A Comparative Error Analysis of Audio-Visual Source Localization

Damien Kelly*, François Pitié, Anil Kokaram, Frank Boland.

Department of Electronic & Electrical Engineering,
Trinity College Dublin, Ireland.
{kelly16, fpitie, anil.kokaram, fboland}@tcd.ie

Abstract. This paper examines the accuracy of audio-video based localization using multiple cameras and multi-microphones. Covariance mapping theory is used to determine the accuracy of audio and video based localization. Both modalities are compared in terms of their ability to provide accurate location estimates of a moving audio-visual source. Relatively, video is found to be significantly more accurate than audio. The problem of audio-video fusion is also examined. The fusion of audio and video location estimates is applied in the audio domain, the video domain and the positional domain. The accuracy of these three fusion strategies for 3D localization are examined from a theoretical basis. The best localization performance is found when fusion is applied in the positional domain. Fusing audio and video data in the video domain is found to exhibit the worst localization performance. This analysis is confirmed by measuring the accuracy of each fusion strategy in localizing a moving audio-visual source.

1 Introduction

There is a current trend in the research community towards the use of multiple modalities in tracking. Particularly, the use of both audio and video in tracking people is gaining significant interest. Literature proposes a variety of Kalman filter based [1,2] and particle filter based [3,4] techniques for joint audio-visual tracking. In general these approaches show improved tracking performance beyond that possible through the use of either modality alone.

This result is mainly due to the complementary nature of the audio and video modalities. The performance of audio-based tracking is dependent on the quality of the received signals at the microphones which is affected by the distance from the speaker to the microphones [5], background noise and most significantly distortion due to reverberation [6]. Such issues do not affect video-based tracking. Similarly, factors affecting video-based tracking such as varying illumination, visual clutter and visual occlusions have no implications on audio-based tracking. In a joint audio-video based system where audio and video fail independently therefore it is reasonable to assume an overall improvement in tracking reliability.

* This research is funded by the Irish Research Council for Science Engineering and Technology (Grant No. RS/2005/95) and Science Foundation Ireland.

There is conflicting evidence in literature that improved tracking accuracy is achieved through a joint audio-video based approach. Specifically, Strobel et al. report improved tracking reliability, but note no clear improvement in accuracy beyond the best available single-modality position estimate [1]. The general approach in evaluating the performance of joint audio-video trackers is to determine the tracking performance against some ground truth (eg. [1,2,4]). Tracking performance however is not only dependent on the accuracy of audio and video-based position estimates but also on the employed motion model. Thus a well chosen motion model can result in undue credit being attributed to a multi-modal approach in improving tracking accuracy.

Furthermore, this measure of performance can only give an indication to the expected accuracy of the resulting fused track. It can not be used to determine how well a system is performing in relation to its best possible performance. Currently, little effort has been made in literature to evaluate the expected performance of joint audio-video based tracking systems. This means that such systems are being proposed without any theoretical basis on which to measure performance.

It is informative to examine the performance of a joint audio-video tracking system by examining the localization accuracy in each domain individually. In this way the contribution of both audio and video in improving localization accuracy through a fused estimated can be determined. Literature proposes techniques for predicting the 3D error associated with localization using multiple cameras [7] and multiple microphones [8]. The incorporation of such theory in the analysis of joint audio-video based tracking to date has not been adequately considered. This work aims at addressing this issue.

In this paper covariance mapping theory is used to examine the error of localizing an audio-visual source using multiple microphones and multiple cameras. This mapping theory is used to determine the 3D localization error associated with audio-based localization using time-delay estimation and video-based localization through triangulation from multiple cameras. Given this, a direct comparison is made between the localization accuracy of both modalities in terms of their ability to provide accurate location estimates of a moving audio-visual source.

Covariance mapping is also used to determine a representation of uncertainty on the time delay estimates in the video domain and similarly to determine a representation for uncertainty on pixel measurements in the audio domain. In essence, audio and video localization uncertainty is examined in the positional domain, audio domain and video domain. Maximum likelihood data fusion is applied in these three domains and a resulting 3D fused localization estimate is obtained in each case. The effectiveness of these fusion strategies is examined from a theoretical basis and their ability to provide accurate location estimates of a moving source is evaluated.

In order to make this analysis tractable the general assumption of Gaussian observation noise is assumed on time delay estimates in the audio domain and also in the video domain on pixel measurements obtained from each camera.

