $I_j, j = 0,...,6$ and then combining the resultant outputs. Figure 3 shows a sample of pixels in image I_0 for which the label $j = 0,...,6$ for each pixel indicates in which of the images $I_j, j = 0,...,6$, the pixel address takes the value 0 *mod* 7. Each pixel in image I_0 may be thus uniquely labeled.

4. Performance evaluation

For initial testing of the proposed eye tremor framework, application of a 7-point hexagonal edge detection operator (previously developed in [11]), with mask values shown in Figure 4, is applied to a video sequence of resolution 160x120 pixels, sampled at 30fps. The video comprises of a man walking across the camera view from right to left. For comparison, the same operator is then applied to the same video sequence using the standard spiral convolution framework and the spiral

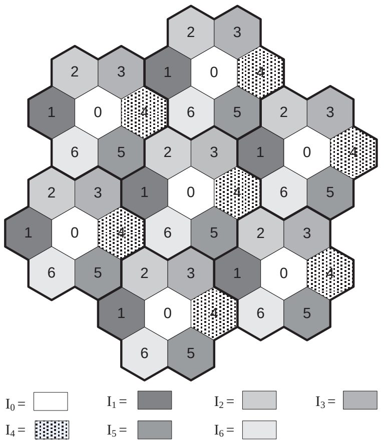


Figure 3. Pixel positions in image I_0 corresponding to pixels $I_j, j = 0,\dots,6$ with address 0 *mod* 7.

convolution framework that utilises a look-up table to compute neighbourhood pixel addresses. The look-up table (LUT) which stores the pixel neighbour addresses takes 153ms to generate, but is significantly faster than using standard hexagonal addressing, which requires *mod* 7 arithmetic. The LUT is an alternative to computing the nodal addresses within a neighbourhood by using hexagonal arithmetic, which is very computationally expensive. The LUT approach effectively pre-computes and stores the indices for all of the 7-pixel neighbourhood clusters.

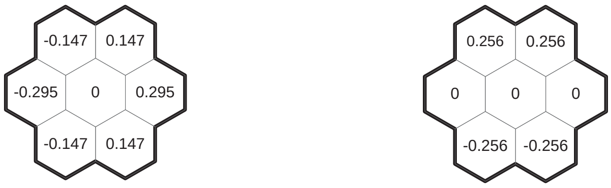


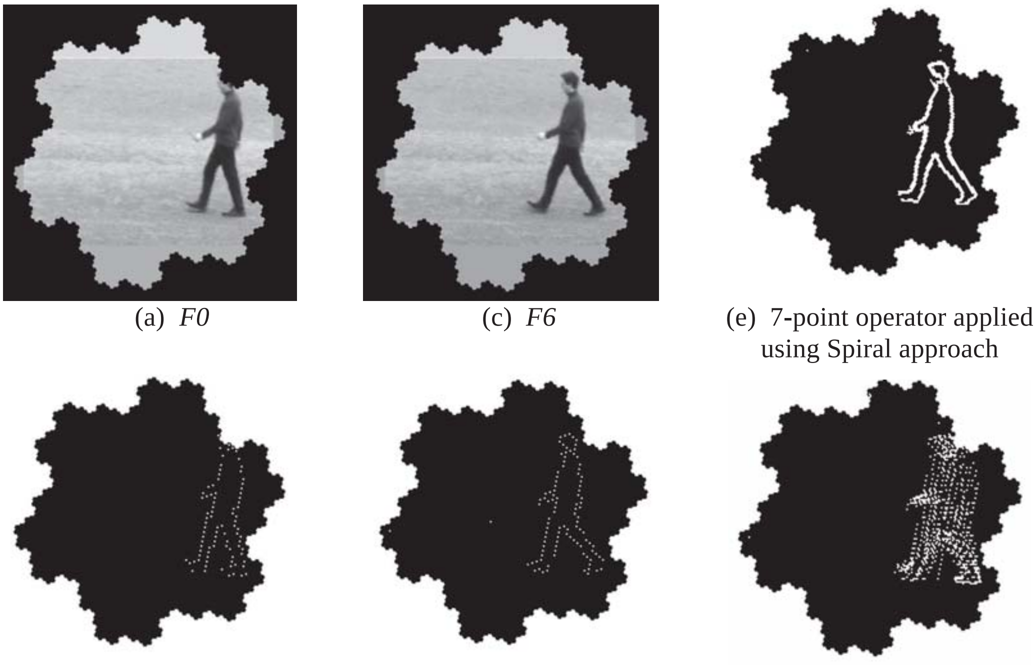
Figure 4: *x*- and *y*-components of hexagonal operator developed in [11]

In Table 1 we present run times for application of the feature extraction operator using the different frameworks. Processing times are computed on a Pentium dual-core workstation using unoptimised C++ code. Each image is processed 100 times and the average runtime is determined. Using the eye tremor approach, the time taken to apply the operator to the first 7 frames and combining these partial edge maps to generate a complete edge map takes 1.114ms. It is important to note that this process need be computed only once for any video sequence. The time taken to apply the operator to one additional frame and combine this with the 6 previously obtained partial edge maps to generate the next complete edge map takes 0.190ms. When comparing these results with the other approaches in Table 1, we see that our biologically motivated approach is approximately 10 times faster than spiral LUT convolution, and 480 times faster than standard spiral convolution.

Table 1: Algorithm runtimes for feature extraction

Method	Runtime
Standard spiral convolution	91.278ms
Spiral convolution using LUT	1.862ms
Biologically motivated “eye tremor” approach	0.190ms

Figures 5 (a) and (c) show two sample frames, the first frame F_0, the seventh frame F_6 respectively. Corresponding sparse edge maps obtained by the proposed eye tremor framework are presented in Figure 5 (b) and (d). Figure 5(e) shows the resultant edge map when the operator is applied conventionally to F_6 in the image sequence, and Figure 5(f) shows the combined partial edge maps for the first 7 frames using the eye tremor approach. Visual results demonstrate promise that applying object tracking algorithms based on the proposed framework will perform well whilst



(a) *F0* (c) *F6* (e) 7-point operator applied using Spiral approach

(b) *S0* (d) *S6* (f) 7-point operator applied using eye-tremor approach

Figure 5: Frames and corresponding edge maps represented on spiral hexagonal structure

significantly reducing computational overhead for processing video data in real time. Having applied the core 7-point operator to each of the images, we may combine the outputs to form a complete edge map. To reiterate, this process need be computed only once for any video sequence. It is only necessary to apply the operator to one additional frame and combine this with the 6 previously obtained partial edge maps to generate the next complete edge map, which takes 0.190ms. In terms of implementation using the one-dimensional vector structure for the images $I_j, j = 0,\dots,6$, each output response $D_\lambda^j, j = 0,\dots,6$ is stored in a one-dimensional vector with non-empty values corresponding to the array positions with indices 0 *mod* 7. These one-dimensional vectors may then be assembled according to the "shifted" structure as illustrated in Figure 6: $\forall s_0 \in \{s | s = 0\, mod\, 7\}$, $E_\lambda(s_0 + k) = D_\lambda^k(s_0)$ for $k = 0,\dots,6$ to yield the consolidated output image $E_\lambda(I_0) = H_\lambda \otimes I_0$ as shown in Figure 7.

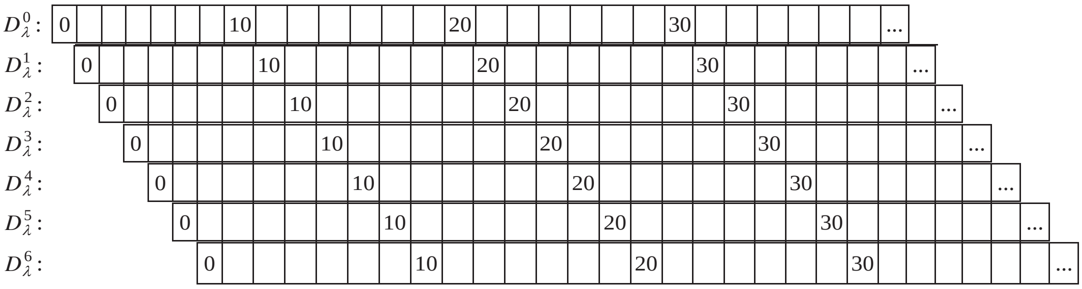

D_λ^0: 0 10 20 30 ...
D_λ^1: 0 10 20 30 ...
D_λ^2: 0 10 20 30 ...
D_λ^3: 0 10 20 30 ...
D_λ^4: 0 10 20 30 ...
D_λ^5: 0 10 20 30 ...
D_λ^6: 0 10 20 30 ...

Figure 6. Assembly of the one-dimensional vectors $D_\lambda^j, j = 0,\dots,6$

$D_\lambda^0(0)$	$D_\lambda^1(0)$	$D_\lambda^2(0)$	$D_\lambda^3(0)$	$D_\lambda^4(0)$	$D_\lambda^5(0)$	$D_\lambda^6(0)$	$D_\lambda^0(10)$	$D_\lambda^1(10)$	$D_\lambda^2(10)$	$D_\lambda^3(10)$	$D_\lambda^4(10)$	$D_\lambda^5(10)$	$D_\lambda^6(10)$	$D_\lambda^0(20)$	...	...

Figure 7. Consolidated output image resulting from assembly of the vectors in Figure 6

5. Summary

In this paper we present a biologically inspired approach to fast video processing. Using spiral addressing within a hexagonal framework, each frame in a sequence can be off-set slightly from its adjacent frames in a cyclic pattern and a non-overlapping convolution can be applied. Once the first seven *a-trous* frames are processed to generate a complete edge map, the addition of each subsequent *a-trous* frame will generate a new complete edge map from the previous 6 processed frames. Hence each subsequent edge map will be generated in one seventh of the time. The results presented in this paper demonstrate that with only minimal degradation of accuracy, this biologically motivated framework is significantly faster than both the standard and LUT spiral convolution for video processing.

References

[1] Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: IEEE CVPR. pp.798–805, 2006.

[2] Agravante, D.J., Cherubini, A., Bussy, A., Gergondet, P., Kheddar, A., “Collaborative Human-Humanoid Carrying Using Vision and Haptic Sensing”, Int. Conf on Robotics and Automation, 2014.

[3] Bay, H., Tuytelaars, T., Gool, L.: Surf: Speeded up robust features. In: ECCV, pp. 404–417, 2006.

[4] Coleman, SA, Gardiner, B, and Scotney, BW, “Adaptive Tri-Directional Edge Detection Operators based on the Spiral Architecture” IEEE ICIP, pp. 141-144, 2010.

[5] Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: IEEE CVPR. pp. 2142–2146, 2000.

[6] Curcio, C., Sloan, K., Kalina, R., & Hendrickson, A.. Human Photoreceptor Topography. The Journal of comparative neurology , Vol. 292 (4), pp. 497-523, 2004.

[7] Dacey DM, Packer OS, “Receptive Field Structure of h1 Horizontal Cells in Macaque Monkey Retina” Journal of Vision, 2(4), 279-292, 2002.

[8] Dalal, N., Triggs, B.: “Histograms of oriented gradients for human detection” CVPR, pp.886–893, 2005

[9] Descartes, R. (2001). Discourse on Method, Optics, Geometry, and Meteorology. Hackett Publishing.

[10] Dietterich, T., Lathrop, R., Lozano-P´erez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89, pp. 31–71, 1997.

[11] Gardiner, B, Coleman, SA and Scotney, BW (2007) A Design Procedure for Gradient Operators on Hexagonal Images. In: International Machine Vision and Image Processing Conference (IMVIP), Maynooth, Ireland. IEEE Computer Society, 2007.

[12] Helmholtz, H. V., & Southall, J. (1962). Helmholtz's treatise on physiological optics. Dover Publications.

[13] Hirsch, J., & Miller, W. (1987). Does Cone Positional Disorder Limit Resolution? Journal of the Optical Society of America A: Optics, Image Science, and Vision , Vol. 4 (8), pp. 1481-1492.

[14] Huang, C., & Lin, C. Bio-inspired Computer Fovea Model based on Hexagonal-type Cellular Neural Network. IEEE Transactions on Circuits and Systems I: Regular Papers , Vol. 54 (1), pp. 35-47, 2007.

[15] Lampert, C., Blaschko, M., Hofmann, T.: Efficient subwindow search: A branch and bound framework for object localization. IEEE PAMI 31, pp. 2129–2142, 2009.

[16] Lau, D., & Ulichney, R. Blue-noise Halftoning for Hexagonal Grids. IEEE Transactions on Image Processing , Vol. 15, pp. 1270-1284, 2006.

[17] Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans.

[18] Lowe, D.: Object recognition from local scale-invariant features. In: ICCV, 1999.

[19] Middleton L, Sivaswamy J, “Hexagonal Image Processing; A Practical Approach”, Springer 2005.

[20] Mollon, J., & Bowmaker, J. The Spatial Arrangement of Cones in the Primate Fovea. Nature , Vol. 360 (6405), pp. 677-679, 1992.

[21] Nema, Rajni, and A. K. Saxena. "Modified Approach for Object Detection in Video Sequences." American Int Journal of Research in Science, Technology, Eng & Mathematics, pp. 122-126, 2013.

[22] O ̈zuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. IEEE Trans. Pattern Anal. Mach. Intell. 32, pp. 448–461, 2010.

[23] Roka A, et al., “Edge Detection Model Based on Involuntary Eye Movements of the Eye-Retina System”, Acta Polytechnica Hungarica, 4(1), 31-46, 2007.

[24] Scotney, BW, Coleman, SA and Gardiner, B, “Biologically Motivated Feature Extraction Using the Spiral Architecture” IEEE ICIP, pp. 221-224, 2011.

[25] Shimonomura, K., Kushima, T., & Yagi, T. Neuromorphic Binocular Vision System for Real-time Disparity Estimation. IEEE Int Conference on Robotics and Automation, pp. 4867-4872, 2007.

[26] Tai, Jen-Chao, et al. "Real-time image tracking for automatic traffic monitoring and enforcement applications." Image and Vision Computing 22.6: 485-501, 2004.

[27] Viola, P., Platt, J., Zhang, C.: Multiple instance boosting for object detection. In: NIPS., 2005.

[28] Wang, Xiaogang. "Intelligent multi-camera video surveillance: A review." Pattern recognition letters 34.1: pp. 3-19, 2013

[29] Wu, Q., He, X., & Hintz, T. (2005). Bi-Lateral Filtering Based Edge Detection on Hexagonal Architecture. Acoustics, Speech, and Signal Processing, Proceedings , Vol. 2, pp. 713-716.

Modelling and Analysis of Retinal Ganglion Cells with Neural Networks

Dermot Kerr, Sonya Coleman, Martin McGinnity
Intelligent Systems Research Centre, School of Computing and Intelligent Systems,
Faculty of Computing and Engineering, University of Ulster at Magee,
Northern Ireland, BT48 7JL.
{d.kerr, sa.coleman, tm.mcginnity}@ulster.ac.uk

Abstract

Modelling biological systems is difficult due to insufficient knowledge about the internal components and organisation, and the complexity of the interactions within the system. At cellular level existing computational models of visual neurons can be derived by quantitatively fitting particular sets of physiological data using an input-output analysis where a known input is given to the system and its output is recorded. These models need to capture the full spatio-temporal description of neuron behaviour under natural viewing conditions. At a computational level we aspire to take advantage of state-of-the-art techniques to accurately model non-standard types of retinal ganglion cells. Using neural network techniques we model the highly complex neuronal structures of visual processing retinal cells and represent the mapping between perception and response automatically.

Keywords: Retinal Ganglion Cells, Linear-Nonlinear Model, Neural Network

1 Introduction

Modelling biological systems is difficult due to insufficient knowledge about the internal components and organisation, and the complexity of the interactions within the system. System identification has emerged as a viable alternative to classical hypothesis testing for the understanding of biological systems and was first used to understand the responses of auditory neurons [De Boer, 1968]. Using white noise stimuli as input, the output responses were recorded and inferences made on mapping the stimulus to the response. White noise stimulation is often selected to model biological vision systems [Sakai, 1988, Chichilnisky, 2001] as it is mathematically simple to analyse. However, it is unlikely that white noise stimuli would test the full function of a neuron's behaviour [Talebi, 2012]. Thus, any model developed with this stimulus could only be considered a subset of the biological model under certain conditions.

In the work by [Marmarelis, 1972], the Wiener theory of nonlinear system identification was applied to study the underlying operation of the three stage neuronal structures in the catfish retina. Following from this work, the Volterra-Wiener method has been used extensively to model nonlinear biological systems [Victor, 1977, 1979, Marmarelis, 2004, Korenberg, 1996]. However, computational effort increases geometrically with the kernel order and in interpretation of higher order kernels [Herikstad, 2011]. Marmarelis and Zhao [Marmarelis, 1997] presented a way of overcoming these limitations by developing a perceptron type network with polynomial activation functions.

Block-structured [Giri, 2010] or modular models in the form of cascaded or parallel configurations have been used to overcome the limitations of Volterra-Wiener models. Cascade models may take various forms such as linear-nonlinear [Ostojic, 2011], nonlinear-linear, linear-nonlinear-linear, etc. In particular, linear-nonlinear models have been used to describe the processing in the retina [Pillow, 2005]. The generalised modular model proposed by [Korenberg, 1991] employed parallel linear-nonlinear cascades generating spike outputs with a threshold-trigger function. To model specific neuron responses such as burstiness, refractoriness and gain control, [Pillow, 2008] amended the linear-nonlinear models with feedback terms. Correlated neuron activity was modelled through the use of coupling filters [Pillow, 2008] to couple multiple linear-nonlinear models of individual cells.

Neural network approaches have also been used to model biological aspects of the vision system. For example [Lau, 2002] used a two layer neural network with the backpropagation training algorithm to model the nonlinear responses of neurons in the visual cortex to visual stimuli. Similarly, [Prenger, 2004] used a multilayer feed-forward neural network to model the nonlinear stimulus-response relationship in the primary visual cortex using natural images. In this paper we show how time-delay neural networks may be used to model the retinas early visual processing system by modeling the biological input-output coupling.

2 Neuronal Data

Recordings were obtained from isolated mice retinas under full field stimulation. The stimulus used is a temporal sequence consisting of an intensity value pseudorandomly selected every 20 ms using a Gaussian white noise sequence. An example temporal sequence is illustrated in Figure 1.

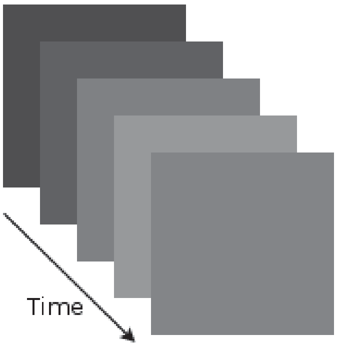


Figure 1. Full-field temporal stimuli sequence generated with Gaussian white noise.

The isolated retina was placed on a multi-electrode array, which recorded spike trains from many ganglion cells simultaneously. Stimuli were projected onto the isolated retina via a miniature cathode ray tube monitor. Spikes were sorted off-line by a cluster analysis of their shapes, and spike times were measured relative to the beginning of stimulus presentation. In the experiments presented in this paper we analyse the response from a temporal retinal ganglion cell (RGC).

3 Neural Network Structure and Experiments

Among all the available techniques and methods to model input-output relationships Artificial Neural Networks (ANN) offer a desirable solution in terms of accuracy. Artificial neural networks can be applied to time series modelling without assuming a priori function forms of models. Many varieties of neural network techniques including Multilayer Feed-forward Neural Network [FFNN], Recurrent Neural Network (RNN), Time delay Neural Network and Nonlinear Autoregressive eXogenous Neural Network (NARX) have been proposed, investigated, and successfully applied to time series prediction. Multilayer FFNN is the most common Neural Network used in prediction and RNN is basically a FNN with a recurrent loop, where the output signals are fed back to the input. NARX are a combination of FFNN, RNN, and time delays.

Artificial Neural Networks have been previously used to model biological aspects of the vision system [Lau 2002, Prenger 2004]. ANNs have the advantage over other techniques of a fast and simple implementation. However, this advantage has to be balanced against the weakness that the obtained mapping is opaque, and not easily analysed. Hence, such an opaque ANN model of the neurons stimulus-response relationship becomes less useful for understanding the underlying neuronal architecture and structure. Even so, we use such techniques to model neuronal behaviour with artificial visual scenes, and to represent the mapping between stimulus and response.

In particular, we use Time delay Neural Network (TDNN) to model the stimulus-response relationship as these have tuneable nonlinearities, interconnectivity structures and additive/multiplicative synapses that may represent the functional behaviour of complicated retinal circuitry. We use a fully recurrent network that is a network of artificial neurons, each with a direct connection to every other artificial neuron. Each neuron has a time-varying real-valued activation function and each connection has a modifiable real-valued weight. Some of the artificial neurons are called input nodes, some output nodes, the remainder are hidden nodes. In supervised training of TDNNs, one starts with teacher data (or training data): empirically observed or artificially constructed input-output time series, which represent examples of the desired model behaviour. The training data is used to train a TDNN such that it approximately reproduces the training data so that the TDNN then generalizes to novel inputs.

3.1 Data Pre-processing

The overall goal of the pre-processing stage is to manipulate the data so that they form a regression dataset, i.e. input-output corresponding to the stimulus-response. In this case the dataset will be single-input single-output. The Gaussian white noise stimulus is a stochastic highly interleaved stimuli spanning a wide range of visual inputs, is relatively robust to fluctuations in responsivity, avoids adaptation to strong or prolonged stimuli and is well suited to simultaneous measurements from multiple neurons. Examples of stimuli are presented in Figure 1 where each image in the sequence is presented sequentially to the isolated retina. As the stimulus has uniform intensity there is no need to extract the stimulus in the region of the receptive field.

Recordings of the ganglion cell neural response (spikes) to the full-field stimulation were supplied for two different ganglion cells in the case of this dataset. Each file contains the recorded times of spikes in seconds. For example, [1.76304, 1.76912, 1.78504,…,546.63776]. Using these recorded spike times we compute a continuous temporal spike rate using the standard method of binning and convolution with a window function. Using this method we then have a continuous valued input-output dataset. For example, in Figure 2 we have illustrated the input data (stimulus intensity), recorded spikes and also computed the spike rate using a number of different window functions for 1000ms of a retinal recording session.

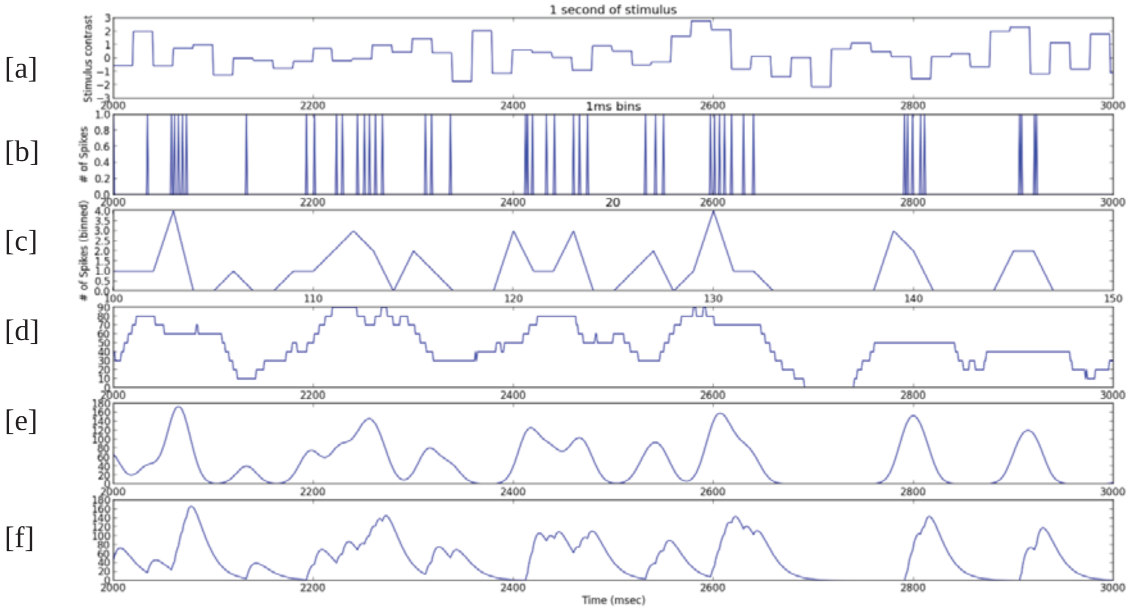


Figure 2. (a) Temporal stimulus intensity, (b) recorded spikes, and (c-f) computed spike rate

Figure 2(a) illustrates the stimulus intensity, Figure 2(b) illustrates the recorded spikes, Figure 2(c) illustrates the spike rates computed using fixed binning, Figure 2(d) illustrates the spike rate computed using a sliding rectangular window, Figure 2(e) illustrates the spike rate computed using a sliding Gaussian window, and Figure 2(f) illustrates the spike rate computed using a half wave rectified α function.

After this pre-processing stage we can use the TDNN method to obtain a predictive model that models the ganglion cells spike rate based on the input stimulus intensity where the spatially uniform stimulus intensity is used as input (Figure 2(a)) and the computed spike rate (Figure 2(f)) in this case) is used as output.

3.2 Experiments and Results

Here we illustrate how the various neural networks within the layered retina structure can be modelled using a TDNN that incorporates the neuron's nonlinear behaviour and dynamics. The proposed approach represents a decisive departure from current methods of generating retina models such as the Linear-Nonlinear model. We propose to model the neuron’s behaviour with artificial visual scenes, and to represent the mapping between perception and response automatically, using the TDNN approach. Using the pre-processed dataset described in 3.1 we then create a time delay neural network with 10 neurons in the hidden layer. The network incorporates a time delay considering the previous 10 time-steps, i.e. 200ms (see Figure 3). This neural network was implemented in Matlab using the neural network toolbox. The network is trained with a dataset that contains neuronal recordings from a retinal ganglion cell recorded over a period of approximately 9 minutes. Training the network take approximately 10 minutes on a high-powered workstation.

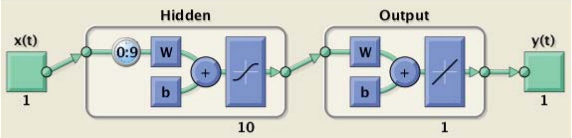


Figure 3. Time Delay Neural Network Structure

Once the neural network has been trained we then use a novel test stimulus sequence to evaluate the performance of the time-delay TDNN and compare to the actual neuronal response to the test stimulus. Results are presented in Figure 4. The plot in the top-left illustrates the actual spike rate (blue) and the TDNN model predicted spike rate (green) compare visually well. The RMSE shows the overall performance for the test stimulus.

To provide further comparison for the NARMAX models we evaluate against a standard benchmark by computing the Linear-Nonlinear (LNL) model [Ostojic, 2011]. The first stage in computing the Linear-Nonlinear model is to compute the spike triggered average (STA) which is the average stimulus preceding a spike. The second stage in the Linear-Nonlinear model is used to re-construct the ganglion cells nonlinearity by plotting the actual response against the STA predicted response, binning the values and fitting a curve using a cumulative density function. For full details of this process please see [Ostojic, 2011]. Next, we apply the STA and nonlinearity to the same test stimulus used previously and compute the response. Results are presented in Figure 5.

Visual comparison illustrates that the TDNN approach results in a closer fitting model when compared to the LNL approach. This can be measured quantitatively by measuring the RMSE which is 32.29 and 49.98 for the TDNN and the LNL model respectively.

4 Discussion

Modelling biological systems is difficult due to insufficient knowledge about the internal components and organisation, and the complexity of the interactions within the system.

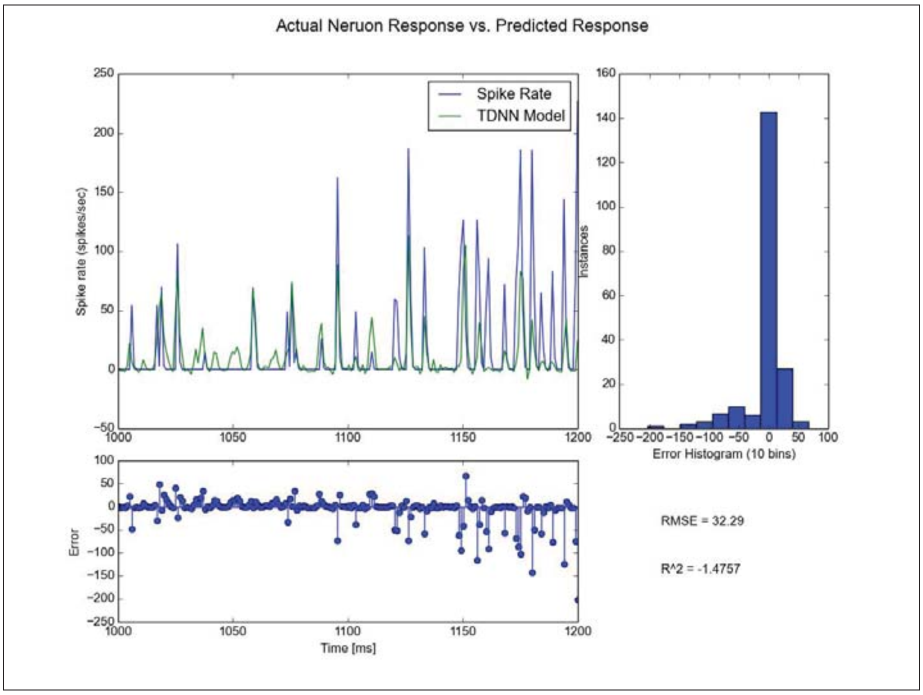


Figure 4. Comparison of TDNN model and actual neuron response to novel test stimulus sequence.

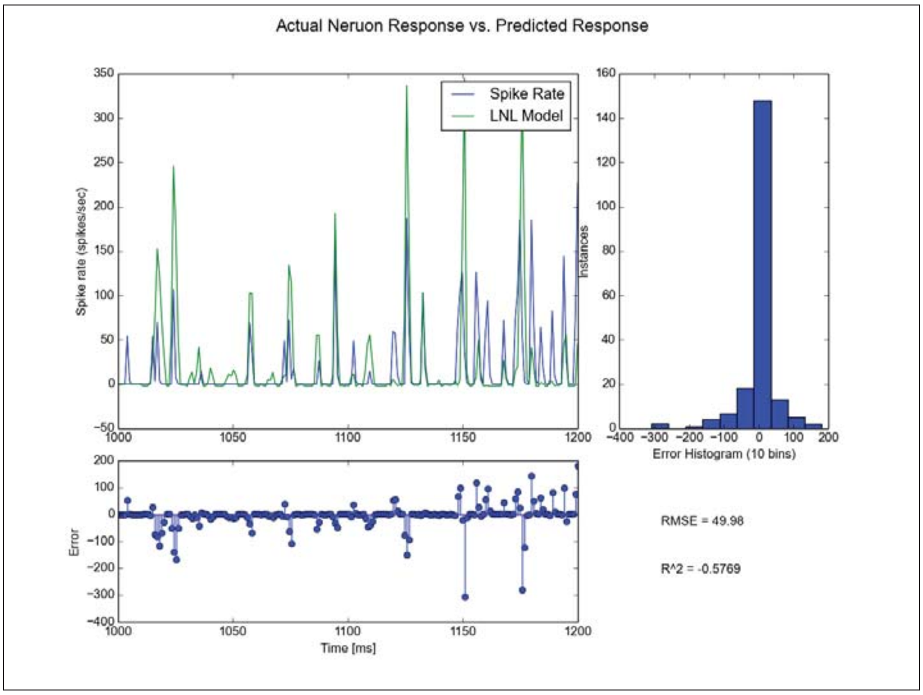


Figure 5. Comparison of LNL model and actual neuron response to novel test stimulus sequence

Existing computational models of visual neurons can be derived by quantitatively fitting particular sets of physiological data using an input-output analysis where a known input is given to the system and its output is recorded as illustrated in the Linear-Nonlinear approach. At a computational level we have presented the use of TDNN methods to accurately model individual retinal ganglion cells as shown in Figure 4. We have presented a comparison of the actual neuronal response and the predicted neuronal response and a comparison with the Linear Nonlinear approach (see Figure 5).

Using TDNN to express the biological input-output coupling mathematically we have modelled highly complex neuronal structures, and modelled ganglion cell behaviour with visual scenes. The next stage in this work will be to increase the complexity of the stimulus by having spatially varying stimuli; we have already started to test the effectiveness of this using the natural image sequences.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7-ICT-2011.9.11] under grant number [600954] [“VISUALISE"]. The experimental data contributing to this study have been supplied by the “Sensory Processing in the Retina" research group at the Department of Ophthalmology, University of Göttingen as part of the VISUALISE project.

References

[Herikstad, 2011] Herikstad, R., Baker, J., Lachaux, J.-P., Gray, C. M., & Yen, S.-C. [2011]. Natural Movies Evoke Spike Trains with Low Spike Time Variability in Cat Primary Visual Cortex. Journal of Neuroscience, 31[44], 15844-15860. doi:10.1523/JNEUROSCI.5153-10.2011

[DeBoer, 1968] De Boer, ., Kuyper, P. [1968]. “Triggered Correlation”. Biomedical ngineering, vol.BME-15, no.3, pp.169-179. doi: 10.1109/TBME.1968.4502561 Transactions on ,

[Sakai, 1988] Sakai,H.M., Naka K.I., Korenberg, M.J. [1988] "White-noise analysis in visual neuroscience". Visual Neuroscience, 1, pp 287-296 DOI: 10.1017/S0952523800001942

[Chichilnisky, 2001] Chichilnisky EJ [2001] A simple white noise analysis of neuronal light responses. Network 12[2]:199-213.

[Talebi, 2012] Talebi, V., Baker, C.L. [2012]. "Natural versus Synthetic Stimuli for Estimating Receptive Field Models: A Comparison of Predictive Robustness". The Journal of Neuroscience, Vol. 32, No. 5., pp. 1560-1576, doi:10.1523

[Marmarelis, 1972] Marmarelis, P.Z., Naka, K.I. [1972]. White-noise analysis of a neuron chain: An application of the wiener theory. Science 175, 1276–1278

[Victor, 1977] Victor, J., Shapley, R., Knight, B. [1977]. Nonlinear analysis of cat retinal ganglion cells in the frequency domain. Proc. Natl. Acad. Sci. U.S.A. 74[7], 3068–3072

[Victor, 1979] Victor, J. [1979] Nonlinear systems analysis: comparison of white noise and sum of sinusoids in a biological system. Proc. Natl. Acad. Sci. U.S.A. 76[2], 996–998

[Marmarelis, 2004] Marmarelis, V. [2004] Nonlinear Dynamic Modeling of Physiological Systems. Wiley Interscience, Hoboken.

[Korenberg, 1996] Korenberg, M., Hunter, I. [1996]. The identification of nonlinear biological systems: Volterra kernel approaches. Ann. Biomed. Eng. 24[2], 250–268.

[Marmarelis, 1997] Marmarelis VZ, Zhao X. [1997]. Volterra models and three-layer perceptions. IEEE Trans Neural Networks 8:1421.

[Giri, 2010] Block-oriented Nonlinear System Identification [2010], Lecture Notes in Control and Information Sciences, Springer Berlin / Heidelberg, Vol. 404. Giri, F. and Bai E.W. Eds

[Ostojic, 2011] Ostojic S, Brunel N [2011] From Spiking Neuron Models to Linear-Nonlinear Models. PLoS Comput Biol 7[1]: e1001056. doi:10.1371/journal.pcbi.1001056

[Pillow, 2005] Pillow JW, Paninski L, Uzzell VJ, Simoncelli EP, Chichilnisky EJ. [2005]. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. J Neurosci.23;25[47]:11003-13.

[Korenberg, 1991] Korenberg MJ. [1991]. Parallel cascade identification and kernel estimation for nonlinear systems. Ann Biomed Eng 19:429.

[Pillow, 2008] Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, Simoncelli EP. [2008] Spatio temporal correlations and visual signaling in a complete neuronal population. Nature 454: 995-999

[Lau, 2002] Lau B, Stanley GB, Dan Y [2002] Computational subunits of visual cortical neurons revealed by artificial neural networks. Proc Natl Acad Sci U S A 99: 8974–8979.

[Prenger, 2004] Prenger, R., Wu, M.C.K., David, S.V., Gallant, J.L., [2004] Nonlinear V1 responses to natural scenes revealed by neural network analysis, Neural Networks, Volume 17, Issues 5–6, Pages 663-679, 10.1016/j.neunet.2004.03.008.

IMVIP 2014

VIDEO PROCESSING

Comparison of Bit-Depth and Resolution Sampling Effects in Video-based Driving Simulation

Michael Brogan
Department of Engineering
IT Blanchardstown
Dublin 15, Ireland
michael.brogan@itb.ie

Charles Markham
Department of Computer Science
NUI Maynooth
Co. Kildare, Ireland
charles.markham@nuim.ie

Sean Commins
Department of Psychology
NUI Maynooth
Co. Kildare, Ireland
sean.commins@nuim.ie

Catherine Deegan
Department of Engineering
IT Blanchardstown
Dublin 15, Ireland
catherine.deegan@itb.ie

Abstract

This paper presents a comparison of the effect of varying both the bit-depth and resolution of a real-world route video sequence with a control video of 1920x1080 resolution and 24-bit bit-depth. A video was acquired of a regional road in Ireland, using an off-the-shelf witness camera. The acceleration and deceleration effects of the acquisition vehicle's speed were removed using the GPS data acquired by the camera. The video was then processed to generate four variants; two variants using a sampled bit-depth, and two variants using a sampled resolution. These videos were then integrated with a driving simulator, allowing the user to control the speed at which a video was played back to the user using the simulator's control pedals. The videos describe a route that is approximately 4 km long. The average driver speed associated with each of these sampled videos is then compared with the average driver speed of the control video. Previous research has shown that the type of simulator display setup can have an effect on long-term post-simulation accident statistics. The results of this paper show that both the bit-depth and resolution of the video in a video-based driving simulator can affect driver speed.

.

Keywords: Video-based driving simulation, bit-depth, resolution, speed

1 Introduction

Fidelity is defined as the faithfulness with which something is reproduced [1]. The auditory, proprioceptive and vestibular elements that a driver experiences have each been replicated to realistic levels of fidelity [2][3]. A high-fidelity representation of a visual cue stream can be acquired using a video camera, although the use of videos in driving simulation is rare. Even though the visual cue stream is the element that delivers the greatest amount of environmental information to a driver, the replication of this is limited to graphical recreations of roads and environments in driving simulators [4]. Other research into introducing a photo-based visual cue stream into driving simulation has produced systems capable of delivering photo-textured three-dimensional environments, although, to date, no data has been presented on the effects of increased visual fidelity on driver behavior. The amount of visual information delivered to a driver that must be processed by the driver's cognitive abilities can be defined as the visual cognitive load. A scene with a large amount of visual information can therefore be considered to be of a higher cognitive load than a scene with a lower amount of visual information [5].

The effect of video quality on user engagement and quality of experience are well-documented, particularly with reference to Internet streaming and downloading [6][7][8]. It has been shown that bit-rate; frame-rate and audio quality can be degraded with little effect on the quality of user experience [9], although such studies are concerned more with the qualitative measures as opposed to the quantitative effects of video quality sampling.

The type of driving simulation display has been shown to have an effect on the performance of driver training, with research showing that drivers trained using a triple-monitor display are involved in fewer road traffic accidents when compared to those trained on a single display [10].

Changing the display setup when using graphical models is straightforward, as the resolution of the virtual world can be updated to reflect the new setup. When dealing with video however, the resolution of the visual stream can change dependent on the new setup. For example, a 1920x1080 resolution video can playback at its native resolution on a single screen, but is sampled when played back on a triple 1920x1080 resolution setup, or on a 1280x1024 resolution display.

This paper describes a comparison on the effect on driver response of sampling the bit-depth of the video, with the effect of sampling the resolution of the video. The control video was acquired along a rural route using an off-the-shelf witness camera, that acquires video data in 1920x1080 resolution at a frame rate of 29.97003 frames per second (fps), alongside global positioning data at a rate of 1 Hz [11].

This paper is divided into seven sections; section 1 gives an introduction to the topic. Section 2 provides an overview of the effects of reduced cognitive loads on drivers in real vehicles. Section 3 details the implementation of the driving simulator and integration of the video datasets. Section 4 describes the bit-depths and resolutions at which the original video was sampled. Section 5 presents the testing and results. Section 6 details the future work being undertaken with the driving simulator. The paper concludes with section 7, which draws conclusions based upon the results of the paper.

2 Effects of Reduced Visual Cognitive Loads on Driver Behavior

The most obvious scenario where a driver encounters a reduced visual cognitive load is during night-time driving. Night-time driving is inherently more dangerous than day-time driving, due to the decreased visibility associated with reduced illumination [12][13]. Even when temporal factors such as driver fatigue are accounted for, road traffic accident statistics indicate that the number of accidents is similar between day and night, even though three-quarters of road use occurs during day-light hours [12][13]. Increased levels of night-time accident rates were observed when research into the effects of reduced speed limits was conducted, with the night-time accident rate being 173% of the corresponding day-time rate on two-lane urban roads [14]. It has been noted that even subtle levels of illumination can play a significant role in the reduction of accidents. It has been reported that fatal accidents involving pedestrians reduce by up to 22% on nights when a full-moon occurs [15]. Other research has shown that increase in fog density results in increased speed when testing is undertaken in driving simulators [16].

3 Driving Simulator Implementation and Video Integration

Previous research used data that were acquired using a Mobile Mapping System (MMS). Data acquired by this system were processed to generate a synchronized graphical model and video sequence. Strong correlations were recorded across the video, model and ground-truth (driver speed during data acquisition) data sets. Due to the low frame rate however, video playback was somewhat uneven [17][18]. The video acquired for the purposes of this paper was in 1920x1080 resolution at 29.97003 fps, increasing the smoothness of the playback significantly.

The driving simulator used in the experiments described in this paper consists of an Ubuntu-based PC, triple 1280x1024 resolution monitors, unified to a single 3840x1024 display using a Matrox TripleHead2Go, and a Logitech G27 gaming steering wheel and pedals [19][20]. As the vehicle carrying the witness camera was driven at non-constant road speeds, the geo-tags were processed and interpolated such that a series of evenly-spaced co-ordinates were generated. This allowed the acceleration/deceleration affect caused by the non-constant acquisition speed to be discounted [17]. The equalized video contains video data as if acquired at a constant acquisition speed of 80 km/h.

3.1 Even Spacing of Video

The witness camera acquired High-Definition frames at a rate of 29.97003 fps, and positional data at a rate of 1 Hz. To space the video frames evenly and remove any acceleration affects introduced by driving the acquisition vehicle at normal road speeds, the distance between each GPS sample was calculated and divided by the frame rate to estimate the distance traveled between frames.

Each frame was assigned a geo-tag based on this method. The route was then divided into equal distances, and the nearest frame to each equal distance was selected using a Look-Up Table that related each video frame's geo-tag to its nearest neighbor along the equally-spaced route. This removed the acceleration effect.

