LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

**Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

**Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

**Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

**Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# Plausible Crowd and Group Formations

by

**Cathy Ennis, B.Eng, M.Sc**

**Dissertation**

Presented to the

University of Dublin, Trinity College

in fulfillment

of the requirements

for the Degree of

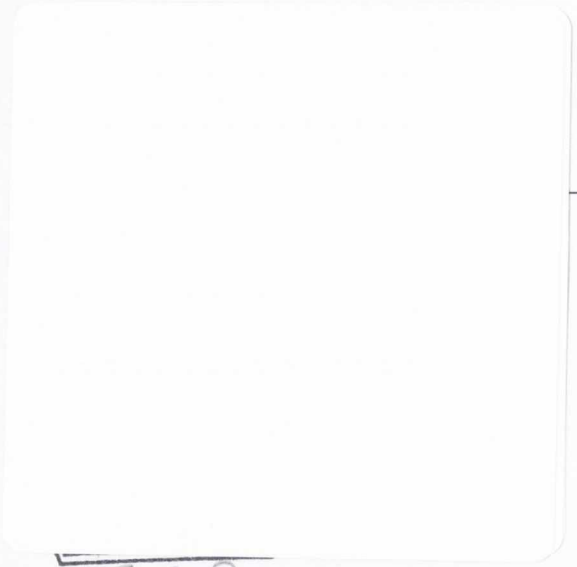
**Doctor of Philosophy**

**University of Dublin, Trinity College**

March 2011

## Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work. I agree that Trinity College Library may lend or copy this thesis upon request.



THESIS
9306

# Acknowledgments

First of all, I would like to thank my supervisor, Carol O'Sullivan. I am hugely grateful for the opportunities, guidance and encouragement she has given me over the past three years.

I would also like to thank my colleagues in the GV2 lab and on the Metropolis team for helping make my time here so enjoyable. In particular, my collaborators Rachel McDonnell and Chris Peters, for being a pleasure to work with, and for their invaluable contributions to this work.

Also, thanks to Fearghal and the girls, for all the chats, tea and sympathy during the more stressful times!

Finally, special thanks to Mam, Dad, Jenny and Ruth for their constant support and the enthusiasm they have for everything I do.

CATHY ENNIS

*University of Dublin, Trinity College*
*March 2011*

**Plausible Crowd and Group Formations**

Publication No. _____

Cathy Ennis, Ph.D.
University of Dublin, Trinity College, 2011

Supervisor: Carol O'Sullivan

# Abstract

Applications using real-time virtual crowds can be found in many areas, including the entertainment industry, for urban planning and health and safety purposes. In some of these applications, where the goal is to replicate as accurately as possible the movement of the crowd, less effort is spent in attempting to make the crowd characters appear convincing to the viewer. Crowds that are used for entertainment have a different purpose, however, where the function of such crowds is to increase the user's immersion in the environment. However, real-time crowds are limited by the need to reduce computational efforts to maintain an adequate frame-rate. For example, more background characters are increasingly being introduced to video games to increase realism for the user.

This research addresses the problem of simulating a perceptually plausible crowd in real-time. Specifically, we address the behaviour of the agents in the crowd. The framework for this research is a real-time crowd system, called Metropolis, that is set in and around the Trinity College campus in Dublin. This cross-disciplinary project involves researchers from Computer Science, Engineering and Neuroscience to construct a multisensory, perceptually plausible experience for the user.

We first examine the role that the environment plays in viewers' expectations of the behaviour of the crowd. We derive a set of rules for populating a virtual world, using global and local properties of various prototypical locations. In a series of perceptual experiments, we then validate the plausibility of these rules.

We also investigate the implementation of conversing groups in the world. These types of background characters are still largely missing from video game crowds. After the collection of a wide range of body motions and audio recordings from natural conversations, we identify possible methods to generate new conversations from this finite data-set. We then evaluate these new conversations in a series of multisensory experiments.

The main contribution of this thesis comes from the results of a number of perceptual evaluation studies, where we identify situations and locations where our rules and methods can be applied to improve the realism of the simulation. Using these results, we can recommend specific guidelines for developers who wish to apply our methodologies to virtual crowds.