Relatively, video-based localization is found to be significantly more accurate than audio-based localization. The contribution of audio data in a fused estimate is also found to vary significantly over the track duration. The analysis of audio and video fusion shows that for the given problem the best localization accuracy is achieved when fusion is applied in the positional domain.

The remaining sections of this paper are organized as follows. Sec. 2 examines the process of uncertainty propagation in a multi-camera multi-microphone configuration. Sec. 3 outlines the experimental setup used in the comparative error analysis and also examines the validity of the uncertainty propagation for this tracking environment. A comparative error analysis of audio-video based localization is made in Sec. 4 and finally, conclusions are presented in Sec. 5.

2 Uncertainty Mapping

In this analysis we examine the 3D position $\mathbf{S} \in \mathbb{R}^3$ of an audio-visual source observed indirectly by pixel measurements $\mathbf{x} = f(\mathbf{S})$ in the video domain and time delay estimates $\tau = g(\mathbf{S})$ in the audio domain. Using these measurements we wish to map their respective covariances $\Sigma_{\mathbf{x}}$ and Σ_τ into positional space in order to estimate the associated covariances $\Sigma_{\mathbf{x}}^V$ of a video-based location estimate and Σ_τ^A of an audio-based location estimate. The covariance of a fused maximum likelihood estimate in the positional domain is then obtained by,

$$\Sigma_{\mathbf{S}}^{Pos} = ((\Sigma_{\mathbf{S}}^A)^{-1} + (\Sigma_{\mathbf{S}}^V)^{-1})^{-1}. \quad (1)$$

Fusion in this application could also considered in two additional domains, the audio domain corresponding to time delays and the video domain corresponding to pixel measurements. This requires the transformation of audio and video measurements to equivalent levels of representation. For instance, for fusion in the video domain an equivalent representation of audio-based measurements and uncertainty must be determined in the image plane. Similarly, for fusion in the audio domain video-based localization measurements and uncertainty must be transformed into an equivalent representation in the audio domain. In the video domain, in addition to the covariance of pixel measurements $\Sigma_{\mathbf{x}}$ the covariance $\Sigma_{\mathbf{x}}^A$ of audio-based 3D localization transformed into the image plane is determined. Likewise, in addition to the covariance of time delay estimates Σ_τ the covariance Σ_τ^V of video-based 3D localization transformed into the audio domain can be determined. Within both the audio domain and video domain then fusion is applied using (1). The resulting covariance of the fused estimate is then mapped from each domain into positional space. This enables the covariance $\Sigma_{\mathbf{S}}^{Vid}$ of 3D localization through fusion in the video domain and the covariance $\Sigma_{\mathbf{S}}^{Aud}$ of 3D localization through fusion in the audio domain to be evaluated. The relevant mappings considered are illustrated in Fig. 1.

2.1 Linear Approximation Mapping

The mapping of covariances between the positional domain, audio domain and video domain can be achieved through a first order Taylor series expansion of

the audio and video measurements functions and their inverses. Here the process of mapping the covariance $\Sigma_{\mathbf{S}}$ of the source position $\mathbf{S}$ to obtain a corresponding measure of pixel uncertainty $\Sigma_{\mathbf{x}}$ in the video domain is presented. If the measurement function $f(\mathbf{S})$ has a continuous first order derivative then a first order Taylor series expansion of $f(\mathbf{S})$ enables the mean and covariance of $\mathbf{x}$ to be approximated. The mean of $\mathbf{x}$ is approximated by $E[\mathbf{x}] = f(E[\mathbf{S}])$ and its covariance $\mathbf{\Sigma_x}$ by,

$$\mathbf{\Sigma_x} = \frac{\partial f(E[\mathbf{S}])}{\partial \mathbf{S}} \Sigma_{\mathbf{S}} \frac{\partial f(E[\mathbf{S}])}{\partial \mathbf{S}}^T. \tag{2}$$

The inverse mapping of (2) is difficult to obtain in cases where the inverse measurement function $\mathbf{S} = f^{-1}(\mathbf{x})$ can not be explicitly defined. For an implicitly defined function $F(\mathbf{x}, \hat{\mathbf{S}}) = \mathbf{x} - f(\hat{\mathbf{S}})$ where $\hat{\mathbf{S}}$ minimizes the criterion function $C(\mathbf{x}, \mathbf{S}) = |F(\mathbf{x}, \mathbf{S})|^2$, the first order derivative of the inverse measurement function can be approximated using the implicit functions theorem [9]. This is found to be,

$$\frac{\partial f^{-1}(E[\mathbf{x}])}{\partial \mathbf{x}} \approx -\left(\frac{\partial F}{\partial \mathbf{S}}\right)^{\dagger} \frac{\partial F}{\partial \mathbf{x}}. \tag{3}$$

where † is used to denote the pseudo inverse.