3.2 Driving Simulator Software Development Environment

The driving simulator software was developed in MonoDevelop C#. Video playback was performed through the use of the MPlayer video playback application, with both MPlayer and the control system being assigned a thread. Current speed was relayed to the driver using an onscreen speedometer, itself assigned a third thread. The speed was normalized within the range of 0 km/h and 90 km/h. The lower value was obtained by releasing the accelerator pedal, in which state the video playback would pause, and the higher value was obtained by applying full pressure to the accelerator pedal.

4 Video Sampling

The original video, as acquired by the witness camera, consisted of a three minute segment of 1920x1080 resolution with a 24-bit bit-depth. When equalized for speed this reduced to two minutes and 20 seconds. This equalized video was taken as the control video, and was then sampled four times. The testing videos consisted of the original control video, a 1920x1080 resolution video with a 3-bit bit-depth, a 1920x1080 resolution video with a 9-bit bit-depth, a 672x378 resolution video with a 24-bit bit-depth, and an 1168x657 resolution video with a 24-bit bit-depth. These resolutions were chosen as they offered a noticeable level of degradation when compared to the control video; the corresponding sized bit-depths were then selected based on the frame size of the sampled resolution videos. Data loss is interpreted as the loss in video frame size occurred through the bit-depth and resolution sampling processes.

4.1 Selection of Bit-Depths and Resolutions

Each frame of the 1920x1080 resolution 3-bit bit-depth video was 0.73 Megabytes (MB) in size, the same as that of the 672x378 resolution video with a 24-bit bit-depth video, representing a data loss across each of 87%. This process was repeated for the 1920x1080 resolution 9-bit bit-depth video (2.22 MB frame size), again, the same as that of the 1168x657 resolution video with a 24-bit bit-depth video, representing a data loss of 63% across each. This resulted in a set of videos that allowed for a comparison on the reduction of similar levels of information from two aspects of visual cognitive load; bit-depth and resolution. The data loss is shown in Table 1.

Table 1: Data loss of the control video and four sampled videos.

Video	Pixels Per Frame	Bit-Depth Per Pixel	Frame Size (MB)	Information Loss (%)
3-bit Bit-Depth	2,073,600	3	0.73	87.5
9-bit Bit-Depth	2,073,600	9	2.22	62.5
672x378 Resolution	254,016	24	0.73	87.8
1168x657 Resolution	767,376	24	2.22	63.0
Control	2,073,600	24	5.94	0.0

4.2 Bit-Depth Sampling and Resolution Sampling Methods

To sample the original 24-bit video, each frame was split into its constituent three channels; Red, Green and Blue (RGB), with each of these channels having 8-bit pixel intensities ranging from 0 to 255. Each 8-bit value was bitwise shifted right the number of required times. For example, a required 1-bit sample would have each pixel intensity shifted right seven times. This shift reduces the 8-bit pixel intensity to an *n*-bit pixel intensity. This sampled value was then rescaled into the 0 to 255 range to allow the sampled images to be written to file as 24-bit format bitmaps. This was undertaken to prevent any compatibility issues between non-standard bit-depth images and the codec used for video decoding in the driving simulator. Once sampled, each channel was recombined into a single RGB image. To sample the resolution of the original 24-bit video, each frame was resized into the desired vertical and horizontal dimensions. Examples of these are shown in Fig. 1.

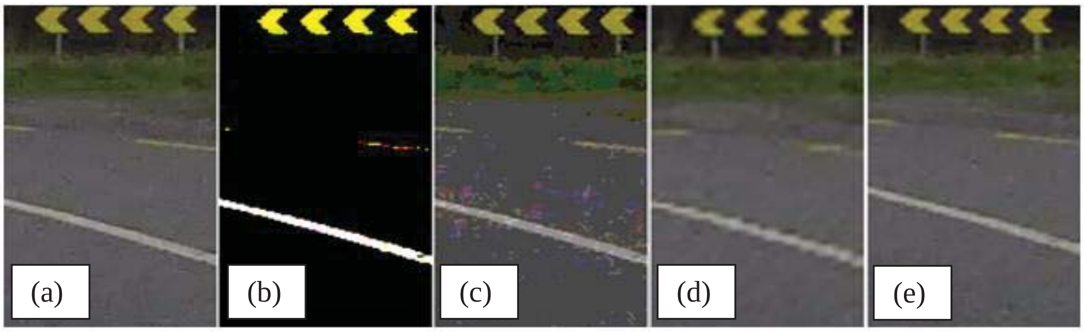


Figure 1: Example frame segment of the (a) control, (b) 3-bit, (c) 9-bit, (d) 672x378 and (e) 1168x657 frames.

5 Testing and Results

Ten participants drove through each of the five videos, presented in a random fashion, with their speed recorded once per frame. The participants consisted of five males and five females, ranging from 21 to 41 years of age. These speed values were averaged using a running calculation to produce 200 values per dataset. Drivers were notified of the control system and the initial speed limit. The only instruction given to the drivers was “*to drive as they would normally*”. The averaged data for each video type are shown in Fig. 2, with the cross-correlation matrix of the five datasets shown in Table 2. It should be noted that, although the speed appears to increase across time, this is, in fact, a reflection of the route’s geometry.

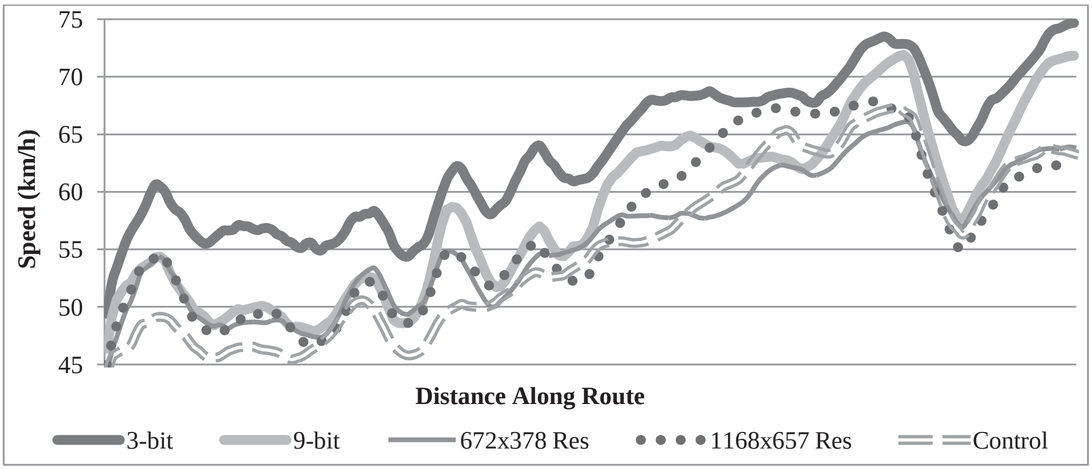


Figure 2: The average driver speeds across the same route using the five different videos.

Table 2: Cross-correlation matrix of average driver speed response (mean = 0.969)

Video	3-bit Bit-Depth	9-bit Bit-Depth	672x378 Resolution	1168x657 Resolution	Control
3-bit Bit-Depth	1.000	0.985	0.971	0.933	0.961
9-bit Bit-Depth	0.985	1.000	0.966	0.912	0.942
672x378 Resolution	0.971	0.966	1.000	0.935	0.975
1168x657 Resolution	0.933	0.912	0.935	1.000	0.958
Control	0.961	0.942	0.975	0.958	1.000

A repeated-measures ANOVA was conducted to compare the means of the five conditions. An overall significant effect was found ($F = 886.6043$, $df = 4, 796$, $p < 0.001$). Subsequent t-tests performed between each sampled video and the control video found significant differences between each. The critical t-value was 1.65, and the calculated t-values were 56.68 (3-bit), 19.45 (9-bit), 6.73 (672x378 resolution) and 12.81 (1168x657 resolution). The mean 3-bit speed was 63 km/h (115.3% of the control's mean), the mean 9-bit speed was 58 km/h (106.2% of the control's mean), the 672x378 resolution mean speed was 56 km/h (101.8% of the control's mean) and the 1168x657 resolution mean speed was 57 km/h (103.4% of the control's mean).

5.1 Discussion of Results

The very strong cross-correlation among the five videos (mean = 0.969), suggests that drivers respond to the geometry of the road in a broadly uniform manner, regardless of the bit-depth or resolution of the video presented to them. The primary difference among the five data sets was the magnitude of speed, where the average speed along the 3-bit video was 115.3% of the control video. The average speeds along the other three sampled videos were between 101.8% and 106.2% of the control video's average speed. The ANOVA test confirmed that the differences observed in the datasets could not be contributed to random sampling error, with an F-value far in excess of the critical F-value.

Previous research has shown a lower cognitive load on a driver in the form of both night-time and fog driving can result in increased speed [14][16]. By sampling a video to produce the same amount of visual data in different forms, this paper has shown that the bit-depth of a video carries less cognitive load for a driver in a video-based driving simulator than the corresponding sampled resolution video, resulting in an increase in speed. A decrease in resolution is also associated with an increase in speed, although not to the same degree.

These results show that when considering the speed values acquired using a video-based driving simulator, the resolution of the video and display system may have some influence on the recorded participant speeds.

6 Future Work

The work presented in this paper has used "on-the-rails" videos (i.e. there was no steering component) for the purposes of measuring one independent variable as an indicator of driver performance; namely speed. Work is underway to introduce a steering element into the video sequence. Once this has been achieved, the experiment will be repeated with the aim of measuring the effect of bit-depth and resolution sampling on the perception of driver steering and positioning.

Future work will also include addressing the limitations of the methods described in this paper; i.e. introducing a comparison of bit-depth with dynamic range, and also a comparison of modified day-time video with an authentic night-time video of the same route.

7 Conclusions

This paper has described a video-based driving simulator, and presented the differences in speed when drivers are presented with videos sampled both in terms of bit-depth and resolution. Drivers still responded to road geometry, with correlations of over 90% across all datasets. The primary difference observed in this work was the overall increase in speed of the drivers when presented with a 3-bit sampled video, where the average speed was in excess of 115% of the control video. The other three sampled videos had average speeds closer to 100% of the control. This supports earlier research suggesting that drivers increase their speed at night and during foggy conditions [14][16]. This is reflected primarily in the 3-bit bit-depth video, and at a lower level again in each of the sampled videos, where a decrease in visual cognitive load resulted in a statistically-significant increase in speed. This indicates that, when measuring driver speed in a video-based driving simulator, the contribution of video resolution and bit-depth on the recorded speeds may have to taken into account.

Acknowledgements

The authors wish to acknowledge the Irish Research Council (IRC) for their continuing support, and also to the participants for their time.

References

[1] Merriam-Webster Dictionary, "*Fidelity Definition*", http://www.merriam-webster.com/dictionary/fidelity [Online, 13 March 2014].

[2] R. W. Allen, G. D. Park, and M. L. Cook, "*Simulator fidelity and validity in a transfer-of-training context*", in Transportation Research Record: Journal of the Transportation Research Board, Washington DC, 2010, pp. 40-47.

[3] J. Greenberg, M. Blommer, "*Physical Fidelities of Driving Simulators*", Handbook of Driving Simulation for Engineering, Medicine, and Pyschology".

[4] D. Kaneswaran, M. Brogan, M. Mulcahy, S. Commins, C. Deegan, C. Markham, "*Replicating reality: Driver assessment using dual-fidelity simulator,*" Signals and Systems Conference (ISSC 2013), 24th IET Irish, pp.1,6, 20-21 June 2013.

[5] J. Engstrom, E. Johansson, J. Ostlund, "*Effects of visual and cognitive load in real and simulated motorway driving*," Transportation Research Part F, Vol. 8, pp 97-120, 2005.

[6] Dialogic Whitepaper, "*Quality of experience for mobile phone users*", [Online, 7 July 2014]. http://www.sintel.com/bibli/telechargement/248/document_Multi.pdf

[7] F. Dobrian, *et al*, "*Understanding the impact of video quality on user engagement*", SIGCOMM '11, August 2011, Toronto, Canada.

[8] A. Vishwanath, *et al*, "*Perspectives on quality of experience for video streaming over WiMAX*", ACM SIGMOBILE Mobile Computing and Communications Review, Volume 13, Issue 4, October 2009, pp. 15-25.

[9] A. Oeldorf, J. Donner, E. Cutrell, "*How bad is good enough? Exploring mobile video quality trade-offs for bandwidth constrained customers*", Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense through Design (NordiCHI '12). ACM, New York, NY, USA, 49-58. doi:10.1145/2399016.2399025.

[10] R. W. Allen, G. D. Park, M. L. Cook, D. Fiorentio, "*The effect of driving simulator fidelity on training effectiveness*", Driving Simulator Conference North America, Iowa City, September 2007.

[11] Mio. (2014) Mio MiVue 388 Witness Camera. [Online, 29 April 2014]. http://eu.mio.com/en_gb/mivue-388.htm#.U1-gu_ldWSo

[12] K. Rumar, "*Night driving accident in an international perspective*", First International Congress Vehicle and Infrastructure Safety Improvement in Adverse Conditions and Night-Driving, 2002.

[13] J. Wood, A. Chaparro, "*Night driving: How low illumination affects driving and the challenges of simulation.*" Handbook of Driving Simulation for Engineering, Medicine, and Psychology", [ISBN: 978-1420061000].

[14] D. R. Herd, K. R. Agent, R. L. Rizenbergs, "*Traffic accidents: Day versus night*" in Transportation Research Record: Journal of the Transportation Research Board, Washington DC, 1980, pp. 25-30.

[15] M. Sivak, B. Schotettle, O. Tsimhoni, "*Moon phases and night-time road crashes involving pedestrians*". Ann Arbor, MI: University of Michigan Transportation Research Institute, 2007.

[16] J. J. Kang, R. Ni, G. J. Andersen, *"Effects of reduced visibility from fog on car-following performance"*, in Transportation Research Record: Journal of the Transportation Research Record, Washington DC, 2008, pp. 9-15.

[17] M. Brogan, D. Kaneswaran, S. Commins, C. Markham, C. Deegan, "*Automatic generation and population of graphics-based driving simulatior using mobile mapping data for the purpose of behavioural testing of drivers*", in Transportation Research Board: Annual Meeting, Washington DC, 2014, Paper Number 14-0473.

[18] Legislation.gov.uk, "*Statutory Instruments 2001 No. 25, The Motor Vehicles (Approval) Regulations 2001*" pg. 37, Item 19. http://www.legislation.gov.uk/uksi/2001/25/pdfs/uksi_20010025_en.pdf [Online, 13 March 2014]

[19] Logitech. (2013) Logitech G27 Racing Wheel. [Online, 13 March 2014]. http://gaming.logitech.com/en-roeu/product/g27-racing-wheel/

[20] Matrox. (2013) Matrox: TripleHead2Go. [Online, 13 March 2014].

Video- and RFID-based Subject Reacquisition in Secure Corridors using Event Reasoning

Fabian Campbell-West, Jianbing Ma, Paul Miller, Weiru Liu
Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
Northern Ireland Science Park, BT3 9DT
f.h.campbellwest@qub.ac.uk

Abstract

This paper describes a novel system for performing subject reacquisition, by reasoning with events from heterogeneous sensors and allowing for decisions to be revised in light of new information. The system uses Radio Frequency IDentification (RFID) readers to build a list of possible subject candidates at a given point and identifies the true subject using facial recognition in video. The system has been evaluated with a data set of 77 subjects, based on a challenging real world scenario. The baseline accuracy of the facial recognition sub-system is 82%, which is increased to 96% by combining RFID information and further increased to 99.5% by applying reasoning techniques. The results illustrate the importance of fusing information from modular components. A practical advantage of the system is that it leverages standard commercial off-the-shelf (COTS) equipment and so can be deployed relatively cheaply using existing infrastructure.

Keywords: Machine vision, event reasoning, application

1 Introduction

Subject reacquisition is a form of subject identification and in particular is a multi-class classification problem. It is the process of identifying an individual at a specific point in space and time given knowledge of one or more previous observations. It is especially important in secure corridors, which are delineated spaces closed to random access where subjects are constantly monitored. Subject reacquisition is one solution to the problem of associating observations across a sensor network, Figure 1.

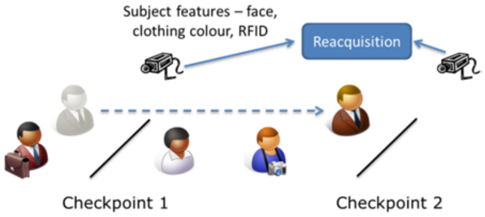


Figure 1: Subject reacquisition, where the subject at checkpoint 2 is identified using features obtained from checkpoint 1.

Subject identification is a common task in sensor networks and existing research covers a broad area of applications. In [1] a novel framework is presented for searching archived video using human attributes as search terms. The authors acknowledge the difficulties of using facial recognition due to illumination variation, changes in face pose and low-resolution imagery. Results are shown using clothing colour, eyewear and hair features in low-resolution video. The precision and recall rates of the detectors vary in quality, with moustaches and hats being difficult to detect. The receiver operating characteristic (ROC) curves show that infrared sensors far outperform visible when using these facial detectors. Bauml et al. [2] present a system for person retrieval in camera networks using facial features that are robust to wide variations in pose. The system can cope with face sizes as small as 18×18. The recognition algorithm uses layered detectors to register face images using eye and mouth features. With operator feedback the system can achieve 86% precision in challenging conditions.

In [3] a watchlist system is presented which is motivated by human perceptual facial recognition. Dedicated neural-network classifiers are trained for each subject on the watchlist using face classifiers to screen probe images before using eye classifiers to make the final decision. The system has been deployed in a live environment and successfully identified 10 people on a watchlist from a total pool of 211 subjects with no error. The computation required for training dedicated classifiers such as these prohibits their use in an online setting. As a counterpoint to watchlist systems, an intruder dection system looks for subjects not on the watchlist. In [4] RFID is used to provide an identity of a subject and facial recognition is used to authenticate. The system is fast in execution, able to search 100 subjects in less than half a second, however it is not clear what level of accuracy is achieved. Jong et al., [5], also combined face recognition and RFID, building an electronic security system for cars to solve the one-to-one authentication problem.

Commercial solutions for tracking across a sensor network include Snap and IntelliVid, which both use time-of-flight between sensors as cues to infer the topology of the camera network. Snap automatically calculates the overlapping fields of view of cameras in the network and overlays this information onto the camera feeds. This makes the task of tracking a subject across the network easy for a human operator, even one without any knowledge of the camera network. IntelliVid is a video intelligence system that can detect suspicious activity, such as loitering or theft, and bundle evidence in alert-based notifications. These systems rely on continuous operator interaction and lack automated intelligent processing.

Räty, [6], identified that the two most substantial factors restricting the deployment of surveillance systems in real world scenarios are real-time performance and cost. Current state of the art methods for face recognition are L1-minimisation algorithms, [7], which traditionally are unsuitable for large-scale systems, but are improving all the time. The need for performance and the suitability of older methods, such as PCA or LDA, is discussed in [8]. Face recognition using correlation filters has been shown to outperform PCA in terms of both accuracy and computation, [9].

Reasoning has been applied to the problem of subject reacquisition in [10] using time-of-flight between known checkpoints as a source of information. The specific challenges of subject reacquisition are not addressed widely in the literature.

The system described in this paper performs automatic, contactless identification of all subjects moving through a secure corridor. A combination of facial recognition from video with RFID sensors is used to identify subjects and a reasoning scheme improves performance and ensures consistency. An authentic data set of 77 subjects, captured to reflect a challenging operational scenario, is used for the evaluation. The data set is partitioned into segments, one for each subject, which allows thousands of unique sequences to be generated with different permutations of subject ordering. The overall system can operate in real-time, is scalable for large installations, and requires no specialist equipment. As a result, it is an effective and low-cost solution to the problem of subject reacquisition in secure corridors.

2 Subject reacquisition

Subject reacquisition can be formulated as a closed-set matching problem. Given an observation of subject X at a checkpoint with heterogeneous sensors, the task is to match with observation(s) at previous checkpoint(s). In this work the sensors are an RFID reader, which can identify tags within a limited radius, and an IP video camera. Each tag is linked to a subject's appearance model, obtained through face detection in video, at the first checkpoint. Given a set of possible identities, i.e. owners of tags within a certain range, facial recognition is used to identify the subject at subsequent checkpoints.

2.1 Face detection and recognition

2.1.1 Face image normalisation

Faces are detected in each video frame using the Viola-Jones algorithm [11], which is popular due to its real-time performance as well as good precision and recall characteristics [12]. Faces are extracted as greyscale intensity images.

Illumination compensation involves adjusting the pixel intensities of a face image so images with different illumination profiles can be directly compared. The method used in this paper is taken from [13], and is based on the Retinex illumination model [14]. The advantages of this method are its computational simplicity and good performance. Given an intensity image I, the logarithm image L is calculated by taking the logarithm of each pixel. The low frequency information is extracted by convolving a 3×3 maximum filter with the logarithm image to produce U. The final illumination compensated image, F, is calculated as F=L-U.

2.1.2 Learning a face model

As a subject walks within the field of view of the video camera, face images are extracted and illumination normalised. In this system the face is modelled by a set of minimum average correlation energy (MACE) filters, [9]. The advantages of using a MACE filter are that it produces a compact feature representation and is computationally efficient so both training and testing can be completed in real-time.

Every normalised face is stored in a buffer with a maximum capacity *B*. When the buffer is full, the faces are processed to create a single face model, then the buffer is cleared and the process continues. A single subject can therefore have many models associated with it.

The buffer contains a set of normalised greyscale images F_i. Each image F_i is d×d pixels. The model, **h**, produced from the images in the buffer is a MACE filter. To construct the MACE filter, for each facial image F_i its 2D Fourier transform Z_i is calculated and stored in the training matrix **X**, which is a d^2×B matrix. The i^{th} column of **X** is a lexicographically re-ordered version of Z_i. The MACE filter can be calculated from the closed form equation [9]:

$$\mathbf{h} = \mathbf{D}^{-1}\mathbf{X}(\mathbf{X}^{*}\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{u} \tag{1}$$

where **D** is a d^2×d^2 diagonal matrix containing the power spectrum of the training images, **h** is a d^2×1 column vector containing a lexicographically re-ordered form of the 2-D correlation (MACE) filter and **u** is a B×1 column vector containing the linear constraints on the training images. We train the filter with positive samples, so **u**=1. $\mathbf{X}^*$ denotes the complex conjugate transpose of **X**.

As the size of the training buffer increases the time required to build the model increases and there is a chance that the subject will change pose. Pose changes are modeled by training multiple MACE filters, but each should contain a single pose. The MACE filter performs well when trained with a small number of samples, so in our experiments we use $B = 7$. As d increases the amount of detail in the face feature increases, at the cost of increased computation time. In our experiments we set $d = 64$.

2.1.3 Evaluating a probe face image

When a subject is detected at checkpoint 2 the steps described in section 2.1.1 are followed to detect, extract and illumination normalise the face images of the unknown subject. Each of these probe face images is compared to the set of known face models.

To compare a probe image with a face model the peak-to-sidelobe (PSR) ratio is calculated [9]. The comparison can be represented by a function which produces a real-valued score given a model and a probe image. The global score matrix, *M*, contains the PSR value for every probe image and every known model. The global set of models *G* contains an element for every observed subject at checkpoint 1. Each element, S_i, contains one or more MACE filters. The maximum value obtained from comparing the probe image with the models in S_i is stored in *M*. If there are *P* probe images and *N* subjects in *G* then *M* contains *P* rows and *N* columns and is constructed by

$$M_{i,j} = \max_{k:\mathbf{h}_k\in S_i}\left(PSR\left(F_j,\mathbf{h}_k\right)\right)\forall F_j, i\in[1,N], j\in[1,P] \tag{2}$$

For example, the score in the third column, second row is the output of the facial recognition algorithm when applied to the third test image using the second candidate as a model for matching.

Once this matrix has been constructed it is possible to determine which of the models S_i best matches the unknown subject *X*.

2.2 Reacquisition decision methods

Given an unknown subject *X* and a closed set of candidate subjects, *G*, the reacquisition task is to identify which of the candidates, i.e. which model in *M*, is the true subject. Described below are three different methods for making this decision.

2.2.1 Method 1 – Dempster-Shafer belief function

The first decision method is based on Dempster-Shafer combination theory. Each column of the score matrix, *M*, is normalised to produce the matrix $\widehat{M}$ as

$$\widehat{M_{\iota,J}} = M_{i,j}\,\frac{1}{\sum_{k=1}^{N}M_{k,j}} \tag{3}$$

For each column of $\widehat{M}$ the confidence scores are sorted in descending order. Each row of $\widehat{M}$ is assigned an initial confidence score of 1. For each column in $\widehat{M}$ the highest confidence score is left unchanged, but all other confidence scores are multiplied by the ratio p_2/p_1, where p_1 is the highest confidence score and p_2 is the second highest.

Once all columns have been processed there is a set of confidence scores, one for each row of $\widehat{M}$. The subject that corresponds to the largest of these scores is chosen as the reacquired subject and the confidence in the decision is the normalised confidence value. This scheme rewards a subject whose model consistently matches the probe images and penalises the other subjects.

2.2.2 Methods 2 and 3 – Normalised mean scores

The second method involves averaging the scores over all probe images, i.e. the average over the columns of $M_{i,j}$. The resultant vector, *V*, calculated by

$$V_i = \frac{1}{P}\sum_{j=1}^{P} M_{i,j} \tag{4}$$

is then normalised, producing the unit vector $\hat{V}$. Each element of $\hat{V}$ corresponds to a model and the value is the normalised mean response of that model to all the probe images. The reacquisition decision is to choose subject S_i, where i is the index of the largest element of V. The confidence score is given by the largest element of the unit vector $\hat{V}$.

The final method is similar to the second, but is more sensitive to differences between models for each probe image. Like method 1, the normalised matrix $\widehat{M}$ is calculated, then a vector, V, is created by averaging the rows of $\widehat{M}$:

$$V_i = \frac{1}{P}\sum_{j=1}^{P} \widehat{M}_{i,j} \tag{5}$$

Note that V is also normalised, since it is the average of normalised values. The reacquisition decision is to choose subject S_i, where i is the largest element of V. The confidence score is given by the maximum value of V. The difference between methods 2 and 3 is that in the former the normalisation is performed after the average is calculated, but in the latter the normalisation is performed beforehand. The methods are similar but are capable of yielding very different results depending on the input values.

The decision methods presented require all face models to be compared with all probe images. As the number of subjects increases, it becomes increasingly difficult to achieve real-time performance. Therefore it is necessary to reduce the search space when looking for the correct subject.

2.3 Comparison sets from RFID

An active RFID tag transmits a signal at regular intervals and an RFID reader can extract the unique identifier of the tag and the signal strength. Due to interference, particularly contact with the human body, the received signal strength at the reader can exhibit large variations. It is possible to determine whether a tag is within a certain radius of the reader, but not to pinpoint its exact position, even with multiple readers.

In this system, the RFID readers at each checkpoint are able to determine which tags are nearby and therefore which tag the current subject may be holding. This means that for a subject X the system does not have to compare the probe images against all N models, improving both precision and scalability.

A comparison set, C, is a subset of the global model set, G. When an unknown subject X appears at a checkpoint the comparison set is constructed. The face-based decision method described in section 2.2 then only needs to be used over the smaller comparison set, as opposed to the global set. An ideal comparison set is the smallest subset of G that is guaranteed to contain the subject's true identity. At worst, $|C|$ is the total number of subjects, N. It is assumed that the true identity of X is contained in C. In secure corridors, such as airports, where the set is closed and subjects carry identification tags this is a reasonable assumption.

The score matrix, M, described in section 2.2 is still constructed, but only rows for subjects within C need to be calculated. This significantly reduces the number of computations required but otherwise the process for making a decision remains the same.

2.4 Belief revision

Belief revision is the process of changing a decision in light of new information. In the context of this system, the change of decision is prompted by a duplicate reacquisition decision. The secure corridor is a closed-set matching problem, therefore each subject must appear once and only once at each checkpoint. If a subject is reacquired twice, it means that another subject has not been reacquired at all.

The solution to this problem presented here is a process called one-step revision, based on [10]. This allows at most one change to be made given new information and prevents cascading of errors through the system as a result of a bad decision. Suppose two different subjects, X and Y have both been reacquired as subject S_1 from checkpoint 1 and that a third subject, Z, has been reacquired as S_2. First, the confidence values of the decisions for X and Y are compared. The decision with the greater confidence score is retained, suppose it is X. The comparison set for Y is then modified to remove S_1. The next best match is chosen from the comparison set. If the new choice does not result in conflict then the change stands and the system is stable. If the new choice for Y is S_2, however, there is then a conflict with Z. If the confidence value for X is less than the confidence value for Z, then Y is left as S_1 and no change is made. If the confidence value for X is greater than the confidence value for Z, then Y is changed to S_2.

Note that this scheme can result in inconsistency, as it is possible for two subjects to be reacquired with the same identity. It is preferable to keep inconsistency in the presence of uncertainty than to force a change which may be incorrect and introduce further errors. Inconsistency can be flagged to a human operator with oversight of the system.

3 Data set

To evaluate the system a simulated secure corridor environment was constructed that contained two checkpoints, one at each end. Both checkpoints consisted of an IP video camera, a Panasonic WV-NP240 with zoom lens, and an RFID reader, an RF Code M250 that operates with 433 MHz active tags. The tags were configured to transmit every 2 seconds.

The tests involved a total of 77 subjects. Each subject moved through the corridor while holding their own RFID tag. The subjects were cooperative, and looked directly at the video cameras but were not forced to adopt a specific pose. Example face images from one subject are shown in Figure 2, where the change in illumination caused by ambient lighting between checkpoints is clear. The data was captured over a period of five days. The data for each subject was manually delineated so that random simulated sequences could be generated using playback of the original data.

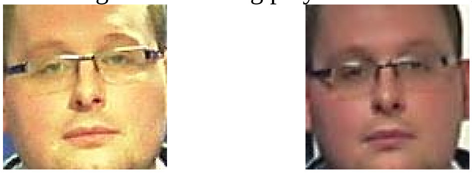


Figure 2: Example face images of the same subject at different checkpoints.

The use of cooperative subjects is justified in a secure corridor scenario because people expect to cooperate when it is in their best interest. For example, people currently tolerate long queues to pass security in major airports. Current automated identification systems deployed in airports rely heavily on cooperation, with passengers being funneled into corridors or cubicles for a sequence of scans, [15]. Using contactless identification technology, such as RFID, can alleviate some of the bottlenecks inherent in these systems.

4 Evaluation

The system evaluation was divided into two parts. First, the whole data set of 77 subjects was considered in its entirety to give the baseline performance of the facial recognition component. The three decision metrics presented in section 2.2 were evaluated in terms of precision versus rank using a cumulative match curve (CMC). In the second part of the evaluation a large number of random simulations based on the real data were performed to show the benefits of using comparison sets and belief revision.

4.1 Baseline facial recognition evaluation

Traditional precision and recall metrics, [16], when applied to evaluate this system will give the same result. When a subject is detected in the field of view of the video sensor, the system will always give a reacquisition result. A false positive (FP) and a false negative (FN) occur simultaneously since to incorrectly classify one subject as ‘positive’ necessarily means the real subject has been incorrectly classified as ‘negative’. The baseline precision of the system for all three decision methods is shown in Table 1, the maximum is 82%.

To put the results in context, the subject reacquisition problem can be viewed as a linear assignment problem (LAP). Given N subjects at checkpoint 2 and scores, or weights, for the same subjects at checkpoint 1 the task is to optimally match subjects in pairs. One solution to the LAP is the Hungarian Method, [17]. Applying this method to the baseline results gives a maximum precision of 94%. This can be viewed as an upper bound on the precision that can be obtained when viewing the whole data set at once. Note that the LAP requires all the data to be available, so it cannot be used as subjects appear at checkpoint 2 in a live setting.

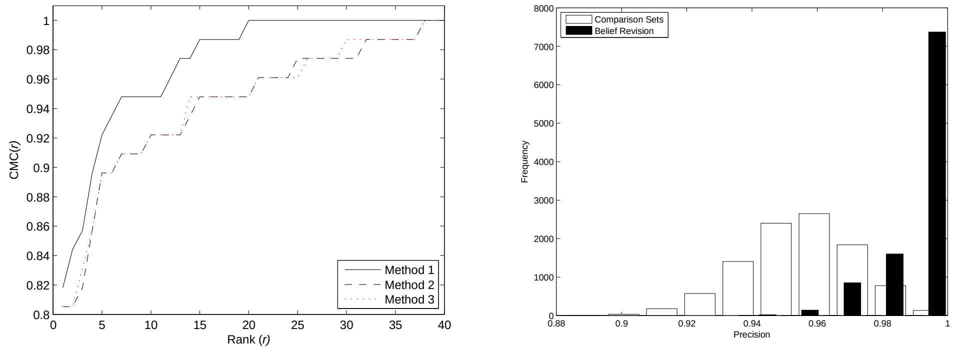
The CMCs, plotting cumulative precision versus rank, for the three decision methods are shown in Figure 3a. Decision method 1 gives the best performance, while methods 2 and 3 produce nearly identical CMC curves. To achieve 90% precision using facial recognition alone it is necessary to consider the top five ranked subjects. The remaining part of the evaluation will show the system-level improvement, compared to the recognition performance at the component level, obtained by using comparison sets and belief revision.

4.2 Simulation results

The data set described in section 3 can be used to create random sequences using different ordering of the subjects. To create a simulated sequence, each subject is assigned two random time values. The first determines their time of arrival at the first checkpoint, the second determines the time-of-flight between checkpoints. The random values are sampled from a Poisson distribution, which is commonly used in queuing theory. For the results presented here the two values, in seconds, were generated with means $\lambda_1 = 60, \lambda_2 = 600$. This simulates one minute between passengers to pass through security then 10 minutes to walk to the gate. With the data set of 77 subjects there are 77! possible unique orderings. Since there are no restrictions on a subject being considered multiple times there are 2^{77}

possible comparison sets and so 2^{5929} meaningfully distinct simulations. For the results here, 10,000 random sequences were generated and the precision results averaged over all simulations, Table 1.

The results show that all three decision methods perform equally well and that limiting the size of the comparison set has a large positive impact on performance. Allowing mistakes to be identified and corrected increases the precision to nearly maximal levels. The histograms in Figure 3b show that for decision method 1 70% of the simulations resulted in 100% precision. The lowest recorded precision when using belief revision is 94%.



(a) *(b)*

Figure 3: CMCs for the three decision methods when applied to the entire data set, (a). Histograms showing the frequency of precision scores from 10,000 simulations using decision method 1 with comparison sets and belief revision, (b).

System Component	Decision Method 1	2	3
Baseline Precision	82%	81%	81%
Hungarian Method	90%	94%	94%
With comparison sets	96%	96%	96%
With comparison sets and belief revision	99.5%	99%	99%

Table 1: Summary of precision results for the data set of 77 subjects.

5 Conclusions

A novel system has been described for subject reacquisition using event reasoning in secure corridors. Facial appearance features are used to train MACE filters for facial recognition, supported by RFID information that reduces the search space for the classification decision. When classification errors occur, a one-step belief revision scheme allows the system to change decisions in an attempt to resolve the inconsistency.

The system was evaluated using a data set with 77 subjects. The baseline precision of the facial recognition system is 82%. Different conditions were simulated by replaying the video and RFID readings in a different order. In total 10,000 different simulations were run using this real data. The use of comparison sets increases the average precision to 96% and the belief revision scheme increases the average precision to 99%.

Several improvements to this prototype are planned. For effective facial recognition, especially with uncooperative subjects, spatial registration is necessary, [2]. Full-body colour and texture features will be used as a third source of information for the reacquisition reasoning. This will help detect uncooperative subjects. The system described here is part of a larger third generation security system (3GSS), [6], being developed with a multi-agent architecture. Each checkpoint in the secure corridor will be monitored by an agent, which sends messages to other agents in the form of events. This architecture allows the system to scale up to large facilities, with each agent capable of autonomous processing. Key decision-making agents collate events from their child agents to infer final decisions from observed events. Future work in the intelligent decision making is to incorporate more sophisticated belief revision rules that include detecting missing subjects and leveraging time-of-flight information.

6 References

[1] D.A. Vaquero, R.S. Feris, D. Tran, L. Brown, A. Hampapur and M. Turk, “Attribute-based people search in surveillance environments”, Workshop on Applications of Computer Vision, 2009

[2] M. Bauml, K. Bernardin, M. Fischer, H.K. Ekenel and R. Stiefelhagen, “Multi-Pose Face Recognition for Person Retrieval in Camera Networks”, Proceedings of AVSS, 2010

[3] B. Kamgar-Parsi, W. Lawson and B. Kamgar-Parsi, “Toward development of a face recognition system for watchlist surveillance”, IEEE Trans. PAMI, 2011

[4] M. Sekar, “A real time surveillance system using wired and wireless sensor networks by multi-agorithmic approach”, Proceedings of Digital Image Computing Techniques and Applications, 2011

[5] G. Jong, P. Peng and G. Horng, “Multi-recognition combined security system for intelligent car electronics”, Int. Journal of Innovative Computing, Information and Control, 8 (4), April 2012

[6] T. Raty, “Survey on Contemporary Remote Surveillance Systems for Public Safety”, IEEE Trans. On Systems, Man and Cybernetics, 40 (5), Sept. 2010

[7] A. Yang, Z. Zihan, A. Balasurbramanian, S. Sastry, "Fast L1-Minimization Algorithms for Robust Face Recognition", IEEE Trans. Image Processing, 22, 8, pp. 3234-3246, 2013

[8] A. Bansal, K. Mehta and S. Arora, "Face Recognition using PCA & LDA Algorithms", 2nd Int. Conf. Advanced Computing and Communication Technologies, 2012

[9] M. Savvides, B. Vijaya Kumar, P. Khosla, “Face verification using correlation filters”, Proceedings of IEEE Automatic Identification Advanced Technologies, pp. 55-61, 2002

[10] J. Ma, W. Liu, P. Miller and F. Campbell-West, "An improvement of Subject Reacquisition by Reasoning and Revision", Proc. Scalable Uncertainty Management, pp. 176-189, 2013

[11] P. Viola, M.J. Jones, “Robust real-time face detection”, Intl. Journal Computer Vision, 57(2), pp. 151-173, 2004

[12] M. Castrillon, O. Deniz, D. Hernandez and J. Lorenzo, “A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework”, Machine Vision and Applications, 22(3), pp. 481-494, 2011

[13] Nabatchian, A., Abdel-Raheem, E., Ahmadi, M., "An Efficient Method for Face Recognition under Illumination Variations", Int. Conf. High Performance Computing and Simulation, pp. 432-435, 2010

[14] S. Shan, W. Gao, B. Cao and D. Zhao, “Lightness and retinex theory”, Journal of Optical Society of America, 61(1), pp. 1-11, 1971

[15] European Patent Application 08291258.5, December 2008

[16] D. Olson and D. Delen, "Advanced Data Mining Techniques", Springer, 2008

[17] R. Burkard, M. Dell’Amico and Silvano Martello, “Assignment Problems”, SIAM, 2012

Frontal Detection of Backpacks in Surveillance Videos

Ian Beatty-Orr, and Kenneth Dawson-Howe
School of Computer Science and Statistics
Trinity College Dublin
beattyoi@tcd.ie, Kenneth.Dawson-Howe@scss.tcd.ie

Abstract

This paper presents a method to detect backpacks worn by individuals when they are facing towards the camera in a wide variety of environments, illumination conditions and with low contrast between the backpack straps and the clothing of the individuals. Previous work has only looked at this problem in well-lit indoor environments typically with high contrast between the backpack and the clothing worn. A combination of background subtraction and the histogram of oriented gradients is used to isolate the upper torso region of individual subjects. Colour clustering is used to extract backpack like features from this region and statistical analysis is used (over the frames in which the person is visible) to classify whether each individual is wearing a backpack. A database containing test videos of individuals (designed to be representative of the conditions real security cameras would encounter) both with and without backpacks walking towards the camera was constructed. Accuracy of 79.5% was achieved.

Keywords: Video Surveillance, Luggage Detection, Frontal Backpack Detection, k-Means

1 Introduction

Backpack detection in surveillance videos is an important and difficult problem, which is motivated largely by security concerns relating to terrorist events. It is important to be able to detect when people abandon baggage and there is a growing body of literature in this field (e.g. [Tian et al., 2011, SanMiguel et al., 2012]). However all of this research assumes that the abandoned objects are visible to some surveillance camera, an assumption which will often be invalid (e.g. if an area is not visible to a camera or if the object is placed under/inside/behind some other object). Hence, to increase the chances of locating abandoned objects, it is necessary to analyse what people are carrying and hence identify when they stop carrying these objects. Part of this problem is the determination of whether a person is carrying a backpack.

In Section 2 we look at the background literature relating to backpack detection. Section 3 presents the method proposed in this paper. Section 4 presents the new dataset and the results of applying the proposed method to this dataset.

2 Background

A number of researchers have looked at the problem of backpack detection but this has been done primarily from consideration of a side-on silhouette view where the backpack creates an asymmetrical protrusion [Haritaoglu et al., 1999, Cutler and Davis, 2000, DeCann and Ross, 2010, Damen and Hogg, 2012]. The most influential method, known as “Backpack, was developed by Haritaoglu *et al.* [Haritaoglu et al., 1999] and relied upon the natural symmetry of the human silhouette. Any asymmetrical protrusions (such as backpacks) are identified by looking at the periodicity of these asymmetrical regions over a full gait cycle, compared to the periodicity of the overall silhouette. Using a slight variation on this approach, BenAbdelkader and Davis [BenAbdelkader and Davis, 2002] used a method that calculated only the periodicity of the upper torso region. This method detected the change in gait due to walking while encumbered by an item such as a backpack. Tao *et al.* used Gabor wavelets to enhance this method [Tao et al., 2006].

Only one research paper to date by Chua *et al.* [Chua et al., 2013] has considered backpack detection from a frontal viewpoint, and in that case the technique presented required that there was a high contrast between the backpack straps and the clothing worn. However the test data used was only representative of the local conditions encountered in an airport terminal in a warm climate where most of the people were wearing bright shirts with dark coloured bag straps which resulted in a high level of contrast. The scene was indoors and the illumination was bright and consistent. Two methods were used to detect straps. The first used background subtraction to extract the upper torso and then applied binary thresholding to this region. The height to width ratio of connected components within this area was examined to find components with the proportions of straps. The second method applied canny edge detection and used a probabilistic Hough transform to find parallel edges. If a pair was found within 15 pixels of each other and a length greater than 20 pixels it was considered a backpack. The comparable accuracy rate for backpacks from all orientations was around 88% (assuming an equal number of positive and negative samples). Unfortunately this figure includes side-on and rear views (as well as frontal views) of backpacks, with no indication of the specific success rates for (or portion of) each viewpoint type. The authors report that false detections were mainly due to segmentation errors due to insufficient contrast between the bag and clothing (or background), and this low contrast is typical in the dataset used in this paper. Hence a lower success rate could be anticipated.