**Relevant Publications:**

1. **Perceptual Effects of Scene Context And Viewpoint for Virtual Pedestrian Crowds**: Cathy Ennis, Christopher Peters and Carol O'Sullivan; ACM Transactions on Applied Perception 8(2), To Appear 2011

2. **Seeing is believing: body motion dominates in multisensory conversations**: Cathy Ennis, Rachel McDonnell and Carol O'Sullivan; Transactions on Graphics (SIGGRAPH 2010) 29(4), pp. 91:1–91:9, 2010

3. **Plausible Methods for Populating Virtual Scenes**: Cathy Ennis, Anton Gerdelan and Carol O'Sullivan; Crowd Simulation Workshop, St. Malo, France, 2010

4. **Talking bodies: Sensitivity to desynchronization of conversations**: Rachel McDonnell, Cathy Ennis, Simon Dobbyn and Carol O'Sullivan; ACM Transactions on Applied Perception 6(4), pp. 22:1–22:8, 2009

5. **Modelling Groups of Plausible Virtual Pedestrians**: Christopher Peters and Cathy Ennis; IEEE Computer Graphics and Applications, Special Issue on Virtual Populace, 29(4), pp. 54–63, 2009

6. **Perceptual Evaluation of Position and Orientation Context Rules for Pedestrian Formations**: Cathy Ennis, Christopher Peters and Carol O'Sullivan; ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization (APGV'08), pp. 75–82, 2008

7. **Crowds in Context: Evaluating the Perceptual Plausibility of Pedestrian Orientations**: Christopher Peters, Cathy Ennis, Rachel McDonnell and Carol O'Sullivan; Eurographics 2008 Short Papers, pp. 227–230, 2008

8. **Virtual Crowds in Context**: Cathy Ennis and Carol O'Sullivan; IRLOGI GIS Conference Student Presentations, 2009

# Contents

**Acknowledgments** iv

**List of Figures** xii

**List of Tables** xviii

**Chapter 1 Introduction** 1

1.1 Motivation . . . . . 1
1.2 Research Question . . . . . 2
1.3 Scope . . . . . 4
1.3.1 Properties of the Metropolis Crowd . . . . . 4
1.4 Contributions . . . . . 6
1.5 Summary of Chapters . . . . . 8

**Chapter 2 Background and Related Work** 10

2.1 Introduction . . . . . 10
2.2 Flocking and Herding . . . . . 10
2.3 Human Crowds . . . . . 12
2.3.1 Urban Planning Research . . . . . 14
2.4 Models for virtual human behaviour . . . . . 17
2.4.1 Macroscopic models for virtual human behaviour . . . . . 17
2.4.2 Path planning and Navigation . . . . . 20
2.5 Virtual Crowds . . . . . 21
2.5.1 Hierarchical models of agents for crowd simulations . . . . . 22
2.5.2 Navigation and path planning for crowds . . . . . 24
2.5.3 Variety and roles in crowd behaviour . . . . . 26
2.6 Informed Environment . . . . . 27
2.7 Perception . . . . . 30

2.7.1 Perceptually Guided Graphics . . . 31
2.7.2 Perceptually Guided Crowds . . . 33
2.8 Discussion . . . 34

**Chapter 3 Overview of Group Behaviour** **35**
3.1 Introduction . . . 35
3.2 Dynamic Groups . . . 35
3.2.1 Pedestrian Dynamic Groups . . . 36
3.2.2 Dynamic Groups In Graphics Applications . . . 36
3.3 Conversational Characters and Groups . . . 39
3.3.1 Non-Dynamic Groups . . . 40
3.3.2 Communication And Gestures . . . 41
3.3.3 Studies on Human Communicative Behaviour . . . 43
3.3.4 The Role Of Audio In Conversations . . . 45
3.3.5 Conversational Agents in Virtual Applications . . . 47
3.4 Discussion . . . 53