The concept in using this mapping of uncertainty between the positional, audio and video domains in illustrated in Fig. 1.

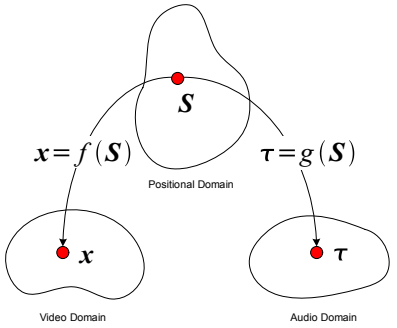
2.2 Audio-based Measurement Function

Time-delay based acoustic source localization infers the position of an active source from time delay estimates between spatially separated microphones. These time delays arise due to the spatial separation of the microphones whereby the source signal is received at each microphone at different points in time. The audio measurement function in this case therefore describes the vector of time delays τ associated with the 3D point $\mathbf{S}$. This function is completely described by the positions of the microphones, the sampling frequency and the speed of sound.

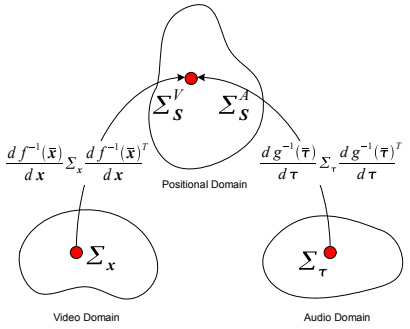
Let $\mathbf{m}_{ij} = [X_{ij}, Y_{ij}, Z_{ij}]^T \;\; j = 1, 2$ denote the positions of the microphones of the ith microphone pair configuration and $\tau = [\tau_1, .., \tau_i, ..., \tau_N]^T$ be the vector of time-delay estimates for N microphone pairs. For the ith microphone pair, the expected time delay τ_i given the source position $\mathbf{S}$ may be determined by,

$$\begin{aligned} \tau_i = \left(\frac{f_s}{c}\right) [((X_{i1} - X)^2 + (Y_{i1} - Y)^2 + (Z_{i1} - Z)^2)^{\frac{1}{2}} \\ - ((X_{i2} - X)^2 + (Y_{i2} - Y)^2 + (Z_{i2} - Z)^2)^{\frac{1}{2}}] \\ = g_i(\mathbf{S}). \end{aligned} \tag{4}$$

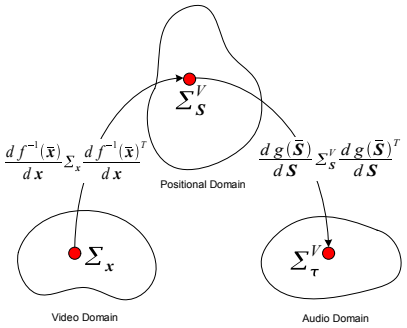
where $f_s = 48kHz$ is the sampling rate and $c = 343m/s$ is the speed of sound. The time delays referred to therefore are in units of audio samples. Using (4) and (2), 3D positional uncertainty can be propagated into the domain of time-delay



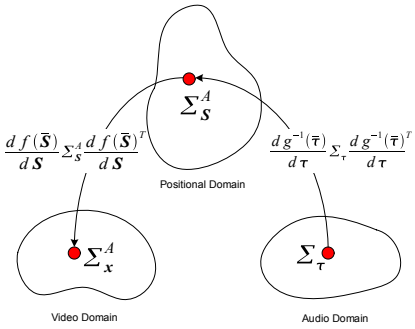
(a) Audio and video measurement functions.



(b) Mapping audio and video uncertainty into 3D space. Here $\Sigma_{\mathbf{S}}^{V}$ represents the covariance of a video-based location estimate and $\Sigma_{\mathbf{S}}^{A}$ denotes the covariance of an audio-based location estimate.