3 Proposed Method

This paper proposes a method that attempts to automatically locate backpacks when looking at the front of an individual. Only the backpack straps will be visible from this angle presenting us with the challenging task of differentiating them from the underlying clothing. Bear in mind that the goal is to locate backpacks in video sequences where the contrast between the backpack and the clothing may be quite low. For example consider the dataset samples shown in Figure 1. We determine if a backpack is present in a video sequence of a person through the following steps:

1. Upper Torso Location determination, using Histograms of Oriented Gradients (HOGs) and a Gaussian Mixture Model (GMM).
2. Strap identification through colour clustering on each image row of the upper torso region.
3. Statistical Analysis of the rows of each image (and of the frames in the sequence) to confirm or reject the presence of a backpack.

3.1 Upper Torso Location determination.

The straps will always lie within the upper torso region which is roughly bounded by the bottom of the head, the bottom of the arm pits and the outer edges of the arms. We locate this region by analysing the intersection of the rectangular *person region* detected by the HOG technique [Dalal and Triggs, 2005] and the foreground pixels identified using a GMM [Stauffer and Grimson, 2000] to distinguish moving objects from the background. See Figure 1. We analyse only those foreground pixels in the top half of the *person region* and take the highest foreground pixel to be the top of the head. We look to each side of the head to find the top of the shoulders, which should be significantly lower than the top of the head. The upper torso location is taken to be a region directly below the shoulders which is one tenth of the height of the person. Again, see Figure 1.

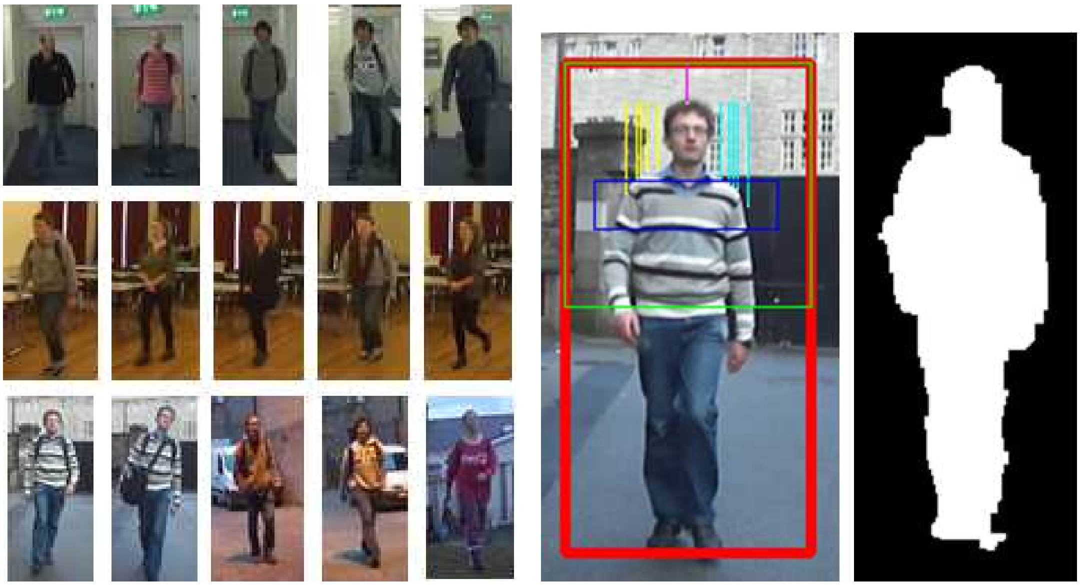


Figure 1: Fifteen sample frames from the dataset (left) together with an illustration of the processing used to find the torso region (in the two larger images on the right). In the first of the larger images a red box is shown around the person detected using the HOG, and a green line is shown half way up this box indicating that only the top half of this region is processed when searching for the torso. Using the foreground pixels (shown in the binary image on the right) a magenta line traces down to the highest foreground pixel in the center column at the top of the head, and yellow and cyan lines trace down to the highest foreground pixels detected in columns placed above the likely shoulder locations. These allow us to select the torso region (highlighted by a dark blue box).

3.2 Strap identification.

K-mean clustering is used to distinguish the straps from the clothing within the upper torso region. It is applied to each row separately using $k = 3$, as this was found to typically provide good segmentation of the straps. It is noted that some clothing can cause a problem for such a low value of k (and hence the value of k should probably be somewhat more adaptive). A sample application of k-means to a single image row (and to the entire upper torso region) is shown in Figure 2. Each row is examined for chains of pixels of the same colour. Pairs of chains (which might represent the backpack straps) are identified which

- have the same colour,
- whose widths are within 180% of each other, and
- whose distances from the centreline of the upper torso region are within 180% (i.e. are positioned roughly symmetrically around the centre line of the torso0.

The relative chain length (and position around the centre line) directly constrain how parallel to the image plane the individual needs to be. As the allowed difference in chain length is increased, the angle of the person with respect to the camera that the system can tolerate also increases. At the same time, though, the rate of false detections also increases.

3.3 Statistical Analysis.

We classify a frame as containing a backpack for an individual if:

- The number of rows between the top and bottom most rows labelled as having a potential strap is greater than 64% of the total number of rows in the upper torso region. This represents the length of the straps relative to the upper torso region.

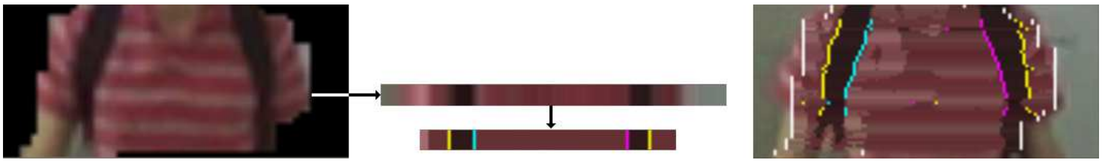


Figure 2: Torso region as detected using HOG and background subtraction (left), the processing for one particular row (centre) and the result of processing all rows (right). In the centre and right images the outer pixels of the person are highlighted in white, the outer edges of the straps are highlighted in yellow and the inner edges of the left and right straps are highlighted in cyan and magentra respectuflly.

- The percentage of rows between the topmost and bottommost rows which were labelled as having a potential strap is greater than 57%. Note that these rows may be well inside the upper torso region.
- The location of backpack straps on all the positive rows line up, as measured by the average horizontal distance between the centre of the straps on each row (for the left and right straps separately. The average horizontal distance must be less than 4 pixels.
- The width of the backpack straps do not vary too much, as measured by the standard deviation of each strap width. The standard deviation must be less than 80%.

We classify a person in a video sequence as wearing a backpack if either

- N sequential frames are positive. The lower N is the less time a backpack needs to be visible before triggering detection. The higher N is the more confidence the detection has. N is a location specific parameter, as, for example, a camera at the end of a long hallway can have it set higher than one positioned near a bend. In the results presented N was 3 frames.
- or, a local percentage of frames are detected as positive, for instance three out of the past five. This catches cases where the method is not triggering detection for every frame. This can be the case when the contrast of the strap is very low relative to the underlying garment. In the results presented the percentage required was 25% out of the previous 16 frames.

4 Results

The test data used consisted of 22 test pairs. Each pair contained two videos clips shot in identical circumstances with the main difference being the presence of a backpack in one sequence and the absence of the backpack in the other sequence. This test set was created using students and locations in our university. This was necessary as no suitable test data could be found in available databases. In total nine different subjects in four outdoor and three indoor locations were recorded with numerous garment and strap combinations. This resulted in a test set that was representative of the conditions likely to be encountered by a surveillance camera in our university. This included clips that tested the ability of the system to detect straps with a low level of contrast relative to the underlying garment, as well as clips with a varying level of illumination including several scenes under a streetlamp in twilight. See Figure 3 for sample successful detections.

The results of applying the new technique to the dataset are shown in Table 1, giving an accuracy of 79.5%. In order to compare with the research described in [Chua et al., 2013] we implemented a technique based on parallel edges, which gave us an accuracy of only 68% (See Table 1). Note that the thresholding approach described in [Chua et al., 2013] does not work on this dataset due to the complexity of the underlying clothing.

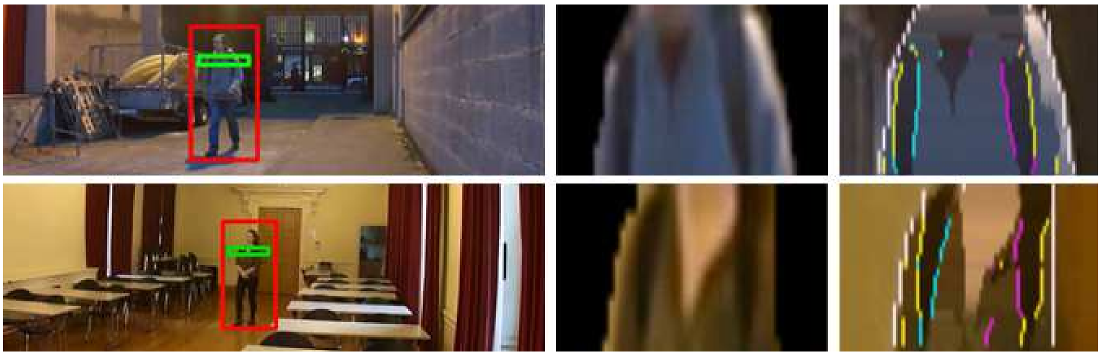


Figure 3: Successful detections of backpacks in individual frames in twilight conditions (top row) and indoor with low level of contrast relative to the underlying clothing (bottom row).

Technique	TP	TN	FP	FN	Precision	Recall	Accuracy
Our approach	16	19	3	6	84.2%	72.7%	79.5%
Parallel edges	17	13	9	5	65.4%	77.3%	68.2%

Table 1: Classification of subjects as carrying a backpack or not. This table shows the number of true positives (TP - a backpack in the scene and detected as such) and negative (TN - no backpack in the scene and none detected) detections as well as the false positive (FP - no backpack in the scene but one detected) and negative (FN - a backpack in the scene but none detected) detections for each method.

The false positive (FP) errors from the evaluation of our technique were due to (1) jacket lapels being detected as straps due to a partial background subtraction failure, (2) the border between scarf and underlying garment being detected as a strap, and (3) strap shaped shadows in the upper torso region. The false negative (FN) errors from the evaluation of our technique were caused twice by (1) scarfs occluding straps, once by (2) a background subtraction failure, once by (3) a two-toned garment causing incorrect clustering of the straps, and twice by (4) too low a number of frames triggering a detection due to low contrast of straps. Two sample failures are shown in Figure 4.

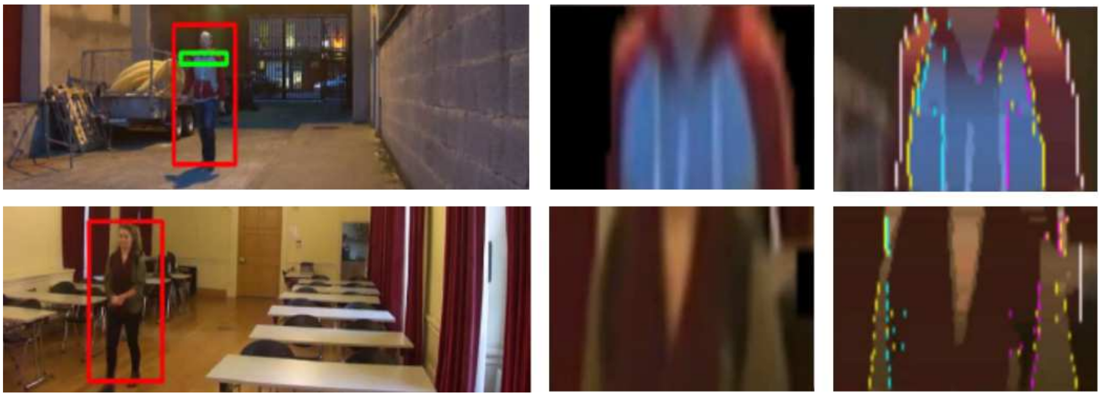


Figure 4: Unsuccessful detections. False positive caused by clothing pattern (top row) and a false negative due to partial occlusion by a scarf (bottom row).

5 Conclusions

The method presented in this paper is reasonably successful at detecting backpacks on individuals when only the front straps are visible. This has been achieved in a variety of conditions including poor illumination and low contrast such as shown in Figure 3. This is an area where edge detection based techniques often fail due to the weak response of low contrasting edges.

There are limitations to the work presented. The dataset used is too small and as it was constructed by the authors there is clearly a question of the independence of the dataset. There is a clear need for a sizeable dataset of videos of individuals wearing backpacks (and not wearing backpacks), viewed from a variety of angles, in a variety of environments and in situations where there are single and multiple subjects. The videos used were taken from a low angle (in comparison to many surveillance cameras) and this raises a question over how the technique presented would respond given data taken from a higher camera.

References

[BenAbdelkader and Davis, 2002] BenAbdelkader, C. and Davis, L. (2002). Detection of People Carrying Objects: a Motion-based Recognition Approach. In *5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2002)*, pages 378–383.

[Chua et al., 2013] Chua, T., Leman, K., Hee Lin, W., Nam Trung, P., Chang, R., and Dinh Duy, N. (2013). Sling bag and backpack detection for human appearance semantic in vision system. In *Intelligent Robots and Systems (IROS)*, pages 2130–2135.

[Cutler and Davis, 2000] Cutler, R. and Davis, L. S. (2000). Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:781–796.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893.

[Damen and Hogg, 2012] Damen, D. and Hogg, D. (2012). Detecting Carried Objects from Sequences of Walking Pedestrians. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1056–1067.

[DeCann and Ross, 2010] DeCann, B. and Ross, A. (2010). Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In *Biometric Technology for Human Identification VII*.

[Haritaoglu et al., 1999] Haritaoglu, I., Cutler, R., Harwood, D., and L.S., D. (1999). Backpack: detection of people carrying objects using silhouettes. In *Seventh IEEE International Conference in Computer Vision*, volume 1, pages 102–107.

[SanMiguel et al., 2012] SanMiguel, J., Caro, L., and Martinez, J. (2012). Pixel-based colour contrast for abandoned and stolen object discrimination in video surveillance. *Electronics letters*, 48(2):86–87.

[Stauffer and Grimson, 2000] Stauffer, C. and Grimson, W. E. L. (2000). Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:747–757.

[Tao et al., 2006] Tao, D., Li, X., Wu, X., and Maybank, S. (2006). Human carrying status in visual surveillance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, pages 1670–1677.

[Tian et al., 2011] Tian, Y., Feris, R., Liu, H., Hampapur, A., and Sun, M. (2011). Robust detection of abandoned and removed objects in complex surveillance videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 41(5):565–576.

Video Adaptation for the creation of advanced intelligent content for Conferences.

Atul Nautiyal, Eamonn Kenny, Kenneth Dawson-Howe
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
nautiyaa@tcd.ie, eamonn.kenny@cs.tcd.ie, kenneth.dawson-howe@scss.tcd.ie

Abstract

The paper presents initial investigations into the development of a virtual conference system which will allow users to locate and view relevant presentations or parts of presentations after a conference has been held. It is part of a much larger investigation into the use of advanced intelligent content. The work presented in this paper details how information from slides can be automatically extracted and used.

Keywords: Video adaptation, Text Segmentation, Optical character recognition, Video Processing Application.

1 Introduction

Information is becoming more and more readily accessible. Data availability, data access and transmission rates have been increasing at exponential rates. Around two thirds of the data now transmitted on the internet is video data. However, unless a user just wants to watch a single complete video, it is very difficult to interact with video to obtain the information that is wanted by the user. Video adaptation is the process of adapting video footage to the needs of a particular individual. This could be in the form of dubbing, sub-titling or extracting smaller, more relevant, clips from a video [2]. For example, if a user wants to learn about a specific topic, video adaptation must identify the complete videos, or clips from videos, which deal with the specific topic.

The Centre for Global Intelligent Content (CNGL) has been working on providing ways for people to interact seamlessly with content by embedding additional knowledge into the content to create *advanced intelligent content*. CNGL is a collaborative effort situated across 4 Irish universities and funded by industry and the Irish government. Its function is to provide an end-to-end value chain of multilingual content which is optimised, adapted, monitored and analysed from multimodal sources such as video, audio and text. The centre is currently split into six themes covering creation and curation concerned with text normalisation and sentiment analysis, delivery and interaction concerned with multimodal input/output delivery, interoperability and analytics concerned with correct multilingual web-based markup, workflows and data-mining, personalisation and adaptivity concerned with adapting the content to the preferences of the end user is an explicit or implicit way, search and discovery concerned with multimodal search, translation and localisation concerned with machine translation to commonly used languages and languages of emergent countries (See www.cngl.ie for further details).

CNGL is now turning its attention to video data and looking at ways in which data can be extracted from video so that video can be included as part of the *advanced intelligent content*. The intention is to allow video to be indexed in such a fashion that relevant sections of videos can be extracted and delivered to users to satisfy their specific needs. The initial domain of study has been limited to video lectures and presentations. For example, Coursera offers 641 courses to nearly 7.5 million users with several hours of video lectures and presentations for each course.

1.1 Demonstration System for Conferences

With rising travel costs, expensive hotel stays, delayed visa applications and large conference fees it's becoming more desirable to attend conferences virtually. Particularly in academic circles the workload of the attendee is increasing. Professors and students alike are expected to attend a large number of conferences per year containing multiple track to keep abreast of the state-of-the-art in their areas of interest. Sometimes the fields of interest of the participant lie in more than one track meaning that they either miss out on sessions that they might attend, or bring along a student to attend the sessions they miss giving them feedback on that session. Ideally it would be less time consuming if a attendee were presented with 10% of the complete conference in a recommender system rather than wading through the complete set of sessions.

A nice virtual solution where a user logs into the conference management system, views the talks of interest to them and is automatically recommended talks that are potentially of interest to them is useful and conceivable. A recommender system can be produced if each session track is videoed with a fixed camera on the slideshow and moving camera on the speaker. Also, good microphone arrays are required to capture the speech of the speaker as audibly as possible.

Aligning the speakers slides to their speech, it is then possible to use optical character recognition (OCR) and automatic speech recognition (ASR) to extract timestamped subtitles for each session. In fact, given that the slideshow of the presenter is timestamped means that the words appearing on the slides of the presenter may also be possibly spoken by the presenter. This means that the OCR can actually be used to inform the language models of the ASR, producing domain specific language models and hence improving the accuracy of the ASR [3].

Since the ASR and OCR are timestamped the information can also be used to inform the ranking scheme of a search engine on the conference management system. Usually the text within search engines is ranking according to how often it occurs, however in this case the search engine can be informed according to size of fonts in the slides obtained from the OCR. Essentially the rank can be increased if the text appears in headings rather than bullet points in slides. Additionally, the duration of time that a sentence appears in the slides also obtained from the OCR can also be used to increase the rank of sentences of text. Since, the OCR produces timestamps, it is possible to create a webservice that produces search results allowing the user to jump to the exact position in the video where a particular phrase was spoken/presented by the presenter. Again this type of technology saves time in perusing important information.

We therefore see that obtaining the timestamped, font-size specific text from the OCR as a key part of the tool from the point of view that it informs three parts of the conference management system: the automatic speech recognition, the search engine and the recommender system.

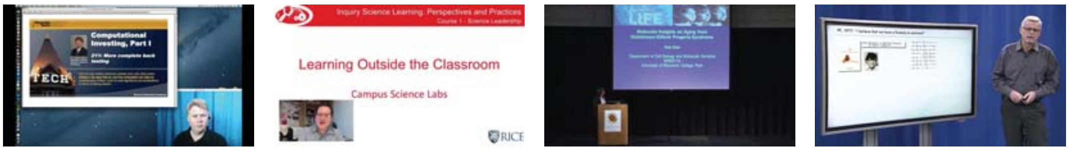
1.2 Background

Our main task (in this paper) is to locate the slides in videos and to recognise the characters on those slides, generating a structured mark-up of what is present (and when). The detection of slide transitions has been addressed by looking at the difference in text layout [4] and by considering the differences with respect to a background template [5]. The extraction and recognition of text from images is a well addressed task (e.g. [8]) and has led to the development of quite powerful and freely available engines such as Tesseract [7].

Perhaps the most similar work to ours is the TalkMiner system [1] which automatically locates lecture presentations on the web, extracts the slides and builds a search index from the words on the slides. They process 1 frame per second from a video and use the global pixel differences to identify slide transitions. They found that this created problems however where the audience and/or speaker are visible in the shot. As a result they enhanced their approach by considering the central part of the frame separately, by considering the size of the bounding box around any changes, by locating any faces in the images, and by explicitly identifying the build up of information on slides.

2 Approach

This paper presents the initial results of our video adaptation work on video lectures and presentations. The two main components in such videos are the slides and the presenter (See Figure 1). The goal is to provide as much indexed information as possible for the video. For example, if we can reliably obtain the words spoken by the presenter we can index these words to the times at which they are spoken. We have used speech recognition to extract what the presenter is saying but in general have obtained very low success rates in terms of words recognised correctly. Some video content provides subtitles as meta data and this provides a much more reliable data stream.


Computational Investing, Part I
TECH
Inquiry Science Learning: Perspectives and Practices
Course 1 - Science Leadership
Learning Outside the Classroom
Campus Science Labs
RICE
LIFE

(a) (b) (c) (d)

Figure 1: Example of videos lectures with slides.

Our method for extracting slides from video frames is presented in Section 3. Section 4 presents our algorithm for extracting text present in slide frames and the structure and creation of XML data from the slides is presented in Section 5. Finally we present initial results and future plans in Section 6.

3 Extracting slide frame from video frame

Generally in a video sequence of a lecture presentation, the total number of slides presented is very small in comparison to the total number of video frames. Therefore, to save both computational power and time it is important to detect slide change in the video and process video frames only if slide has changed. In video frames, we assume that most pixels belonging to the slide will be brighter than most of the remaining pixels. Ostu's method is used for detection of the brighter pixels [6]. Ostu's method returns a binary mask where pixels brighter than an optimal threshold are marked as foreground. It is possible that a video frame may have several small bright areas other than the slide. As a result, the binary mask may also contain several foreground regions. In the first video frame, the slide normally takes maximum part of the frame. Therefore, in the binary mask of the first frame foreground region belonging to the slide should be bigger than all other foreground regions.

The biggest foreground region present in the binary mask of the first video frame is used for extracting the pixels belonging to the slide in the first frame. Pixels belonging to the slides are saved in a separate image file called *slide frames*. Figure 2(a) shows slide frame extracted from the frame shown in figure 1(a). Information about slide location in the previous frame can be used to detect the slide location in the current frame. Therefore, after the first frame it is not necessary that slide take up the maximum part of frame. In some cases slides may present in a distorted (geometrical transform) form due to the projection. Using the boundaries of foreground region belonging to the slide and the boundaries of a minimum bounding box holding that region, parameters of projective transformation can be learned. These parameters can further be used for removing the projective distortion.

Our method starts from extracting the text information from the first frame. A new frame is processed in case of a slide transition. For each processed frame the OCR result are saved in a XML file (See Section 5).

4 Text recognition in slides

After extracting slides from the video we use the Tesseract OCR engine [7] for recognising the text in the slides. Tesseract results are significantly improved if text is present in a single column or block and the font size is the same throughout. If text is in multicolumn, Tesseract returns garbled output. Therefore the text layout (number of blocks and font sizes) needs to be determined for successful recognition.

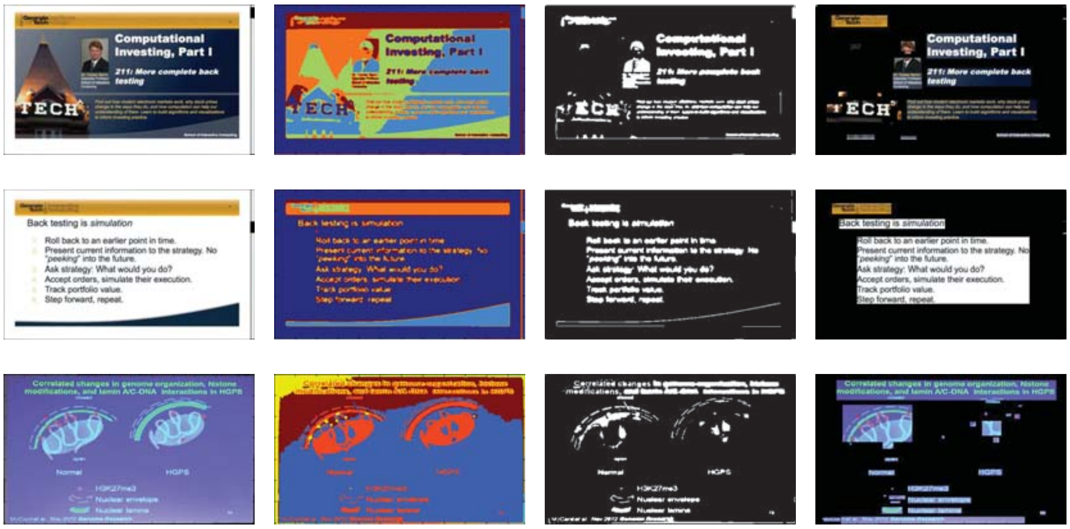
In this section, we present our two stage algorithm for understanding the layout of text in slides. In the first step, pixels of slides are classified as either background or foreground. Pixels belonging to text should be part of the foreground. Generally, all characters of similar font size present in close proximity belong to a single column block of text. Therefore they should be passed to the OCR engine together. In the second step, the algorithm determines font size of each foreground region and then merges neighbouring foreground regions of similar font size to get individual block of text. Once layout is determined, the various parts of the video frame representing individual blocks of text are passed separately to Tesseract OCR engine for text recognition.

4.1 Foreground and background segmentation

Alphabets and words present in slides can be seen as group of pixels of similar colours. To separate text from the background we use colour based segmentation using *kmeans*. There typically will only be a limited number of colours present in slide so a small values of k is used. Figure 2(b), shows results of *kmeans* clustering for $k = 5$, where k is the number of cluster centres. The k colours are then further segmented into regions using the standard connected components technique. Each of these regions should either correspond to foreground (individual letters or words) or background. Normally, regions corresponding to text will be smaller than the regions corresponding to the background. Therefore, a size based metric is used to separate foreground and background regions. All pixels p belonging to any region c are classified as follows:

$$c(p) = \begin{cases} foreground & if\ c_{area} < \frac{1}{\gamma} \times I_{area} \\ background & otherwise \end{cases} \qquad (1)$$

where, γ is constant, i.e, $1 \leq \gamma \leq I_{area}$; c_{area} and I_{area} represent number of pixels in region c and slide frame I respectively. Figure 2(c) shows foreground (white) and background (black) regions where $\gamma = 40$ is used for all three slides.


Computational Investing, Part I
211: More complete back testing
TECH
Back testing is simulation
Roll back to an earlier point in time.
Present current information to the strategy. No "peeking" into the future.
Ask strategy: What would you do?
Accept orders, simulate their execution.
Track portfolio value.
Step forward, repeat.
Correlated changes in genome organization, histone modifications, and lamin A/C-DNA interactions in HGPS
Normal
HGPS
H3K27me3
Nuclear envelope
Nuclear lamina

(a) (b) (c) (d)

Figure 2: Text location. Slides (a), colour based segmentation using *kmeans* (b), foreground and background segmentation (c) and separated text sections (d).

4.2 Joining foreground regions based on font size and proximity

Foreground regions created in the previous section can represent either alphabets or words or sentences or noise. We used median height of each foreground region as its font size. Foregrounds regions of same font size in close proximity within $(n \times font\ size)$ distance from region boundaries are merged together. After merging all possible foreground regions, each joint region holds all characters of a single block of text. Bounding boxes of final foreground regions are used to mark the block of texts in slide frames. Figure 2(d) shows the foreground regions located in the slide. These block of texts are passed one by one to OCR engine. In the slides some sections of images also were marked as blocks of text (See Figure 2). These false foreground regions do not effect the accuracy of the algorithm because for these regions the OCR engine does not return any text data.

5 Creating XML from detected text

OCR data generated by the OCR engine for whole video sequence is saved in a single XML file. A XML element contains:

```
<Sentence>
        <Text>. . . </Text>
        <UID>. . . </UID>
        <StartTime>. . . </StartTime>
        <EndTime></EndTime>
</Sentence>
```

where *Text* element contains text present inside a foreground region which appears and disappears together. Therefore, text inside a single foreground region in a frame can be represented by different *Sentence* elements in the XML file. *UID* is a unique identification number, *StartTime* is the time when text present inside *Text* element appeared in video and *EndTime* is *StartTime* plus length of time for which text is continuously present in the video.

Video adaptation tool uses this XML file to create customised video content based on user requirements. For example, if user wants to learn about a particular topic, the video adaptation tool searches for the topic in the *text* elements. If topic is present then customised video can be created by extracting frames from video using start and end time information.

6 Initial Results and Future Plans

Our initial tests have been restricted to videos from the Coursera dataset such as those shown in Figure 1. Four videos similar to the one in Figure 1(a) were evaluated which contained roughly 2800 words in 169 slides. Words in the body of the slides were detected with a precision of 93% and a recall of 99%. One video shown in Figure 1(b) was also evaluated giving a precision of 98% and a recall of 94% for the 570 words in the body of the 19 slides.

These initial tests are simply a precursor to more thorough testing to be conducted following a conference which is being held during May 2014, where all sessions will be recorded. At this event it is intended that the original presentations will also be kept and hence our task may be changed from recognising the text to recognising the slides, as well as determining when one presentation has finished and when another is starting. A demonstrator system will be built based on this conference.

There is other information which can be extracted from lecture presentations, such as images of the presenter and the audience. In addition we need to be able to identify (and extract) any demonstrations (e.g. videos or animations) from the slide presentations and not regard these as slide transitions. The ultimate goal of this work is not simply to extract text information but rather to allow users to access video data efficiently, obtaining the relevant part(s) as easily as possible.

References

[1] Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., and Rowe, L. (2010). TalkMiner: a lecture webcast search engine. In *Proceedings of the ACM international conference on Multimedia (MM '10)*, page TalkMiner: a lecture webcast search engine.

[2] Chang, S.-F. and Vetro, A. (2005). Video adaptation: Concepts, technologies, and open issues. *Proceedings of the IEEE*, 93(1):148–158.

[3] Ianeva, T., Boldareva, L., Westerveld, T. H. W., Cornacchia, R., Hiemstra, D., and de Vries, A. P. (2005). Probabilistic approaches to video retrieval. In *TREC Video Retrieval Evaluation Online Proceedings (TRECVID 2004), Gaithersburg, MD, USA*, TREC Video Retrieval Evaluation Online Proceedings. National Institute of Standards and Technology (NIST).

[4] Mukhopadhyay, S. and Smith, B. (1999). Passive capture and structuring of lectures. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 477–487, New York, NY, USA. ACM.

[5] Ngo, C.-W., Pong, T.-C., and Huang, T. (2002). Detection of slide transition for topic indexing. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 2, pages 533–536 vol.2.

[6] Otsu, N. (1979). A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66.

[7] Smith, R. (2007). An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, pages 629–633, Washington, DC, USA. IEEE Computer Society.

[8] Smith, R. (November 1987). *The Extraction and Recognition of Text from Multimedia Document Images*. PhD thesis, University of Bristol.

IMVIP 2014

IMAGE PROCESSING

Constant colour matting with foreground estimation

Guillaume GALES
National University of Ireland Maynooth
Department of Computer Science
guillaume.gales@nuim.ie

John MC DONALD
National University of Ireland Maynooth
Department of Computer Science
johnmcd@cs.nuim.ie

Abstract

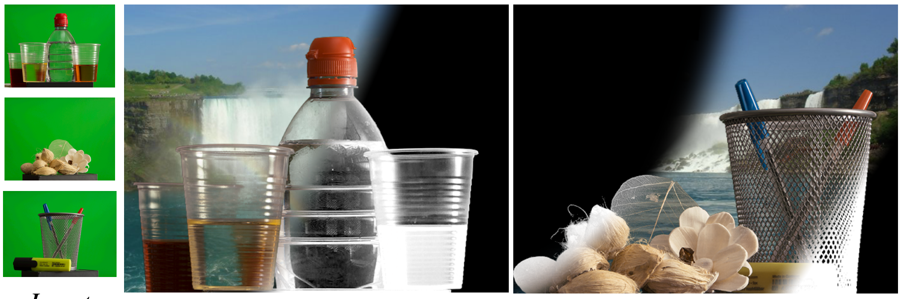
Constant colour matting consists of estimating for each pixel of an image the proportion α of an unknown foreground colour with a known constant background colour. The α-matte is then used to replace this background with another image. Existing approaches approximate α directly but post-processing is required to remove spill of the background colour in semi-transparent areas. Instead of estimating α directly, we propose 3 methods to estimate the unknown foreground colour, and then to deduce α. This approach leads to high quality mattes for transparent objects and allows spill-free results (see Fig. 1). We show this through an evaluation of the proposed methods based on a ground truth dataset.

Keywords: constant colour matting, foreground colour estimation, α estimation.

1 Introduction

Matting is a classic problem that consists of creating a *matte* to mask an unwanted area from video footage. This area is then replaced by content from separate footage to create a *composite*. Cinema, television and web TV use this technique extensively for visual special effects. Although the principle of this technique is simple, it is often difficult to achieve a realistic seamless result. This is particularly true where the observed colour is a blend from foreground and background. The proportion of foreground colour F, and background colour B for one pixel is called α. This typically occurs at the boundary between both areas, with semi-transparent foreground, or blur. We distinguish between two types of matting techniques: *constant colour matting* techniques, where the background colour is known, or *natural image matting*, working with an arbitrary background. The former methods are widely used in visual productions but post-processing is usually required to clean up the matte and to deal with spill (background colour reflecting in the foreground). The latter methods can give impressive results but require additional prior information and are more computationally expensive.

In this paper, our goal is to estimate the F and α, knowing B, to obtain high quality results in difficult areas exhibiting fine and semi-transparent details. Unlike others constant colour methods, we start by estimating F, inspired by natural image matting approaches, but without



Input

Figure 1: Examples of final results obtained with the proposed method.

the need to input additional information. Our contributions are three methods for the estimation of F (and consequently α) that give mattes requiring little or no cleaning, and produce spill-free results. Furthermore, our algorithm is highly parallelisable (working independently on each pixel) with a low computational complexity. We also provide a ground truth dataset which we use to demonstrate and quantitatively evaluate the performance of our technique.

After a brief description of the state of the art, we detail our approach. Then, we provide an evaluation of the three methods based on a a proposed ground truth dataset.

2 Previous work

2.1 Constant colour matting

Constant colour matting aims at estimating a matte from images where the background is assumed to be a constant colour (usually blue or green).

In [Smith and Blinn, 1996], the authors formalise the problem of constant colour matting as follows. For each pixel, we express the observed colour (O, known) as the proportion $\alpha \in [0; 1]$ (unknown) of foreground (F, unknown) and background (B, known) colours. The *matting equation* is given by:

$$\underbrace{\begin{bmatrix} r & g & b \end{bmatrix}^\top}_{O} = \alpha \underbrace{\begin{bmatrix} F_r & F_g & F_b \end{bmatrix}^\top}_{F} + (1-\alpha) \underbrace{\begin{bmatrix} B_r & B_g & B_b \end{bmatrix}^\top}_{B} \tag{1}$$

There are four unknowns for three equations and therefore the problem is underdetermined, i.e. it exists no or many solutions. The authors identify three cases where a solution can be found: no blue in the observed colour, the observed colour is gray or two different shades of background are known. However, these cases do not usually occur in practice.

We can distinguish between the following types of method to estimate α with a constant background colour:

– *Colour difference technique* – α is based on differences between the red, green and blue components. This method (a.k.a *Ultimatte*®), invented by [Vlahos, 1964], is a legacy of an optical multistep process where colour filters are placed in front of an optical printer to filter out the background colour (an interesting history of matting in filmmaking is given in [Filmmaker IQ, 2013]). This process is usually summarised by $\alpha = 1 - \max(0, b - \max(r, g))$ where B is assumed to be blue. Spill is then removed by changing the blue component to $b \leftarrow \min(b, kg)$ where k is a user control parameter.

– *Colorspace segmentation* – The colorspace is partitioned into background, foreground and semi-transparent regions. In the semi-transparent region, α is based on the distance between the other two. In [Ashikhmin, 2001, Jack, 1996], they use the YC_bC_r colorspace to separate the luminance (Y) and the chrominance (C_bC_r). The segmentation is performed in the 2D chrominance space. A classic approach, called *Hue Saturation Luminance keying* and described in [Schultz, 2006], consists of segmenting the HSL colorspace to isolate the background and foreground regions. These segmentations are done manually by selecting the center and the dimensions of basic shapes (simple polyhedron or sphere) that encapsulates the different regions. The *Primatte*® algorithm by [Mishima, 1992], is based on this principle, however it automatically adjusts a 128 face polyhedron to obtain an fine segmentation of the colorspace.

2.2 Natural image matting

Natural image matting aims at estimating a matte from images with arbitrary background. These methods use optimisation techniques to estimate the best combination of α, F and B that minimises an objective function based on spatial statistical models. To build these models, they require a pre-segmentation (called *trimap*) in three classes: foreground, background and unknown. For example, in [Chuang et al., 2001], the algorithm marches inward from known to unknown regions. It uses the colour distribution in a weighted window of the known (or already computed) neighbouring regions to estimate the most likely combination of α, F and B.

In [Rother et al., 2004], an adapted iterative graph cut optimisation method is used to minimise an objective function that takes into account a fitting term (how close the solution is to what is observed) and smoothness term (to prevent abrupt changes of α between two neighbours). These methods use local sampling for the estimation of the foreground and background colours. In [He et al., 2011], the authors obtain good results with a global sampling approach for the estimation of F and B. A detailed survey of such methods is provided in [Wang and Cohen, 2007].

3 Our approach

Our approach starts by building, for each pixel, a set of candidates for the foreground colour F. These candidates are computed from a set of predefined possible colours $\mathcal{C}$ for the image. We propose three independent methods for this computation. Then, we assign to F the best candidate according to a distance measure between the candidate and the observed colour. Finally, assuming F known, we can calculate α.

First, we describe how we obtain the set of possible colours $\mathcal{C}$. Then, we give the general structure of the algorithm. Finally, we present the three methods to estimate F.

3.1 Set of possible foreground colours for the image

We reduce the solution space for the foreground colour F using the following constraints:
(i) F should be already present in the image. – As in natural image methods, the first assumption is that F, is already present in the image. To obtain a set of possible colours for F, we calculate the modes of the colour distribution of the image. To do so, we use the *mean-shift* algorithm on the image histogram, [Comaniciu and Meer, 2002] using the *Lab* colorspace as it is perceptually uniform. The output is a set $\mathcal{C}_0$ of colour clusters.
(ii) F should be distinct from B. – We need to remove from the set $\mathcal{C}_0$ the clusters C_i that are too close to B: C_i is removed from $\mathcal{C}_0$ if $\|\overrightarrow{BC_i}\| < t_1$. Let $\mathcal{C} = \mathcal{C}_0 \setminus \mathcal{C}_B$ with $\mathcal{C}_B = \left\{ C_i \middle| \|\overrightarrow{BC_i}\| < t_1 \right\}$.
(iii) If $B \neq O$, F lies on the line passing through B and O. – By definition of Eq. (1).

3.2 Algorithm

Algorithm 1 gives a high-level description of the approach used to estimate F and α from a set of possible colour clusters $\mathcal{C}$. It starts by looking at the distance between B and O. If $B = O$, the given pixel is a background pixel. On the other hand, if this distance is large enough (above a threshold th_1), one can be confident that the given pixel is a foreground pixel. For the other pixels, each colour cluster $C_i \in \mathcal{C}$ is used to estimate F_i and a cost c, as described in the Section 3.3, depending on the chosen method m. This cost evaluates how "close" C_i is to the line BO . If it is too "far away" (above a threshold th_2) from BO (not satisfying the condition (iii)), the estimated F_i is rejected. Finally, if no candidate can be found, we assume the given pixel belongs to the background. If more than one candidate are found, we chose the closest one from the observed colour O.

3.3 Estimation of F

Ideally, if we can find exactly one colour cluster C_i lying on the line BO, satisfying the condition (iii), we could assume $F = C_i$. In practice, this alignement does not occur because of noise and because the colour clusters correspond to the mode of the colour distribution for F. To deal with this issue, we propose three independent methods to estimate F_i from a cluster C_i (see Table 1):

(a) This method minimises the sum of squared residuals of the matting equation system where F is replaced by C_i. Geometrically this solution minimises $\|\overrightarrow{O'_iO}\|$ where O'_i is the orthogonal projection of the observed colour O on BC_i. Thus, F_i is given by the intersection of the line BO and the plane passing through C_i orthogonal to the line BC_i. The cost c_i is given by: $c_i = \|\|\overrightarrow{C_iF_i}\|\|$.

Algorithm 1: Algorithm for one pixel.

```
Data: $O, B, \mathcal{C}, th_1, th_2, m$
Result: $F, \alpha$
if $\|\overrightarrow{BO}\| = 0$ then
    $F \leftarrow \mathbf{0}$ ; $\alpha \leftarrow 0$                    /* This is a background pixel. */
else
    if $\|\overrightarrow{BO}\| > th_1$ then
        $F \leftarrow O$ ; $\alpha \leftarrow 1$                    /* This is a foreground pixel. */
    else
        $\mathcal{L} \leftarrow \varnothing$                    /* Initiate a list of candidates. */
        for each $C_i \in \mathcal{C}$ do
            $F_i \leftarrow$ estimate F with a method $m$ using ($O$,$B$,$C_i$)
            $c_i \leftarrow$ cost for this $F_i$
            if $c_i < th_2$ then
                $\mathcal{L} \leftarrow \mathcal{L} \cap F_i$                    /* $F_i$ is a candidate. */
        if $|\mathcal{L}| = 0$ then
            $F \leftarrow \mathbf{0}$ ; $\alpha \leftarrow 0$                    /* No candidate → background. */
        else
                                                  /* Get the best candidate. */
            $F \leftarrow F_i \in \mathcal{L} \,\big|\, \forall F_i, F_j \in \mathcal{L}^2, \|\overrightarrow{OF_i}\| \leq \|\overrightarrow{OF_j}\|$ ; $\alpha \leftarrow \frac{\|\overrightarrow{BO}\|}{\|\overrightarrow{BF}\|}$
```

(b) In this method, F_i is given by the orthogonal projection of the C_i onto the line BO. The cost is given by the Euclidean distance between C_i and F_i: $c_i = \|\overrightarrow{C_iF_i}\|$.

(c) This method rotates C_i around B with an angle $\theta_i = \cos^{-1}\left(\frac{\overrightarrow{BC_i}\overrightarrow{BO}}{\|\overrightarrow{BC_i}\|\|\overrightarrow{BO}\|}\right)$ so that B, O and C_i are collinear. The cost is given by the rotation angle: $c_i = \theta_i$.