**Chapter 4 System and Methods** **54**
4.1 Introduction . . . 54
4.2 Metropolis Crowd System . . . 54
4.2.1 Objectives . . . 54
4.2.2 Background . . . 55
4.2.3 System Overview . . . 56
4.2.4 Research Areas . . . 59
4.3 Motion Capture Method . . . 63
4.3.1 Motion and Audio Capture Set-up . . . 64
4.3.2 Pipeline from Motion Capture to Metropolis . . . 66
4.4 Discussion . . . 70

**Chapter 5 Plausible Configurations** **73**
5.1 Introduction . . . 73
5.2 Static Crowd Formation Experiments . . . 74
5.2.1 Data Collection Phase . . . 74
5.2.2 Annotation Phase . . . 75
5.2.3 Reconstruction Phase . . . 75
5.2.4 Orientation Experiment Modification . . . 76
5.2.5 Position and Orientation Experiments Modification . . . 78

5.3 Orientation Experiment . . . . . 82
5.3.1 Method . . . . . 84
5.3.2 Results . . . . . 85
5.4 Position and Orientation Experiment . . . . . 87
5.4.1 Method . . . . . 89
5.4.2 Results . . . . . 90
5.5 Real Scene Variation Experiment . . . . . 94
5.5.1 Method . . . . . 95
5.5.2 Results . . . . . 96
5.6 Camera Viewpoint . . . . . 98
5.6.1 Method . . . . . 98
5.6.2 Experiment . . . . . 100
5.6.3 Results . . . . . 100
5.7 Dynamic Groups . . . . . 101
5.7.1 Corpus Analysis . . . . . 102
5.7.2 Method . . . . . 103
5.7.3 Experiment . . . . . 105
5.7.4 Results . . . . . 105
5.8 Discussion . . . . . 106

**Chapter 6 Conversational Groups** **108**
6.1 Introduction . . . . . 108
6.2 Body Motion Desynchronisation Experiments . . . . . 110
6.2.1 Experiment Design . . . . . 110
6.3 Exp 1: 2 Alternative Forced Choice Experiment . . . . . 112
6.3.1 Results . . . . . 114
6.4 Exp 2: Single Stimulus Experiment . . . . . 116
6.4.1 Results . . . . . 118
6.5 Multisensory Desynchronisation Experiments . . . . . 120
6.5.1 Experiment Design . . . . . 120
6.6 Exp 3: Omni-Directional Audio Experiment . . . . . 125
6.6.1 Results . . . . . 126
6.7 Exp 4: Localised Audio Experiment . . . . . 129
6.7.1 Baseline . . . . . 129
6.7.2 Localised Audio Experiment . . . . . 131
6.7.3 Results . . . . . 132

6.8 Audio Experiments Cross-Analysis . . . . . 134
6.9 Three Experiments Cross-Analysis . . . . . 135
6.10 Discussion . . . . . 137
6.10.1 Alternative Approaches . . . . . 138
6.10.2 System Implementation . . . . . 139

**Chapter 7 Conclusions and Future Work** **140**
7.1 Summary of Contributions . . . . . 140
7.1.1 New Methodologies . . . . . 140
7.1.2 New Experimental Results . . . . . 141
7.1.3 Motion Capture Database . . . . . 143
7.2 Guidelines . . . . . 143
7.3 Future Work . . . . . 144
7.3.1 Conversational Model . . . . . 144
7.3.2 Dynamic Groups . . . . . 145
7.3.3 Awareness of Environment . . . . . 146

**Bibliography** **147**

# List of Figures


1.1 Overview of the work carried out for this thesis; outlining our methodology of using real world data against which to compare our rules for evaluation purposes. . . . . 7