(c) Mapping video uncertainty into the audio domain. The covariance of a video based 3D location estimate in the audio domain is denoted Σ_{τ}^{V}.



(d) Mapping audio uncertainty into the video domain. The covariance of an audio based 3D location estimate in the video domain is denoted Σ_{τ}^{V}.

Fig. 1: The mapping of uncertainty between the audio, video and positional domains through a first order Taylor series expansion of the measurement functions and their inverses. The video measurement function is denoted $f(\mathbf{S})$ and the audio measurement function is denoted $g(\mathbf{S})$. The corresponding inverse measurement functions are denoted $f^{-1}(\bar{x})$ and $g^{-1}(\bar{\tau})$ respectively.

estimates. This linearization of the audio-based measurement function follows in principle the theory of extended Kalman filtering [10].

Given (4) a set of implicit functions can be defined, one for each microphone pair as,

$$G_i(\tau, \hat{\mathbf{S}}) = \tau - g_i(\hat{\mathbf{S}}), \tag{5}$$

where $\mathbf{S}$ is the 3D location estimate which minimizes the criterion function $C_g(\tau, \hat{\mathbf{S}}) = \sum_i^N G_i(\tau, \hat{\mathbf{S}})^2$. Using (5) and (3) enables uncertainty on time delay estimates to be propagated into the 3D positional domain. This approach to predicting the error region associated with a time-delay based location estimate is not readily available in literature therefore a complete derivation of this result is presented in Appendix A.

2.3 Video-based Measurement Function

The video measurement function relates the 3D point $\mathbf{S}$ to a vector of pixel measurements $\mathbf{x} = [\mathbf{x}_1, .., \mathbf{x}_i, .., \mathbf{x}_N]^T$, where $\mathbf{x}_i = [x_i, y_i]^T$ is the 2D pixel measurement corresponding to the *ith* camera view. This function is dependent on the camera matrices of the multi-camera views and their associated distortion parameters. Assuming that the distortion characteristics are known and distortion therefore can be corrected, the projection of a 3D point to the *ith* image plane is described by [11],

$$\mathbf{x}_i = f_i(\mathbf{S}) = \begin{bmatrix} \frac{p^i_{11}X + p^i_{12}Y + p^i_{13}Z + p^i_{14}}{p^i_{31}X + p^i_{32}Y + p^i_{33}Z + p^i_{34}} \\ \frac{p^i_{21}X + p^i_{22}Y + p^i_{23}Z + p^i_{24}}{p^i_{31}X + p^i_{32}Y + p^i_{33}Z + p^i_{34}} \end{bmatrix} \tag{6}$$

where p^i_{uv} is the (u, v) entry in the camera matrix corresponding to the *ith* camera view. Using (6) and (2) 3D positional uncertainty can be propagated into the video domain.

Generally, $\mathbf{S}$ is determined as the point $\hat{\mathbf{S}}$ which satisfies the implicit function,

$$F_i(\mathbf{x_i}, \mathbf{S}) = \begin{bmatrix} X(p^i_{31}x_i - p^i_{11}) + Y(p^i_{32}x_i - p^i_{12}) + Z(p^i_{33}x_i - p^i_{13}) + p^i_{14} \\ X(p^i_{31}y_i - p^i_{21}) + Y(p^i_{32}y_i - p^i_{22}) + Z(p^i_{33}y_i - p^i_{23}) + p^i_{24} \end{bmatrix} \tag{7}$$

such that some criterion function $C_f(\mathbf{x}, \hat{\mathbf{S}}) = \sum_i^N F_i(\mathbf{x}, \hat{\mathbf{S}})^2$ is minimized. Using (7) and (3) enables uncertainty on pixel measurements to be propagated into the 3D positional domain.

3 Evaluation of Audio-Visual Source Localization

Three 720×576 resolution cameras and six microphones were used to record an audio-visual source moving along a 3D path. The cameras were calibrated using 37 point correspondences from known 3D points within a $5.33m \times 6.98m \times 2.45m$ room. The 3D positions of the points were measured using a measuring tape, a square and a level using a single wall as a datum plane. A linear estimate of

the camera matrices was used to initialize a bundle adjustment minimization procedure which minimized the reprojection error of the 37 points in the three views [11]. The overall reconstruction error was found to be 5mm and over an additional set of 37 test points it was found to be 13mm.