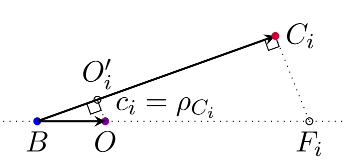
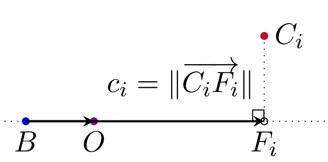
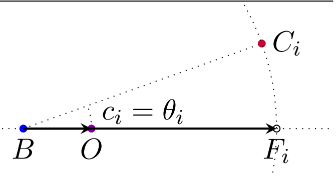
(a)	(b)	(c)
 C_i O'_i $c_i = \rho_{C_i}$ B O F_i	 C_i $c_i = \|\overrightarrow{C_iF_i}\|$ B O F_i	
$F_i = \frac{\overrightarrow{BO}\|\overrightarrow{BC_i}\|^2}{\overrightarrow{BO}\cdot\overrightarrow{BC_i}} + B$	$F_i = \frac{\overrightarrow{BO}(\overrightarrow{BO}\cdot\overrightarrow{BC_i})}{\|\overrightarrow{BO}\|^2} + B$	$F_i = \frac{\overrightarrow{BO}}{\|\overrightarrow{BO}\|}\|\overrightarrow{BC_i}\| + B$

Table 1: Illustration of the 3 methods (a), (b) and (c) proposed to estimate F_i in the Algorithm 1.

4 Evaluation and results

Ground truth dataset To provide a quantitative evaluation of our three methods, we created a ground-truth dataset. As explained in [Smith and Blinn, 1996], if at least two different shades of background are known, Eq. 1 becomes overdetermined and we can estimate α and F. We took pictures of six different objects ($bath$, $bottle$, $muppet$, $pot1$, $pot2$ and $spider$, see Table 3) in front of five different backgrounds (blue, green, black, yellow and red) for further overdetermination. Then, we calculated a least squared solution for αF and α:

$$\begin{bmatrix} -B_{blue} & -B_{green} & \dots \\ \mathbf{I}_3 & \mathbf{I}_3 & \dots \end{bmatrix}^\top \begin{bmatrix} \alpha \\ \alpha F \end{bmatrix} = \begin{bmatrix} (O_{blue} - B_{blue}) & (O_{green} - B_{green}) & \dots \end{bmatrix}^\top \quad (2)$$

where $\mathbf{I}_3$ is the 3×3 identity matrix.

Parameters Each method requires two parameters for the initial clustering: the size of the bins s_b, and the size of the mean-shift window s_{ms}. It also requires th_1, the minimum $\|\overrightarrow{BF}\|$

distance, and th_2 the threshold on the cost depending on the method employed, see § 3.2. According to an initial experiment, (s_b, s_{ms}) can be fixed to $(2, 4)$ for the method (a) (giving about 60 clusters) and to $(2, 2)$ for the methods (b) and (c) (giving about 400 clusters). The best results (according to the criterion described below) are obtained with $th_1 \approx 40 \pm 5$ and (a) $th_2 \approx 2$; (b) $th_2 \approx 10$; (c) $th_2 \approx 0.4$ (see Tables 3 and 2). These values can then be fine tuned interactively.

Measure of error For each of the three methods m, we evaluate the results obtained with a different set of values for th_1 and th_2. We chose to evaluate the methods using the images having a green background. As we are interested in F and α, we compare the estimated values with the ground truth. The mean squared error for an image is given by:

$$MSE_{th_1,th_2,m} = \frac{1}{hw} \sum_{i,j=0,0}^{h-1,w-1} \|(\alpha F)_{i,j} - (\overline{\alpha F})_{i,j}\| \tag{3}$$

where (h, w) are the dimensions of the image, $(\alpha F)_{i,j}$ is the estimated value for the pixel at coordinates (i, j) and $(\overline{\alpha F})_{i,j}$ is the ground truth value for the pixel of same coordinates.

Tables 2 and 3 show the results. With the appropriate set of parameter values, the three methods can achieve results with a low error score. The best average values for each method are $MSE_{45,3,(a)} = 3.339$, $MSE_{50,5,(b)} = 2.814$ and $MSE_{50,0.4,(c)} = 2.811$ showing that (b) and (c) perform slightly better than (a). In some scenes, (b) and (c) clearly outperform (a).

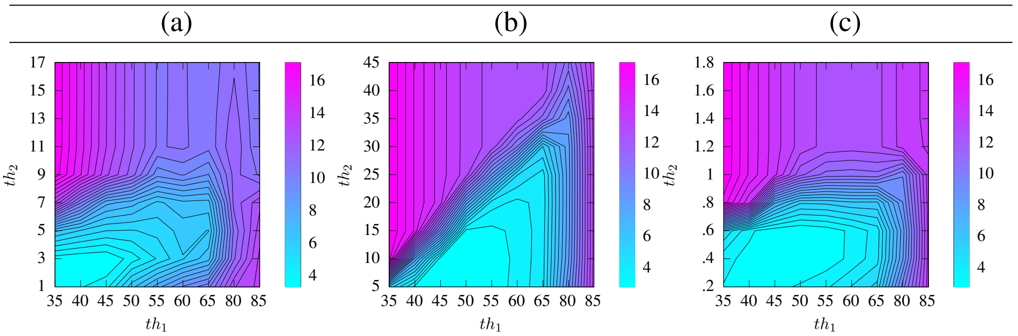
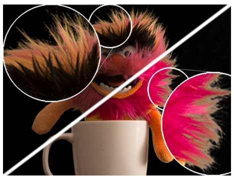


Table 2: Contour plots of the averaged $MSE_{th_1,th_2,m}$ over the 6 images for each of the three methods in function of th_1 and th_2.

5 Discussion and conclusion

We proposed three methods to estimate F and α in a constant colour background matting problem. According to our evaluation, these methods give very encouraging results.

A comparison with commercial methods would be interesting. However, source code is not available, executables are not free and they may include extra post processing to give a visually appealing, but not mathematically accurate, result. These issues make a fair and rigorous comparison difficult. But, to permit a qualitative evaluation of how our algorithm may compete, we visually compared our result (left) with one obtained using *Apple Motion HSL Keyer* (right). The latter one has more spill. We also noticed one inconsistent ground truth data in *Pot1* where the opaque fluorescent marker is considered to have some transparency. Although the variances in the observed values are relatively small (due to noise and background indirect illumination), the residual of Eq. 2 is quite large. We propose to re-estimate the ground truth using a robust



(a) *Our method*

(b) *Apple Motion HSL Keyer*

Figure 2: Visual comparison with a commercial solution.

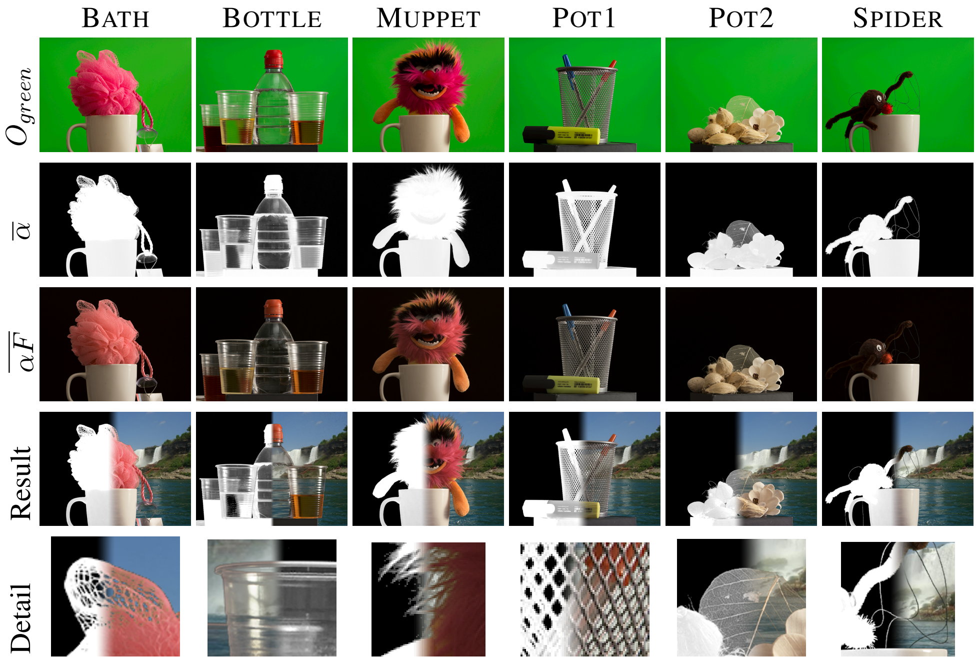


Table 3: Input (green background), ground truth dataset ($\overline{\alpha F}$, $\overline{\alpha}$) and best result with our method. Also available `www.cs.nuim.ie/research/vision/data/imvip2014/`

estimation method (e.g. LTS). We propose in a future work to refine our evaluation with more detailed criteria (examining errors in difficult areas only, sensitivity to noise), cross-validation for the choice of parameter values and provide a GPU-based implementation.

References

[Ashikhmin, 2001] Ashikhmin, M. (2001). Hight quality chroma key http://www.cs.utah.edu/ michael/chroma/. Technical report.

[Chuang et al., 2001] Chuang, Y.-Y., Curless, B., Salesin, D. H., and Szeliski, R. (2001). A bayesian approach to digital matting. In *CVPR*, volume 2, pages 264–271. IEEE Computer Society.

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *PAMI*, 24(5):603–619.

[Filmmaker IQ, 2013] Filmmaker IQ (2013). Hollywood history of faking it. the evolution of greenscreen compositing http://filmmakeriq.com/lessons/hollywoods-history-of-faking-it-the-evolution-of-greenscreen-compositing/.

[He et al., 2011] He, K., Rhemann, C., Rother, C., Tang, X., and Sun, J. (2011). A global sampling method for alpha matting. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*.

[Jack, 1996] Jack, K. (1996). *Video Demystified*. Newnes, Newton, MA, USA, 5th edition.

[Mishima, 1992] Mishima, Y. (1992). A software chromakeyer using polyhedric slice. In *NICOGRAPH*.

[Rother et al., 2004] Rother, C., Kolmogorov, V., and Blake, A. (2004). "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314.

[Schultz, 2006] Schultz, C. (2006). Digital keying methods. Technical report, University of Bremen Center for computing Technologies TZI.

[Smith and Blinn, 1996] Smith, A. R. and Blinn, J. F. (1996). Blue screen matting. In *SIGGRAPH*, pages 259–268, New York, NY, USA. ACM.

[Vlahos, 1964] Vlahos, P. (1964). Composite color photography. US Patent 3,158,477.

[Wang and Cohen, 2007] Wang, J. and Cohen, M. F. (2007). Image and video matting: A survey. *Found. Trends. Comput. Graph. Vis.*, 3(2):97–175.

Combination method: Photometric stereo with Shadows

Amina Dulac, Sandy Martedi, Hideo Saito
Graduate School of Science and Technology
Keio University
3-14-1 Hiyoshi Kohoku-ku, Yokohama 223-8522, Japan
amina.dulac@centrale-marseille.fr, sandy@hvrl.ics.keio.ac.jp, saito@hvrl.ics.keio.ac.jp

Kouichi Tezuka, Masayoshi Shimizu
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa 211-8588, Japan
ktezuka@jp.fujitsu.com, shimizu.masa@jp.fujitsu.com

Abstract

This paper presents an algorithm allowing to perform photometric stereo with high accuracy results in the presence of shadows. First, we choose a combination of three images suited to reveal the presence of shadow and use it to calculate the normal vector on each pixel. This normal vector is then used to recalculate the intensity in order to evaluate the distance with the real intensities and detect the shadow pixels on each image. If the error between the calculated intensities and the real intensities are superior to a certain threshold, the pixel is determined to be in the region of a shadow. The algorithm runs in loop to identify a valid set of images, that is to say without shadow, for each pixel. The method is designed to limit the computational cost. The results show that the accuracy is maintained as compared to much heavier computational cost algorithms.

Keywords: photometric stereo, shadows, 3D reconstruction.

1 Introduction

Photometric Stereo is a method used to recover the 3D shape of an object with high detail accuracy. For that it uses a set of images of the object from the same point of view but under different illuminations. In case of calibrated photometric stereo, which is the one being presented in this paper, the light sources directions are known. The approximation of Lambertian surface is also made. One remaining issue in photometric stereo is how to handle shadow areas. Some papers proposed algorithms to deal with the shadows in photometric stereo, however all the proposed methods require heavy calculations and are not suitable for real time processing. In this paper we describe a new algorithm with reduced time of calculation and producing results with as good accuracy as heavier methods. This method was developed in order to be used for 3D shape reconstruction of the human skin by photometric stereo technique for medical control.

This paper is separated into four sections. In the first section, we provide an overview of previous work. In section 2 we introduce our methods and in section 4 we conclude with results.

2 Related work

Photometric stereo is considered as an interesting intensity-based 3D shape recovering technique because of the precision of the results it can achieve on objects with Lambertian reflectance. Methods have also been developed to handle specular surfaces [7].The theory called calibrated photometric stereo requires known the light sources directions. Additionally, the principles of this method have now been extended to uncalibrated photometric stereo where the light sources directions are unknown [3]. This paper however uses calibrated photometric stereo since our application uses known light sources and focuses on the accuracy of the results as opposed to overcoming such constraints.

The main difficulty in the accuracy of the results lies in the presence of shadows. In order to have results reliable enough, especially for a medical use, the shadow areas need to be detected and handled properly. The traditional way to remove the shadows of an image is to apply an intensity threshold. However this simple technique only works on objects with constant surface albedo coefficient. Chandraker et al. [6] proposed a method allowing shadow labeling in photometric stereo with changing albedo and mutli-light sources using energy minimization where the "data term" is based on photometric stereo and the "smoothness term" supports the spatial continuity. Due to the exponential number of possible label configurations, minimizing this type of energy requires the use of a fast graph cuts algorithm from Boykov et al.[10]. However, even with this type of approximation the computational cost stays very high. Sunkavalli et al. [8] also proposed a method dealing with shadows in uncalibrated photometric stereo. Instead of reasoning about per-pixel intensity, their approach is reasoning about illumination subspaces using a RANSAC type algorithm, which needs about 1000 iterations thus requiring a large amount of time to compute.

Our work is a per-pixel approach but using targeted combinations to decrease the amount of calculation required. We show that by not considering the totality of the possible labels, by choosing the right combination for each pixel, we can achieve a similar degree of accuracy and allow for high-speed execution.

3 Identifying shadows

For clarity we will briefly describe notations. We consider a set of n images of a same object from the same point of view. Each image is illuminated by a light source j and has m surface points. This section begins by describing the photometric stereo theory. We then explicitly show the impact of the shadows in the equations and propose an algorithm to detect pixels corrupted by shadow in each image.

3.1 Calibrated Photometric Stereo

Photometric Stereo is a technique used to recover the 3D shape of an object from 2D images of it. The objects considered in this paper are assumed to have Lambertian reflectance. Given a set of images of the same object, from the same point of view but illuminated by different known light sources, the intensity of the pixel i in the image j is expressed as:

$$c_{ij} = \rho l_j^T n_i \, , \qquad (1)$$

where l_j is the light source direction vector, n_i the normal vector at the pixel i and ρ the albedo coefficient.

For m surface points and n>3 light sources, the concatenation of all pixels on all images leads to the following system:

$$I = L^T N \ , \qquad (2)$$

where I indicates a $n \times m$ intensity matrix, L denotes a $3 \times n$ light source matrix and N is a $3 \times m$ matrix containing the product of the albedo coefficients and normal vectors. For a number of linear independant light sources superior to 3, L is at least of rank 3. The normal matrix N can be recovered using the pseudo inverse of L as:

$$N = (L^T)^+ I \ . \qquad (3)$$

The relative depth of each point can then be calculated by integrating the normal vectors.

3.2 Shadows

One limitation of this method comes from the presence of shadows. When a point of the surface of an object is not reached by the light the Photometry stereo theory is not applicable anymore. Areas of shadows can produce corrupted results in the normal map and thus lead to an inaccurate recovery of the depth. If we write equation (3) for one pixel we have:

$$\begin{cases} N_x = \sum_{j=1\ldots n} c_j R_x(L) \\ N_y = \sum_{j=1\ldots n} c_j R_y(L) \ . \\ N_z = \sum_{j=1\ldots n} c_j R_z(L) \end{cases} \qquad (4)$$

The coefficients $R_x(L), R_y(L)$ and $R_z(L)$ only depends on the lights source directions, which are known in the case of calibrated photometric stereo. When a pixel is in the shadow, only the term c_j representing its intensity is corrupted in the equation: the intensity value will be lower than it should be because of the shadow. Thus we can naturally infer that when the percentage of images touched by shadow for a pixel is low, the influence on the calculated normal vector for this pixel is almost undetectable. If the percentage of shadow images is however too important the calculation leads to a highly inaccurate normal vector. Based on this simple observation we can deduce that if we apply photometric stereo on three images among which one contains shadows, the normal vector of the pixels in the shadow will be highly corrupted, making it easier to detect. Consequently, a method based on combinations of three images to detect the shadows can be used.

3.3 Algorithm

We consider a set of n images containing some shadows. We also make the hypothesis for each pixel that there is at least three images without shadow. If that is not the case, Photometric Stereo theory is not applicable. We assume that the hypothesis is respected by choosing a number of images important enough and light source directions offering a good coverage of the object. The previous observations lead us to the following per-pixel algorithm:

1. Sort the images by intensity. Notice that we first convert the images into grayscale images. We organize this group of sorted images from the brightest to the darkest $\{im_1, im_2 \ldots im_n\}$.
2. Apply three sources photometric stereo using the darkest image im_n and two of the three brightest images in order to recover the normal vector N. The image choice relies on the fact that we consider that the three brightest images do not contain shadows and that if the pixel is in the shadow in some images then the darkest image will contain shadows. Thus we get a combination of three images, two of which without shadows and one potentially containing shadows.

3. Use the recovered normal vector N to recalculate the intensity $I_j = L_j{}^T N$ on each image j and calculate the global relative error with the real intensities $E_{tot} = \sum_{j=1}^{n} \left|\frac{I_j^{real} - I_j}{I_j^{real}}\right|$.
4. If $E_{tot} < \alpha$, we can regard the pixel as non-affected by shadow because the calculated intensities I_j are almost equal to the real intensities I_j^{real}. Then we attribute the set of images as a valid set to this pixel. If $E_{tot} > \alpha$, which mean that the calculated intensities I_j are too different from I_j^{real}, then we label this pixel as a shadow pixel in image n and repeat step one to three on the set of images $\{im_1, im_2 \dots im_{n-1}\}$. The threshold α is chosen according to experimental results.

The loop runs until every pixel is attributed a valid set of images or until there is only three images left in the set of images, so there is maximum $n - 3$ loops and the number of combinations calculated is a $O(\binom{n}{3})$. As a comparison, the method from [6] requires the calculation of a $O(\sum_{k=3}^{n} \binom{n}{k})$ combinations and the fast graph cut algorithm needs about 100 iterations to reach a good minimum. The method developed in [8] does not operates on a per-pixel basis but still has a high computational cost because of its RANSAC type procedure. If we consider that the image contains h visibility subspaces, each occupying the same proportion of space then for m pixels the number of operations is $\sum_{k=1}^{h} 1000 \times (m - (k-1)\frac{m}{h})$.

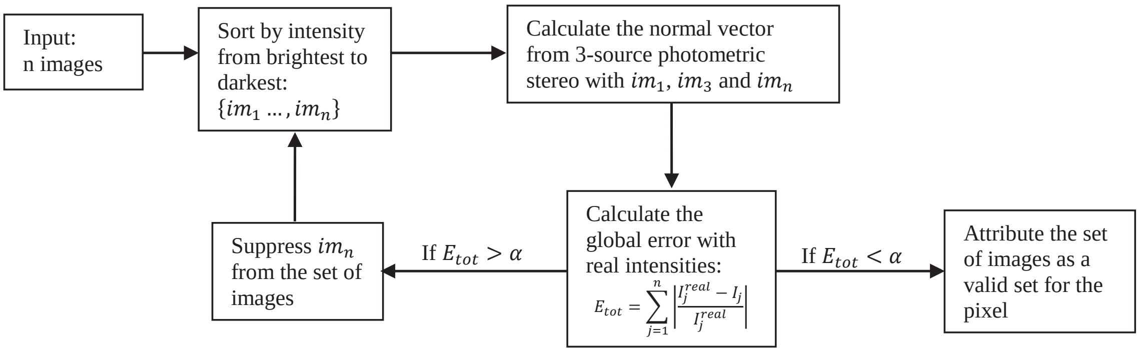

Input:
n images
Sort by intensity from brightest to darkest: $\{im_1 \dots, im_n\}$
Calculate the normal vector from 3-source photometric stereo with im_1, im_3 and im_n
Calculate the global error with real intensities: $E_{tot} = \sum_{j=1}^{n} \left|\frac{I_j^{real} - I_j}{I_j^{real}}\right|$
If $E_{tot} > \alpha$
Suppress im_n from the set of images
If $E_{tot} < \alpha$
Attribute the set of images as a valid set for the pixel

Figure 1. schema of the algorithm.

4 Results

In this section we present the results obtained with our method on synthetic data and real data. The first example we show is a synthetic set of images of a simple sphere on a plane (Figure 2). This dataset contains both attached and cast shadows and the albedo coefficient differs from the sphere to the plan. The multiple shadows are also overlapping, creating a more complex map of valid sets but the symmetry of the problem makes it easy to judge on the accuracy of the results. We can notice the difference of the normal map obtained by using classic photometric stereo where the influence of shadows is explicitly corrupting the result and with our shadow handling method where the marks left by the shadows do not appear anymore.

The second dataset presented contains eight images of a real object with a complex shadow map (Figure 3). Each color on the shadow map represents a different valid set of images, this means pixels sharing the same color are in the shadow in the same images. Please note that the color is decided arbitrary for each valid set of images. We can notice the produced shadow map seems very close to the true shadow map of the object. For comparison we show the shadow map obtained using the algorithm from [8], the accuracy of the shadows area is about as high as the one we achieve. Additionally, this is performed with a smaller computation time. The last example is a set of twelve pictures of a horse head shaped object (Figure 4). If we first look at the horse's ear we can see that the area is highly affected by shadows. As a consequence, the 3D reconstruction obtained from classic photometric stereo contains some artifacts. In this case a peak along the ear area is formed. Our method however avoids the formation of such inconsistencies by effectively detecting the shadows projected by the ear.

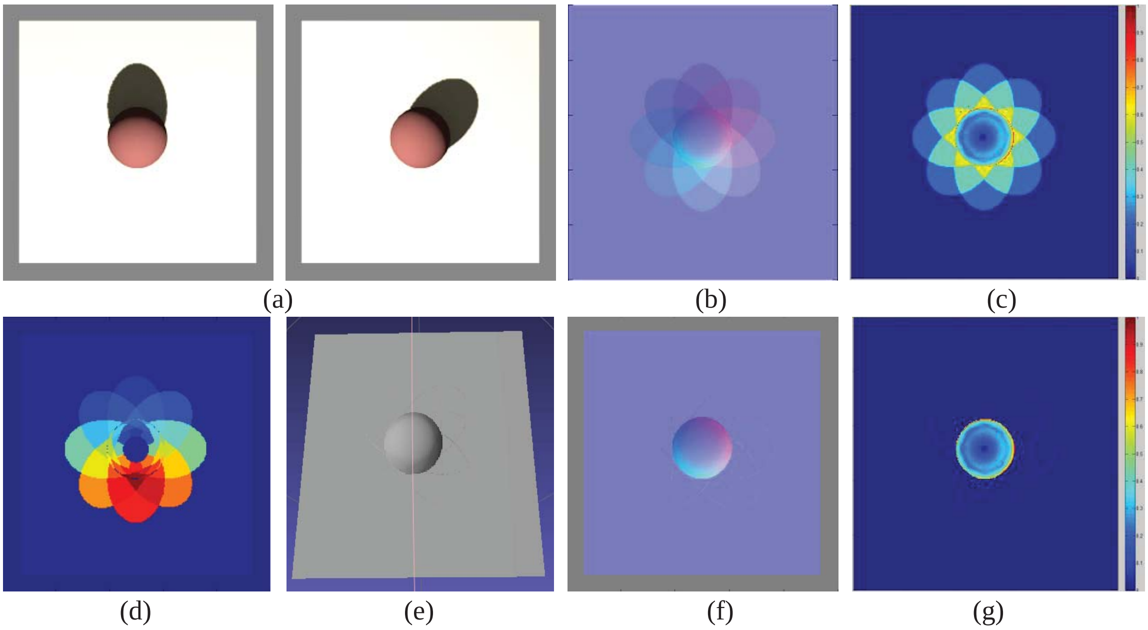


Figure 2. (a) 2 of the 8 input images of synthetic sphere on a plan. (b) normal map without applying our method. We can see the impact of the shadows. (c) error between the true normal map and the normal map (b) . (d) shadow map obtained with our method. (e) 3D reconstruction of the object after applying our method. (f) normal map obtained with our method, you can notice the disappearance of the marks previously made by the shadow areas. (g) error between the true normal map and the normal map obtained with our method (f).

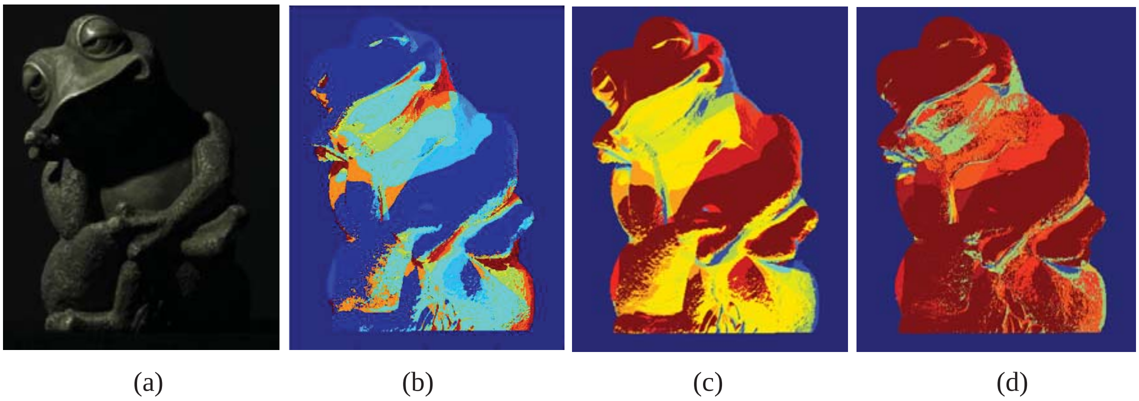


Figure 3. (a) sample input images of the statue of a frog. (b) shadow map obtained with our method. Pixels sharing the same color are in the shadow in the same images. Please note that the color of each valid set of images is chosen arbitrary. We can see how similar it is to the true shadow map (c) true shadow map taken from [8]. (d) shadow map obtained with method from [8].

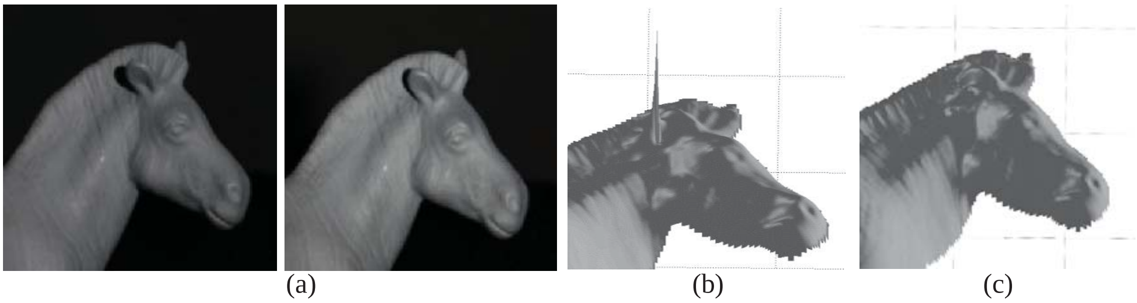


Figure 4. (a) sample input images of the statue of a head of a horse. (b) 3D reconstruction obtained from classic photometric stereo. We can notice the pic the area of the ear due to cast shadows. (c) 3D reconstruction obtained from our method. The artifact created by the shadows of the ear have been removed. Note that image (b) and (c) look more pixelated than the input images due to the lost of definition caused by the passage from 2D to 3D.

5 Conclusion

We have presented in this paper a new method to detect both cast and attached shadows in photometric stereo with albedo variation and complex shapes. Using well-chosen combinations of three images to calculate the intensity error on a per-pixel basis allow us to avoid the heavy computational cost needed in the previous existing methods while conserving the high degree of accuracy in the results.

This paper only focuses on calibrated photometric stereo because it offers the best accuracy for medical applications such as human skin reconstruction, but one possible direction of work could be to extend the method to uncalibrated photometric stereo in order to use it in a wider range of applications.

References

[1] Horn, B. (1986) *Robot Vision*.

[2] Belhumeur, P.N., Kriegman, D.J., Yuille, A.L. (1999). The bas-relief ambiguity. Int. Journal of Computer Vision 35(1), 33-44.

[3] Basri, R., Jacobs, D., Kemelmacher, I. (2007). Photometric Stereo with general unknown lighting.Int. Journal of Computer Vision 72(3), 239-257.

[4] Barsky, S., Petrou, M. (2003). The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. PAMI, 25(10), 1239-1252.

[5] Hertzmann, A., Seitz, S. (2005). Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. PAMI, 27(8), 1254-1264.

[6] Chandraker, M., Kahl, F., Kriegman, D. (2007). Shadowcuts: Photometric stereo with shadows. Inc: Proc. IEEE conf. Computer Vision and Pattern recognition.

[7] Ikeushi, K. (1981). Determining surface orientations of specular surfaces by using the photometric stereo method. IEEE Trans. Pattern Anal. Mach, Intell, 3(6), 661-669.

[8] Sunkavalli, K., Zickler, T., Pfister, H., (2010). Visibility Subspaces: Uncalibrated Photometric Stereo with Shadows. Inc: ECCV 2, volume 6312 of Lecture Notes in Computer Science, page 251-264. Springer.

[9] Woodham, R. (1978). Photometric stereo: A reflectance map technique for determining surface orientation from image intensity. In SPIE, volume 155.

[10] Boykov, Y., Vexler, O., Zabih, R. (2001). Fast approximate energy minimization via graph cuts. PAMI, 20(12), 1222-1239.

Combining Detectors for Robust Head Detection

Henrik Brauer, Christos Grecos and Kai von Luck
Living Place - HAW Hamburg
Berliner Tor 11
20099 Hamburg, Germany
Henrik.Brauer@HAW-Hamburg.de

Abstract

This paper, addresses the problem of detecting heads in crowded real world scenes, by combining a human head, an upper-body and a body detector to create a robust head detector. The idea is not to rely on a single detector. Instead, a head, an upper-body and a body detector, are used for decision making by combining their individual opinions to derive a consensus decision. The combined classifier is tested on the town centre dataset, and results show an 18% reduction in log-average miss rate of our combined classifier and illustrate that combining classifiers may perform better than a single head detector.

Keywords: Pedestrian Detection, Head Detection

1 Introduction

Detecting humans is an important task for a wide range of applications, like surveillance, smart environments, or ambient assisted living. In crowded scenes such as shown in Figure 1, only some humans are fully visible; for many others, only the upper-body is visible, or even just the head. Such impediments led previous works such as [Benfold and Reid, 2011] and [Rodriguez et al., 2011] to rely on head detection and ignore the rest of the body. However, robust head detection is difficult to achieve, and our experiments indicate that head detection is not as reliable as full body detection (see Section 3), this is a significant drawback. These observations motivated us to combine detectors to create a more robust head detector. The idea is not to rely on a single detector. Instead, a head, an upper-body and a body detector, are used for decision making by combining their individual opinions to derive a consensus decision.

Person detection is a well-studied problem in computer vision with many methods and evaluation benchmarks available [Dollár et al., 2012]. Most of the methods consider full-body (pedestrians) or upper-body detection. In theory, the same algorithms can be used for head detection, but in practice, these algorithms do not achieve a satisfactory result. In order to overcome this problem, several authors have proposed strategies to exploit addition features.

Zhang et al. [Zhang et al., 2009] constructed a categorical model for hair and skin, and trained the models in four categories of skin representing the different illumination conditions (bright, standard and dark) to increase pedestrian detection rates during an occlusion event. Head detection using a skeleton graph is proposed in [Merad et al., 2010]. The skeleton graph is extracted from the foreground mask obtained using background subtraction.

In [Venkatesh et al., 2012], interest points are detected using gradient information in order to approximately locate top of head regions to reduce the search space. The interest points are then masked using a foreground region that were obtained by background subtraction. A sub-window is then placed around the interest points, and it is classified as a head or non-head region using an AdaBoost classifier. Xie et al. [Xie et al., 2012] detected heads using the histogram of gradients (HoG) feature. To improve the detection result, motion and appearance features are extracted and then the Bayesian posterior is used to represent the probability of detected region belonging to actual human head regions.

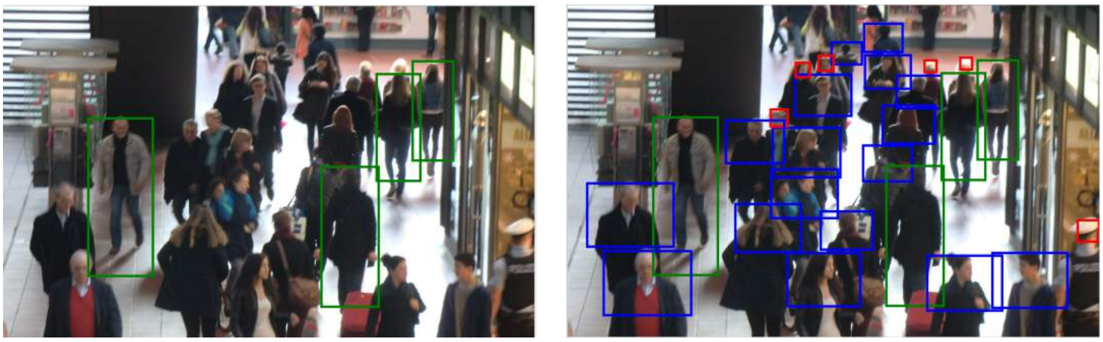


Figure 1: In a crowed scene only some humans are fully visible (left image), for many others (right image), only the upper-body is visible, or even just the head.

Marin-Jimenez et al. [Marin-Jimenez et al., 2014] proposed a two-level pipeline in which an upper-body detector is applied and then heads are detected within upper-body detection areas. For both the upper-body and the head detector, they trained a part-based model.

In this paper, a novel approach is proposed that combines different detectors for head detection. Three new detectors for head, upper-body, and body detection were trained based on the Aggregated Channel Features (ACF) detector framework of Dollár et al. [Dollár et al., 2014]. For each detector, the head is defined as a point of reference that allows a naive and obvious geometric approach to combining the detectors, which only takes geometric properties into account. The main principle of combination consists of estimating the head location of each part detector and then group detections by partition into disjoint subsets. For confidence score combination, the maximum posterior probability over all parts is computed. In order to validate the findings, the combined detector is tested on the town centre dataset, and results show an 18% reduction in the log-average miss rate of our combined classifier and illustrate that combining head detectors may perform better than a single detector.

2 Combined Head Detection

The proposed approach consists mainly of three steps. The first step is to train each part detector separately. The second step is to apply the part detectors to a test image and to group the resulting detections. The final step is to compute the combined detection score. In the following section, each step is described in detail.

2.1 Detector

The proposed detector is based on three part detectors: a head, an upper-body and a body detector. A separate detector is trained for each part using the Aggregated Channel Features (ACF) detector framework of Dollár et al. [Dollár et al., 2014], which has shown higher accuracy on the related task of full-body pedestrian detection.

The ACF detection framework first smooths an input image I with a [1 2 1]/4 filter and then computes several channels $C = \Omega(I)$. Then the channels are divided into blocks, and pixels in each block are summed. Finally, the resulting channels are smoothed again with a [1 2 1]/4 filter. Features are single pixel lookups in the aggregated channels. Boosting is used to train and combine decision trees over these features (pixels) to distinguish objects from the background. In order to allow multi-scale detection, a feature pyramid is build. At each scale, a sliding-window approach is then used to detect objects.

In order to train the detector a novel dataset was created that contains 650 overhead person images (plus horizontal mirror images) from different indoor and outdoor locations. The people are usually standing, but appear in any orientation and against a wide range of backgrounds.

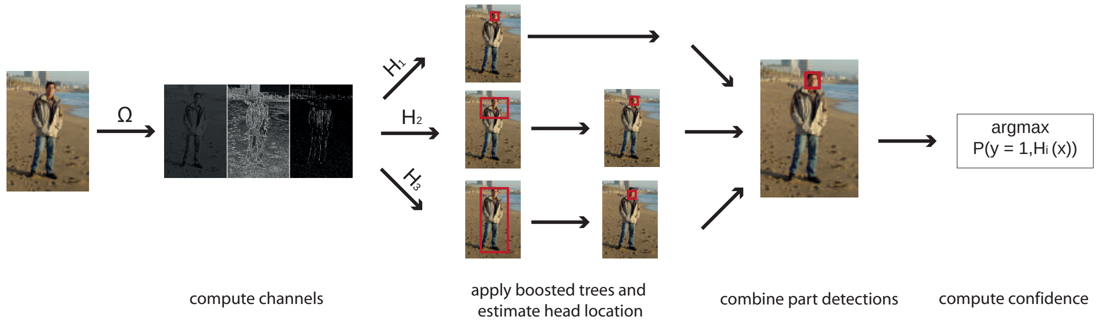


Figure 2: Overview of the combined head detector. Given an input image I, several channels are computed $C = \Omega(I)$. Boosted trees ($H_1 - H_3$) for head, upper-body and body are used to distinguish objects from the background. Head locations are computed for the upper-body and body detections using a fixed ratio and then all part detections are grouped to the resulting detections. Finally, a combined confidence score is computed.

The head bounding box (BB) were manual annotated, the upper-body and body BB were estimated by extending the head BB by fixed ratio. Head images have a size of 16×16 without and 32×32 with padding, upper-body images a size of 30×41 without and 60×64 with padding, and body images a size of 41×100 without and 64×128 with padding. As negative training set, background images from the INRIA dataset [Dalal and Triggs, 2005] were used. By estimating the upper-body and body locations based on the head location, the resulting detectors are aligned at the head location. That allows more accurate head location estimation based on this detector output than if the training set were centred at the body or upper-body location as is commonly done [Dalal and Triggs, 2005, Dollár et al., 2012].

For each detector, the same configuration was used. Ten feature channels were used: normalised gradient magnitude, histogram of oriented gradients (6 channels), and LUV colour channels; the block size was set to 4×4. AdaBoost was used to train and combine 2048 depth-two trees over the candidate features (channel pixel lookups) in each window. The step size of the detectors was set to 4 pixels and 8 scales per octave.

2.2 Combining Part Detectors

The final part detectors are then applied across a test image. The same feature pyramid is used for each detector, which speeds up the detection process (see Section 3). Then the head location is estimated for the upper-body and body detections, using the same ratio that was used to build the training set 2.1. Detections from multiple parts will usually occur around each head in the scanned image. In order to return one final detection it is useful to combine overlapping detections into a single detection.

Part-based models such as [Felzenszwalb et al., 2010], model the geometric relationship between parts explicitly. In the proposed case, this is not necessary since the part detectors are aligned at the head location, so the detection can be combined in a very simple way.

The set of all part detections is first partitioned into disjoint subsets. Two detections are in the same subset if their bounding regions overlap more than 0.5 %. Each partition yields a single final detection. The final bounding box is the bounding box of the most confident head detection (see Section 2.3), and if the partition does not include a head detection, then the bounding box of the most confident detection of the remaining parts is used. In order to be able to compute the combined confidence for each detection (see Section 2.3), the most confident detection of each part in a partition is saved. It is worth nothing that it is not required that all parts exist. If a partition does not include detections of all parts, the confidence of the missing parts is set to zero.

2.3 Confidence Combination

The confidence scores of a part detection can be obtained from the boosted classifier H, which consist of K weak classifier:

$$H(x) = H_K(x) = \sum_{j=1}^{K} \alpha_i h_i(x) \tag{1}$$

where each h_j is a weak classifier (with output -1 or 1) and α_i is its associated weight; x is classified as positive if $H(x) > 0$ and $H(x)$ serves as a score. When a person is not occluded, our experiments have shown that a body detector is more reliable than a head detector. However, if a person is partly occluded, a head detector is significantly more reliable than a full body detector. In order to address this problem, occlusion information is inferred from the scores of the part detections by selecting the part, which maximises the detection score:

$$score(x) = \arg\max_{1 \leq i \leq 3} P(y = 1, H_i(x)) \tag{2}$$

where $P(y = 1, H_i(x))$ is the posterior probability of the i-part being a true positive and y is the class label $y = \{1, -1\}$. In this work the posterior is defined as a sigmoid function of the score $H_i(x)$:

$$P(y = 1, H_i(x)) = \frac{1}{1 + exp(A_i H_i(x) + B_i)} \tag{3}$$

The sigmoid model is equivalent to assuming that the detection score is proportional to the log odds of a positive example. The parameters A and B are learned for each part separately on the training set (see Section 2.1) by the sigmoid fitting approach proposed in [Platt, 1999].

3 Experimental Setup and Evaluation

In order to evaluate the ability of the detector to distinguish between heads and all other objects, experiments were done on the town centre dataset [Benfold and Reid, 2011], which is a high definition video (1920x1080/25fps) of a shopping street that has a ground truth consisting of 71500 hand labelled head and body locations. Following the methodology of [Dollár et al., 2012], the performance is summarised using the log-average miss rate (MR), computed by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range 10^{-2} to 10^{0}. The log-average miss rate is similar to the performance at 10^{1} FPPI but in general gives a more stable and informative assessment of performance [Dollár et al., 2012]. A detected bounding box and a ground truth bounding box form a potential match if they overlap sufficiently. Because head regions are considerably smaller than full body regions, any error in the location has a much more significant impact on the performance measures, which is why the measure of Benfold and Reid [Benfold and Reid, 2011] is employed for heads, which states that their area of overlap must exceed 25%, not 50%, as used in [Dollár et al., 2012] for full body. Tests are performed on all three detectors separately and on the combined detector; in addition, the two detectors provided by Dollár et al. [Dollár et al., 2014] that were trained on the INRIA [Dalal and Triggs, 2005] and the Caltech [Dollár et al., 2012] dataset are tested.

Results are reported for head and body regions in Table 1, some examples of detection results are shown in Figure 4. In addition, in Figure 3 the detectors are compared by plotting the MR against FFPI (using log-log plots) by varying the threshold of detection confidence (details can be found in [Dollár et al., 2012]). The body region for the combined, the head and the upper-body detector as well as the head region for the upper-body, the body, the AcfInria and the AcfClatech detector are estimated using the same ratios used in Section 2.1 to create the training set.