2.1 "Boids" steering behaviours: (a) separation, (b) alignment and (c) cohesion [Rey99]. . . . . 11
2.2 (a) Intention generator for predator fish and (b,c) resulting behaviours of stalking and eating prey [TT94]. . . . . 13
2.3 Demonstration of stripe formation in an intersection of two dynamic groups [HJ09] . . . . . 17
2.4 Results from Helbing and Molnar's social force model for pedestrian dynamics showing: (left) bottleneck streams and (right) flow lanes. Black circles are travelling from left to right and white circles are travelling right to left. The size of the circle indicates velocity [HM95]. . . . . 18
2.5 Sensing objects: (a) examining map cells along the rasterized eye ray, (b) fan of eye rays used for broader view and (c) perceiving dynamic objects with a tiered sensing fan [ST05]. . . . . 19
2.6 Overview of the level-of-detail method for reactive and proactive real-time behaviour described in [NG05]. . . . . 20
2.7 Overview of architecture to represent agents in a crowd with high level behaviours and low-level physical simulators [Tha01]. . . . . 23
2.8 Overview of evacuation navigation algorithm in Hi-DAC [PB06]. . . . . 28
2.9 Hierarchical structure of a town according to Thomas and Donikian [TD00]. 30
2.10 Example motion used in [RP03] . . . . . 33

3.1 Comparison between observed and simulated average group formations for different group sizes as found by Moussaïd et al. [MPG+10] . . . . . 37

3.2 Group motion using advised tiles [Mar02]. . . . . 39
3.3 Group formation changes for individual joining the group (a-c) and leaving the group (d-f) [Ken90]. . . . . 41
3.4 Example of the stimuli presented to participants in Rose and Clarke [RC09] 44
3.5 Example of the stimuli presented to participants by Giorgolo and Verstraten [GV08] . . . . . 47
3.6 Example of the output from the BEAT system [CVB01] . . . . . 49
3.7 Sample scene utilising role-passing LOD [OCV+02] . . . . . 52

4.1 Screenshots of Virtual Dublin, the precursor to Metropolis. . . . . 56
4.2 Screenshot of Metropolis. This scene is in the front square of the Trinity College Dublin campus and in the crowd, conversational groups as well as dynamic agents can be seen. . . . . 57
4.3 Screenshot of Metropolis. This is a close-up of one of our conversational groups. . . . . 57
4.4 Screenshot depicting the capacity of our system to render a large crowd in real-time. . . . . 58
4.5 Screenshot of Morpheme animation engine, showing animation selection area, real-time animation viewer and FSM editor. . . . . 60
4.6 Part of our environment (Trinity College) subdivided according to CDT (left), and the informed environment showing obstacles, destination and generation locations (right). . . . . 61
4.7 Predictive geometric collision avoidance for an agent for three time steps, showing the allowed space the agent can use at each time step in blue. . . . 62
4.8 Screenshots of traffic simulation in Metropolis, showing navigable demarcations rendered (left) and a congested scene outside Trinity College (right). 62
4.9 Noisemap showing audio levels at different locations around Trinity College. 63
4.10 Showing marker placement for our motion capture skeleton (left and centre) and a photograph of an actor in one of our motion capture suits, with a motion capture camera in the background (right). . . . . 65
4.11 Photograph of our research lab, showing four of our Vicon cameras. . . . . 66
4.12 Interface of Vicon IQ, our motion capture software showing data for a 3 person conversation. . . . . 67
4.13 Example of a Vicon skeleton in T-Pose. . . . . 68

4.14 Illustration of data processing pipeline in Vicon: (a) shows a circle fit of each subject, (b) shows a skeleton fit and trajectory label operation performed for each subject, (c) shows 2 subjects removed to manually label one at a time and (d) shows a kinematic fit applied to the body after labelling and filling gaps. . . . . 69
4.15 Screenshots of Asset Manager workspace, where events (shown at bottom of image) for the FSM are chosen: (left) shows a conversation clip of the talking motions for the character and (right) shows a event with a listening motion. . . . . 71
4.16 Screenshots of Network workspace: (left) shows the FSM structure for a character conversation and (right) shows inside the IdleFSM, with links between idle and conversational states for 6 events. . . . . 72