The six microphones were arranged into two 3-element microphone arrays. The array geometry used was that of a vertical equilateral triangle with the relative spacing between the microphones set to 340mm.

The audio-visual source followed a known 3D trajectory in space. The source was localized in each camera view through intensity thresholding and blob extraction. The center of mass of the blob was then taken as the object's location in the frame and its 3D location was determined through triangulation. Recursive least squares using the Levenberg-Marquardt algorithm was used to localize the source in the acoustic domain. An audio sampling frequency of $48kHz$ was used and the time-delay estimates were obtained using GCC-PHAT [12] from $40ms$ audio data frames i.e. a single audio-based location estimate for each video frame. The video frame rate was 25 fps and the total track duration was 2594 $frames$.

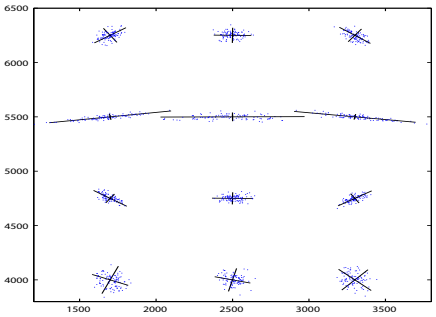
Given both the audio-based and video-based tracking results, further calibration procedures were applied so as to optimize the relative calibration between the audio and video tracking spaces. This was necessary so as to ensure no bias existed between the tracking spaces.

Sufficient reconstruction accuracy was achieved from multi-view visual reconstruction such that any error in localization relative to audio based localization was deemed negligible. The visually reconstructed track therefore was taken as the true track's 3D position. Gaussian noise was added synthetically to pixel measurements to simulate noisy visual localization. In the audio-domain the variance of time-delay estimates was measured empirically using a running variance estimate.

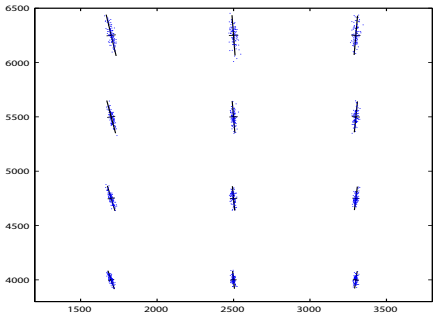
3.1 Validity of First Order Error Propagation

Of particular concern in the application of the error propagation techniques presented in Sec. 2.1 is the validity of using a first order Taylor series expansion. From Fig. 2 however it can be seen that under the assumption of Gaussian noise first order uncertainty propagation is sufficient for the propagation of uncertainty into each domain.

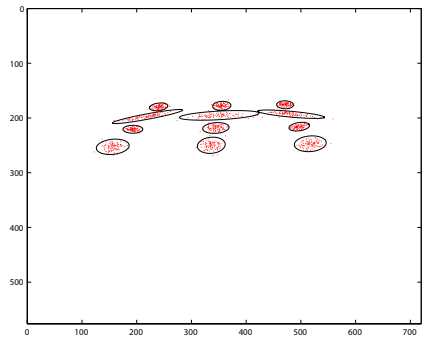
Shown in Fig. 2a and Fig. 2b are the results of 100 Monte Carlo simulations for the localization of 12 positions in the room from noisy time-delay estimates and noisy pixel measurements respectively. Also shown in these figures are the predicted 95 percentile error regions. The variance of time-delay estimates in this case was set to 1 audio sample. The variance of pixel measurements was set to 5 pixels in both x and y. Shown in Fig. 2c are the audio based localization estimates mapped into the image plane. Also shown are the predicted error ellipses in the image plane. In Fig. 2d the result of mapping the video based location estimates to time delays in the audio domain is given. Also seen in this figure are the predicted error bars representing the 95 percentile error regions associated with the mapped time delays. In all of the described cases the predicted error regions are seen to match the localization results.



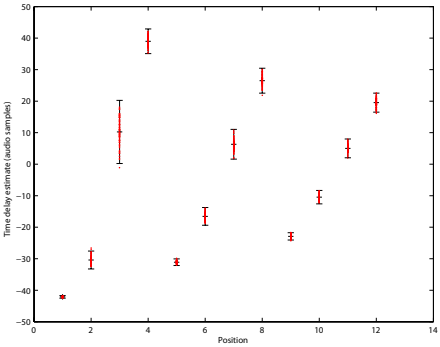
(a) Audio-based localization error in xy-plane.