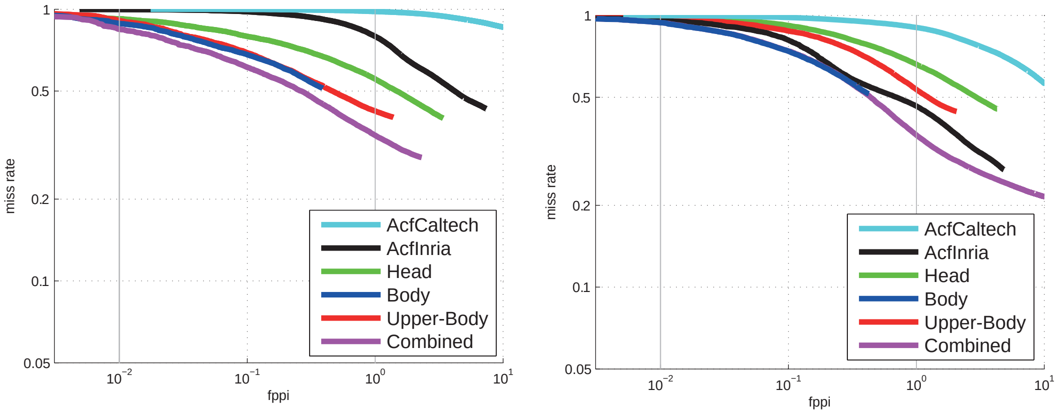


Figure 3: Log-log plots miss rate against false positives per image. Left: Head region. Right: Body region.

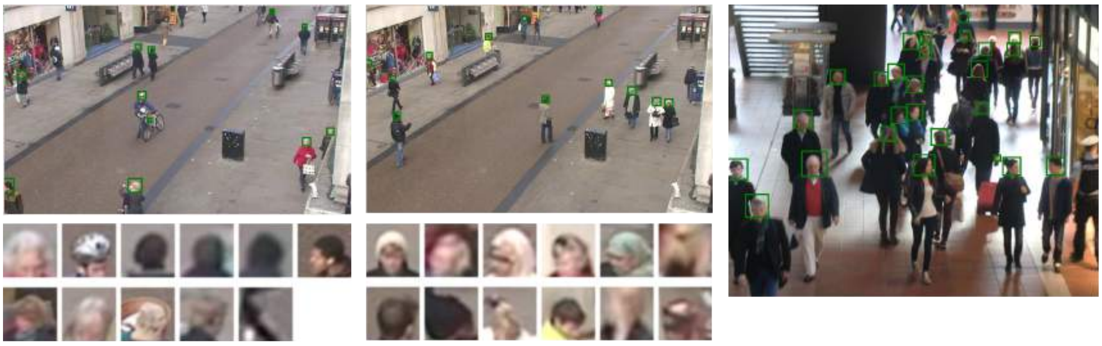


Figure 4: Some examples of detections on test images (1 + 2 town centre, 3 test image from the training set) for the final person detector.

The results show (see Table 1) that the combined head detector outperforms all other detectors and reduced the MR for head detections from 76% (Head-Detector) to 58% (Combined-Detector). Even in the case of the full body region, the combined detector achieves the best result, particularly remarkable, because using the head is often not discriminative in various tasks. In case of the body location, the proposed body detector and the AcfInria detector achieve similar results, but in case of the head location the proposed body detector archives a 28 % better result. This a result of defining the head as point of reference for the trainings image instead of the full body.

Speed comparison The combined detector needs 360ms to process a 1920x1080 image on the test machine, a desktop computer with an Intel Core i5-3470 CPU with 3.2 GHz and 8GB RAM, a single detector needs 260ms. That the combined detector is only 38% slower is due to the fact that the most time-consuming process, the features computation, only has to be done once.

4 Conclusion

In this paper, a method was developed to combine different detectors for head detection. Three separate detectors for head, upper-body, and body detection were trained based on the ACF detector framework of Dollár et al. [Dollár et al., 2014]. An algorithm was proposed to combine part detections that first estimates the head location of each part detector and then groups detections by partitioning them into disjoint subsets. The final confidence score is then calculated by

Method	MR - Head	MR - Body
AcfClatech-Detector [Dollár et al., 2014]	99	96
AcfInria-Detector [Dollár et al., 2014]	95	72
Head-Detector	76	87
Body-Detector	67	70
Upper-Body-Detector	66	81
Combined-Detector	58	66

Table 1: Performance on the town centre dataset

maximising the detection score over all parts. In order to validate the findings, the performance of the detection systems was examined on the town centre dataset. The results showed that combing a head, an upper-body and a body detector gives very good result for head detection, by reducing the MR by 18%.

References

[Benfold and Reid, 2011] Benfold, B. and Reid, I. (2011). Stable Multi-Target Tracking in Real-Time Surveillance Video. *CVPR*, pages 3457–3464.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. CVPR, pages 886–893.

[Dollár et al., 2014] Dollár, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast Feature Pyramids for Object Detection. *PAMI*.

[Dollár et al., 2012] Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian Detection: An Evaluation of the State of the Art. *PAMI*, 34.

[Felzenszwalb et al., 2010] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

[Marin-Jimenez et al., 2014] Marin-Jimenez, M. J., Zisserman, A., Eichner, M., and Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296.

[Merad et al., 2010] Merad, D., Aziz, K.-E., and Thome, N. (2010). Fast people counting using head detection from skeleton graph. AVSS '10, pages 233–240, Washington, DC, USA. IEEE.

[Platt, 1999] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

[Rodriguez et al., 2011] Rodriguez, M., Sivic, J., Laptev, I., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. *ICCV*.

[Venkatesh et al., 2012] Venkatesh, B. S., Descamps, A., and Carincotte, C. (2012). Counting people in the crowd using a generic head detector. In *AVSS*, pages 470–475. IEEE.

[Xie et al., 2012] Xie, D., Dang, L., and Tong, R. (2012). Video based head detection and tracking surveillance system. In *FSKD*, pages 2832–2836. IEEE.

[Zhang et al., 2009] Zhang, Z., Gunes, H., and Piccardi, M. (2009). Head detection for video surveillance based on categorical hair and skin colour models. In *ICIP*, pages 1137–1140.

Multi-view Face Detection using Flipping Scheme

Mengbo You and Takuya Akashi
Graduate School of Engineering
Iwate University
4-3-5, Ueda, Morioka, Iwate, 020-8551 Japan
you@scv.cis.iwate-u.ac.jp, akashi@iwate-u.ac.jp

Abstract

This paper extends the generic frontal face detection framework using SURF cascade classifier to handle profile views and rotated faces by the addition of minimized consuming time with only the frontal face detector. We proposed a novel flipping scheme for multi-view face detection making use of frontal face detector's over-representation, rather than building different detectors for different views. Our flipping scheme is actually a sliding window scheme for searching the hypothetic axis of symmetry and classification of left profile, right profile, or rotated profile. Our experimental results proved SURF cascade based flipping scheme was able to detect faces of almost all variations over the pitch and yaw angles while processing online.

Keywords: multi-view face detection, flipping scheme, profile face detector, frontal face detector

1 Introduction

We are interested in automatic detection of human face captured by a single camera from different view angles. Theoretically, an ideal multi-view face detector under real-world scenarios has to be capable of handling any possible human head rotations with the least time cost. Nowadays, frontal faces can be detected accurately [Osuna et al., 1997], [Rowley et al., 1998]. While profile face detection methods are always not reliable which inspires some methods [Li et al., 2002], [Schneiderman and Takeo, 2000] to specifically address the profile problem. To achieve fast processing of frontal face detector, [Viola and Jones, 2001] proposed a boosted cascade classification framework using Haar-like features, which is also called Viola-Jones (VJ) framework. SURF cascade framework which was proposed in 2011 [Li et al., 2011] and developed in 2013 [Li and Zhang, 2013] is derived from the VJ framework and adopt not Haar-like features but multi-dimensional SURF features [Herbert et al., 2008] to describe local patches.

Both SURF cascade and VJ frameworks outperform other methods particularly on the processing time. SURF cascade based frontal face detector is faster than the face detector which applies VJ framework [Li and Zhang, 2013]. To extend the frontal face detector to a multi-view one while preserving the quality of efficiency, the common idea is to assemble many detectors of different view angles each of which is trained individually to address a small specific view angle range [Jones and Viola, 2003]. We proposed a novel method to make a multi-view detector by making use of only one trained detector, which is the frontal face detector. In other words, our method can combine with either SURF cascade or VJ framework for multi-view detection. The detector trained with SURF cascade or VJ framework will be denoted by the cascade detector below, because they both adopt the cascade classification framework.

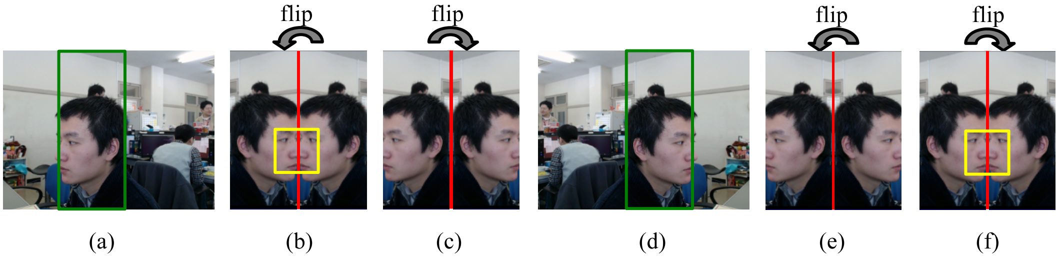


Figure 1: Two cases of flipping profile faces to make new target images; (a), (d) ROI contains right profile face or left profile face; (b), (e) flipping ROI to left; (c), (f) flipping ROI to right.

2 Flipping Scheme

The cascade detector is able to correctly locate frontal faces of the image with an acceptable true-positive rate. However, when we input an imitated frontal face which is made by flipping a profile face horizontally, the detector can still recognize it as a frontal face, see Figure 1 (b) and (f). The interesting part is that the imitation is composed of a pair of mirrored profile views. Tests of flipping different views are made and the results show that it is possible to detect profile faces using a frontal face detector. We made an achievement to extend the frontal face detector trained with SURF cascade framework and another trained with VJ framework to multi-view detectors respectively. We call this extension "Flipping Scheme".

2.1 Profile Face Detector

To identify profile face, the ROI (region of interest) is selected and flipped to the left or the right to get the left flipping ROI or the right flipping ROI, see Figure 1. The original ROI and either flipping ROI compose a new target image. Once a frontal face is detected in the new target image correctly, there is a half-face in the original ROI. In this way, the flipped profile face sharing the same symmetry axis with the new target image can be detected. The coordinate results in the new target image can be restored to be corresponding ones in the original image. The next section will describe how to detect a profile face in the image.

2.2 Sliding ROI

The detection is started by sliding the ROI from left to right. For each ROI, the common cascade detector is used to judge whether there are frontal faces. The coordinates will be recorded if the detector reports a "positive" which means there are faces found. Section 3.1 discusses about best face selection among candidates. Otherwise, the profile face detector which is described above will be activated. As for the result, the positive on the left flipping ROI indicates the left profile face is detected. The left or right are based on the observer's view. Similarly, the positive on the right flipping ROI indicates the right profile face is detected. If there are still no faces found after checking flipping ROIs, the searching comes to an end with the judgment of no faces for this ROI and slides to another ROI until all the possible regions are checked.

The frontal face detector is made replaceable. Hence its algorithmic complexity (denoted by D) will not be affected by the flipping scheme. D is determined by classification tree generated by the cascade classifiers which cannot be altered once trained out. The total complexity equals D if frontal faces are firstly detected. If there are profile faces found, the complexity will be $O(n){\cdot}D$. The worst case is that neither frontal faces nor profile faces are found, the complexity will be $n{\cdot}D$. The complexity D cannot be reduced other than turning to another faster frontal face detectors. We use some techniques to speed up computation by minimizing n like choosing the best width and height for ROI.

2.3 Specify ROI's Width and Height

The ROI will be used to make the flipping ROIs which may contains a frontal face. ROI's width determines the maximum size of the face the detector can handle. The best way is choosing image width as ROI width initially and reducing automatically to fit the distance from the sliding vertical symmetry to the final possible position. The size of profile faces which could be able to be detected ranges from the minimum size defined by the cascade classifier to the maximum which may cover the whole image. However, ROI's width should be flexible in case that the detector is targeted at certain face size range and a higher speed.

There is another reason that we choose the ROI which is larger than the true-positive face region. In order to use sliding ROI method to find the best symmetry for frontal face construction, the minimum ROI size will make the fastest searching, considering the cascade detector is much faster in the small image than in the big one. However, the cascade classifier will not work if input only the positive face region without surroundings because that will block the rectangle grouping. Hence, a width larger than the target face width range is necessary.

The common detection method loops ROI's height from the minimum to the maximum. For the flipping scheme, there will be many overlapped ROIs sharing the same symmetry axis after flipping. Therefore building ROI with the the image's height once would be more efficient.

3 Algorithm

Profile face detector is developed to address head rotation along the yaw direction. Instead of estimate head pose accurately, we simply differentiate profile problem into left profile and right profile. There are two versions of detection depends on how to specify ROI's width. The first version (Algorithm 1) is capable to detect profile faces while the face size varies from the minimum (same with the normalized minimum feature size which is 32 $\times$ 32) to the size of the whole image, but costs more computation time.

Profile Face Detection Algorithm 1

- Initialization: Initialize ROI R with (x=0, y=0, $width$=w, $height$=h) in image I whose width and height is denoted as w_I and h_I. R's width and height is set same as I's initially: w=w_I, h=h_I.
- Loop: while x<w_I
  - Construct new target image with the left flipping ROI $f_l(R)$:
    I_l=[$f_l(R)$, R].
  - Construct new target image with the right flipping ROI of region outside ROI $f_r(I-R)$ if large enough:
    if $I-R>minSize$, I_r=[$I-R$, $f_r(I-R)$].
  - Perform frontal face detection on both I_l and I_r, and record detected face region with best confidence.
  - Move ROI horizontally to new positions with user-defined shift $step$ and update ROI's width to fit the image:
    x=x+$step$, w=$w-x$.
- Output: face region with best confidence if detected.

The other version (Algorithm 2) provides an option for user to define the maximum half face width w' which could lesson ranges of searching and process faster. w' is used as a fixed sliding ROI's width and moves to a new position without changing on width. A fixed sliding ROI's width would miss some cases: left profile faces when ROI locates at the initial position and right profile faces when ROI locates at the final position which need additional process. Finally, to perform multi-view detection, profile face detector is activated when frontal face detector cannot find any faces.

Profile Face Detection Algorithm 2

- Initialization: Initialize R with (x=0, y=0, $width$=w, $height$=h) in I. R's width and height is set as follows: w=w', h=h_I. w' is the user-defined maximum half face width.
- Perform algorithm 1 described above within initial position of R when treating R region as a new target image
- Loop: while x<w_I−w'
    - Construct new target image with the left flipping ROI $f_l(R)$:
      I_l=[$f_l(R)$, R], I_r=[R, $f_r(R)$]
    - Perform frontal face detection on I_l and I_r, and record detected face region with best confidence
    - Move ROI horizontally to new positions with user-defined shift $step$:
      x=x+$step$
- Perform algorithm 1 within the last possible position while treating R region as a new target image
- Output: face region with best confidence if detected

3.1 Best Face Selection

There is an evaluation needed when the cascade detector output several positive candidates. The usual method is to count the number of overlapped rectangles to evaluate one candidate's confidence. And to judge whether two rectangles is overlapped, we apply the well-known over 50% ratio rule which label two rectangles as overlapped when the overlapped area account for more than half of their area's sum. By counting overlapped rectangles the system retains the best candidate every loop which results in single-target detection. Experimental results prove the detection is fairly reliable in Section 4.

3.2 In-plane Rotation

The sliding window method can also handle in-plane rotation caused by head pose changes in the tilt direction of 3D space. Suppose the point (x, y) in image $I(w, h)$ is rotated around image center to be (x', y') in image $I'(w', h')$ with an angle α. The range shift of image center can be estimated by $(0.5 \times (w' - w), 0.5 \times (h' - h))$. The correspondence can be calculated using trigonometric function:

$$x' = (x - 0.5 \cdot w)\text{cos}\alpha - (y - 0.5 \cdot h)\text{sin}\alpha + 0.5 \cdot w' \tag{1}$$

$$y' = (x - 0.5 \cdot w)\text{sin}\alpha + (y - 0.5 \cdot h)\text{cos}\alpha + 0.5 \cdot h' \tag{2}$$

The formulas can be used to get the rotated image I'. But detection result on I' need rotating back to get coordinates on I which follows an inverted rotation whose rotation angle is $-\alpha$. To speed up the rotation module, we are supposed to put computation out of loop as much as possible, such as store sine function values ahead of time and plus coordinates' increment instead of computing every time. Because SURF cascade has been trained with rotating positive images ±10°, we can set the step value as 20°.

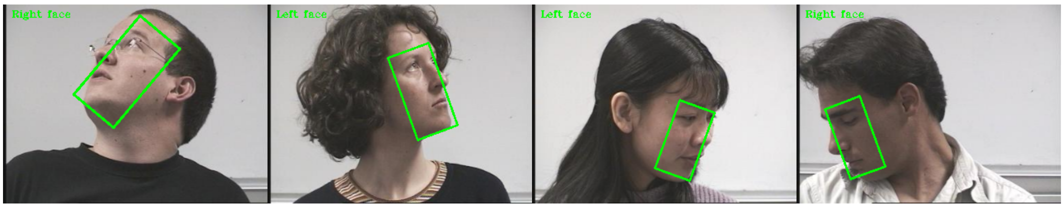


Figure 2: Examples of successfully detected profile faces on images of Pointing'04.

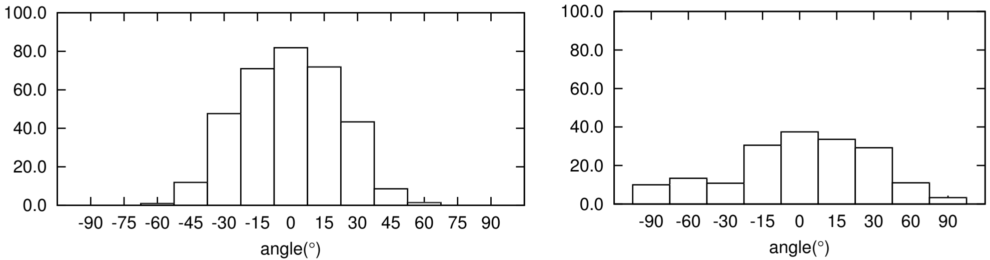
4 Experimental results

An automatic multi-view detection system is built in which the cascade detector is used to detect frontal faces, the flipping scheme is added to detect profile face rotating along yaw and pitch direction, and in-plane rotation module is made to handle profile face rotating along roll direction. Because SURF cascade is faster than the OpenCV face detector on frontal face detection, we choose SURF cascade to compare with SURF cascade based flipping scheme for evaluation of flipping scheme. We define rectangles containing three points of the left eye center, the right eye center and the mouth center as positive frontal face detection results and rectangles containing one eye's center, a part of nose and a part of mouth as positive profile face detection results.

Figure 3 displays distribution over pitch and yaw angles for the faces detected by SURF cascade on database Pointing'04 [Gourier and Crowley, 2004]. Figure 4 shows results using the SURF cascade based flipping scheme. The low true-positive rates of Figure 3 does not mean SURF cascade is bad because this is experimented with only the frontal face detector and the database contains various head poses. For example, when pitch angle equals to $+0°$, yaw angle varies from $-90°$ to $+90°$ with shift step of $15°$ and most images in this range are not frontal face image. Conversely, almost no false-positives shows SURF cascade detector's excellent performance. In Figure 4, results over yaw angle indicates that for all possible values of pitch angle and yaw angles on the range of $[-45°, +45°]$, true-positive rates are all above 80%. By comparison, we proves that the flipping scheme is able to handle almost all variations over the pitch and yaw angles with an average processing speed of 41 fps on the images with normalized size of 320×240 with optimized parameters (3.4GHz Core-i5 CPU and 8GB RAM).

5 Conclusion

A novel half face scheme is proposed to extend cascade classifier's frontal face detection to multi-view detection. Experimental results proves that SURF cascade outperforms the OpenCV default face detector in efficiency. After integrating SURF cascade frontal face detector into flipping scheme, almost all variations over the pitch and yaw angles can be handled perfectly for the single-target detection problem. Although the half face scheme brings with



(a) (b)

Figure 3: Distribution over the pitch and yaw angles for the heads detected by SURF cascade only. White shows true-positive rate (%), and gray shows false-positive rate; (a) yaw angle; (b) pitch angle.

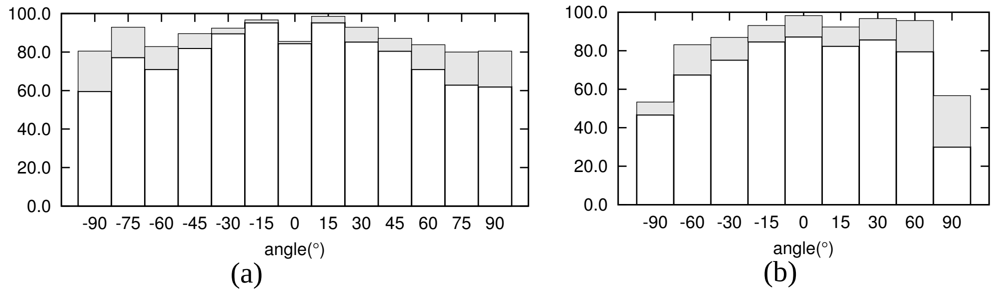


Figure 4: Distribution over the pitch and yaw angles for the heads detected by SURF cascade based flipping scheme. White shows true-positive rate (%), and gray shows false-positive rate; (a) yaw angle; (b) pitch angle.

it more possibility for the flipping ROI to contain false-positives, our experiments shows the single target detection is fairly reliable which would become weak for multi-target detection. We will take measures to address this problem for multi-target detection in the future.

References

[Gourier and Crowley, 2004] Gourier, N. and Crowley, J. L. (2004). Estimating Face orientation from Robust Detection of Salient Facial Structures. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK.

[Herbert et al., 2008] Herbert, B., Tuytelaars, T., and Gool, L. V. (2008). SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359.

[Jones and Viola, 2003] Jones, M. and Viola, P. (2003). Fast multi-view face detection. Technical report.

[Li et al., 2011] Li, J., Wang, T., and Zhang, Y. (2011). Face detection using SURF cascade. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2183–2190.

[Li and Zhang, 2013] Li, J. and Zhang, Y. (2013). Learning SURF Cascade for Fast and Accurate Object Detection. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3468–3475.

[Li et al., 2002] Li, S. Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H., and Shum, H. (2002). Statistical Learning of Multi-view Face Detection. *ECCV 2002 Lecture Notes in Computer Science*, 2353:67–81.

[Osuna et al., 1997] Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Rowley et al., 1998] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE transactions on pattern analysis and machine intelligence*, 20:22–38.

[Schneiderman and Takeo, 2000] Schneiderman, H. and Takeo, K. (2000). A statistical method for 3D object detection applied to faces and cars. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 746–751. IEEE Comput. Soc.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1:I–511–I–518.

IMVIP 2014

POSTERS

Human trajectory tracking using a single omnidirectional camera

Atsushi Kawasaki, Dao Huu Hung and Hideo Saito
Graduate School of Science and Technology
Keio University
3-14-1, Hiyoshi, Kohoku-Ku, Yokohama, Kanagawa, Japan
{kawasaki, hungdaohuu, saito}@hvrl.ics.keio.ac.jp

Abstract

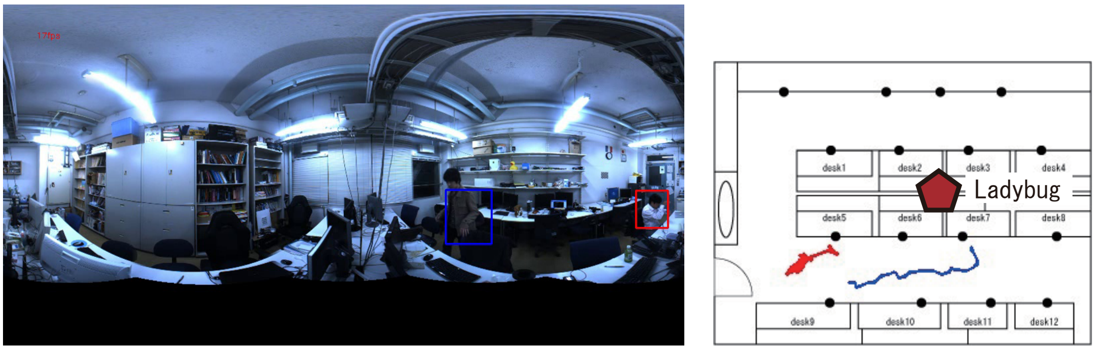
We propose a method to detect and track humans from an omnidirectional camera image, and to visualize human trajectories in the room plan. There are two problems to solve to achieve this issue. One is a robust algorithm of human detection in the case that humans are moving or sedentary in complicated background such as a office room. Dynamic background subtraction is suitable for detection in the complicated background but is not working to detect of a object having a little movement. In this paper, we propose a detection method based on the combination of static background subtraction, dynamic background subtraction, and Histogram of Oriented Gradients(HOG). The second problem is the way to visualize human trajectories in the room plan from distorted panoramic image. Therefore, we propose a method to create a correspondence relation by sandwiching a perspective image between a room plan and a panoramic image. The panoramic image is divided into multiple areas, and each area is converted into a perspective image. It is feasible to calculate the coordinate of human position in the room plan by using a Homography matrix between the perspective image and the room plan. We conducted the accuracy evaluation of human detection and human trajectory in order to ensure the effectiveness of this method. As a result, the proposed method of human detection reduced false positive detection remarkably in comparison with existing method. The experimental results of visualizing human trajectories demonstrated the range of errors of position estimation is about from 16 cm to 55 cm, but it is sufficient for use in data analysis such as head-count and residence time.

Keywords: Human detection, Human tracking, Omnidirectional camera

1 Introduction

A number of surveillance have widely been studied. Among them, tracking of human trajectory is extensively studied for use in the analysis of surveillance videos. The problem of single camera monitoring is limited visual field, while the problem of multi-camera monitoring is geometrical calibration of multiple cameras. An omnidirctional camera can cover 360° direction and enables to monitor broad range without overlapping. However, it is not easy to accurately understand the human position in the distorted panoramic image. This paper proposes a method which performs human detection and tracking from the omnidirectional camera image, and visualizes human trajectories in the room plan.

There are a lot of researches related to a surveillance camera. For example, Oktavianto et al.[1] proposed attendance logging system. Okabe et al.[2] tracked human trajectory using a stereo camera. The key issue in these research is human detection and tracking. Oktavianto et al.[1] proposed a method for human detection by background subtraction, but overlapped humans can not robustly be detected by just using background subtraction. Human detection using Histogram of Oriented Gradients(HOG)[3] is widely used. However, it is difficult to use



(a) (b)

Figure 1: The result of system.(a) shows the result of human detection. (b) shows the trajectories of detected humans.

HOG features in complex backgrounds because a complex background have much gradient information. Vondrick et al.[4] presented the results HOG features cause false positives in some cases. There are many research to approach this problem. Jin et al.[5] proposed a method of tracking in complex background by particle filter and HOG. Bing-bing et al.[6] proposed a detection method that combines HOG and dynamic background subtraction. This approach detects humans with HOG in the target area extracted by background subtraction. However, in this approach humans need to be moving because sedentary humans are included in the background image by using dynamic background subtraction.

Some researches on omnidirectional camera have been activity conducted in the past few years. Peri et al.[7] proposed a method to generate perspective and panoramic image from omnidirectional image obtained from a parabolic mirror omnidirectional camera. We used an omnidirectional multi-camera system, Ladybug, Point Grey.

In this paper, we propose a method to detect humans robustly in complex background based on the combination of static background subtraction, dynamic background subtraction and HOG. Furthermore, we propose a method to visualize human trajectories in room plan from the distorted panoramic image. Figure.1 shows an example of results of this system.

2 Overview of the system

This section provides an overview of the major steps in our system. Implementation details are presented in Section 3 through Section 5. The system is composed of the combination of on-line processing and off-line processing . In the off-line processing, We manually select corresponding points between the room plan and panoramic image. After dividing the panoramic image based on the corresponding points, each area is converted into a perspective image. The Homography matrix is then calculated between the perspective image and the room plan.

In the on-line processing, the system obtains a panoramic image as an input from the omnidirectional camera. The target areas are then extracted from a input image by background subtraction. In these areas, human is detected using HOG, and the background image is updated based on the detection result. After human detection, human tracking is performed by data association. The trajectory of each detected human is visualized in the room plan by using Homography between the perspective image and the room plan.

3 Human detection based on background subtraction and HOG

We propose the method to detect moving or sedentary humans in the complicated background. This method is based on combination of static background subtraction, dynamic background subtraction and HOG. Basically, by using Real AdaBoost classifier with HOG, upper bodies are detected in the target areas extracted by background subtraction because upper bodies are certainly visible even when lower body is hidden by other object such as desks. And then, the background image is updated based on the detection result and accumulation of background statistics[8]. In this approach, I is pixel intensity of the background area and is modeled as follow:

$$I = \overline{I} + \sigma \sin(2\pi\omega t) + k\zeta \tag{1}$$

where $\overline{I}$ is the time average of pixel intensity, σ is the amplitude of intensity, ω is the frequency of intensity, t is time, $k\,(-1 \leq k \leq 1)$ is coefficient, and ζ is the maximum value of the noise depending only on the camera.

Based on the detection result and accumulation of background statistics, one of the following three processes is performed in each pixel. One is in the case that I is not included in detected rectangles and $\overline{I} - \sigma - \zeta \leq I \leq \overline{I} + \sigma + \zeta$ is satisfied. In this case, the pixel is identified to exist in the background area and $\overline{I}$ and σ is updated based on follows:

$$\overline{I}' = \frac{(n-1)\,\overline{I} + I}{n} \tag{2}$$

$$\sigma' = \frac{(n-1)\,\sigma + \sqrt{2\left(I - \overline{I}\right)^2}}{n} \tag{3}$$

where n is the parameter of the update speed.

Another is in the case I is not included in the rectangles and $\overline{I} - \sigma - \zeta \leq I \leq \overline{I} + \sigma + \zeta$ is not satisfied. In this case, the pixel is identified to exist in the area of a moving object except humans, and $\overline{I}$ is not updated but σ is updated as:

$$\sigma' = \frac{(m-1)\,\sigma + \sqrt{2\left(I - \overline{I}\right)^2}}{m} \tag{4}$$

where $m\,(m \geq n)$ is the parameter of the update speed in object areas. By doing this, moving object other than human will be included in the background area gradually. Updating the background image over time prevents undetected error of the detector.

The other is in the case I is included in the rectangles. In this case, the pixel is identified to exist in human areas, and $\overline{I}$ and σ in this areas is not updated. By keeping these values, detected areas are not included in the background image and will always be the subtraction areas. From the empirical results n is $1/50.0$, m is $1/300.0$ and ζ is 10.0 in our implementation.

4 Human tracking by data association

In this system, human tracking is performed by data association that compares the Euclid distances and the distances of the color histograms between the detected rectangle in the current frame and the ones in the last frame. Data association is suitable for omnidirectional multi-camera system whose frame rate is low because data association is simple and does not take processing time. Firstly, the Euclid distance d is calculated. Bhattacharyya distance d_H is then calculated between histograms after making color histogram of each rectangle. Combined distance D is calculated in all combination of rectangles as follow:

$$D = d + \alpha d_H \tag{5}$$

where α is a parameter. When there is a one-to-one correspondence between the rectangle of current frame and the rectangle of the last frame, the combination of the minimal sum of D becomes the tracking result.

However, it is not sufficient to compare the histogram of only the previous frame to make an accurate tracking possible even if human is overlapping or occlusion occurs. In this paper we classify the K clusters from N histograms from N frame before to the current frame by using

K-means. Each cluster's center of gravity is adopted as representative histograms of each human. We use the smallest d_H of K representative histograms when calculating Bhattacharyya distances. From the empirical results N is 100 and K is 3 in our implementation.

5 Visualizing human trajectory

We consider distortion of the panoramic image when visualizing human trajectories. In this paper this problem is solved by converting perspective images from divided panoramic images.

In the off-line processing, we first select corresponding points between the room plan and the panoramic image. The area between a desk and the other desk is utilized because the floor is hidden by desks in the environment such as a office. (In the case that the aisle is broad like figure.4(b), we select corresponding points on the floor.) We create some perspective images which respectively contain an area constituted by every four input points. Assuming there is the sphere with a focus central on omnidirectional camera, Figure.2 shows a relationship between the panoramic image and the sphere. This relationship enables creating a perspective image by projecting the pixel intensity of the spherical surface on a plane including four input points. The Homography is calculated for each area utilizing the corresponding points between the room plan and the perspective image.

In the on-line processing, human trajectories are visualized using the Homography and the coordinate value of the detected rectangle. Assuming the lower side of the rectangle exists on a plane between a desk and the other desk, we can estimate the coordinate value of the human position in the room plan by multiplying the center of the lower side on the perspective image by the Homography. Figure.3 shows the flow of visualizing.

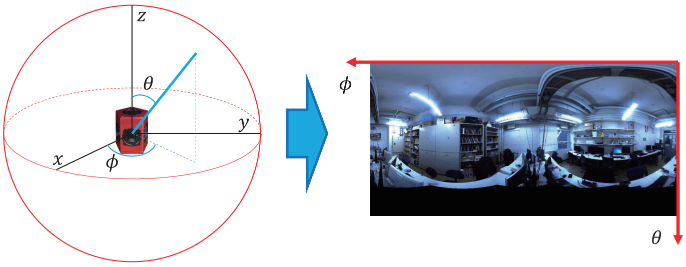


Figure 2: Assuming the sphere of the constant radius exists with a focus central on omnidirectional camera, pixel intensities are projected to the spherical surface. When the latitude is ϕ and the longitude is θ, we can make a panoramic image by using polar coordinate.

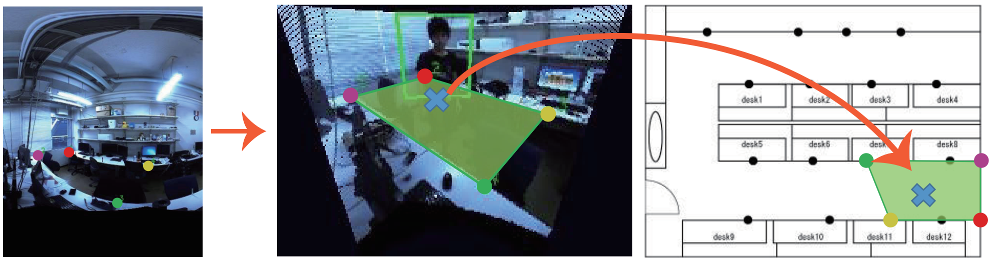
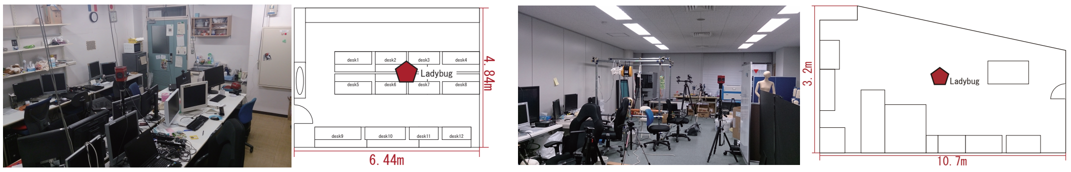


Figure 3: In off-line process, After inputing the corresponding points between the panoramic image(left image) and the room plan(right image), the perspective image(center image) is created. In on-line process, human position in the room plan is calculated by multiplying the center of the lower side of the rectangle on the perspective image by the Homography between the perspective image and the room plan.

6 Experiment

We conducted the experiments for evaluating the human detection and trajectory tracking. Figure.4 shows two environments of the experiments. Figure.4(a) is the environment where the aisle is narrow, and Figure.4(b) is one where the aisle is broad. Training images were cropped from panoramic image to minimize the loss of accuracy of detection by distortion.



(a) The room interior

(b) The room interior

Figure 4: The environment of the experiment

6.1 Evaluation of human detection

We compared our method with the combination method of static background subtraction and HOG. In this experiment, a video of 8800 frames was prepared. The ROC curve is given in Figure.5[a], in which the x-axis is False Positives Per Image(FPPI), and the y-axis is the detection rate. Figure.5[a] shows FPPI of our method is markedly lower than the comparative method. Since our method takes advantage of dynamic background subtraction, the false positive detection of our method is lower in subtraction area generated by movement of objects and illumination changes. However, the maximum value of the detection rate is also lower. The cause is possibly human areas included in the background image when human is not detected.

6.2 Evaluation of trajectory tracking

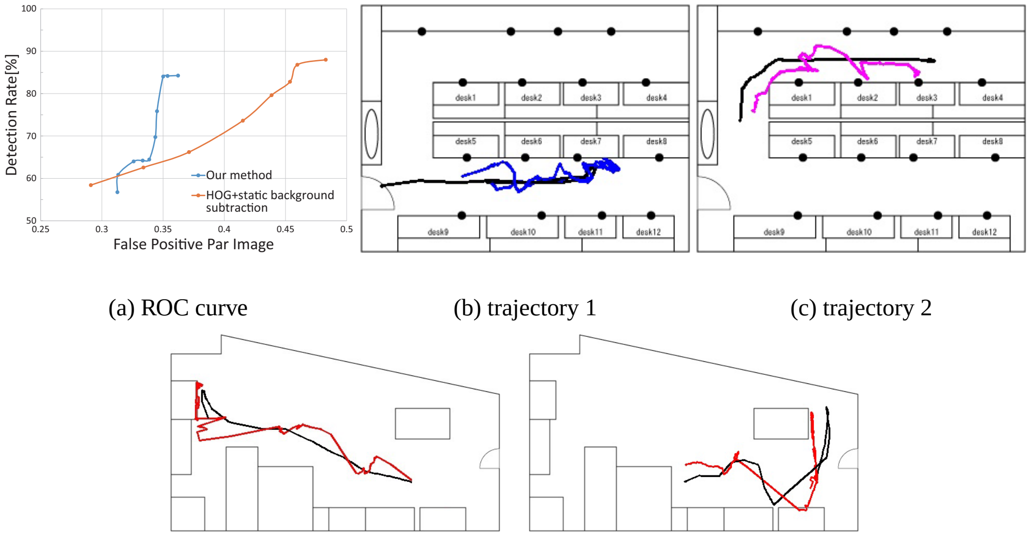
We measured the error of the Euclid distance between the visualizing result and the ground truth in all frame, and calculated the average, the standard deviation, and the maximum of the error. In this experiment, four videos of 200 frames were prepared. The ground truth is the coordinate on the room plan obtained by manually plotting human position on the room plan. The results of the trajectories are given in Figure.5 (b)-(e). The trajectories of black lines indicate the ground truth. Table.1 shows the evaluation index of the error of the trajectories. The range of distance error is about from 16 cm to 55 cm. The assumed cause of the distance error is as follows. Since the classifier can detect only upper bodies, the amount of information is not enough to decide human areas. When the rectangle can not surround a human accurately, the coordinate multiplied by the Homography is misaligned.

Table 1: The error of trajectory

	(b)	(c)	(d)	(e)
Average[cm]	16.50	36.75	50.91	55.31
Standard deviation[cm]	10.98	22.91	28.88	26.23
Maximum[cm]	47.41	110.29	141.61	135.79

7 Conclusion

In this paper, we presented the method to detect and track humans using the omnidirectional camera, and to visualize trajectories in the room plan. Our method which is based on combination of static background subtraction, dynamic background subtraction, and HOG, can detect moving or sedentary human in the complicated background. This method overcomes the issue that dynamic background subtraction is not working to detect a object having a little movement.



(a) ROC curve (b) trajectory 1 (c) trajectory 2

(d) trajectory 3 (e) trajectory 4

Figure 5: The result of the experiments

Human tracking is performed by data association which is suitable for this system because the frame rate of omnidirectional multi-camera system is low. On the other hand, in order to visualize trajectory from distorted panoramic image, we proposed the method that the panorama image is divided into multiple areas and each area is converted into a perspective image. As an experimental result, the proposed method of human detection reduced false positive detection remarkably. Furthermore, the result of visualizing human trajectories showed trajectory error of position estimation is about from 16 cm to 55 cm.

References

[1] H. Oktavianto, H. Gee-Sern, and C. Sheng-Luen. Image-based intelligent attendance logging system. *System Science and Engineering (ICSSE), 2012 International Conference on*, pages 1–6, 2012.

[2] A. Okabe, M. Ambai, and S. Ozawa. In-store human detection based on abs method for flow line analysis. *IEEJ Transactions on Electronics, Information and Systems*, 127:506–512, 2007.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 1:886–893, 2005.

[4] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hog-gles: Visualizing object detection features. 2013.

[5] L. Jin, J. Cheng, and H. Huang. Human tracking in the complicated background by particle filter using color-histogram and hog. pages 1–4, 2010.

[6] W. Bing-bing, C. Zhi-xin, W. Jia, and Z. Liquan. Pedestrian detection based on the combination of hog and background subtraction method. pages 527–531, 2011.

[7] V. Peri and Shree K N. Generation of perspective and panoramic video from omnidirectional video. 1:243–245, 1997.

[8] S. Morita, K. Yamazawa, and N. Yokoya. Networked video surveillance using multiple omnidirectional cameras. 3:1245–1250, 2003.

An Improved JPEG Image Compression Algorithm

Samruddhi Kahu
Department of Electronics and Communication
Visvesvaraya National Institute of Technology
Nagpur, Maharashtra, India
samruddhikahu@gmail.com

Reena Rahate
Department of Electronics and Communication
Shri Ramdeobaba College of Engineering and Management
Nagpur, Maharashtra, India
reenapalashn@rediffmail.com

K. M. Bhurchandi
Department of Electronics and Communication
Visvesvaraya National Institute of Technology
Nagpur, Maharashtra, India
bhurchandikm@ece.vnit.ac.in

Abstract

Image compression is still an open area of research due to high memory and bandwidth requirements for storage and transmission. A novel scheme to store the Huffman coded words in JPEG algorithm is proposed in this work to improve the Compression Ratio while maintaining the same perceptual image quality. The proposed work also introduces a technique to reduce the number of intensity levels using novel regrouping of the zigzag ordered 63 'AC' Coefficients. Image quality metrics such as Compression Ratio, SSIM, MSE and PSNR are used for quantitative benchmarking of the results with JPEG algorithm.

Keywords: JPEG, Image Compression, Quantization.