5.1 Reconstructing the scene. From (a) initial photograph to (d) a virtual scene. 76
5.2 Example of the basic orientation rules. Reconstruction (a) of the original photograph, (b) random rule, (c) uniform rule and (d) even rule applied to all pedestrians. . . . . 77
5.3 View of MetroPed for the corridor location: (a) design of the grid with out-of-bounds and flow directions; (b) examples of output for random position; (c) random position and orientation and (d) context modified position and orientation rules. . . . . 80
5.4 Generation of pedestrian transformations for the open location: (a) design of the grid with out-of-bounds and flow directions; (b) examples of output for random position (c) random position and orientation and (d) context modified position and orientation rules. . . . . 81
5.5 Position images for the open location: (a) the original still image; (b) virtual Real image; (c) virtual Random image and (d) virtual Context image. . . . 82
5.6 Orientation images for the corridor location: (a) the original still image; (b) virtual Real image; (c) virtual Random image and (d) virtual Context image. . . . . 83
5.7 Main effect of orientation. On the Y axis of the graph, 1 corresponds to a 'real' response and 0 corresponds to 'synthetic'. . . . . 86
5.8 Interaction between orientation and location. On the Y axis of the graph, 1 corresponds to a 'real' response. . . . . 87

5.9 Position and orientation images for the open scene: (a) Photograph; (b) Real; (c) Random; (d) Position Context; (e) Orientation Context and (f) Both Context. . . . . 88
5.10 Graphs showing position results for Corridor and Open locations. . . . . 90
5.11 Graphs showing orientation results for Corridor and Open locations. . . . . 91
5.12 Graphs showing position and orientation results for Corridor and Open locations. . . . . 93
5.13 Showing example of real scenes used in Experiment 2 for (a) the open location and (b) the corridor location. . . . . 94
5.14 Showing (a) random and (b) context random scenes for the constrained and unconstrained locations respectively. . . . . 96
5.15 Graphs showing results for corridor and open locations for 4 real scenes. . . 96
5.16 Graphs showing all results for corridor and open locations (Note: for the corridor location, random scenes received a zero rating averaging across all participants). . . . . 97
5.17 Showing each viewpoint (eye-level (1), canonical (2), isometric (3) and top-down(4)) for the corridor location (a) to (d), and open location (e) to (h). . 99
5.18 Shows responses for both locations (Cor/Op), both formations (Con/Ran) and each viewpoint (1-4). . . . . 101
5.19 Shows participant responses for each viewpoint for context and random formations, across each location. . . . . 102
5.20 Showing frequencies of numbers of single pedestrians in a scene for 20 random frames of video, numbers of pairs and numbers of triples. . . . . 103
5.21 Example scenes from a sample of animations that were rated by participants in the perceptual study. These images corresponding to group ratios 1:1:1 (left), 2:1:1 (center) and 1:1:2 (right). . . . . 104
5.22 Graph showing 'real' ratings for each group ratio (singles:pairs:triples), with 1 representing a rating of 'Real' and 0 representing a rating of 'Synthetic'. 106
5.23 Graph showing 'real' ratings for the three main conditions, with 1 representing a rating of 'Real' and 0 representing a rating of 'Synthetic'. . . . . 107

6.1 Example of talker/listener (T/L) annotation of two conversation recordings (C1 and C2) for three actors (A1, A2, A3). Each T/L clip is greater than 10 seconds in length. . . . . 111
6.2 Examples of synchronised dominant-speaker conversations used in the experiment. . . . . 111