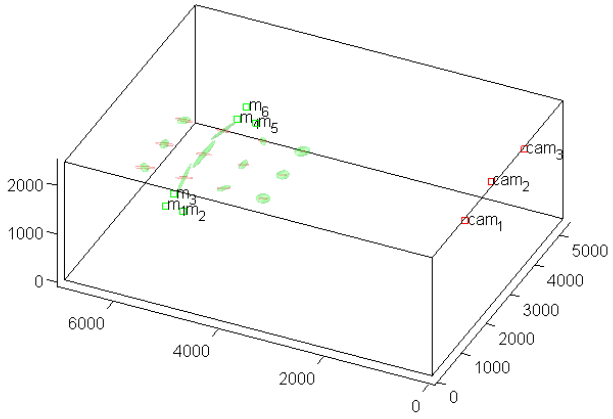
(b) Video-based localization error in xy-plane.



(c) Audio-based localization error mapped into the image plane of camera cam_1.

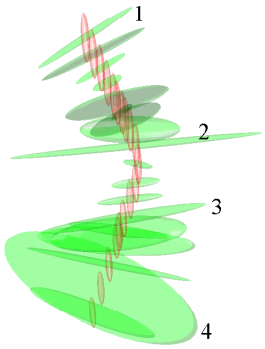


(d) Video-based localization error mapped to time delay domain of microphone pair $m_1 - m_2$.



(e) 3D Error ellipsoids for audio (green) and video (red) based localization.

Fig. 2: *Validity of first order error propagation.*



(a)

Location	1	2	3	4
$tr(\Sigma_{\mathbf{S}}^{A})$ (cm^2)	178	717	328	639
$tr(\Sigma_{\mathbf{S}}^{V})$ (cm^2)	29	21	17	14
$tr(\Sigma_{\mathbf{S}}^{Aud})$ (cm^2)	2	17	13	9
$tr(\Sigma_{\mathbf{S}}^{Vid})$ (cm^2)	18	20	16	13
$tr(\Sigma_{\mathbf{S}}^{Pos})$ (cm^2)	2	1	2	7

(b)

Fig. 3: *(a) Error ellipsoids for audio-based localization (green) and video-based localization (red). (b) Comparison of localization error at the numbered locations.*

4 Comparative Error Analysis and Discussion

Using the covariance mapping techniques presented in Sec. 2.1 and given the covariance of time delay estimates, the error associated with audio-based localization can be determined. Similarly, given the covariance of pixel based measurements the error associated with video-based localization can be determined. Shown in Fig. 3a are the resulting 95 percentile error ellipsoids for points along the track as described in Sec. 3. The variance of pixel measurements is assumed to be 5 pixels in both the x and y image axes and the variance of time delay estimates is measured empirically using a running variance estimate over the track duration.

From this it can be seen that in each of the cases of both audio and video-based localization the error associated with a location estimate is non-uniform in space. Also, the orientation of the error regions is dependent on the configuration of the sensors. Direct comparison of the associated 3D localization covariance matrices therefore can not be made. The trace of the covariance matrix defining the error ellipsoids was chosen as a performance measure for the accuracy of localization.

Shown in Fig. 3b is a table quoting this performance measure for the numbered points along the track as shown in Fig. 3a. Included in this table are the expected values of this performance measure of 3D localization accuracy for the three fusion strategies as outlined in Sec. 2. From these results it can be seen that as expected the theoretical performance of 3D localization using audio-video fusion is greater than the use of audio or video alone. It can also be seen that the best overall performance for 3D localization is expected where fusion is applied in the positional domain. Of the fusion strategies considered the worst 3D localization accuracy is expected where fusion is applied in the video domain. In comparing the results between localization using video data only to localization

through fusion in the video domain it can be seen that little improvement is achieved. This suggests that in the image plane the contribution of audio data in improving localization accuracy is small.

Using the trace of the 3D localization covariance matrix as a measure of localization accuracy, audio-based localization was compared to simulated video-based localization whereby noise was added to the true pixel measurements. The percentage of frames for which audio-based localization was found to be more accurate than video-based localization for varying pixel measurement noise is shown in Fig. 4a. From this it can be seen that even with a pixel measurement noise of 20 $pixel^2$ the percentage of frames for which audio localization is more accurate over the track duration is less than 40%. Although this is the case, for 5% of the track duration audio-based localization was found to be more accurate than video-based localization for a pixel measurement variance of 1 $pixel^2$. This reveals significant variation in audio-based localization accuracy over a typical track duration.