1 Introduction

In recent years, with the advent of 3G and 4G, the available bandwidth has increased many-folds. But due to HD High Definition (HD) technology, the quality of multimedia signals has also increased further increasing the memory and bandwidth requirements. This has again resulted into a need for higher compression ratios without compromising with perceptual quality. While compressing images, we are taking advantage of the redundant information present in them. Redundancies present in images may be;
i) Spatial (between neighboring pixels)
ii) Spectral (correlation between color components)
iii) Psychovisual (due to human visual system)
In JPEG, spatial and spectral redundancies are exploited using DCT which has high energy compaction in frequency domain. This technique is effective since the image is first divided into 8x8 blocks and DCT is applied on each 8x8 block separately. The mathematical expression for 2D-DCT for an M x N matrix, in general, is given an

$$F(u,v) = C(u)C(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\cos(\frac{(2x+1)u\pi}{2M})\cos(\frac{(2y+1)v\pi}{2N}) \quad (1)$$

$$\begin{aligned} \textit{where} \quad & C(u), C(v) = 1/\sqrt{2} && \textit{for} \quad u, v = 0; \\ & C(u), C(v) = 0 && \textit{otherwise}; \end{aligned}$$

After this step, each block is quantized using a specific quantization matrix (Q matrix). This step makes JPEG a lossy image compression. The above process may be represented as;

$$D(u,v) = round\left[\frac{F(u,v)}{Q(u,v)}\right] \quad (2)$$

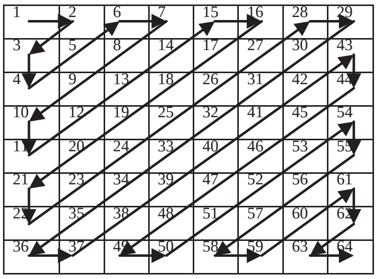
where F(u,v) is the DCT coefficient matrix, Q(u,v) is the standard JPEG quantization matrix. JPEG uses two standard quantization matrices, one for quantization of the luminance component of an image and one for chrominance component. The luminance and chrominance quantization matrices are shown in fig. 1 and fig. 2 respectively. Step sizes of the quantization matrix are defined based on luminance and chrominance perception of the human eye i.e. more importance is given to lower frequency coefficients than higher frequency coefficients. Quantization removes psycho visual redundancy in an image. The quantized matrix is then zigzag ordered using the scanning order shown in fig. 3.The linearly arranged coefficients are then run length coded and entropy coded using variable or fixed length codes.

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Fig. 1. Luminance Quantization Matrix

17	18	24	47	99	99	99	99
18	21	26	66	99	99	99	99
24	26	56	99	99	99	99	99
47	66	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99

Fig. 2. Chrominance Quantization Matrix



1	2	6	7	15	16	28	29
3	5	8	14	17	27	30	43
4	9	13	18	26	31	42	44
10	12	19	25	32	41	45	54
11	20	24	33	40	46	53	55
21	23	34	39	47	52	56	61
22	35	38	48	51	57	60	62
36	37	49	50	58	59	63	64

Fig. 3. Zigzag Scanning Order

Since JPEG is both an ISO standard and CCITT recommendation [7], it has been a universal image compression standard with its operability spanning a wide range of devices. The standard JPEG encoder and decoder are shown in [7]. JPEG is used for both color as well as gray scale image compression. For gray scale images, only one plane i.e. intensity (luminance) plane is coded. For color image coding, image is first converted to YCbCr color space where Y is the luminance component and Cb and Cr are the blue and red chrominance components respectively. As human eye is less sensitive to chrominance information than luminance, chrominance subsampling [9] is used without any perceptual loss in image quality. Later the luminance and the two subsampled chrominance planes are coded.

Since the invention and application of JPEG to image compression, work has been done on each block of JPEG to improve compression ratio and/or image quality. In [1], author has used triangular and trapezoidal blocks according to the shape of the objects in the image instead of rectangular 8x8 blocks. Instead of DCT, DHT (Discrete Hartley Transform) has been used in [2] and the quantization and scanning order has been changed accordingly. A few authors have optimized the Quantization table and Run-length coding [3] and [4] making the compromise between compression and image quality an optimum. Improving image compression through hiding binary information and the study of Human Visual System (HVS) is discussed in [5] and [6].

This paper deals with storing of the coded words and regrouping of AC jpeg coefficients. If the coded words are efficiently stored, higher compression ratio can be achieved with negligible or no loss of image quality. After zigzag ordering, we have prioritized different groups (bins) of 63 linearly arranged AC coefficients and analyzed its effects on the quality of the reconstructed image.

Organization of this paper is as follows. In section II, JPEG basics and JPEG algorithm is discussed in detail. Section III deals with the proposed enhancements in the JPEG algorithm followed by results in section IV and concluding remarks in section V.

2 Proposed Algorithm

2.1 Efficient Storage

Huffman encoder outputs variable length code words. If these code words are stored in individual memory locations (one code word in one memory location), a lot of memory space is wasted. For example, for a 3 bit code word 16 bits are used if each memory location is of size 16 bits (2 bytes). In such a case very less compression is achieved or in some cases compression even may not be achieved. For our experimental image, compression ratio of 5.48/1 is achieved. Now a lot of memory space is saved if coded bits are saved without any space. Suppose we have two code words (010 and 0100) which are to be stored. If each code word is stored in one memory location, 25 bits (≈ 3 bytes) of memory space is wasted. To avoid this, both the code words can be concatenated and stored as one (0100100) code word. But then the decoder will not be able to identify the actual code words transmitted. For example, the concatenated code word can be interpreted as either 0100 & 100 or 010 and 010 and 0 or in many more ways. Thus along with the code words, their lengths are also needed to be stored. Lengths can also be concatenated and stored but they are concatenated as decimal numbers as against the code words which are concatenated as binary numbers. Concatenation of lengths is possible as no code word has length less than 3 and more than 26. So, if a 1 or a 2 (decimal) is encountered in the length array, two consecutive digits are taken as length for example 15, 19, 25, etc. With this technique compression ratio is increased to 26.05/1 for our experimental image shown in fig. 6 (d) and without any further deterioration in the image quality. Results for this step are not quoted separately as they are same as that of the results of the complete modified algorithm with Quantization Factor of 1. Hence, fig. 7 (d) and fig. 8 (d) also show the results of this step for experimental images shown in fig. 7 (a) and fig. 8 (a) respectively.

2.2 AC Quantization after zigzag ordering

Compression ratio can be further increased if the linearly ordered 63 AC coefficients are further quantized using a certain quantization factor (QF) or a quantization ratio (QR). For this the 63 AC coefficients are divided into different groups as follows:

Groups	Coefficients
1	1,2,3
2	4,5,6,7
3	8,9,10,11
4	12,13,14,15
5	16,17,....,63

Table I. AC Coefficients' Groups

Since these coefficients are ordered from lower to higher frequencies, the groups contain increasingly higher frequency coefficients with group 1 containing the low frequency coefficients and group 5 all the high frequency coefficients [8]. In other words, these groups contain increasingly higher amounts of energies of the image sub-blocks. Thus, one possible way of quantization would be to divide each coefficient of a certain group by the group number itself. This means that each group is prioritized according to the amount of energy it contains. This results in a quantization ratio of 1:2:3:4:5. After experimentations it has been observed that groups 1, 2 and 5 contain most of the energy of the sub-block [8]. So, a quantization ratio of 1:2:3:4:1 can also be used. One possible reason for the high energy content of group 5 even though it contains all the high frequency (low energy) coefficients is that it contains a large number of AC coefficients as compared to other groups as is evident from Table I. A variety of other quantization ratios have been tried and the results are analyzed and tabulated in Table III.

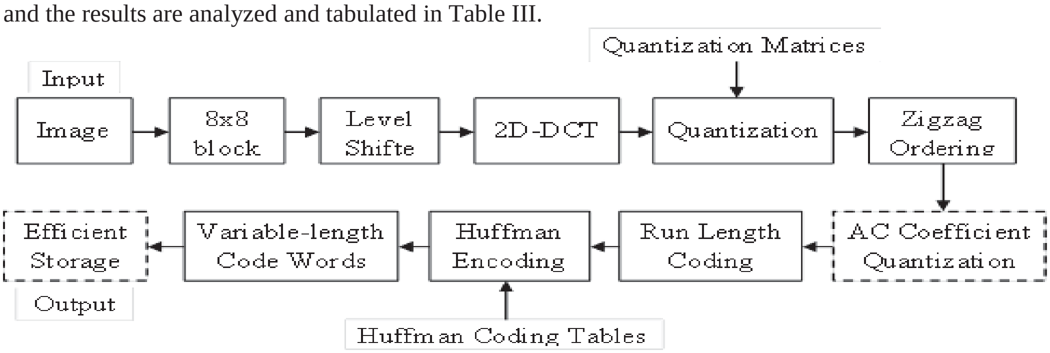


Fig. 4. Modified JPEG Encoder

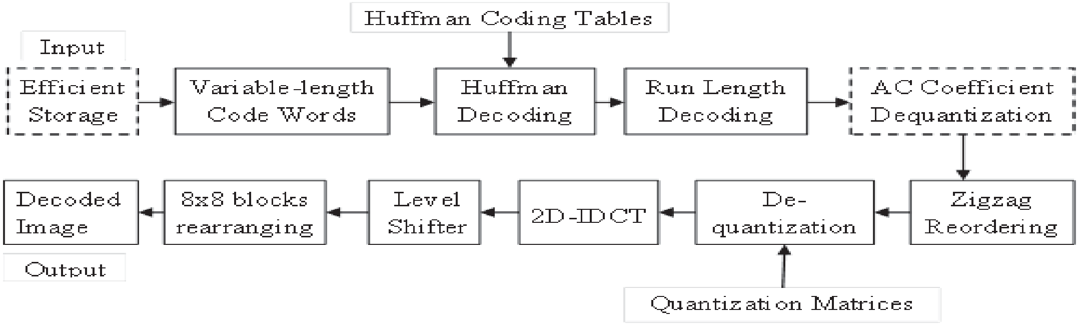


Fig. 5. Modified JPEG Decoder

Now, instead of quantizing each group using a different number, all the groups can be quantized using a same quality factor as it affects quality of the image. Thus, this quantization using a quantization ratio can be called as variable quantization and that using a quality factor, constant quantization. Constant QFs of 1 to 5 have been used to quantize the AC coefficients and the results are tabulated in Table II. The 63 AC coefficients can also be quantized using a combination of the above two methods such that all the AC coefficients are first quantized by a constant QF and then using a variable quantization ratio.

Thus, we are adding two additional blocks (dotted) to the standard JPEG encoder and decoder as shown in fig. 4 and fig. 5.

3 Metrics for Performance Evaluation

Following quality metrics are used in this paper to compare original and reconstructed images:

3.1 Compression Ratio

It is defined as the ratio of the number of bits required to store the original image file to that required by the compressed image file.

3.2 Mean Square Error (MSE)

It is a measure of distortion of the reconstructed image as compared to the original image. Mean Square Error is defined as the sum of squares of the difference between pixel values of original and reconstructed image averaged over the complete image.
Mathematically,

$$MSE = \sum_{x=1}^{M}\sum_{y=1}^{N}[(g(x,y) - g'(x,y))]^2 \ / \ (\mathrm{M^*\,N}) \qquad (3)$$

where, g(x,y) is the intensity at pixel location (x,y) of the original image and g’(x,y) that of the reconstructed image.

3.3 Peak Signal to Noise Ratio (PSNR)

It is defined as the ratio of peak signal power of an image to the noise power that corrupts it. In an image, peak signal power refers to the maximum intensity value that an image pixel can have. Thus for an n-bit image, peak signal power will be 2^n-1 and noise power for an image is represented by MSE. Hence, PSNR is given as:

$$PSNR = 10\log_{10}\left[\frac{(2^n - 1)^2}{MSE}\right] \qquad (4)$$

MSE and PSNR are the most popularly used performance metrics for comparing original and reconstructed images but they may not give faithful results in some cases as they do not take into consideration spatial and temporal correlation between pixels. Thus, we are using a performance metric called Structural Similarity Index Metric (SSIM) which is more consistent with human eye perception.

3.4 SSIM

Mathematically, SSIM is defined as in

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x{}^2 + \mu_y{}^2 + c_1)(\sigma_x{}^2 + \sigma_y{}^2 + c_2)} \qquad (5)$$

Where x and y are two windows of same size, μ_x and μ_y are the means of x and y respectively, σ_x and σ_y are standard deviations of x and y respectively, σ_{xy} is the covariance of x and y, $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ where L is the dynamic range which is 2^n-1 for an n-bit image, $k_1 = 0.01$ and $k_2 = 0.03$ by default.

4 Results and Discussion

Performance of the proposed algorithm was tested on several real life and test images. Out of these, results of implementation on one real life image, one nature image and one test image are shown in fig. 6, fig. 7 and fig. 8 respectively. Table III shows the variation of compression ratio, MSE, PSNR and SSIM for different Quantization Ratios (QRs) implemented on fig. 6 (a). Images compressed and reconstructed using some of these QRs are shown in fig. 6 (b), (c), fig. 7 (b), (c) and fig. 8 (b), (c). Highest compression ratio is obtained when QR 5:4:3:2:1 is used but MSE in this case is least i.e. 23.69. However, as seen prominently in fig. 9 (c) which is a zoomed in part of fig. 8 (c), we get blocking effect in the background. This proves that group 1 contributes for the background information (low frequency) in the image and since in this case it is getting quantized by the highest factor (i.e. 5), we see that a lot of blocks appear in the background. Blocking effect in the background is reduced slightly if quantization ratio of 1:4:3:2:1 is used. In this case while the compression ratio gets reduced to 33.69/1, MSE and SSIM are improved. For quantization ratios of 1:2:3:4:5 and 1:2:3:4:1, background is reconstructed faithfully but a lot of blocks appear on the edges (fig. 9 (b)). In spite of all these facts, as seen from fig. 8 (b) to (d), visually all the images are more or less the same.

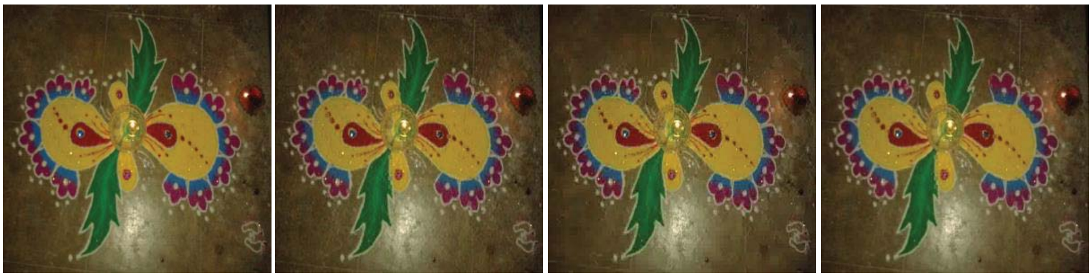


Fig. 6. (a) Original Image (Real life Image) Fig. 6. (b) Image coded using QR 1:2:3:4:5 (CR = 32.44). Fig. 6. (c) Image coded using QR 5:4:3:2:1 (CR = 41.89). Fig. 6. (d) Image coded using QF 1 (CR = 26.05).

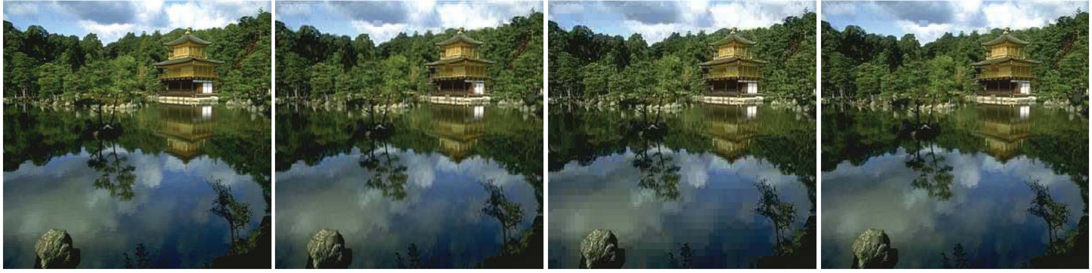


Fig. 7. (a) Original Image (Nature Image) Fig. 7. (b) Image coded using QR 1:2:3:4:5 (CR = 26.60). Fig. 7. (c) Image coded using QR 5:4:3:2:1 (CR = 24.25). Fig. 7. (d) Image coded using QF 1 (CR = 17.42).

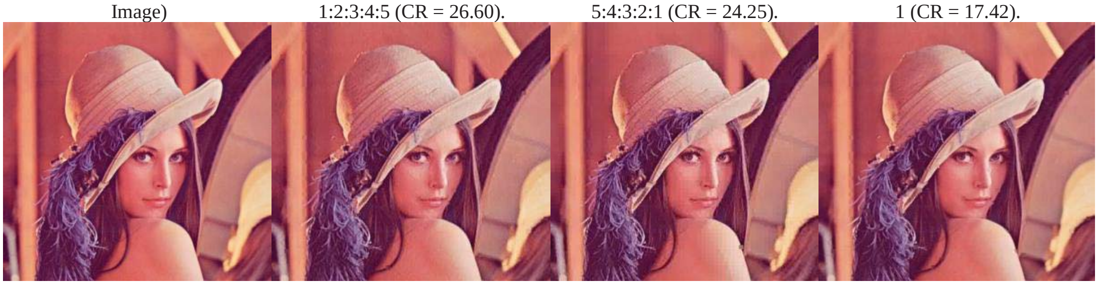


Fig. 8. (a) Original Image (Std. test Image) Fig. 8. (b) Image coded using QR 1:2:3:4:5 (CR = 31.13). Fig. 8. (c) Image coded using QR 5:4:3:2:1 (CR = 36.46). Fig. 8. (d) Image coded using QF 1 (CR = 24.56).

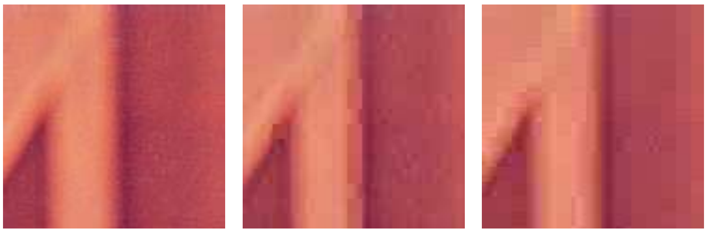


Fig. 9. (a) Zoomed section of the Original Image shown in Fig. 8. (a) Fig. 9. (b) Zoomed section of Fig. 8. (b) with QR 1:2:3:4:5 Fig. 9. (c) Zoomed section of Fig. 8. (c) with QR 5:4:3:2:1

Table II shows the effect of increasing Quantization Factor (QF) on compression ratio, MSE, PSNR and SSIM. As the Quantization Factor goes on increasing, compression ratio goes on increasing but image quality deteriorates as is evident from the values of MSE, PSNR and SSIM. Fig. 6 (d), fig. 7 (d) and fig. 8 (d) show the images compressed and reconstructed using QF 1.

Quantization Factor	Compression Ratio	MSE	PSNR	SSIM
1	26.05	13.81	36.72	0.9180
2	28.37	21.10	34.88	0.8658
3	46.11	19.72	35.18	0.8741
4	46.98	22.51	34.61	0.8529
5	55.44	23.98	34.33	0.8377

Table II. Constant Quantization Results

Quantization Ratio	Compression Ratio	MSE	PSNR	SSIM
1:2:3:4:5	32.44	18.13	35.54	0.8853
1:2:3:4:1	31.02	17.9	35.6	0.8864
5:4:3:2:1	41.89	23.69	34.38	0.8445
1:4:3:2:1	33.92	19.77	35.17	0.8795
1:2:3:2:1	29.05	18.46	35.45	0.8849

Table III. Variable Quantization Results

Results tabulated in Tables II and III are graphically represented in fig. 10 and fig. 11. Fig. 10 and fig. 11 show the plots of Compression Ratio and SSIM versus Quantization Factor for different Quantization Ratios respectively. Each curve in fig. 10 and fig. 11 represents the values of Compression Ratio and SSIM for a single Quantization Ratio obtained by progressively increasing the Quantization Factor from 1 to 25.

As seen in Fig. 10, QR 5:4:3:2:1 gives the highest Compression Ratio but Fig. 11 shows that its image quality is the lowest. Fig. 10 also proves that quantization of the 63 linearly arranged AC coefficients by any of five above mentioned QRs always gives better compression than that by constant QF.

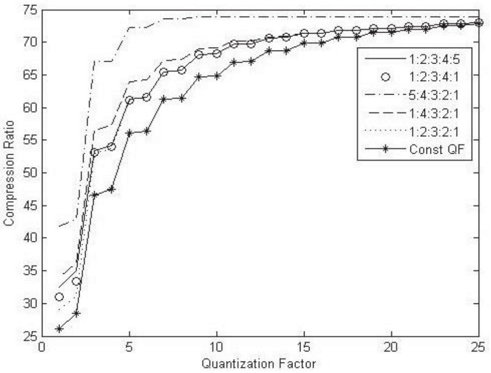


Fig. 10. Compression Ratio vs Quantization Factor for different Quantization Ratios.

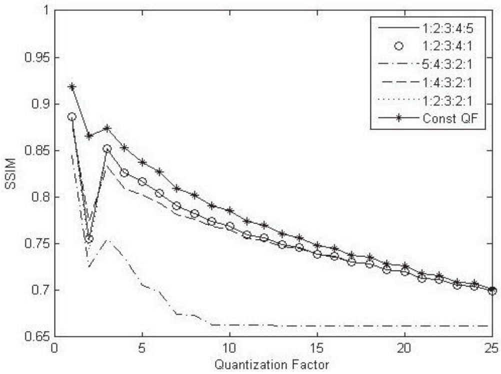


Fig. 11. SSIM vs Quantization Factor for different Quantization Ratios.

5 Conclusion

The proposed modifications to the JPEG standard drastically improve the compression ratio. The presented experimentation shows that the efficient storage of Huffman coded words and the variable quantization of the zigzag ordered (linearly arranged) AC coefficients are resulting in the improved compression compared to the conventional JPEG technique. This is also evident from the fact that the images compressed using the proposed algorithm required on an average 86.306 % less storage space as compared to those compressed using the standard JPEG algorithm and the average time required for compression and reconstruction of these images is 50.316 seconds. The perceptual quality of the images is maintained even with the achieved higher compression ratios.

References

[1] Ding, J. J., Huang, Y. W., Lin, P. Y., Pei, S. C., Chen, H. H., Wang, Y. H. (2013). Two dimensional orthogonal DCT expansion in trapezoid and triangular blocks and modified JPEG image compression. *IEEE transactions on image processing*, vol. 22, no. 9, 3664- 3675.

[2] Pattanaik, S. K., Mahapatra, K. K. (2006). DHT based JPEG image compression using a novel energy quantization method. *IEEE Intl. Conference on Industrial Technology*, 2827- 2832.

[3] Jiang, Y., Pattichis, M.S. (2011). JPEG image compression using quantization optimization based on perceptual image quality assessment. *Conference Record of the 45th ASILOMAR-2011*, 225-229.

[4] Akhtar, M. B., Qureshi, A. M., Qamar-ul-Islam (2011). Optimized Run Length coding for JPEG image compression used in space research program of IST. *Intl. Conference on Computer Networks and Information Technology*, 81-85.

[5] Jafari, R., Ziou, D., Mammeri, A. (2011). Increasing compression of JPEG images using Steganography. *IEEE Intl. Symp. on Robotic and Sensors Environments*, 226-230.

[6] Sreelekha G., Sathidevi, P.S. (2007). An improved JPEG compression scheme using Human Visual System model. *14th Intl. Workshop on Systems, Signals and Image Processing*, 98-101.

[7] Wallace, G. K. (1992). The JPEG Still Picture Compression Standard. *IEEE trans. On Consumer Electronics*. vol. 38, no. 1, 18-34.

[8] Chaddha, N., Agarwal, A., Gupta, A., Meng, T. H. Y. (1994). Variable compression using JPEG. *Proc. Intl. Conference on Multimedia Computing ans Systems*, 562-569.

[9] Dumic, E., Mustra, M., Grgic, S., Gvozden, G. (2009). Image quality of 4:2:2 and 4:2:0 Chroma Subsampling formats. *51st Intl. Symp. ELMAR-2009*, 19-24.

[10] Gonzales, R. C. and Woods, R. E. (2012), *Digital Image Processing*, 3rd ed., Pearson Ed., 525-614.

Auto-Generation of Runner's Stroboscopic Image and Measuring Landing Points Using a Handheld Camera

Kunihiro HASEGAWA, Hideo SAITO
Graduate School of Science and Technology
Keio University
3-14-1, Hiyoshi, Kohoku-Ku, Yokohama, Kanagawa, Japan
{hiro, saito}@hvrl.ics.keio.ac.jp

Abstract

This paper discusses a method for automatically generating a stroboscopic image and measurement of landing points from a video image using a handheld camera. The purpose of this method is training of amateur runners. For the training, generating a stroboscopic image which has an arbitrary frame's runner and measurement of landing points are important to measure stride length and speed of the runner from the stroboscopic image. The proposed method for generating a stroboscopic image has two steps. First, a stitched background image is generated from the input video image. Then, an arbitrary frame's runner image is overlapped to the stitched background image. Thus, this method can generate a stroboscopic image including the runners in arbitrary frames without overwriting them in some frames. These processes can be executed automatically. The method for measurement of landing points uses a homography which is a matrix to project and transform a plane to other plane. We demonstrate the effectiveness of the proposed method by showing some experiments under various environments. The Result of the measurement as an input to the stroboscopic image obtained the measurement result whose accuracy was sufficient for analysis.

Keywords: Sports Vision, Stroboscopic Image, Runner, Overlap, Handheld Camera

1 Introduction and Motivation

In various sports, a variety of scientific analysis for them has been performed. These results contribute to players of sports for improving their skills. The information given by the scientific analysis is also useful to make a TV program interesting. Computer vision has also been used for analyzing sports. For example, Hamid et al.[1] visualized an offside line by tracking of multiple persons in football. Lu et al.[2] proposed the method for motion analysis for basketball. Atmosukarto et al.[3] performed the study of the recognition method for a formation of American Football. Beetz et al.[4] proposed the system it creating an analysis models by tracking the ball and multiple players from a video image assumed to be used in various sports.

Athletics is one of the targets of an analysis as described above. Athletics have a huge variety of items to be analyzed. A number of studies for these analyses have also been performed. For example, Yang et al.[5] proposed the method of discrimination a runner's running state by recognizing the difference of the motion of the foot from the input video image. Furthermore, studies using variety of sensors have also been performed. For example, Strohrmann et al.[6] analyzed the kinematic change with the fatigue in running. This method measures the displacement of the position and angle etc. of a hip and each joint using 12 wearable orientation sensors in the whole body of the runners initially. Analysis results are obtained by comparing these measured data. Oliveira et al.[7] visualized the change in a heart rate of runners during running with a heartbeat sensors.

Such scientific analyses have been used only for professionals. However, amateur runners' interest in such analyses is also growing with the growing popularity of a running in recent year. In targets of athletics, a runner's stride length and speed have advantages as ease visualization and clarity of the results. Thus, these items are suitable to support amateur runners. Existing methods for analyzing them are using instrument by laser or wearing GPS sensors and so on. However, using a special instrument or a trouble to wear sensors are a problem in these methods. In addition, the system, such as above study[4], does not have a simplicity of use for amateur runners. Further, it does not fit for the analysis of an athletics stride length and so on.

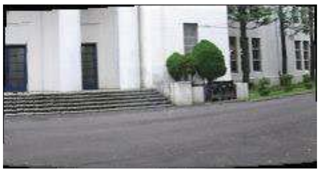
For these problems, we have decided to develop the method to measure a runner's stride length and speed using only one handheld camera. For realizing this goal, we need to automate a judging the landing and measuring landing points. In this paper, we discuss to create follow images from the video image captured by handheld camera and measure landing points as first step for the goal.

1. The stitched image extracted only background(Figure 1 (a))

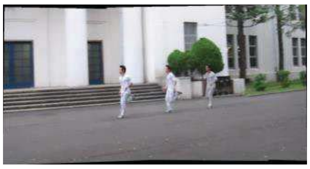
2. The stitched image projected runners in each frame(Figure 1 (b))

In above items, creating the stroboscopic image which has an arbitrary frame's runner shown as Figure 1 (b) and measurement landing points are main outcome. These images are necessary for an automation to judge landing timings. In addition, a measurement of landing points uses the stroboscopic image. We can create the image extracted only the runner by taking the difference between them. In addition, we think that we can solve the problem to judge landing timings.

In this paper, Section 2 details our proposed method. The result of experiments are described in section 3. Section 4 discuss the result. Finally, conclusion is described in section 5.



(a) Stitched image extracted only background



(b) Stitched image projected runners in each frame

Figure 1 Creating result images

2 Proposed Method

2.1 Overview of the System

Figure 2 is the overview of the system for measuring runner's speed and stride length using only one hand-held video camera. As shown in this figure, this system has three elements, generating stroboscopic image of the runner, a judgment of the landing timing and a measurement of the landing points. Among them, we developed the method of generating the stroboscopic image and a measurement of landing points in this paper. As described above, a judgment of landing timings will uses the stroboscopic image generated by the method proposed in this paper. For measurement, we input landing points manually this time because a judgment of landing section have not been complete yet.

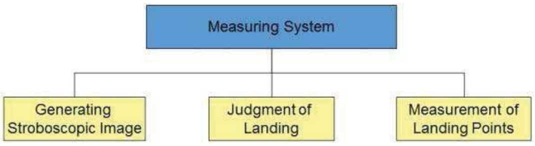


Figure 2 The overview of measuring system

2.2 Overview of Method

The goal of this method is to generate a runner's stroboscopic image and measure landing points using this stroboscopic image. For former section of this goal, generated stitched image needs to have runners of arbitrary frames. However, existing method[8] erases runners in some frames shown as Figure 3. In this figure, the runner in the second frame of input images is erased in the output stitched image. It is because the current frame overwrites the area which have the runner in a previous frame if this area have an overlap with a current frame regardless of whether it has a runner or not. Thus, we tried to solve the problem in this study. This problem is solved by two steps. The first step is generating of the stitched image extracting only background. The next step is a projection of a runner to this stitched image. The first step uses the generating method of a background image using Mean-Shift proposed by Cho et al.[9]. The second step uses the extraction of runners using HOG descriptor[10] and a projection of this runners to the stitched image.

For latter section of this goal, we developed the method using a homography which is a matrix to project and transform a plane to other plane between an image coordinate system and a world coordinate system.

2.3 Generation of Background Stitched Image

The first step is a generation of the stitched image extracting only background. We use the method proposed by Cho et al.[9] as described above. The first step is extraction of feature points from each frame of an input video. Then, a calculation of a homography between each frame uses these feature points. Finally, a stitched image is generated by this homography. Feature points are SURF[11] in this method.

The next step is a generation of the stitched image extracting only background. The extraction of a background uses Mean-Shift[12][13] for each pixel of the stitched image. Each pixel of original images projected to the same pixel in the stitched image are sample points for Mean-shift in this step. Pixels of the background area projected to the stitched image more than that of a runner. This method bases the assumption that the period existing runners on a certain location is shorter than the period there are no runner.

In the other words, pixels of the background area projected to the stitched image more than that of runner. Thus, using Mean-shift can extract a background.

In Mean-Shift processing, a pixel value of each original image is assigned as the initial value. The bandwidth is 10. The kernel formula is Epanechinikov kernel just like Cho et al. In addition, the weight for decision the result is calculated for each candidate value. Finally, the value whose weight is biggest is a background pixel value. The weight is the Gaussian in this calculation for easy processing.

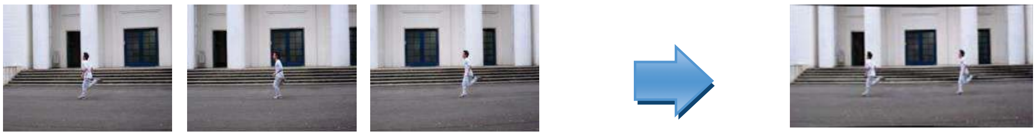


Figure 3 The result of generating a stitched image by the existing method

2.4 Overlapping of Runners

This section explains how to overlap runners in each frame to stitched image as the former section's goal of our proposed method. This process generates a runner's stroboscopic image which is main outcome of this method. Figure 4 is flowchart of this process. First, HOG descriptor[10] extracts the area of runner in every frame from input video image.

Then, this area is projected to background stitched image generated in 2.3. Calculation for this projection uses the homography of each frame used for generating stitched image. The correspondence of projection position between the stitched image and original image is also same after generating background stitched image. Thus this calculation can also use this homography. Figure 5 is the image of this process. This figure shows the example of generating the stitched image from an input video image which has n frames. If an existing method creates a stitched image, runners in some frames are overwritten all or partially shown as the lower left area of Figure 5. Our proposed method extracts the area of a runner first. The area surrounded by a frame is it in this figure. The projected area of a runner is overwritten on the stitched image. Finally, this method can create the stroboscopic image of a runner shown as the lower right area of Figure 5.

2.5 The Method of Measuring Landing Points

This section explains how to measure a runner's landing points as the latter section's goal of our proposed method. Figure 6 is the image of this method. A runner runs between two rows markers placed in a straight line at regular intervals. For measurement, we calculate the homography between the image coordinate system and the world coordinate system using these markers as known feature points. Landing points inputted manually are projected to the world coordinate by this homography. Only 4 feature points set at four corners need for calculate the homography at least. Thus, other feature points are for landmark for runner to run straight.

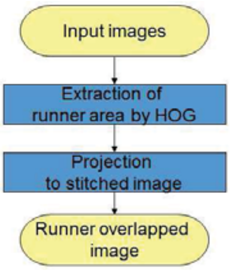

Input images
Extraction of runner area by HOG
Projection to stitched image
Runner overlapped image

Figure 4 The flowchart of overlapping runners

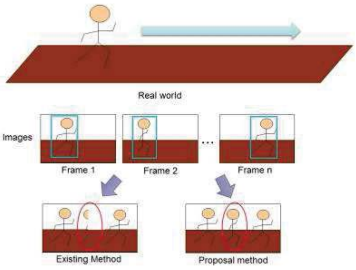


Figure 5 The image of overlapping runner process

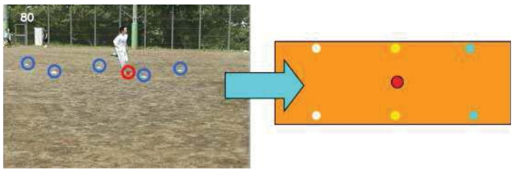


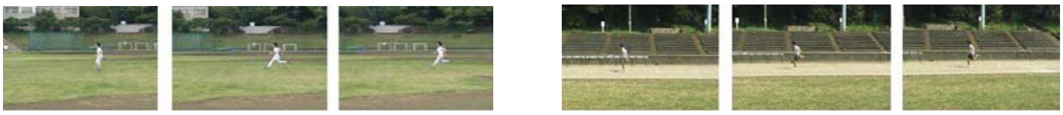
Figure 6 The image of measuring landing point

3 Experiment and Result

3.1 Overview of Experiments

This section explains the experiments about three methods described in the previous section. We experimented three methods separately to clarify each process. We used some frames extracted from the video image capturing the person is running by a video camera or digital still camera having video capture function as input for each case. As mentioned initially, the person who uses a camera holds it without using a tripod and so on and only moves a camera so as to follow runner. Further, we captured video images in three scenes that the ground surface is a concrete, dirt and grass for a generation of the background and overlapping of the runner in order to show that this method can be used in various environments. Figure 7 is examples of input images. In addition, Figure 3 is example of input images that a ground surface is a concrete.

In the experiment of measurement, Figure 8 is the examples of input image. We use the stroboscopic image created from these images for input. We needed the ground truth of landing points to confirm the accuracy of our method. For this purpose, we use the scene that the runner ran 30m on the ground whose surface is a dirt. We could get footprints made by spike shoes. We used Microsoft Visual Studio 2010 as the IDE, C++ as the programing language, OpenCV 2.4.6 as the image processing library for implementation



(a) Grass (b) dirt

Figure 7 Examples of input image for two experiments

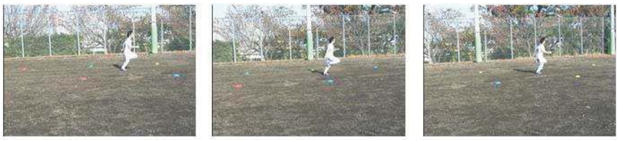
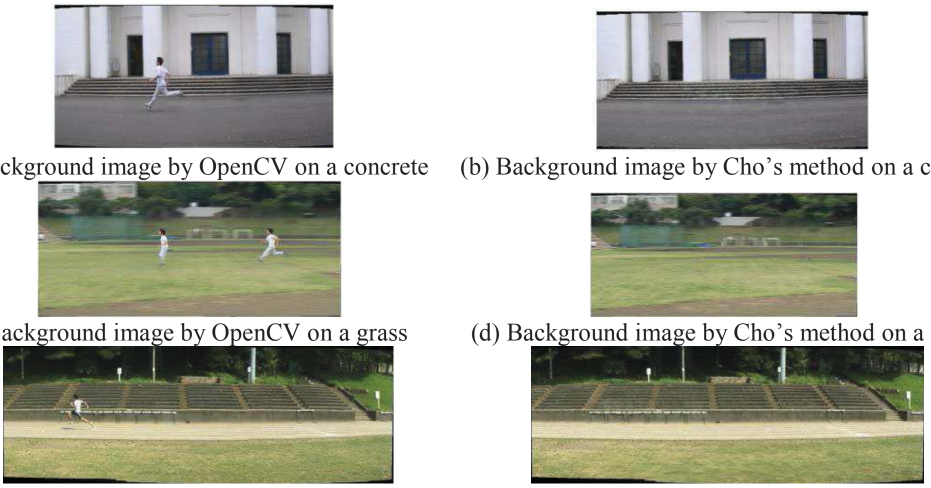


Figure 8 Examples of input image for measurement

3.2 Generation of Background Stitched Image

In this section, we describe the result of generating the stitched image extracting only background. Figure 9 is the result of this experiments. Figure 9(a), (c), (e) are the stitched images on a concrete, grass and dirt scene respectively using "stitch" function in OpenCV. These results disappear some runners. This reason is same as Figure 3. Figure 9(b), (d), (f) are the results of using Cho's method in same scenes. However each image has a few noises influence of calculation for projection, they have no runner. Thus, we confirmed that this process is successful.



(a) Background image by OpenCV on a concrete (b) Background image by Cho's method on a concrete

(c) Background image by OpenCV on a grass (d) Background image by Cho's method on a grass

(e) Background image by OpenCV on a dirt (f) Background image by Cho's method on a dirt

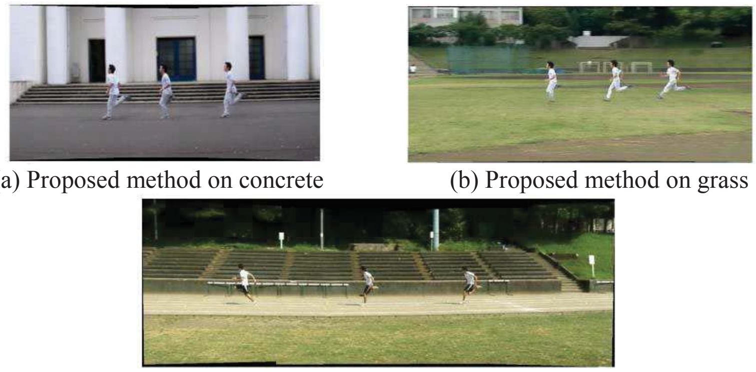
Figure 9 Stitched images by extracting only background

3.3 Overlapping of Runners

We describe the result of overlapping of runners in this section as the former section's goal of our proposed method. Figure 10 is the result of this experiments. We used frames in Figure 3 and Figure 7 for overlapping of runners. In addition, the stitched image generated by the existing method for comparison are same as the previous section. Figure 10(a), (b), (c) are the results of using the proposed method on a concrete, grass and dirt scene respectively. Stitched images generated by the existing method show only runners in a part of frames. In contrast, images generated by our proposed method show all runners without disappeared. From this results, we confirmed that our proposed method can generate the stroboscopic image of runners.

3.4 Measurement of Landing Points

Finally, we describe the result of measuring landing points with the stroboscopic image as the latter section’s goal of our proposed method. Figure 11 is visualized landing points as footprints of runner generated by this experiment. We can get an appearance of landing. From this result, we can also create the graph showing a change of stride length. Figure 12 is the graph. According to this graph, for example, we can understand that this runner’s stride length extends gradually. The errors of the measurement results obtained by this method ware that the average was 0.14m, maximum was 0.61m, minimum was 0.01m, and the standard deviation was 0.14m. Some errors occur by the accuracy of the click of the landing points. However, according to preliminary experiments, the error of approximately pair of the shoes length (≑0.30m) does not affect for analyses. Thus, we confirmed that our method can get a sufficient accuracy of a measurement.



(a) Proposed method on concrete

(b) Proposed method on grass

(c) Proposed method on dirt

Figure 10 Runner's stroboscopic images synthesized by the proposed method

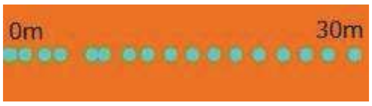


Figure 11 Footprint image generated by our method

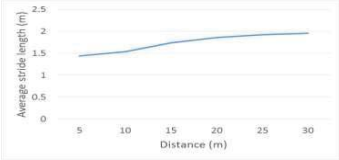


Figure 12 The graph showing stride length

4 Discussion

We confirmed that our proposed method solved the problem described in section 3 from the results of experiments. By these results, generating of a stroboscopic image part in the system for measuring a runner’s stride length and speed was completed. Measuring runner’s landing points part can measure with an accuracy which does not affect the analysis. Then we think that an automatic judgment of landing timing can be realized by using these stroboscopic images and background stitched images. As a point that should be noted, this proposed method has failure cases. The concrete examples are following three cases.

I. Too fast to move a video camera.
II. Rate of a runner on the image is too large.
III. Rate of a runner on the image is too small.

We explain briefly each case. I is that feature points are not taken by blurring occurs in the image, or not appearing a same area between frames because a camera is moved too fast. However, use cases of this method are the case that the person who stops at a certain place captures a runner. Thus, we can consider not to occur this failure. II is that feature points not obtained between frames to generate the stitched image because the area of the runner in image is too large. On the other hand, III is that area of the runner is not extracted by the HOG descriptor because the area of the runner is too small. II and III are scenes that can occur depending on the environment. Simultaneously, a technical solutions are difficult. Thus, we need to be careful at capturing.

Further, the frame rate of a video image is also important because we cannot get enough measurement accuracy for analyses if that is too small. For this point, we can use the result of preliminary experiments for to develop measurement system. In this experiments, the frame rate of video is 14.985fps and we could get the good result of analyses. Thus, we consider that there is no problem if the frame rate of video image is 14.985fps which is a half of 29.97fps (a frame rate of NTSC) or more. We used the video image deactivated interlace mode. So, the frame rate of this video image is 14.985fps.

Finally, we describe the theory of how to obtain landing timings using the result of this proposed method. First, we take the difference between the background stitched image overlapped the area of a runner and background stitched image. Only overlapped runner on each frame remains as a trajectory of the runner by this process. Of course, the existing method for extracting the human contours have been proposed are various. However, it is possible to extract a runner easily and more accurately using this method. Further, it is possible to obtain ground timings by analyzing this trajectory of runner. We think that there are to look the up-and-down motion of the runner or overlap of the runner between a current frame and a previous one for the analysis method. For example, in the former case, the change timing from descending to rising of runner's movement can be regarded as the landing timing. For this purpose, the changing point is searched in the trajectory of runner. In the latter case, we use that the runner's feet does not move almost at the landing timing. From this fact, the landing timing can be regarded as the timing that the overlapped area of the runner's feet is large between frames. We consider that these methods are applicable to judge the landing timings except failure cases described above. We are going to consider their detail in the future.