6.3 Examples of synchronised debates used in the experiment. . . . . 112
6.4 Examples of desynchronised dominant speaker conditions, where 10-second talker or listener (T/L) clips are chosen at random from the annotated recordings. . . . . 114
6.5 Main effect of conversation type. A correct response is when a user correctly identifies the real conversation. . . . . 115
6.6 No main effect of number of speakers in desynchronised dominant speaker conversations was found. A correct response is when a user correctly identifies the real conversation. . . . . 116
6.7 Results for our single stimulus body motion experiment for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition. . . . . 118
6.8 Results for our single stimulus body motion experiment for debates, showing mean 'real' ratings and standard error bars for each condition. . . . . 119
6.9 Our audio and motion capture setup with three female actors. . . . . 122
6.10 Examples of stimuli used in our experiments: (L) a real female debate and (R) a real male dominant speaker conversation. . . . . 123
6.11 Participant in Localised Audio experiment. . . . . 125
6.12 Results for our Omni-directional Audio experiment for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition. . . . . 127
6.13 Results for our Omni-directional Audio experiment for debates, showing mean 'real' ratings and standard error bars for each condition. . . . . 128
6.14 Comparison for matching conditions across our Debates (blue series) and Dominant Speaker conversations (orange series). . . . . 129
6.15 Results for our Audio localisation Baseline experiment showing participant accuracy in localising conversational audio well above chance (33%) for each of our three male and three female actor voices. . . . . 131
6.16 Results for our localised Audio experiment for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition. 133
6.17 Results for our localised Audio experiment for debates, showing mean 'real' ratings and standard error bars for each condition. . . . . 133
6.18 Comparison of results for Omni-directional (blue) and localised Audio (red) experiments for debates, showing mean 'real' ratings and standard error bars for each condition. . . . . 135

6.19 Comparison of results for Omni-directional (blue) and localised Audio (red) experiments for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition. . . . . 136
6.20 Results across 3 experiments for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition. Note: for No Audio (green) ignore $A_x$ label, all are $A_0$. . . . . 136
6.21 Results across 3 experiments for debates, showing mean 'real' ratings and standard error bars for each condition. Note: for No Audio (green) ignore $A_x$ label, all are $A_0$. . . . . 137
6.22 Screenshot of our real-time crowd system with dynamic agents and conversing groups. . . . . 139

# List of Tables


5.1 Group ratios used for the *groups* category of animations. The third column shows the numbers of singles (S), pairs (P) and triples (T) when 6 is chosen as the base value, giving the resulting total number of characters column 4. In each case, we hypothesised that the ratios would be Plausible or Implausible . . . . . 104

6.1 Notation used for describing experimental conditions. $B_x$ denotes body motion condition and $A_0$ denotes that no audio was used for these experiments. 113
6.2 Mean 'real' ratings and standard deviations for No Audio experiment. . . . 119
6.3 Notation used for describing experimental conditions. . . . . 121
6.4 Experimental design for both Omni-Directional and localised Audio experiments, showing total number of trials (50% male actors, 50% female). . . 126
6.5 Mean 'real' ratings and standard deviations for Omni-directional Audio experiment. . . . . 128
6.6 Mean 'real' ratings and standard deviations for localised Audio experiment. 134

# Chapter 1

# Introduction

## 1.1 Motivation

Virtual crowds are used in many applications across different industries. In the entertainment industry, virtual crowds are often employed for a variety of purposes. For example, to reduce costs in movie production, a large number of extras could be replaced by virtual actors. Background crowds are used in video games, particularly in urban environment games such as Grand Theft Auto or The Sims. Other applications employ models of crowd behaviour to help plan urban developments, architectural prototypes, emergency evacuation procedures, airport security methods and for other safety/security purposes. These crowds are needed for accurate representation of real crowd dynamics, while crowds used for entertainment purposes designed with a view to being aesthetically pleasing to the viewer or player. Such crowds will need to have high quality animation and be rendered to an adequate level of detail, as well as behave in a way that the user can believe in.

With increasing demand for realism in video games, the need for increased more convincing behaviour of individuals and the crowd as a whole is becoming more important. A good example of the current state of the art in crowd simulation and behaviour can be seen in Assassin's Creed [BTM+08], where a four layered AI system maintains cohesion of the agents, giving the player a feeling of presence in the crowd, while allowing the individuals in the crowd to make decisions about the main character based on his actions in the environment. However, in many real-time crowd applications like video games, the notion of social agents, or agents that give the impression that they are interacting with each other, is still largely missing. Many games have small numbers of these groups, but they

tend to have repetitive actions and are quite sparse throughout the environment.