The percentage of frames where fusion resulted in improved tracking accuracy can be seen in Fig. 4b. This figure reveals that the overall best improvement in accuracy over the track duration was obtained through fusion in the positional domain. The worst performance occurred for fusion in the video domain. These results are in accordance with the analysis presented in Fig. 3.

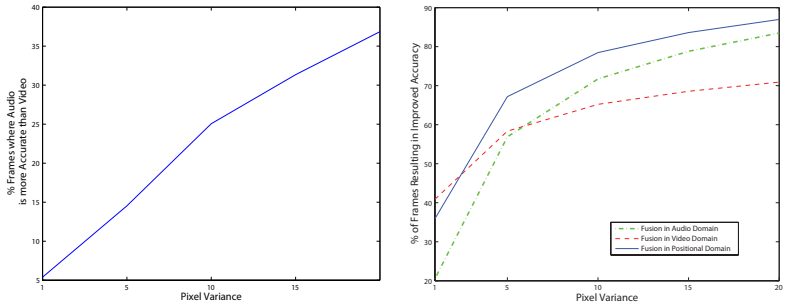
5 Conclusions

The use of error propagation was presented in this paper as a useful means of evaluating the performance of a joint audio-video based tracking system. First order error propagation was shown to adequately map audio and video measurement uncertainty across the positional, video and audio tracking domains. In the comparison of audio-based and video-based localization accuracy video-based localization was found to consistently outperform that of audio-based localization in terms of accuracy and consistency. Maximum likelihood fusion of location estimates in the audio domain, video domain and positional domain was examined. The best 3D localization accuracy for the described multi-camera and multi-microphone setup was found to be achieved where fusion is applied in the positional domain. Little contribution from audio in improving localization accuracy through a joint audio-video estimate was found where fusion is applied in the video domain.

The analysis presented in this paper can also be seen to provide a convenient basis for determining optimal sensor placement configurations [13].

References

1. N. Strobel, S. Spors, R. Rabenstein: Joint Audio-Video Signal Processing for Object Localization and Tracking. In M. Brandstein and D. Ward, ed.: Microphone Arrays: Signal Processing Techniques and Applications. Springer-Verlag (2001) 203–225



(a) Audio Accuracy vs. Video Accuracy

(b) Performance of Data Fusion

Fig. 4: *Comparison of audio-based, video-based and fusion-based localization performance.*

2. N. Katsarakis, G. Souretis, F. Talantzis, A. Pnevmatikakis, L. Polymenakos: 3D Audiovisual Person Tracking using Kalman Filtering and Information Theory. In: CLEAR Evaluation Workshop, Southampton U.K. (2006) 45–54
3. D. N. Zotkin, R. Duraiswami, L. S. Davis: Joint Audio Visual Tracking Using Particle Filters. EURASIP Journal on Applied Signal Processing **11** (2002) 1154–1164
4. N. Checka, K. W. Wilson, M. R. Siracusa, T. Darrell: Multiple Person and Speaker Tracking with a Particle Filter. IEEE Int. Conf. on Acoustics, Speech and Signal Processing **5** (2004) 881–884
5. T. Gustafsson, R. D. Bhaskar, M. Trivedi: Source Localization in Reverberant Environments: Modeling and Statistical Analysis. IEEE Transactions on Speech and Signal Processing **11** (2003)
6. J. Chen, Y. A. Huang, J. Benesty: A Comparative Study on Time Delay Estimation in Reverberant and Noisy Environments. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2005) 21–24
7. G. Olague, R. Mohr: Optimal Camera Placement for Accurate Reconstruction. Pattern Recognition **35** (2002) 927–944
8. M. Brandstein and J. Adcock and H. Silverman: Microphone Array Localization Error Estimation with Application to Sensor Placement. J. Acoust. Soc. Am. **99** (1996) 3807–3816
9. O. Faugeras: 5. In: Three-Dimensional Computer Vision: A Geomteric Viewpoint. MIT Press (1993)
10. T. Gehrig, J. McDonough: Tracking Multiple Speakers with Probabilistic Data Association Filters. In: CLEAR Evaluation Workshop, Southampton U.K. (2006) 137–150
11. R. Hartley, A. Zisserman: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press (2004)
12. C. Knapp, G. Carter: The Generalized Correlation Method for Estimation of Time Delay. IEEE Trans. on Acoustics Speech and Signal Processing **ASSP-24** (1976) 320–327

13. D. Kelly, F. Boland: Optimal Microphone Placement for Active Speaker Localization. In: 8th IMA International Conference on Mathematics in Signal Processing, Cirencester, England UK (2008) Accepted for publication.