5 Conclusion

In this paper, we proposed the method of automatic generating stroboscopic image of runner and measuring landing points that they are necessary for the realization of the system for measuring runner's speed and stride length. To realize the method, we develop three sub methods, the generation of stitched image extracting a background using the Mean-Shift, overlapping of the extracted runners region using the HOG descriptor to the stitched image and measurement of landing points using a homography. The usefulness of our proposed methods is confirmed by results of experiments. Finally, we are going to develop the method of judgment the landing timings automatically using the stroboscopic image generated by this proposed method in the future.

References

[1] R. Hamid, R. K. Kumar, M. Grundmann, K. Kim, I. Essa and J. Hodgins (2010). Player localization using multiple static cameras for sports visualization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 731-738.

[2] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy (2013). Learning to Track and Identify Players from Broadcast Sports Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(7), 1704-1716.

[3] I. Atmosukarto, B. Ghanem, S. Ahuja, K. Muthuswamy, and N. Ahuja (2013). Automatic Recognition of Offensive Team Formation in American Football Plays. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 991-998.

[4] M. Beetz, N. v. Hoyningen-Huene, B. Kirchlechner, S. Gedikli, F. Siles, M. Durus and M. Lames (2009). ASpoGAMo: Automated Sports Game Analysis Models, International Journal of Computer Science in Sport, 8(1).

[5] Yang J. Y., Xu G. L. and Li Y. (2013). Running state recognition in videos via frames' frequency and positions of two feet. Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 310-313.

[6] C. Strohrmann, H. Harms, C. Kappeler-Setz and G.Troster (2012). Monitoring Kinematic Changes With Fatigue in Running Using Body-Worn Sensors. IEEE Transactions on Information Technology in Biomedicine, 16(5), 983-990

[7] G. Oliveira, J. Comba, R. Torchelsen, M. Padilha and C.Silva (2013). Visualizing Running Races through the Multivariate Time-Series of Multiple Runners. 26th SIBGRAPI - Conference on Graphics, Patterns and Images, 99-106.

[8] Microsoft Corporation (2011). Microsoft Research Image Composite Editor (ICE), http://research.microsoft.com/en-us/um/redmond/groups/ivm/ICE/

[9] Cho, S.-H. ; Kang, H.-B. (2011). Panoramic background generation using mean-shift in moving camera environment. Proceedings of the international conference on image processing, computer vision, and pattern recognition (IPCV), 829-835.

[10] N. Dalal and B. Triggs (2005). Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 886-893.

[11] H. Bay, A. Ess, T. Tuytelaars and L. V. Gool (2008). SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU), 110(3), 346-359.

[12] K. Fukunaga and L. Hostetler (1975). The estimation of the gradient of a density function. IEEE Transactions on Information Theory, 21(1), 32-40.

[13] D. Comaniciu and P. Meer (2002). Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), 603-619.

Tactile approach to Material Classification - Evaluated with Human Performance

Emmett Kerr*, T.M. McGinnity, Sonya Coleman
Intelligent Systems Research Centre,
University of Ulster, Magee,
BT48 7JL
ep.kerr@ulster.ac.uk*

Abstract

Knowledge of the physical properties of objects is a requirement to enable effective robotic grasping. Identifying the material from which an object is made, is one such physical property. Characteristics of the material can be retrieved using different sensors; vision-based, tactile based or sound based. Both visual inspection and physical contact with materials can enable the retrieval of detailed information about the material, e.g. colour, compressibility, surface texture and thermal properties. This paper describes a system to classify a wide range of materials based on their thermal properties and surface texture. This research seeks to develop a combined system using both tactile sensing and vision based sensing. Following acquisition of data from a sophisticated tactile sensor, the system uses principal component analysis (PCA) to extract features from the data, which are then used to train a two stage Artificial Neural Network (ANN) to classify materials, first into groups and then as individual materials. The system is compared with human performance and the results demonstrate that the proposed system can outperform humans.

Keywords: Material Classification, Tactile sensing, Neural Networks, PCA

1 Introduction

Humans can quickly learn a lot about an object or material by viewing it from different angles and can estimate how it might feel to touch and how heavy it could be. This is due to our highly sophisticated visual capabilities and based on adapting knowledge learned from known objects in the past which may appear similar. However there are some properties which may be difficult to detect by vision alone, for example surface texture or determining what material an object is made from. Manipulating the object by hand enables us to learn a vast amount more about the object, which in turn will help to determine how the object should be grasped.

Due to their sophisticated tactile perception, humans can inherently perform complex manipulation tasks, such as squeezing (to assess compressibility), adjusting the size and strength of their grasp to securely hold an object, and distinguishing between objects of different textures and different temperatures. Completion of general exploratory movements outlined by experimental psychologists [Lederman and Klatzky, 1987] is very fast for humans, leading to rapid evaluation and possible identification of the object.

In this paper we focus on the use of an artificial fingertip to acquire data on the thermal properties and surface texture of materials, that are subsequently analysed using a two stage approach, to initially identify which group the material belongs to (e.g. wood, metal, plastic etc.) and subsequently the specific individual material (e.g. aluminium, copper, pine etc.). Building on the work presented in [Kerr et al., 2014] the contribution of this work is the introduction of a two stage ANN approach. This alternative approach is evaluated against the previous approach in [Kerr et al., 2014] to see if an approach with increased accuracy and speed could be achieved. The remainder of this paper is organised as follows: Section 2 presents an overview of related research in material classification using both vision and tactile sensors. Section 3 describes the learning algorithm used for the robot system and the experimental set-up, including an overview of the equipment used. Section 3.4 explains the set up of the human evaluation experiments while Section 4 presents an evaluation and discussion of the performance of the artificial system and the use of human subjects. Conclusions and plans for future work are given in Section 5.

2 Background and Related Research

There are many texture classification methods using images in the literature. Sharma and Singh [Sharma and Singh, 2001] presented a performance evaluation of five feature extraction tools used in image analysis for texture. The five that were tested are auto-correlation, edge frequency, primitive length, Law's method, and co-occurrence matrices. These five methods and combinations of them were tested on images from the Meastex database [Meastex, 1997], containing samples of asphalt, concrete, grass and rock images. For performance evaluation of the feature extractors, Linear Discriminant Analysis (LDA) and two modified k-nearest neighbour (*k*-NN) were compared. The best results for the *k*-NN methods were obtained with co-occurrence matrices whereas for LDA a combined set of features produced the best results.

Three critical characteristics that must be known about an object if performing a grasping action are surface texture, compressibility and thermal properties. Some methods reported in the literature are capable of achieving high classification rates of materials based on surface texture, e.g. Chathuranga et al. [Chathuranga et al., 2013] achieved a rate of 85%. However, many methods struggle to distinguish between quite different materials of similar roughness [Jamali and Sammut, 2010, Decherchi et al., 2011]. Xu et al. [Xu et al., 2013], used the BioTAC$^{\text{TM}}$ finger tactile sensor from Syntouch® and present an algorithm which considers the aforementioned three key properties of the material to enable classification. Classification rates of 99% across the ten test materials were achieved. The only failure in the system was due to a damp sponge being identified as a feather due to the similar compliance. Although the approach in [Xu et al., 2013] achieved 99% classification this approach would be slow and computationally expensive if it was used to explore a full object for identification as it requires the analysis of a large quantity of datasets.

A method using only the thermal properties of a material was presented in [Kerr et al., 2013]. Materials were classified into groups initially in one experiment and classified individually in a second experiment. PCA was performed on the raw thermal conductivity and static temperature data in order to extract the relevant features of the data and these features were used to train an ANN. The artificial system was found to outperform human performance, when the human was restricted to use of the same thermal properties. This work was extended in [Kerr et al., 2014] by introducing a further modality, namely surface texture (vibration). Furthermore the classification system was redesigned to be much more efficient by reducing the number of principal components, hidden layer neurons and training epochs required. The work presented in this paper further extends this previous work and proposes a two stage ANN approach to classify materials into groups initially and then use the output from this network to feed a second set of networks for each material group in order to identify each individual material, within a specified group. Knowing the exact material could allow a robot system to estimate the weight of an object using its shape and estimated volume of the identified material.

3 Methodology

3.1 Data Collection and Pre-Processing

To classify the materials, an experiment was designed to enable a BioTAC fingertip to perform two actions on the test materials. Both actions replicate, to some extent, the actions of a human when inspecting an unknown material for the first time.

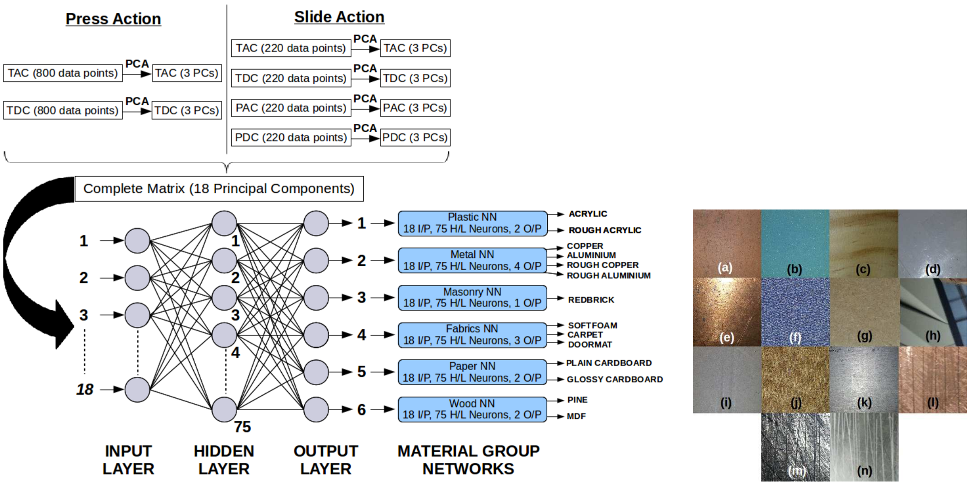
The BioTAC fingertip is a tactile sensor which is shaped like a human fingertip and is liquid filled, giving it similar compliance to a human fingertip [Lin et al., 2009, Kerr et al., 2014]. Using the fingertip, the thermal flow rate (TAC) and absolute temperature (TDC) values can be used to determine the thermal properties of the material with which the fingertip is in contact. The AC pressure vibration signal (PAC) and DC pressure signal (PDC) values can be used to determine the vibration (of the internal conductive fluid) caused by the surface texture of the material when the fingertip is slid along it. Similar to the experiments carried out in [Kerr et al., 2013, Xu et al., 2013], the BioTAC fingertip is allowed 15-20 mins to reach its steady state temperature (approximately 31°C, 10°C above ambient) after being first powered on. To produce the thermal exploratory movement, the fingertip is then pressed onto the material with a constant force of 3N. All data for the press action were collected from the fingertip for 20 seconds after initial contact was made. This allowed time for the heat flow to settle after contact. To produce the slide action, the fingertip was pressed down on to the surface of the material again using a constant weight applied to the fingertip, equating to a force of 1.59N, and slid along the surface for a distance of approximately 5 cm. The data from the BioTAC fingertip are extracted using MATLAB.

To reduce the dimensionality of the data inputs, Principal Component Analysis (PCA), using Eigendecomposition of the sample's covariance to calculate the principal components, was applied to the modalities' datasets. PCA was applied to each individual modality first. This allows combinations of the different modalities to form a matrix of the principal components for each material to suit each experiment (i.e. the principal components calculated from TAC and TDC for the press experiment and principal components calculated from TAC, TDC, PAC, PDC for the slide experiment). These combinations of principal components are then used to train the ANN.

3.2 Classifier

The two stage approach used is a series of back propagation ANNs, each with one hidden layer, as shown in Figure 1.1. The first ANN is used to classify the materials into groups. Various versions of the ANN were trained and retrained until the optimal ANN was found. Variations of the number of neurons in the hidden layer and the number of training epochs were evaluated. It was found that the best performing structure for the ANN consisted of 75 neurons in the hidden layer after training for 1500 epochs. Various trails demonstrated that the optimal value of principal components (PCs) for each dataset using PCA was found to be three. Therefore, as there are six modalities in total that represent each material at each experiment (two from the press action and four from the slide action) there are a total of 18 inputs (three PCs per modality, 12-bit values) to the ANN for each training sample, as shown in Figure 1.1. The second stage of the ANN is comprised of six individual ANNs for each material group (i.e. plastic, metal, masonry, fabrics, paper and wood). These ANNs were all trained individually with their respective outputs relating to how many materials there are in each group, for example plastic has two outputs for the two plastic materials whereas metal had four outputs for the four metal materials in that group. After training these networks were saved and then used for classifying the resulting output test material from the first ANN. In order to avoid false positives being identified there was a threshold put in place for the testing of materials within the group. If the output of the neuron fired for the material sample being tested is not greater than 0.3 then it was eliminated and counted as a failed classification.

For all evaluations of the ANN, 5-fold cross validation was carried out. The data were split into 5 subsets, then 80% (4 of the 5 subsets) of the data was used for training and the other 20% for testing (1 of the 5 subsets). The subsets used for training and testing were alternated until all combinations were utilised. This produced five training accuracies and five testing accuracies for first ANN and each of the individual material ANNs. The five accuracies for both training and testing were then averaged, giving the average training and testing accuracies for the classification of the materials into their groups. The test outputs from each of the individual group ANNs were compiled in a matrix and evaluated to calculate an average classification



1.1 1.2

Figure 1: (1.1) Diagram showing the two stage ANN used for material classification (1.2) Fourteen test materials used for classification based on thermal conductivity and surface texture. Materials are (a) Red Brick, (b) Soft Foam, (c) Pine Wood, (d) Acrylic Plastic, (e) Copper, (f) Carpet, (g) MDF Wood, (h) Glossy Finish Cardboard, (i) Plain Cardboard, (j) Doormat and (k) Aluminium Metal, (l) Rough Copper, (m) Rough Aluminium, (n) Rough Acrylic.

accuracy of the fourteen individual materials.

3.3 Materials to be Classified

Fourteen materials are used in both the press and slide experiments. Some of the materials are quite similar (i.e. two types of wood, two types of metal and two types of cardboard) in order to test if the approach is capable of not only classifying the type (group) of material but also distinguishing between materials within a group. Examples of each material and the list of materials used can be seen in Figure 1.2. Although similar to three of the other materials, the rough materials had a very rough surface in comparison to their smooth counterparts. This roughness was apparent when dragging a finger across the surface. Fifteen trials of each material were completed.

The first experiment aims to classify each material in terms of its group type, for example MDF and pine are both in the wood group. The materials tested were split into six groups; these groups and the materials belonging to each group are shown in Table 1. The second experiment evaluates how accurately the artificial system can classify each material individually within the classified grouped, by only having to consider the materials in the respective group.

Table 1: Table showing the groups of materials and their members.

Group	Materials
Plastic	Acrylic, Rough Acrylic
Metal	Copper, Aluminium, Rough Copper, Rough Aluminium
Masonry	Redbrick
Cardboard	Cardboard Glossy, Cardboard Plain
Fabrics	Soft Foam, Carpet, Doormat
Wood	MDF, Pine

3.4 Evaluation Set-up

To evaluate the artificial system, it is compared with human performance using the same set of fourteen materials. The participants consisted of 12 healthy humans, two female and 10 male, all aged between 23-56 years and all participants used their right hand. The participants were instructed to perform a ’press’ and ’slide’ procedure on each material using their finger; a training and testing phase was completed. In the training phase, the participants are informed what each material is. In the testing phase the materials were presented to the participants again, but in a random order and the participant was required to identify the material. They could select a material more than once and also had the option of saying they didn’t know. For the press experiment, the participants were instructed to press down on the material with their index finger and to leave it on the material for a maximum of 20 seconds (similar to the artificial fingertip). They were also instructed that it was prohibited to slide or rotate their finger, or move it laterally at any time when in contact with the material, for this initial phase of the experiment. They were then instructed to lift their fingertip off and reapply it to the material and this time to slide their fingertip along the material.

4 Evaluation Findings

Initially, the participants were evaluated for identification of material groups. It is found the average accuracy was 79.76%. Secondly, the participants were evaluated for identification of individual materials. It is found that the human participants achieved an average of 69.64% accuracy for identifying the individual material from the fourteen test materials. A breakdown of the results of the human evaluations from two experiments can be seen in [Kerr et al., 2013, Kerr et al., 2014], one experiment where the participants identified the material using their thermal properties only and the other experiment using both the thermal properties of the material and the surface texture. Every participant achieved 100% identification of the soft foam, the carpet and the doormat. It was found that all of the individual material identification accuracies, with the exception of aluminium, either increased or stayed at a maximum when the exploration of the texture, via the sliding action, was introduced to the human experiments. This shows that the surface texture plays a vital part in the identification of materials, and indeed is a critical characteristic that humans use to identify materials. The results also showed that pine and rough copper were the most difficult materials for the human participants to identify. Furthermore, the human participants struggled with the identification of the rough materials, with only 2 out of 12 of the participants being able to identify all three rough materials.

Table 2: Table comparing the experimental results

	Material Group	**Individual Materials**
Human Participants	79.76%	69.64%
One Stage Artificial System [Kerr et al., 2014]	83.81%	79.05%
Two Stage ANN Artificial System	72.86%	70.48%

Human participants did not perform as well as the artificial system on either of the experiments. A comparison of the results from the human experiments, the results obtained from the system presented in [Kerr et al., 2014] and the results obtained by the system presented in this paper can be seen in Table 2. The results for the artificial system are calculated by computing the average of the classification rates using 5-fold cross validation as explained in Section 3.2. When classifying for the individual material, the two stage approach proved to be less computationally expensive and therefore marginally faster because the material had been firstly classified into its group. Therefore there was a maximum of only 4 materials to classify between and materials from all the other groups could be ignored, unlike the system in [Kerr et al., 2014] where there are 14 materials to classify between when attempting to identify the individual material. However, despite this and the fact that it has outperformed the human participants, the dual NN approach did not perform as well as the system presented in [Kerr et al., 2014]. If weighing up the marginal speed advantage of the system against the reduction in accuracy then the system with the more accurate classification, namely that presented in [Kerr et al., 2014], would be chosen as the more efficient system overall.

5 Conclusion and Future Work

A two stage ANN approach was presented for the classification of fourteen individual materials, firstly into their respective groups and secondly as individual materials within their groups. The system outperformed human participants in both stages of experiments however was slightly less accurate than the approach presented in previous work [Kerr et al., 2014], therefore although the presented approached was marginally faster this was outweighed by the reduction in accuracy meaning the approach presented in previous work [Kerr et al., 2014] is the preferred approach. Other training classifiers will be considered for future work.

References

[Chathuranga et al., 2013] Chathuranga, D., Ho, V., and Hirai, S. (2013). Investigation of a biomimetic fingertip's ability to discriminate fabrics based on surface textures. In *Advanced Intelligent Mechatronics (AIM), 2013 IEEE/ASME International Conference on*, pages 1667–1674.

[Decherchi et al., 2011] Decherchi, S., Gastaldo, P., Dahiya, R., Valle, M., and Zunino, R. (2011). Tactile-data classification of contact materials using computational intelligence. *Robotics, IEEE Transactions on*, 27(3):635–639.

[Jamali and Sammut, 2010] Jamali, N. and Sammut, C. (2010). Material classification by tactile sensing using surface textures. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 2336–2341.

[Kerr et al., 2013] Kerr, E., McGinnity, T., and Coleman, S. (2013). Material classification based on thermal properties - a robot and human evaluation. In *2013 IEEE International Conference on Robotics and Biomimetics, December 12-14 2013, Shenzhen, China*, pages 1048–1053.

[Kerr et al., 2014] Kerr, E., McGinnity, T., and Coleman, S. (2014). Material classification based on thermal and surface texture properties evaluated against human performance. In *SUBMITTED TO 2014 IEEE International Conference on Control, Automation, Robotics and Vision, ICARCV 2014, Singapore, Dec 2014*.

[Lederman and Klatzky, 1987] Lederman, S. and Klatzky, R. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19(3):342–368.

[Lin et al., 2009] Lin, C., Erickson, T., Fishel, J., Wettels, N., and Loeb, G. (2009). Signal processing and fabrication of a biomimetic tactile sensor array with thermal, force and microvibration modalities. In *Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on*, pages 129–134.

[Meastex, 1997] Meastex (1997). Meastex database.

[Sharma and Singh, 2001] Sharma, M. and Singh, S. (2001). Evaluation of texture methods for image analysis. In *Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001*, pages 117–121.

[Xu et al., 2013] Xu, D., Loeb, G., and Fishel, J. (2013). Tactile identification of objects using bayesian exploration. In *IEEE International Conference on Robotics and Automation (ICRA) 2013*.

Application of Similarity Measures to Magnetoencephalography data

Richard Gault*, T.M. McGinnity, Sonya Coleman
Intelligent Systems Research Centre,
University of Ulster, Magee, BT48 7JL
gault-r2@email.ulster.ac.uk*

Abstract

Magnetoencephalography (MEG) is a non-invasive neural imaging technique which passively measures the miniscule magnetic fields produced by neuronal activity without any risk to the subject. Its superior temporal resolution has led to its use in multimodal studies alongside neural imaging techniques, such as functional magnetic resonance imaging (fMRI), which provide high spatial resolution. Multivariate analysis (MVA) is a well-established field of statistics and is currently applied to MEG data for the purposes of artefact identification using principal component analysis (PCA) and independent component analysis (ICA). This paper considers how the similarity measures of Frobenius norm, PCA similarity measure (S_{PCA}) and Eros can directly analyse the multivariate data produced from MEG recordings. These techniques are applied to auditory stimuli to evaluate to what extent different stimuli can be distinguished by the corresponding neural activity. The results show that Frobenius norm finds dissimilarity between the neural response to distinct tones while Eros and S_{PCA} show similarity between neural responses to certain pairs of tones. The results will be used to inform future studies where the measures identified in this study can be used as part of classification algorithms as well as provide a basic measure to map the similarities in the conditions to the similarity of the neural responses.

Keywords: Multivariate Analysis; Multivariate Time Series; Magnetoencephalography; Similarity measures; classification

1. Introduction

Magnetoencephalography (MEG) is a non-invasive and passive neural imaging technique which records the magnetic fields emitted from the brain as a result of neural activity. MEG has temporal resolution of sub millisecond order, superior to that of MRI/fMRI, PET and CT, and better spatial resolution than EEG [1]. Thanks to advancements in technology and this high spatiotemporal resolution MEG has become a popular neural imaging technique for researchers. Despite this, the only clinical application for MEG is pre- and post-assessment of patients with epilepsy and brain tumors being considered for neurosurgery [2, 3]. Research has seen MEG used in other areas such as Alzheimer's disease [4], autism [5], schizophrenia [6] and tinnitus [7]. In both clinical and experimental work MEG is often considered as part of a multimodal approach; in particular with fMRI [8]. This requires multiple scans across the different imaging modalities which may not be appropriate for certain subjects; for example those with metallic implants may not be eligible for an fMRI scan but are able to undergo a MEG scan due to its passive nature. MEG alone can produce large amounts of information about neural activity in a short period of time with no known health risks to subjects. The data produced from recordings are multivariate as each sensor can be thought of as a condition or observation across multiple time points. This paper discusses how similarity measures from multivariate analysis (MVA) can be used for comparing experimental conditions in MEG analysis in order to utilise the information gathered from a MEG recording. Although it is highly beneficial to use MEG in a multimodal context, specifically when its temporal resolution is combined with the spatial resolution of fMRI, this paper is designed to consider the potential of MEG as a powerful neural imaging tool in its own right. The outline of the paper is as follows. Section 2 considers what data can be extracted from MEG recordings while Section 3 introduces different similarity measures for multivariate time series data. Key points discussed are illustrated in Section 4 with MEG data representing different auditory evoked responses. The results are presented in Section 5 and final discussion follow in Section 6.

2. Introduction to MEG data

The miniscule magnetic fields produced by the brain, of magnitude 10fT, are vastly different in character to magnetic artefacts such as muscular or cardiac movements as well as environmental noise. It is therefore sufficient to filter data with simple band pass filters, which provide a dichotomy between neural activity and some environmental noise. Further filters such as signal space separation (SSS) or spatio-temporal signal space separation (tSSS) can be used to further remove external interferences and artefacts near the scalp [9]. The data collected from a MEG recording, filtered or unfiltered, can be represented by a matrix $F = [n \text{ sensors} \times t \text{ time points}]$. The rows of F, which are the sensor recordings, can be thought of as observations while the columns, corresponding to the amplitude of the magnetic field emitted across all sensors at a particular time, can be consider features. It is appropriate to describe F as a *multivariate time series* (MVTS) with each row/sensor acting as a univariate time series.

In multimodal studies it is common to overlay the information gathered from MEG recordings onto that of MRI or fMRI recordings [10]. To determine the location of the sources researchers have had to address the inverse and forward modelling problems. This requires assumptions to be made about the potential sources and the propagation of the magnetic fields from these sources to the sensors. Details of different techniques and the assumptions used in gaining a solution to the inverse and forward models are explained in [11]. MEG recordings are simple, safe and comfortable for subjects. Although medical implants can often produce magnetic fields, which distort the recorded signal, there are no known health risks for a subject with such an implant when they undergo a MEG recording and these artefacts can be accounted for using statistical methods. Unlike a MR scanner, MEG scanners are completely silent and subjects can make small movements for comfort which can easily be accounted for in post analysis. This means that the recording process is less stressful than, for example, MRI and is more suitable for children and subjects who are unable to remain stationary during long recordings. Multimodal imaging is limited to the aggregation of the restrictions from individual imaging techniques.

3. Similarity Measures of Multivariate Time Series

Analysing the similarity between pairs of neural responses can provide a co-domain to map the experimental diversity to the differences between the neural responses observed by the MEG scanner. In this section common similarity measures will be compared: Frobenius norm, principal component analysis similarity factor (S$_{\text{PCA}}$) and Eros [12].

3.1. Frobenius norm

Given a $m \times n$ matrix A, the Frobenius norm of A is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}^2} \qquad (1)$$

where the elements of A, $a_{i,j} \in \mathbb{R}$, $\forall\ i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. As this is an extension of the Euclidean norm for vectors, the Frobenius norm is often thought as the ‘distance’ of the matrix from an arbitrary origin. This simple measure could be used to access the similarity between two matrices based on their distances from a common origin. Given the multivariate time series A and B for two experimental conditions, the similarity between the responses can be measured by $\|A - B\|_F$ where 0 would mean the two responses are identical. This measure does not take into account direction, rather only the distance. Also, there is no upper bound to this measure and thus it is difficult to prescribe a tolerance level of similarity. One way to overcome these problems is to represent each MVTS as a vector and evaluate the cosine of the acute angle between two vectors, bounded by 0° corresponding to perfect similarity and 90° representing perfect dissimilarity. Principal component analysis (PCA) is one such method of representing a MVTS as vectors.

3.2. Principal Component Analysis Similarity Factor (S$_{\text{PCA}}$)

Given the k principal components ($k \geq 1$) of two MVTS A and B, the PCA similarity factor between A and B is

$$S_{PCA}(A,B) = \sum_{i=1}^{k}\sum_{j=1}^{k} \cos^2 \theta_{i,j} \qquad (2)$$

where $\theta_{i,j}$ is the angle from the i^{th} principal component of A and the j^{th} principal component of B. If A and B are perfectly similar then $S_{PCA}(A,B) = k$. Alternatively if the components are pairwise orthogonal, i.e. completely dissimilar, $S_{PCA}(A,B) = 0$. A heuristic for a similarity boundary of 5% could be placed upon each pair of principal components and thus if $S_{PCA}(A,B) \geq 0.95 \times k$ then A and B are said to be *similar*.

Although this measure allows for an intuitive assessment of the similarity between two time series it does not account for the covariance between components of the individual MVTS nor the magnitude of the principal components.

3.3. Eros Similarity Measure

Eros is a weighted similarity measure which compares two MVTS through their principal components. Let A and B denote two MVTS both of size $m \times n$ and $V_A = (a_1, \dots, a_n)$ and $V_B = (b_1, \dots, b_n)$ be the two right eigenvector matrices obtained from single value decomposition (SVD) of the covariance matrices of A and B respectively. Formally, the Eros similarity of A and B is

$$Eros(A, B, w) = \sum_{i=1}^{n} w_i |\cos\theta_i | \qquad (3)$$

where $\sum_{i=1}^{n} w_i = 1$ and θ_i is the acute angle between a_i and b_i, $\forall i \in \{1, \dots, n\}$. This accounts for the variance of individual principal components by the choice of the weight. Yang and Shahabi suggest an algorithm for computing w, the weight vector [12]. It should be noted that it is not necessary to compare two MVTS with the same number of rows, however in the case of MEG data it is most sensible to exclude the same channels in the analysis of all data; this is why A and B have been defined to be of equal dimension. Furthermore, it is not necessary for A and B to be compared over the same time interval but when comparing neural responses it is unnecessary to consider otherwise. The interpretation of the result $Eros(A, B, w)$, where $0 \leq Eros(A, B, w) \leq 1$, follows that A and B are similar when $Eros(A, B, w) = 1$ and dissimilar when $Eros(A, B, w) = 0$. Eros has been applied to the real world situations of identifying a person's gait from camera footage and to analyse sign language based on sensor data from the hands of different people signing [12]. In this setting Eros out performed S_{PCA} in a precision-recall assessment as well as being less computationally intensive.

4. Experimental Approach

MEG data were recorded from a female subject 27 years old, right handed with normal hearing using a 306-channel system (Elekta Neuromag Triux) at a sampling rate of 1000 Hz with a low pass filter of 330Hz and high pass filter of 0.03Hz. The subject listened to five distinct pure tones (250Hz, 500Hz, 750Hz, 1000Hz, and 1500Hz) which were randomly presented binaurally and at equal volume. It should be noted that the subject found 250Hz challenging to hear in comparison to the other pitches due to its low tonal quality. Tones were played every second with each tone lasting 0.2secs. The subject was instructed to fixate on a cross on a screen in front of them. After every fourth tone this fixation point changed colour signifying the moment the subject was to blink and press a button. By providing this opportunity for the subject to blink, muscular blink artefacts were suppressed from the stimulus response. The experiment was run three times with each tone presented 120 times.

Data were preprocessed using the temporal extension of signal space separation (tSSS) within MaxFilter software (Elekta Neuromag Oy). Epochs were extracted from the filtered data 100ms before stimulus onset to 700ms after stimulus onset ensuring there was no overlap between two different conditions. All occurrences of eye blinks and button presses were removed before the analysis process. Initial analysis used Brainstorm [13] to calculate an average (mean) response to each condition or tone. The matrix representation for the average response to condition i, given by $F_i = [n \text{ channels} \times t \text{ time points}] = [306 \times 800]$, was extracted from Brainstorm for similarity analysis in Matlab (Matlab R2013a V8.1) using the Frobenius norm, S_{PCA} and Eros. The Frobenius norm was calculated for the difference between the pair F_i and $F_j, \forall\, i, j \in \{1, \dots, 5\}$, that is $\|F_i - F_j\|_F$. The results of the Frobenius norm similarity were normalised with respect to largest distance between all pairs.

5. Results

The similarity between the neural responses for each pair of conditions is presented in a similarity matrix; shown in Tables I-III. For example, the element in row 2 and column 3 of the similarity matrices corresponds to the similarity between the neural responses to 500Hz and 750Hz respectively. Table I shows that the normalised Frobenius norm finds no two neural responses to be similar as the minimum 'distance' between any pair is 79% of the maximum distance. Eros and S_{PCA} (Table II and Table III respectively) both describe the data for 1000Hz and 1500Hz to be similar (>95% similarity). Under this same criterion Eros considers the response to three other pairs of distinct tones to be similar (500Hz and 1000Hz, 750Hz and 1000Hz, 500Hz and 1500Hz). There is an anomalous result in the Eros data of no similarity between the neural responses to 750Hz and 1500Hz despite similarity occurring between the neighboring conditions. Even though similarity is shown to two decimal places, the result is not in keeping with the surrounding information. Eros and S_{PCA} define the relative strength of similarity between pairs in the same way except for the pairs 500Hz and 1000Hz, 750Hz and 1000Hz where Eros found that the response to 500Hz is more similar to that of 1000Hz than the response of 750Hz is to 1000Hz. All three measures found the response to 250Hz to be dissimilar to all other tones. The weaker similarity to other responses can be explained by the subject's difficulty to hear the associated tone during the recording.

The results show that the Frobenius norm finds dissimilarity between the neural responses to different tones. This could be beneficial in classification algorithms with the purposes of differentiating tones through the neural response. However, this measure appears to be poor at finding relations between the neural responses to tones which could be related, for example closer tones in pitch do not show similar neural responses. Eros and S_{PCA} do elude to this as demonstrated with 1000Hz and 1500Hz being similar while the responses to 250Hz and 1500Hz are dissimilar. The anomalous result within the Eros results suggest that Eros could find too many similarities between neural responses which could be misleading.

Table I: Similarity matrix for the normalised Frobenius norm

Freq(Hz)	250	500	750	1000	1500
250	0	1.0000	0.9158	0.9406	0.9560
500	1.000	0	0.9230	0.8817	0.8781
750	0.9158	0.9230	0	0.7867	0.8512
1000	0.9406	0.8817	0.7867	0	0.8234
1500	0.9560	0.8781	0.8512	0.8234	0

Similarity matrices

Table I: The normalised Frobenius norm. Data has been normalised with respect to the largest value 0.4759×10^{-9} representing the similarity between 250Hz and 500Hz.

Table II: Similarity matrix for S_{PCA}

Freq(Hz)	250	500	750	1000	1500
250	1	0.7993	0.7506	0.7707	0.7912
500	0.7993	1	0.8604	0.9107	0.9379
750	0.7506	0.8604	1	0.9279	0.8951
1000	0.7707	0.9107	0.9279	1	0.9571
1500	0.7912	0.9379	0.8951	0.9571	1

Table II: S_{PCA} indicates the largest similarity lies between 1000Hz and 1500Hz.

Table III: Similarity matrix for Eros

Freq(Hz)	250	500	750	1000	1500
250	1	0.9083	0.8656	0.8741	0.8882
500	0.9083	1	0.9447	0.9699	0.9774
750	0.8656	0.9447	1	0.9667	0.9467
1000	0.8741	0.9699	0.9667	1	0.9823
1500	0.8882	0.9774	0.9467	0.9823	1

Table III: Eros determines 4 pairs of distinct tones as having similar neural responses.

6. Discussion

The analysis of the multivariate data obtained from MEG recordings has not been exploited to date. By using MVA on MEG data, one obtains more authentic conclusions as less assumptions have been made during the analysis process in comparison to results obtained when MEG is a constituent of a multimodal approach. The analysis techniques represent a spatiotemporal review of the data. The channels which depict the location of detected activity are represented by the rows of the MVTS matrix while the columns represent the temporal information. One advantage when comparing the neural activity to different experimental conditions is the ability to account for the spatial and temporal information across the entire time window. Alternatively by isolating specific rows or channels the analysis could be made across the same region, here we included all channels in the discussion. Although the largest amplitude of neural activity for an auditory response occurs around 100ms, when investigating speech the information after this peak provides more interesting data than what appears in simple auditory tones [14]. By using MVA on the entire time window this later activity is taken in to account.

The interpretation of the data presented is limited by the single subject and the number of tones played during the experiment. The different pitches cover a small spectrum of the human auditory range. More subjects would be needed to conclusively explain the results. The lack of similarity between the response to 250Hz and all other tones, as seen across the three measures, is easily illustrated with the topography of the response 100ms after stimulus onset (Figure 1). The strength of the response 250Hz is up to 1/4 lower than that of all other responses. This illustrates that similarity measures can only go so far as to describe the relationship between two neural responses. These measures could be used in classification algorithms, as illustrated by Yang and Shahabi [12] with k-nearest neighbors. Cichy et al. [15] used a support vector machine (SVM) to classify 92 images based on the visual response obtained from MEG recordings. Through this MVA technique they revealed the dynamics of the neural processing of objects at various levels of categorization. More research is required into the best classification technique for MEG data, such as the commonly used k-means, SVM or Ward’s method. MVA has already been applied to MEG analysis in the form of artefact separation [16] and source localization [17].

Future work is proposed to use classification techniques to provide a novel approach to tinnitus research. Tinnitus is the phantom perception of a sound in the absence of an identifiable source. It is thought that after the onset of tinnitus, there are plastic changes in the auditory system which alters the cortical organisation of frequencies. However, a definitive answer has been elusive [18, 19]. If there is a microscopic shift in the cortical organisation then the neural responses to frequencies perceived by people with tinnitus then the classifier for people without tinnitus should not classify the neural responses of the tinnitus group correctly; and vice versa.

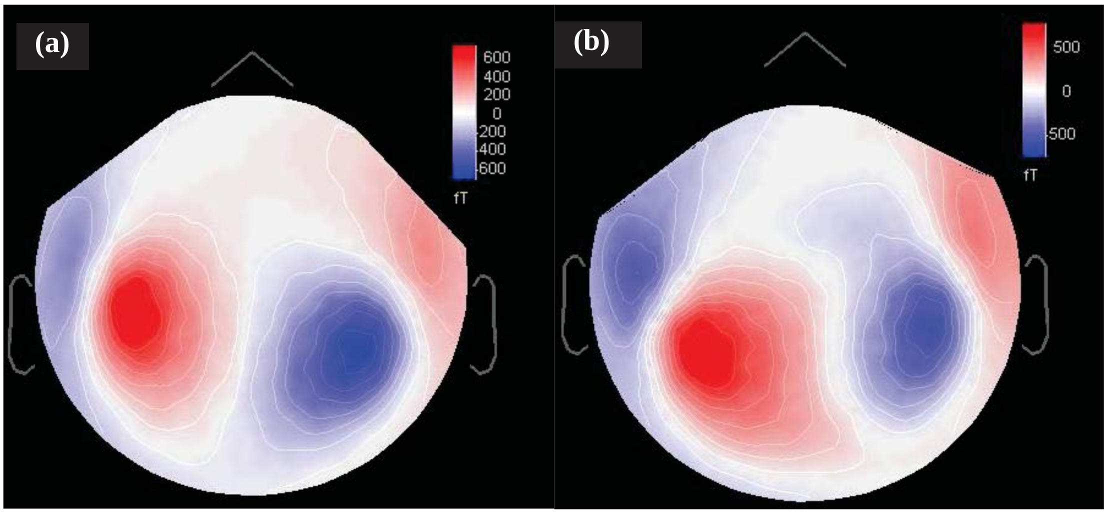


Figure 1: Topography of the neural response for (a) 250Hz and (b) 1500Hz. Both represent 100ms after stimulus onset with a clear difference in amplitude of the responses. Both illustrate clear activity over the auditory cortices.

To conclude, this paper illustrates how MEG data, by its multivariate nature, is ideal for application of MVA techniques. With only one real clinical application of MEG appearing in the form of Epilepsy treatment, more research is required to extract information from MEG data that can clearly and effectively distinguish subtleties in neural responses and provide a composite mapping between experimental conditions, neural response and computational modelling. Kriegeskorte et al. [20] outline the need to unite neural imaging data, behavioral measurements and mathematical modelling quantitatively through dissimilarity measures, or analogously similarity measures. This study for fMRI could be translated to MEG research and emphasizes the need to relate the similarity measures and classification techniques back to the experimental conditions.

7. Acknowledgements

We would like to thank Dr. Dimitrios Pantazis and his colleagues of the MEG Lab at the Martino's Imaging Center (Massachusetts Institute of Technology) for their invaluable advice and assistance and for providing the data sets.

References

[1] V. Walsh and A. Cowey, "Transcranial magnetic stimulation and cognitive neuroscience," *Nature Reviews Neuroscience,* no. 1, pp. 73-80, 2000.

[2] H. Schiffbauer, M. S. Berger, P. Ferrari, D. Freudenstein, H. A. Rowley and T. P. Roberts, "Preoperative magnetic source imaging for brain tumor surgery: a quantitative comparison with intraoperative sensory and motor mapping," *Journal of Neurosurgery,* vol. 97, no. 6, pp. 1333-1342, 2002.

[3] G. L. Barkley and C. Baumgartner, "MEG and EEG in Epilepsy," *Journal of Clinical Neurophysiology,* vol. 20, no. 3, pp. 163-178, 2003.

[4] E. Zamrini, F. Maestu, E. Pekkonen, M. Funke, J. Makela, M. Riley, R. Bajo, G. Sudre, A. Fernandez, N. Castellanos, F. del Pozo, B. W. van Dijk, A. Bagic and J. T. Becker, "Magnetoencephalography as a Putative Biomarker for Alzheimer's Disease," *International Journal of Alzheimer's Disease* , vol. 2011, p. 10, 2011.

[5] S. Braeutigam, S. Swithenby and A. Bailey, "Magnetoencephalography (MEG) as a tool to investigate the neurophysiology of autism," in *Researching the Autism Spectrum: Contemporary Perspectives,* Cambridge, Cambridge University Press, 2011, pp. 156-175.

[6] L. B. Hinkley, J. P. Owen, M. Fisher, A. M. Findlay, S. Vinogradov and S. S. Nagarajan, “Cognitive impairments in Schizophrenia as Assessed Through Activation and Connectivity Measures of Magnetoencephalography (MEG) Data,” *Frontiers in Human Neuroscience,* vol. 3, no. 73, 2009.

[7] N. Weisz, S. Moratti, M. Meinzer, K. Dohrmann and T. Elbert, “Tinnitus Perception and Distress Is Related to Abnormal Spontaneous Brain Activity as Measured by Magnetoencephalography,” *PLOS Medicine,* vol. 10, no. 1371, 2005.

[8] H.-P. Müller and K. Jan, Multimodal Imaging in Neurology: Spectral Focus on MRI Application and MEG, USA: Morgan and Claypool, 2008.

[9] S. Taulu and R. Hari, “Removal of Magnetoencephalographic Artifacts with Temporal Signal-Space Speration: Demonstration with Single-Trial Auditory-Evoked Responses,” *Human Brain Mapping,* vol. 30, no. 5, pp. 1524-1534, 2009.

[10] F. Tadel, S. Baillet, J. C. Moscher, D. Pantazis and R. M. Leahy, “Brainstorm: A User-Friendly Application for MEG/EEG Analysis,” *Computational Intelligence and Neuroscience,* vol. 2011, no. 2011, p. 13, 2011.

[11] S. Baillet, “The Dowser in the Fields: Searching for MEG Sources,” in *MEG: An Introduction to Methods*, Oxford, Oxford university Press, 2010, pp. 83-123.

[12] K. Yang and C. Shahabi, “A PCA-based Similarity Measure for Multivariate Time Series,” *Proceedings of the 2nd ACM international workshop on Multimedia databases.,* vol. 2004, pp. 65-74, 2004.