The research described in this thesis specifically focuses on real-time crowds. A real-time crowd system, called Metropolis, forms the framework for our contributions. This system is based around an urban environment, largely populated by pedestrians, set in and around Trinity College in Dublin. The purpose behind this project is to create a realistic large, multisensory crowd and to maintain real-time frame rates. The project itself is cross-disciplinary and involves researchers for rendering, animation, behaviour, audio and neuroscience. We will discuss the project in more detail in Chapter 4.

Also of interest, and of particular influence on the work carried out in this thesis, is to use our virtual crowd to help answer psychological questions about how crowds in general are perceived. People are very attuned to the behaviour and appearance of humans, and virtual humans, and a great deal of research has been carried out investigating how to create virtual characters that can be related to in a similar way to real humans. However, relatively little is known about how we perceive crowds in general, let alone virtual crowds. We are interested in finding out what people expect from a virtual crowd; how they expect the characters to behave, how interactions between characters are perceived and how multisensory signals are processed when looking at these virtual characters. For instance, if we are presented with both audio and visual cues, how do we process and integrate these for virtual characters? Does the reliability of one of these signals affect how we depend on the other? Do we expect virtual characters to adhere to social cues coming from the environment around them, as we do for real people?

## 1.2 Research Question

Our research is specifically focussed on real-time crowds for entertainment purposes i.e., we do not necessarily want our crowd to be as photo-realistic as possible, but we do want the user to accept the crowd as being plausible. This can be approached by tackling a variety of technical challenges, such as rendering and animation. However, in this thesis we focus in particular on the behaviour of the crowd. Many crowd behaviour methods have been presented for navigation, collision avoidance and interaction with the environment. Our main aim is to determine what users expect from the behaviour of the crowd. We focus specifically on the overall formation of a crowd and, following on from that, social groups within a crowd.

Very little is known about how we expect a virtual crowd to behave. There are many behavioural properties of humans in a crowd, which every person experiences instinctively, without thinking. However, people are less used to observing a crowd from a distance, as we tend to do in video games, or other real-time crowd applications, so less is known about what properties of crowd behaviour need to be accurately simulated in order for viewers to get a sense of realism from the scene.

In order to populate a crowded scene, random placement of agents in the scene is most likely to appear highly unrealistic to the user. Therefore, characters need to be added to an existing environment in a plausible manner. While this task can be done manually, using real life examples to guide the process, it can be a very time consuming process and is not efficient for populating large areas, with different sized crowds. Our aim is to provide methods to populate urban environments with agents in a meaningful and efficient way. We also examine the role of groups in a crowd, and investigate how individuals interact with each other to give the impression of a real conversation. Humans, by nature, are social creatures and every one of us has a number of interactions throughout our days with other people in different capacities. There are roles and rules that need to be followed for these interactions to happen smoothly, and it is only natural that we would expect the same of virtual humans in a virtual interaction.

The main research questions addressed in this thesis are:

1. How to discover what people expect from a crowd:

   If we create new methods to populate our environment, and create ways for our agents to interact in these scenes, how do know if these are realistic to a user? Very little is known about how virtual crowds are perceived and what people expect from them, so it is important to determine the important properties of a crowd that makes it realistic.

2. How to populate our environment:

   If we wish to populate a crowd scene with agents, what is the best way to achieve this goal? Manual placement will be very time consuming and repetitive and is not practical for a large crowd system. Random placement is computationally efficient, but is unlikely to produce realistic formations. Can we identify a set of rules to follow that produce varied, realistic formations? Does the type of scene we are populating make a difference?

3. How to make our agents interact in a believable way:

   Once we have populated our scene with agents, how can we make the behaviour of these agents appear realistic? If we create social groups using a limited data set of body motions, can we reuse data to create new plausible conversations? If we add audio to these conversational groups, will this help make our new conversations appear more realistic?

## 1.3 Scope

The work described in this thesis considers semi-automatic generation of plausible formations for pedestrian crowds over prototypical areas, based on properties of the areas themselves. We also investigate the generation of conversations between virtual characters using body motions from a limited data set. These conversations consist of body motion alone (i.e., no audio) and multisensory conversation types (body motion and audio together). We do not consider facial or lip motion for our characters, and concentrate solely on small conversing groups, rather than larger groups. Both our formations and conversations are designed to be plausible to the user, and are evaluated against *real* representations in order to assess their usability in a real-time urban crowd system.