APPENDIX A.

Given $G(\tau, \mathbf{S})$ and C_g as defined in Sec. 2.2 the implicit functions theorem implies [9],

$$\frac{\partial g^{-1}(\tau)}{\partial \tau} = -\mathbf{H}^{-1}\frac{\partial \Phi}{\partial \tau} \tag{8}$$

where Φ and $\mathbf{H}$ are derived in the following as,

$$\Phi = \left(\frac{\partial C_1}{\partial \mathbf{S}}\right)^T = 2\sum_i^N G_i(\mathbf{S},\tau)\left(\frac{\partial G_i}{\partial \mathbf{S}}\right)^T, \tag{9}$$

$$\mathbf{H} = \frac{\partial \Phi}{\partial \mathbf{S}} = \frac{\partial}{\partial \mathbf{S}}\left(\frac{\partial C_1}{\partial \mathbf{S}}\right)^T \tag{10}$$

$$= 2\sum_i^N G_i(\mathbf{S},\tau)\frac{\partial}{\partial \mathbf{S}}\left(\frac{\partial G_i}{\partial \mathbf{S}}\right)^T + 2\sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{S}}\right)^T \frac{\partial G_i}{\partial \mathbf{S}} \tag{11}$$

$$\approx 2\sum_i^N (\frac{\partial G_i}{\partial \mathbf{S}})^T \frac{\partial G_i}{\partial \mathbf{S}} \tag{12}$$

$$= 2\sum_i^N \begin{bmatrix} \frac{\partial G_i}{\partial X}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial X}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial X}\frac{\partial G_i}{\partial Z} \\ \frac{\partial G_i}{\partial Y}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial Y}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial Y}\frac{\partial G_i}{\partial Z} \\ \frac{\partial G_i}{\partial Z}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial Z}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial Z}\frac{\partial G_i}{\partial Z} \end{bmatrix} \tag{13}$$

$$= 2\left(\frac{\partial G}{\partial \mathbf{S}}\right)^T \frac{\partial G}{\partial \mathbf{S}}, \tag{14}$$

$$\frac{\partial \Phi}{\partial \tau} = \frac{\partial}{\partial \tau}\left(\frac{\partial C_1}{\partial \mathbf{S}}\right)^T \tag{15}$$

$$= 2\sum_i^N G_i(\mathbf{S},\tau)\frac{\partial}{\partial \tau}\left(\frac{\partial G_i}{\partial \mathbf{S}}\right)^T + 2\sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{S}}\right)^T \frac{\partial G_i}{\partial \tau} \tag{16}$$

$$\approx 2\sum_i^N \left(\frac{\partial G_i}{\partial \mathbf{S}}\right)^T \frac{\partial G_i}{\partial \tau} \tag{17}$$

$$= 2\sum_i^N \begin{bmatrix} \frac{\partial G_i}{\partial \tau_1}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial \tau_1}\frac{\partial G_i}{\partial \tau_1} & \frac{\partial G_i}{\partial \tau_1}\frac{\partial G_i}{\partial Z} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \frac{\partial G_i}{\partial \tau_N}\frac{\partial G_i}{\partial X} & \frac{\partial G_i}{\partial \tau_N}\frac{\partial G_i}{\partial Y} & \frac{\partial G_i}{\partial \tau_N}\frac{\partial G_i}{\partial Z}] \end{bmatrix} \tag{18}$$

$$= 2\left(\frac{\partial f}{\partial \mathbf{S}}\right)^T \frac{\partial G}{\partial \tau}. \tag{19}$$

Substituting (14) and (19) into (8), $\frac{\partial g(\tau)}{\partial \tau}$ is obtained through,

$$\frac{\partial g^{-1}(\tau)}{\partial \tau} = -\left(\frac{\partial G}{\partial \mathbf{S}}\right)^\dagger \frac{\partial G}{\partial \tau} \tag{20}$$

where † is used to denote the pseudo inverse.