[13] F. Tadel, S. Baillet, J. C. Moscher, D. Pantazis and R. Leahy, “Brainstorm: A User-Friendly Application for MEG/EEG Analysis,” *Computational Intelligence and Neuroscience,* vol. 2011, p. 13, 2011.

[14] A. C. Papanicolaou, E. Castillo, J. I. Breier, R. N. Davis, P. G. Simos and R. L. Diehl, “Differential brain activation patterns during perception of voice and tone onset time series:a MEG study,” *NeuroImage,* vol. 18, no. 2, pp. 448-459, 2003.

[15] R. M. Cichy, D. Pantazis and A. Oliva, “Resolving human object recognition in space and time,” *Nature Neuroscience,* vol. 17, pp. 455-462, 2014.

[16] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari and E. Oja, “Independent Component Analysis for identification of artifacts in Magnetoencephalographic recordings,” *Advances in neural information processing systems,* pp. 229-235, 1998.

[17] V. Litvak, A. Eusebio, A. Jha, R. Oostenveld, G. R. Barnes, W. D. Penny, L. Zrinzo, M. I. Hariz, P. Limousin, K. J. Friston and P. Brown, “Optimized beamforming for simultaneous MEG and intracranial local field potential recordings in deep brain stimulation patients,” *NeuroImage,* vol. 50, no. 4-3, pp. 1578-1588, 2010.

[18] W. Mühlnickel , T. Elbert, E. Taub and H. Flor, “Reorganization of auditory cortex in tinnitus,” *Proceedings of the National Academy of Sciences, USA,* vol. 95, pp. 10340-1343, 1998.

[19] D. R. M. Langers, E. de Kleine and P. van Dijk, “Tinnitus does not require macroscopic tonotopic map reorganization,” *Frontiers in Systems Neurscience,* vol. 6, no. 2, 2012.

[20] N. Kriegeskorte, M. Mur and P. Bandettini, “Representational Similarity Analysis-Connecting th branches of Systems Neuroscience,” *Frontiers in Systems Neuroscience,* vol. 2, no. 4, p. , 2008.

An Architecture for Social Media Summarisation

Zbigniew Zdziarski, Joe Mitchell, Pierre Houdyer, Dave Johnson
Tapastreet Ltd, Ireland
{zbigniew, joe, pierre, dave}@tapastreet.com

Cyril Bourgès & Rozenn Dahyot
Trinity College Dublin, Ireland
{bourgesc, Rozenn.Dahyot}@tcd.ie

Abstract

Social media traffic and mobile usage is growing at an accelerating rate, and the amount of media that is being uploaded on social media sites (such as Twitter, Facebook and Instagram) is also increasing. The consortium GRAISearch aims at developing tools to merge, to visualise and to present this wealth of data in a comprehensive, compact and user-friendly way. This poster will present a work-in-progress architecture for such a purpose - to provide users with a central point of access to media from the largest social media sites.

Keywords: Social Media, Web Harvesting, Video Summarisation, Visual Saliency

The European Project GRAISearch (Use of Graphics Rendering and Artificial Intelligence for Improved Mobile Search Capabilities, `http://tapastreet.com/GRAISearch`, FP7-PEOPLE-2013-IAPP (612334), 2014-18) is a research collaboration between two universities (Trinity College Dublin, Ireland, and INSA Lyon, France) and the company Tapastreet Ltd (see mobile app at `http://tapastreet.com/`). It aims at providing enhanced visualisation tools for visual content available on social media and an architecture for social media summarisation. Tapastreet has a location based social media search engine platform that, in its current form, returns geo-located video and image media from major social networks for any location and any topic (#hashtags) anywhere in the world. The current platform deals well with images on social media but videos are yet not well tackled. Several challenges exist for videos on social platforms. First, they are too large to all be downloaded when browsing on mobile devices and therefore need to be summarised very efficiently. Second, media on social platforms consist mainly of very diverse amateur recordings with little or no editing rules that also contain many artefacts that alter their quality, such as low lights and motion shakiness when the recording device is hand-held.

Harvesting Social media. Using social network APIs, a Ruby script is used to download all media using a user-defined query (hashtags, GPS location), and links to these images and videos are stored along with their description in JSON format (keywords, GPS location, creation date, etc.) on our server. An image-processing pipeline is currently under development for automatically creating video summaries.

Video Summarisation. A lot of research has been devoted to creating video summarisations (a.k.a. video abstractions) [Truong and Venkatesh, 2007] and many algorithms have been proposed. For example, several authors [Zhang et al., 2003, Kim and Hwang, 2002] suggest processing a video sequentially and marking keyframes as those that are significantly different from previously extracted keyframes. A more computationally demanding method has been proposed by Gibson et al. [Gibson et al., 2002] and Yu et al. [Yu et al., 2004]. They employ a clustering technique where video frames are treated as points in a feature space (e.g. colour histogram) and representative points from each cluster are selected as keyframes of the video. The keyframes extracted by these methods can then be presented to the user as a summary of the video.

For the GRAISearch project, a real-time system is required, hence video summarisation needs to be very fast as well as informative (i.e. representative of the video) and very small in storage size (due to the limitations of mobile devices and wireless networks [Liu et al., 2014]). Several techniques are currently being tested using information theory (e.g. measure of entropy) to select the most diverse frames in a stream, and visual saliency algorithms [Hou and Zhang, 2007] to assist in detecting salient regions in frames and hence improve the keyframe extraction process. A further extension to this project is to develop video summaries suitable for 3D screens and for this 3D visual saliency algorithms [Zdziarski and Dahyot, 2014] will also be investigated.

Examples of scenarios. Tapastreet Ltd currently have a number of outside bodies actively using their app. The Danish Football Association uses the app as a fan engagement tool. Whenever the Danish football team plays, the association channels all images onto their website that were taken at a particular stadium (specified by GPS location) and/or tagged with appropriate hashtags for users to view in real-time. The MET office in the UK uses the app to measure the impact of weather events on human activity. When it knows that a significant weather event is imminent for a particular area, it will mine all media from that area before, during and after the event, allowing the study of the impact of climate change on human behaviour.

References

[Gibson et al., 2002] Gibson, D., Campbell, N., and Thomas, B. (2002). Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 814–817 vol.2.

[Hou and Zhang, 2007] Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

[Kim and Hwang, 2002] Kim, C. and Hwang, J.-N. (2002). Object-based video abstraction for video surveillance systems. *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(12):1128–1138.

[Liu et al., 2014] Liu, Y., Wang, S., and Dey, S. (2014). Content-aware modeling and enhancing user experience in cloud mobile rendering and streaming. *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, 4(1):43–56.

[Truong and Venkatesh, 2007] Truong, B. T. and Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1).

[Yu et al., 2004] Yu, X.-D., Wang, L., Tian, Q., and Xue, P. (2004). Multilevel video representation with application to keyframe extraction. In *Multimedia Modelling Conference, 2004. Proceedings. 10th International*, pages 117–123.

[Zdziarski and Dahyot, 2014] Zdziarski, Z. and Dahyot, R. (2014). Extension of gbvs to 3d media. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pages 2296–2300.

[Zhang et al., 2003] Zhang, X.-D., Liu, T.-Y., Lo, K.-T., and Feng, J. (2003). Dynamic selection and effective compression of key frames for video abstraction. *Pattern Recognition Letters*, 24(9-10):1523 – 1532.

Broadcast Language Identification System (BLIS)

James Connolly[1,2], Kevin Curran[2], Paul Mc Kevitt[1], John Macrae[3], Stephen Craig[4]
[1]School of Arts and Creative Technologies, University of Ulster, Magee
[2]Intelligent Systems Research Centre (ISRC), University of Ulster, Magee
[3]Office of Innovation, University of Ulster, Jordanstown
[4]Maiden Technology Ltd., Larne, Co. Antrim
{jp.connolly, kj.curran, p.mckevitt, j.macrae}@ulster.ac.uk
stephen@maiden.org.uk

Abstract

Virtually all TV transmission systems use highly automated file-based broadcast systems for audio and video. Content management systems automatically deliver programmes ready for broadcast by matching content to a schedule and deliver all ancillary services in their correct format and on time. Within these sub-systems is a growing need to validate that the correct language is delivered to a particular service and/or region. In many cases, a single instance of a programme exists and the automated system merely selects the correct language for a particular service. This process is currently managed manually by operators listening to the audio of each programme and confirming that the accompanying language is correct for its video broadcast. Incorrect language transmission can be caused by system faults or errors in the scheduling workflow. An error can occur at numerous points during the broadcast. The Broadcast Language Identification System (BLIS) will provide a single operator with the ability to monitor multiple services by "dash-boarding" language flags from each service and enable the operator to intervene if an error is detected. BLIS will examine streaming audio from a pre-broadcast to identify spoken language within the broadcast content and compare it with the expected language of the video for broadcast.

Keywords: Audio, Language Identification, Automatic Speech Recognition, Television Broadcast, Broadcast Language Identification System (BLIS)

1. Introduction

In a file-based modern highly-automated transmission environment, unintentional errors can produce mismatches between transmitted video and audio, resulting in a reduction in broadcast Quality of Service (QoS). The common broadcasting technical standards document [1] agreed by the BBC, BskyB, Channel 4, Channel 5, ITV and S4C includes details on video and audio production formats. All audio is encoded within the Pulse Code Modulation (PCM) standard [2] and must have an audio sampling frequency of 48 kHz, 24 bit audio depth. Although such documentation provides audio structure standards, there are several stages within audio integration and delivery processes where errors can occur. Programmes can be broadcast in standard definition (SD) or High Definition (HD) and may be delivered for broadcast in file or tape format. Audio track layout and allocations differ for both platforms and can exist in 4 or 16 track layout. Programmes that contain single language tracks, or are in SD format use 4 track audio. The first 2 tracks of SD and HD formats contain left and right final mix sound. Third and fourth tracks may contain music and effects (M&E), audio description or digital silence. With HD audio, additional tracks provide 5.1 surround sound. Remaining tracks may contain 2 or 3 additional languages. Additional audio tracks can be independently delivered in Broadcast Wave Format (BWAV) [3]. For tape broadcast format, a supplementary language can be allocated to the third and fourth track, or to track 11 and 12 for HD audio. Remaining tracks contain surround sound and M&E. Furthermore, standards for live broadcasts differ between broadcasters, although all are working towards a standardised audio layout.

Metadata contains all information relevant to a file or tape broadcast. It ensures all video and audio content is correctly reconstructed for playback or for various system conversions. Structural metadata is manually added by a broadcast producer and includes title and ID number for the programme and structure of associated audio tracks. Errors can be introduced to the system at several stages. Audio track layout may be accidently mismatched to the definition standard of the accompanying programme. Incorrect use of file, tape or live audio format can cause misalignment between intended and expected broadcast language. Metadata information can contain incorrect reference to primary, secondary and tertiary audio language channels. Section 2 discusses our proposed solution to these problems using BLIS and section 3 concludes with future work.

2. Broadcast Language Identification System (BLIS)

An operator manually examines audio for each broadcast to ensure both are correctly matched. Typically an operator is responsible for correct identification of 8 simultaneous broadcast channels. Broadcast Language Identification System (BLIS) will provide a single operator with the ability to monitor multiple services by “dash-boarding” language flags from each service and enable the operator to intervene if an error is detected. BLIS will examine streaming audio from pre-broadcasts to identify spoken language content and compare it with the expected language of the video for broadcast. BLIS will exist as two units. The first unit will provide front-end functionality to the local broadcast operator through dash-boarding software as shown in Fig. 1(a). The dash-board will deliver error feedback to the operator if a mismatch between pre-broadcast audio and video is identified, and an instant view on the status of programmes throughout each broadcast. Training for new language dictionaries will improve language detection and accuracy. Software functionality may be integrated with existing broadcast software systems that have the capability to call third party functions that exist on the broadcast network. These functions will reside on a broadcast network server.

The second unit will provide back-office functionality to the dash-boarding software as shown in Fig. 1(b). It will reside on the broadcast network on a Windows Server system (Fig. 1(c)). This server may exist within the broadcast network, or function as a separate sub-domain. Communication using REST technology will respond to language query services of the local software system to identify spoken language of a pre-broadcast. The server-based system will interact with cloud-based language models if the local system cannot uniquely identify the spoken language of a pre-broadcast audio stream. The cloud based language model will update locally stored language models.

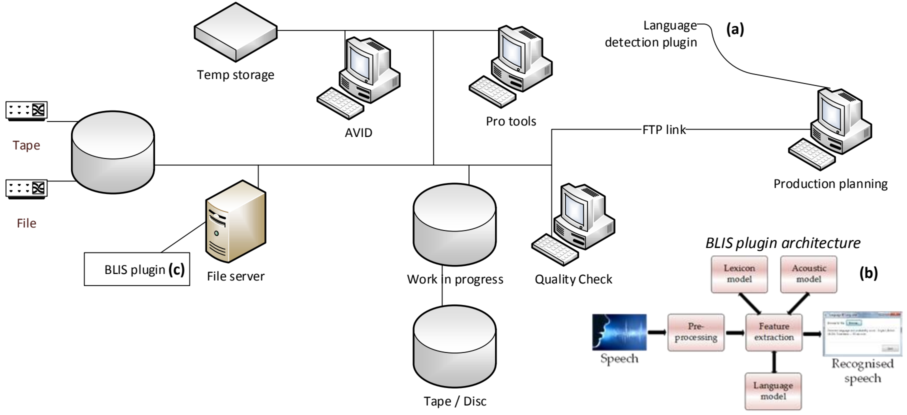


Figure 1: Architecture of typical broadcast system.

3. Conclusion & future work

TV transmission networks use highly automated file-based broadcast systems for audio and video. Human error can unintentionally introduce audio and video mismatches to the system. Manual techniques are currently used for problem identification. BLIS is an automatic language identification system that replaces manual intervention with dash-boarding software. Multiple channels can be automatically monitored and the broadcast operator can be immediately notified of potential problems.

Acknowledgements

BLIS is part financed by the European Regional Development Fund (ERDF) under the EU Sustainable Competitiveness Programme for Northern Ireland 2007-2013.

References

[1] BBC, “Technical Standards for delivery of television programmes to BBC.” pp. 1–41, 2012.
[2] W. M. Waggener, *Pulse Code Modulation Techniques*, 1st ed. New York: Thompson Publishing Inc, 1995.
[3] EBU - UER, “Specification of the Broadcast Wave Format (BWF).” Geneva, pp. 1–20, 2011.

Full body gender Recognition - Feature Combination via a Two Stage SVM approach

Matthew Collins, Paul Miller and Jianguo Zhang
Centre for Secure Information Technologies
Queen's University Belfast,
Northern Ireland Science Park, Queen's Road, Queen's Island, Belfast, BT3 9DT
m.collins, p.miller@qub.ac.uk
jgzhang@computing.dundee.ac.uk

Abstract

A relatively simple, computationally inexpensive 2-Stage SVM based approach to feature combination is applied to the problem of gender recognition of pedestrian images. As expected, performance is boosted over single features alone. Examination of the weightings assigned by the SVM to combined features gives insight into the most influential features in the classification problem.

Keywords: Full Body Gender Classification, Feature Combination, Support Vector Machines

1 Background

Full body gender recognition is a difficult problem due to the high degree of variation in subject appearance. Previous work has established that some feature types will play a much higher role in the recognition of some objects, while other less informative features should be ignored or at least assigned a lower weighting when features are combined. Automatic feature selection techniques and methods of effective feature combination have been greatly focused on in much of the recent research in object classification.

In this paper, we explore the use of a 2-Stage SVM feature combination approach of [Zhang et al., 2007] and apply it to a selection of key feature types shown to be effective for full body gender classification.[Collins et al., 2009, Collins et al., 2010] We examined the weightings applied to each feature vector in comparing combinations and find that consistently, the features which would be expected to perform better individually are prioritised by the learning method and assigned the highest weights.

2 Gender Recognition System

The features used attempt to capture aspects of body shape, appearance. For shape these are Canny Histogram of Gradients (CHOG) and PiHOG. For appearance we use local HSV Colour Histograms to exploit the fact that males tend to wear darker clothing than females. PCA based features were also calculated and ranked according to their encoding power for the gender classification problem using Linear Discriminate Analysis. The top seven identified gender components were concatenated to form a feature vector. [Collins et al., 2010]

Lastly a variation on the the Gist features of [Oliva and Torralba, 2001] was also explored. We experimented with a number of variations of the Gist feature vector applying different levels of spatial constraints to enhance their descriptive power for the task at hand.

For the first stage of the classification, SVM classification was performed for each feature type individually. The outputs from these SVMs where then concatenated to produce new feature vectors and second stage SVM classification was performed on these.

3 Results and Conclusions

The individual standalone classification scores for each feature from the first stage of the 2-Stage SVM showed that PiHOG and the variants of Gist with the stronger spacial constraints applied, had the highest descriptive power. Classification accuracies in the region of 75% were recorded for the dataset of 826 pedestrian images (413 of each gender).

The dataset itself was constructed by merging two pedestrian recognition datasets and uniformly scaling the images. It was noted that Gist based features tended to perform better on one of these source datasets than the other, while the PiHOG feature was consistent across the board.

Various combinations of the features were explored, from a simple pairing on HOG based features and LHSV colour features, to a overall combination of all available features. In all cases the classification accuracy for a combination was higher than any one feature alone. The highest recorded accuracy using the 2-Stage SVM approach was 84.05% for a combination of PiHOG, LHSV, PCA, and 3 variants of Gist.

Furthermore, upon analysing the weightings assigned to the features involved in the combinations, consistently it was the ones which were intuitively more descriptive which were assigned the highest weightings. Complimentary features were higher weighted and those features which didnt contribute. In combinations containing many similar features such as multiple variations of the Gist feature, the variations which intuitively describe the image better are assigned significantly higher weights than their simpler counterparts, which were assigned very low to negligible weightings. For the 2-Stage SVM it was consisently the PiHOG feature which was assigned the overall highest weighting.

Looking at classification accuracies across multiple combinations of feature types, it is clear that while more features will often improve the overall classification accuracy, sometimes, the accuracy may in fact fall from what it would have been if the feature had not been included at all. It is not sufficient to rely on the automatic weight learning framework to assign a zero weight to uncomplimentary features. Generally the weights assigned to the features by automatic learning methods are a good indication of their strength for the task, but should not be relied on as the overall decider. Ultimately the weights assigned to features should be used as a guide in fine tuning the selection of the optimal complementary features for a particular classification task.

References

[Collins et al., 2009] Collins, M., Zhang, J., Miller, P., and Wang, H. (2009). Full body image feature representations for gender profiling. In *IEEE Workshop on Visual Surveillance (ICCV).*

[Collins et al., 2010] Collins, M., Zhang, J., Miller, P., Wang, H., and Zhou, H. (2010). Eigenbody: Analysis of body shape for gender from noisy images. In *IMVIP*, University of Limerick. Cambridge Publications.

[Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175.

[Zhang et al., 2007] Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238.

Investigation into the Automated Detection of Image based Cyber bullying on Social Media Platforms

Dr. Gaye Lightbody, Dr. Raymond Bond, Prof. Maurice Mulvenna, Dr. Yaxin Bi
School of Computing and Mathematics,
University of Ulster, Jordanstown Campus, Northern Ireland, BT37 0QB
g.lightbody, rb.bond, md.mulvenna, y.bi@ulster.ac.uk

Macartan Mulligan
Treze Technology
Carnbane Business Centre, Newry, Northern Ireland, BT35 6QH
macartan@treze.co.uk

Abstract

Access to the Web 2.0 technologies and social networking has become a ubiquitous factor of our lifestyles. Mobile devices, smart phones and other such technologies create a permanent link between the person and their presence on the Internet. Reporting daily activities, feelings, opinions and emotions have become common place, particularly with teenagers and young adults. Circles of friendships can expand rapidly without careful consideration to what is being shared and to whom. Such an environment has become an avenue for cyber bullying, enabling persecutors to breach the safety of the home environment to reach their victims. In order for social networking to be a safe technology in the lives of adolescents and vulnerable adults, measures need to be put in place to monitor and detect potentially threatening online activity. In this paper an overview is given of the possible techniques that could be employed in the detection of negative online interactions. The combination of sentiment analysis with image processing techniques is considered as a suitable platform from which to categorize the textual and visual connotations of the content. This is illustrated as a flow diagram illustrating the key processes envisaged for the detection of potential cyber bullying threats.

Keywords: Image analysis, Sentiment analysis, Cyber bullying, Social networking.

1 Introduction

The detection of negative material within social media sites often centers on the textual content of the communication providing sentiment analysis by looking at key phrases and context. A complexity that arises in the detection of cyber bullying, is that in contrast to spam which is generic and not targeted, the attack is "more personal, varied and contextual" [1]. Dinakar et al. [1] investigate an approach that combines natural language processing with machine learning algorithms and a common sense knowledge base, which provides coverage over a range of situations. They break down the context of the attack as being based on sexuality, race, culture, intelligence, physical attribute etc., and based on these categories they label and define verbal communications, performing machine learning algorithms to classify intent. Sentiment analysis on the textual information is a key aspect to detecting cyber bullying but often images are attached to social media posts (or tweets) and these too require consideration as they open up another dimension as to how a victim may be targeted. Violation could incur through the spread of generic obscene and offensive images or the distribution of images of the victim that may have been tampered or may depict inappropriate behavior. In addition to using sentiment analysis to isolate certain images a further stage can be included to investigate image content through image processing techniques. Examples of features that might signify negativity include level of nudity, evidence of edits within the image and analysis of text within the image.

Vanhove et al. [2] present a platform for automated detection of cyber bullying on social media networks. Their proposed architecture is a modular extendible framework, which can be tailored to a range of scenarios. In their paper they demonstrate two example scenarios, one for the detection of suicidal tendencies, whereby nudity detection and self-harm features, were combined in parallel with textual content analysis, and transgressive sexual behavior, whereby image and video analysis for nudity were combined in parallel with sexual language detection. In this abstract a proposal is made for a similar architecture with a further examples on image analysis.

2 Methods

Figure 1 illustrates a proposed flow of analysis. In particular, textual content associated with the images can be used to help determine not only the risk of negative content present but also to which category it might relate to. For example, text consisting of sexual content would signify a higher risk that the associated image contains nudity [3], thereby analysis on skin tone can give greater confidence to this classification.. Kazemier and Heijkoop [4] provide a review of digital forensic mechanisms such as the analysis of quantization tables used in a JPEG image which can signify the likelihood that the image has been saved from Photoshop. They also discuss techniques such as Principle Component Analysis to specify regions in which an image has been manipulated. Optical character recognition on images and videos [5] can also be used to extract messages allowing for sentiment analysis to be extended to the images themselves. It may not be necessary to fully categorize the negative content but to instead use the gathered information to send an alert (via MMS or Email) to a guardian once a threshold has been met. The person alerted can then visually review the image and judge the negative impact it may have. This can streamline the processing required by the system.

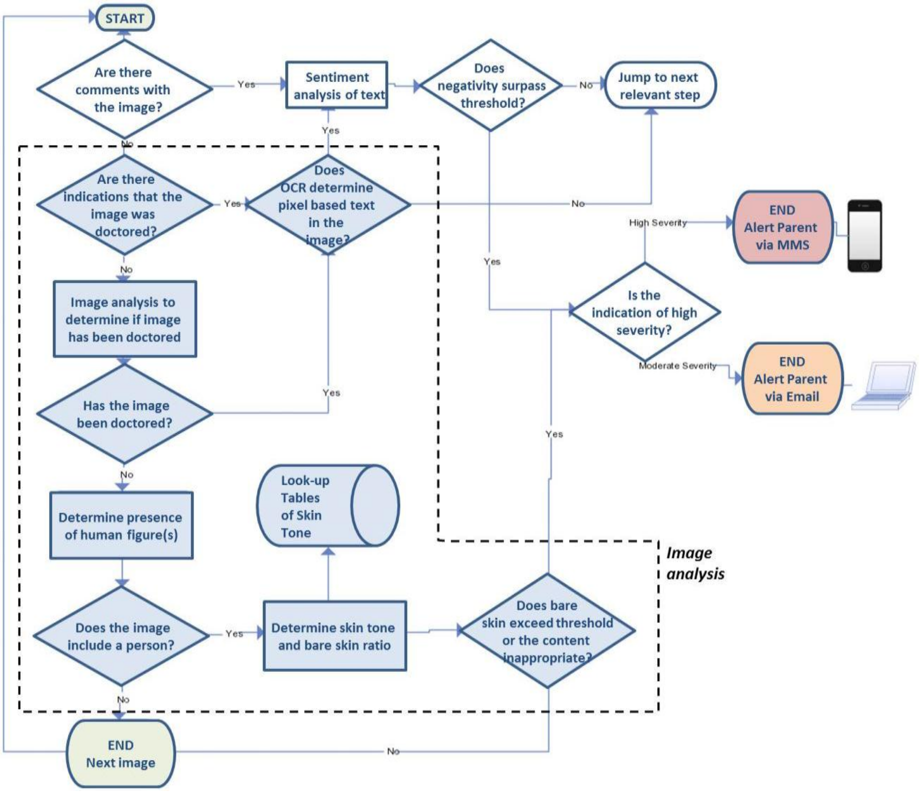

START
Are there comments with the image?
Yes
Sentiment analysis of text
Does negativity surpass threshold?
No
Jump to next relevant step
Yes
No
Are there indications that the image was doctored?
Yes
Does OCR determine pixel based text in the image?
No
High Severity
END Alert Parent via MMS
No
Yes
Image analysis to determine if image has been doctored
Is the indication of high severity?
Moderate Severity
END Alert Parent via Email
Has the image been doctored?
Yes
Yes
No
Look-up Tables of Skin Tone
Determine presence of human figure(s)
Image analysis
Does the image include a person?
Yes
Determine skin tone and bare skin ratio
Does bare skin exceed threshold or the content inappropriate?
No
END Next image
No

Figure 1: Flow chart for determining the risk of negative content in an image and associated message on a social media platform.

3 Discussion

A high-level overview of a monitoring system for the automatic detection of negative content on social media platforms has been presented which focusses on the analysis of images as a useful contributing measure. There is a growing acceptance that intervention and monitoring methods are required for online social networks particularly when in the use by adolescents or vulnerable adults [6], thus the inclusion of such a monitoring system governed by guardians is expected to be a popular ‘add-on’ feature in the near future that will complement existing anti-bullying campaigns (www.nobullying.com/).

References

[1] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, “Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying,” *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–30, Sep. 2012.

[2] T. Vanhove, P. Leroux, W. Tim, and D. T. Filip, “Towards the design of a platform for abuse detection in OSNs using multimedial data analysis,” in *IFIP/IEEE IM2013Workshop: 5th InternationalWorkshop onManagement of the Future Internet (ManFI)*, 2013, pp. 1195–1198.

[3] P. Yogarajah, J. Condell, K. Curran, P. McKevitt, and A. Cheddad, “A dynamic threshold approach for skin tone detection in colour images,” *Int. J. Biom.*, vol. 4, no. 1, p. 38, 2012.

[4] J. Kazemier and M. Heijkoop, “Digital image forensics.,” *Sci. Am.*, vol. 298, no. 6, pp. 66–71, 2008.

[5] K. Jung, K. In Kim, and A. K. Jain, “Text information extraction in images and video: a survey,” *Pattern Recognit.*, vol. 37, no. 5, pp. 977–997, May 2004.

[6] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, “Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability,” *Telemat. Informatics*, Apr. 2014.

Appearance-based SLAM using priors over network structure and trajectory complexity

Padraig Corcoran, John Leonard
Computer Science and Artificial Intelligence Laboratory Massachusetts
Institute of Technology
The Stata Center, Building 32 32 Vassar Street Cambridge, MA 02139.
padraigc@mit.edu

Abstract

This paper proposes a novel appearance-based Simultaneous Localization and Mapping (SLAM) method which integrates priors over network structure complexity and one's trajectory through this network. A quantitative evaluation relative to an existing state-of-the-art method was performed and the corresponding results achieved were very positive.

Keywords: SLAM, Appearance, Network Complexity Prior.

1 Introduction

Simultaneous Localization and Mapping (SLAM) is a fundamental problem in the field of robotics which concerns mapping an environment while simultaneously localizing within this map (Leonard and Durrant-Whyte, 1991). Much of the initial research into this problem focused on solutions which construct maps containing rich metric information (Kaess et al., 2008). Such methods are commonly referred to as *metric* SLAM methods. Recently there has been significant interest in the development *appearance-based* SLAM methods which use appearance or visual information to detect loop closures and in turn use this information to solve the SLAM problem. In this context a loop closure corresponds to one returning to a previously visited location (Cummins and Newman, 2008). Since no metric information regarding the environment is estimated, the resulting maps only represent topological properties such as connectivity between locations. In this paper we present a novel appearance-based SLAM method.

Loop closure detection is subject to the following two types of errors. The first type corresponds to a failure to detect a loop closure and may be caused by a change in a locations visual appearance. The consequence of such an error is the incorrect addition of a new location to the map. The second type of error corresponds to detecting an incorrect loop closure and may be caused by perceptual aliasing. The consequence of such an error is an incorrect localization and/or incorrect inference of a path between locations. Given the adverse consequences of loop closure detection errors, much research has been invested in the development of methods for performing robust inference over these detections (Carlone et al., 2014).

In this work we propose a novel method for performing inference over such detections. This method exploits the fact that SLAM methods are regularly applied in corridor or network type environments and consequently strong priors may be placed on the topological and metric structure of, and one's trajectory through, such environments.

2 Methodology and Results

The map representation used consists of a set of vertices and edges where each edge consists of a sequence of discrete locations along that edge. Each of these locations is represented by appearance information for an individual image and specifically a bag-of-words represen-tation. One's location within this map is specified by an edge and location along that edge. Simultaneously estimating the above map and one's locations corresponds to an instance of the appearance-based SLAM problem.

In this work we propose a solution to this problem which uses a multi-hypothesis tracking Bayesian filter. Let M_t, L_t and Z_t represent the map, one's location and appearance informa-tion respectively at time t. We factor the problem as follows:

$$P(M_t, L_t|Z_t, M_{t-1}, L_{t-1}) \propto P(Z_t|M_t, L_t, M_{t-1}, L_{t-1})P(M_t, L_t|M_{t-1}, L_{t-1}) \quad (1)$$

The term $P\,(M_t, L\,_t|M_{t-1}, L\,_{t-1})$ represents a prior placed over the complexity of the network structure and trajectory through the network. Specifically network structures with fewer vertices and edges, and longer edges are assigned a higher prior probability. Also trajectories of a shorter length and constant velocity are assigned a higher prior probability. Using this factorization the likelihood term $P\,(Z_t|M_t, L\,_t, M\,_{t-1}, L\,_{t-1})$ is conditionally depended on the change in map and localization. This in turn allows a solution to the problem of determining if the current appearance information corresponds to a previously unexplored location. On a conceptual level, the current appearance information is only determined to correspond to a new location if it is not similar in appearance to a previously explored location where the act of traversing to that location does not adversely increase the complexity of the network and/or involve a complex trajectory.

A quantitative evaluation of the proposed SLAM solution was performed on the New College Dataset. Precision and recall values with respect to loop closure detection were used to quantify performance. Results achieved were very positive and actually outperformed current state of the art methods (Cummins and Newman, 2008).

References

Carlone, L., Censi, A., and Dellaert, F. (2014). Selecting good measurements via l1 relaxation: a convex approach for robust estimation over graphs. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Cummins, M. and Newman, P. (2008). Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665.

Kaess, M., Ranganathan, A., and Dellaert, F. (2008). isam: Incremental smoothing and mapping. *IEEE Transactions on Robotics*, 24(6):1365–1378.

Leonard, J. J. and Durrant-Whyte, H. F. (1991). Simultaneous map building and localization for an autonomous mobile robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1442–1447.

Analysis of the Sparse Representation of the Payload Image for Digital Image Watermarking

P. Yogarajah and J.V. Condell
Intelligent Systems Research Centre
University of Ulster, Northern Ireland, UK
p.yogarajah, j.condell@ulster.ac.uk

Abstract

In pixel-based digital image watermarking, the number of pixels in the payload (i.e. secret pixels) image is one of the main concerns. In most cases, the payload image is converted to a lower dimensional compressed signal to reduce the number of pixels. However the reconstruction from the compressed signal to the payload image is a challenging task. The sparse representation technique suggests that a small set of values from a sparsified image is enough to reconstruct the original image. The aim of this short paper is to analyse the sparse representation and reconstruction of the payload image.

Keywords: Image watermarking, L1-Minimization, Sparse representation.

1 Introduction

A sparse representation of a signal is successfully applied in different applications such as radar systems, medical imaging, speech compression and image compression [Baraniuk, 2007]. The sparse representation based compression and reconstruction are explained as follows: Let $\mathbf{x} = [x_1, ..., x_N]^T$ be the column vector that presents the N values of an image column. We assume that $\mathbf{x}$ is a sparse vector and the projection basis $\mathbf{\Phi} = [\Phi_1^T, ..., \Phi_M^T]^T$ is an $M \times N$ matrix, where $M \ll N$. Thus, a compressed vector can be defined as $\mathbf{y} = \mathbf{\Phi x}$, where $\mathbf{y} = [y_1, ..., y_M]^T$, see Figure 1(a). Here the system $\mathbf{y} = \mathbf{\Phi x}$ is under-determined (i.e. $M < N$) thus the solution can be obtained by solving the optimization problem [Baraniuk, 2007], $\hat{x} = \underset{x}{\operatorname{argmin}} ||x||_1 \ \ subject\ to \ \ y = \Phi x$, where $||.||_1$ denotes the ℓ_1-norm. This problem is often known as Basis Pursuit (BP) and can be solved in polynomial time [Chen et al., 1998]. As soon as the signal $\mathbf{x}$ is estimated as $\hat{\mathbf{x}}$, the original signal $\mathbf{x}$ can be reconstructed as $\tilde{\mathbf{x}} = \mathbf{\Phi^T} \hat{\mathbf{x}}$ with small distortion between $\mathbf{x}$ and $\hat{\mathbf{x}}$. In general, the image itself is not sparse. Therefore the sparsifying basis Ψ such as Haar Wavelets (HW), Discrete Cosine Transform (DCT), Hadamard Transform (HT) and Slant Transform (ST) can be applied to sparsify the signal, see Figure 1(b).

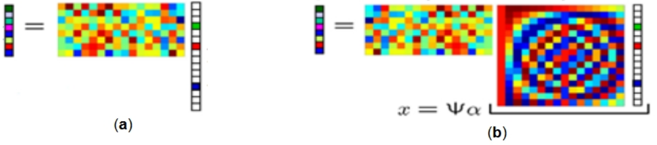


Figure 1: Sparse representation process.

However it is important to choose an appropriate projection basis and sparsifying basis. The sparsifying basis needs to be incoherent with the projection basis. Coherence (μ) is the measure of the highest correlation between any two elements of the projection basis and sparsifying basis. The variable μ is given by [Candes, 2006]: $\mu(\Phi, \Psi) = \sqrt{N} \max_{1 \leq j,k \leq N} | < \phi_j, \psi_k > |$.

In general, if Φ and Ψ have many correlated elements, coherence is high, otherwise it is low. For optimal reconstruction, the system works best with a Ψ that has low coherence with Φ and the μ within $[1, \sqrt{N}]$. In the next section, the optimal sparsifying basis and projection basis are analysed.

1.1 Learning optimal projection basis and sparsifying basis

In this learning, a $M \times N$ random matrix whose entries are independent normal variables with mean 0 and variance $\frac{1}{M}$ (i.e. $\Phi \sim N(0, \frac{1}{M})$) is chosen as a projection basis and HW, DCT, HT and ST are chosen as sparsifying basis. Table 1 shows the calculated values of μ.

Table 1: Calculated μ values using $M = 64$ and $N = 128$.

	HT	DCT	HW	ST
$M \times N$ random matrix, $\Phi \sim N(0, \frac{1}{M})$	5.07	4.40	3.23	3.11

For the optimal solution the value μ should be within $[1, \sqrt{128}]$. It shows that HT, DCT, HW and ST are suitable sparsifying bases but ST outperformed the rest. The next section shows the reconstruction results from the compressed signals.

2 Image reconstruction

A grayscale Lena image of size 128×128 is considered as the payload. The payload is converted to size of $M \times N$ compressed signals using projection and sparsifying bases. The reconstructed payloads from compressed signals using $M = 64$ and 96 are shown in Figure 2. From the results, it can be seen that ST and HW performed well during reconstruction.

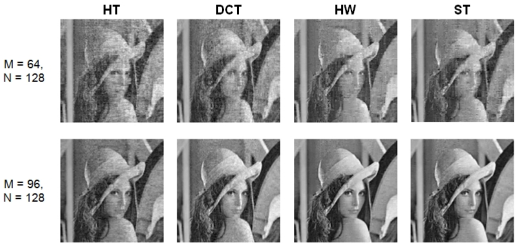


Figure 2: Payload reconstruction: Φ of size $M \times N$ and (HT, DCT, HW, ST) of size $N \times N$.

3 Conclusion

The sparse representation based compression and reconstruction of an image is analysed. Based on the μ values and the reconstruction results, it can be concluded that the Slant, Haar sparsifying bases and the projection basis $\Phi \sim N(0, \frac{1}{M})$ performed well on reconstruction. Therefore, instead of the payload image, the compressed signals can be used in image watermarking.

References

[Baraniuk, 2007] Baraniuk, R. G. (2007). *New Theory and Algorithms for Compressive Sensing*. Final technical report, Rice University Houstan TX.

[Candes, 2006] Candes, E. (2006). Compressive sampling. *Int. Conf. on Congress of Mathematics*, 1:33–52.

[Chen et al., 1998] Chen, S. S., , David Donoho, L., and Michael Saunders, A. (1998). Atomic decomposition by basis pursuit. *CSIAM Journal on Scientific Computing*, pages 33–61.

Gait Analysis and Ageing

P. Chaurasia, Y. Pratheepan and J.V. Condell
Intelligent System Research Centre (ISRC)
School of Computing and Intelligent Systems
University Of Ulster, Northern Ireland, United Kingdom
{p.chaurasia, p.yogarajah, j.condell}@ulster.ac.uk

Abstract

Gait analysis is the process of analysing individual’s walk for the purpose of individual identification. The human gait not only produces distinctive gait silhouettes that can uniquely identify the person but also reflects individual’s physical and health conditions. In a clinical setting gait analysis is typically used to detect problems in individual’s gait and to plan healthcare accordingly.

In biometrics the term ageing usually refers to the gradual degradation in system performance caused by changes in the biometric features suffered by the individual’s trait in long-term. This short paper gives a brief overview of how gait biometric features can be used in the analysis of the gradual drift occurring in an ageing individual.

Keywords: Behavioural biometrics, Gait analysis, and Ageing.

1 Introduction

Gait is a behavioural biometric that can be used to uniquely identify an individual. The use of gait analytics can range beyond the general trend of biometric identifications. The following sub-sections provide an overview of gait analysis.

1.1 Gait Analysis

The human walk, also referred to as the human gait, is a periodic movement of the body that involves repetitive motions (Figure 1) [1]. A normal gait cycle can be divided into different gait phases. The analysis of these phases can define the functional status of different types of motions observed for an individual.

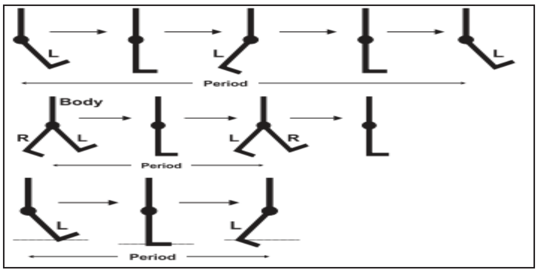


Figure 1: Periodic movement of swinging limbs [1].

1.2 Ageing and Biometrics

A biometric is a measurable characteristic of an individual. Biometric features have been robustly used to build recognition systems that can characterise individuals and recognise them based on the unique features extracted. For such systems, performance is a major issue and one of the most commonly reported issues that can degrade system performance is *ageing* [2]. With ageing biometric features tend to degrade long-term and thus the initial biometric template taken for an individual can differ significantly from the individuals initial biometric samples. Thus, ageing can be considered as a special case of intra-class variability where the individual's own samples differ over time due to inherent transformations caused by body changes or behaviour.

It is necessary to consider the ageing factor in the determination of the time period over which the individual's features are consistent. Analysing long-term biometric data can be useful in predicting the gradual degradation that occurs [2]. In previous research ageing has been studied from a clinical perspective but ageing has rarely been analysed from a biometric perspective [3]. In addition, most of the previous research studies have discussed physical biometrics, such as face and finger modality, to analyse the performance of face and fingerprint recognition systems respectively. In this short paper the analysis of gait behavioural biometrics are proposed for the study of ageing and for the performance of biometric systems.

2 Gait Features in Ageing Measurement

Normal aging changes and health problems frequently show themselves as a decline in the functional status of older adults [4]. Body language information such as gait [5] can be useful aspects to analyse to monitor this deterioration. Each individual has a **unique walking style** which they usually adhere to during their normal walk. Based on this concept it is then possible to create a different gait profile which distinguishes one individual from another in addition to their own previous template.

The variations occurring in the individual can be analysed by studying the individuals' stance position. Double stance (which occurs when the individuals two feet are on the ground) increases with age: from 18% of a total gait cycle in young individuals to approximately 26% in healthy individuals [6]. The learned pattern can provide useful insight into an individual's health and ageing condition. An Ageing Coefficient (AC) can be defined and a possible approach to "ageing detection" protocol can then be followed as described below:

1. Set a suitable AC threshold (∂_{AC}) depending on the level of ageing allowed.
2. With every new access of gait features estimate the last best known gait features mean and variance.
3. Compare the variation of the mean and variance between the old and new set of features.
4. If ∂_{AC} is exceeded, apply suitable feature template update.

The selected value of ∂_{AC} will depend on the different application settings.

3 Conclusion

The proposed gait analysis can be used in addition to or in conjunction with existing analysis for assessing performance of biometric systems with ageing. The technique can also be used to identify various abnormalities in the subject's gait, potentially suggesting injury, sickness, or simply the formation of poor behavioural habits.

References

[1] Tao, W., Liu, T., Zheng, R., Feng, H. (2012). Gait Analysis Using Wearable Sensors. *Sensors*, 12(2): 2255-2283.

[2] Galbally, J., Martinez-Diaz, M., Fierrez, J. (2013). Aging in Biometrics: An Experimental Analysis on On-Line Signature. *PLoS ONE*, 8(7): e69897. doi:10.1371/journal.pone.0069897.

[3] Lanitis, A. (2010). A survey of the effects of aging on biometric identity verification. *International Journal of Biometrics*, 2: 34–52.

[4] Brorsson, B., Asberg, KH. (1984). Katz Index of Independence in ADL: Reliability and validity in short-term care. *Scand J Rehabil Med*, 16:125–132.

[5] Pratheepan, Y., Torr, P. H. S., Condell, J. V., Prasad, G. (2008). Body Language Based Individual Identification in Video Using Gait and Actions. *ICISP*, 368-377.

[6] Abnormal Gait. http://www.patient.co.uk/doctor/Abnormal-Gait.htm. Accessed on 8th May 2014.

AUTHOR INDEX