### 1.3.1 Properties of the Metropolis Crowd

As mentioned earlier, real-time virtual crowds are the main application for the results discussed in this thesis. There are many different types of crowds, with a variety of practical applications. Some of our desired properties of a virtual crowd and how they contribute to our research goals are:

1. Real-time

   We want to generate socially appropriate behaviour for characters that are part of a large real-time crowd. These characters are considered to be background characters, i.e., not hero characters. Therefore we do not employ very high levels of detail such as facial or finger animation, as we are particularly interested in features that are visible from a distance, or when briefly attending to a group (e.g., when passing by). This includes their formations, body motions and voices. Also, while we want our crowd to be as plausible as possible, the most important factor to take into account when trying to achieve this is that the system needs to run in real-time. Therefore,

we need to take short cuts to give the illusion of a crowd behaving in a realistic manner, while keeping the behaviour as simple as possible.

2. Proximity to camera

   While the characters are primarily background characters and will not in general be focussed on for long periods of time, we still want to be able to navigate through the environment, so characters will be in close proximity to the camera at different times. This means that the behaviour of the crowd characters, both in transit and standing still, needs to be plausible for many different camera angles and for different distances to the camera. So we need to ensure that any methods we implement are robust to different angles and distances.

3. Perceptually plausible

   Many existing crowd behaviour methods need to conform to physical or statistical accuracy requirements. Applications such as evacuation simulations, investigations of crowd safety, urban planning or customer navigation in department stores, need to be extremely accurate in representing the movement and paths taken by each agent around the environment, as well as their reactions to certain events. Our approach is different to these, as we want our crowd to *appear* to be real. We do not necessarily need our agents to emulate the exact behaviours performed by real people in real life situations, but focus rather on what people perceive to be plausible and realistic. Very little is known about what we expect from virtual crowds, so we conducted many perceptual experiments to investigate these issues further. We use data from real scenes and people to create our stimuli.

4. Pedestrian areas

   The areas looked at for the purpose of this thesis are fully pedestrianised in a university environment, so we do not consider interactions between our agents and other factors, such as traffic, or road crossings. This is a major objective for our future work however, and we feel that crowd formation rules that take the context of the location into account could be developed for more complex scenes also. In our crowd system, development is near completion on an integrated traffic system, that has the capability to stop at pedestrian crossings when there are agents crossing the road. Once this system is fully integrated, implementing interaction with traffic and road crossings will be straightforward.

5. Path planning

   The majority of research on virtual crowd behaviour has focussed on the development and optimization of path planning methods, i.e., computing how the agents navigate the environment and how they avoid collisions with obstacles and each other. In our crowd system, path planning and collision avoidance algorithms are fully implemented, so research on this type of crowd behaviour is outside the scope of this thesis. We focus instead on the evaluation of population methods where the type of scene affects the behaviour and on social conversing groups, where research remains scant.

6. Gestures

   We focus largely on conversing groups for the second part of this thesis and how to create new conversational body motions using a limited data set. There is a lot of existing work on conversational behaviours and the different types of gestures that humans use to impart information. Many models use these specific gestures to produce conversational body motion for individual agents. We do not look at our conversational agents at such a low level. Instead of breaking down a conversation into specific gesture types, we take the conversation as a whole and focus more on group interactions such as turn taking, talker/listener roles and attention within the group. Every person knows how to conduct a conversation, and therefore how to read and react to social cues when we are interacting with others. We want to know what social cues are important to represent in a *virtual* conversing group, which should form part of a virtual crowd.

## 1.4 Contributions

A number of contributions have been made to the domain of virtual crowds and our knowledge of the perception of these crowds:

1. Creation of methodologies for populating a static crowd scene. We analysed two prototypical locations found in urban pedestrian areas and derived a number of rules for population of these locations with characters, based on properties of the locations themselves. Position and orientation of characters in these scenes were specifically targetted, as they are particularly salient [Publications 6 and 7].

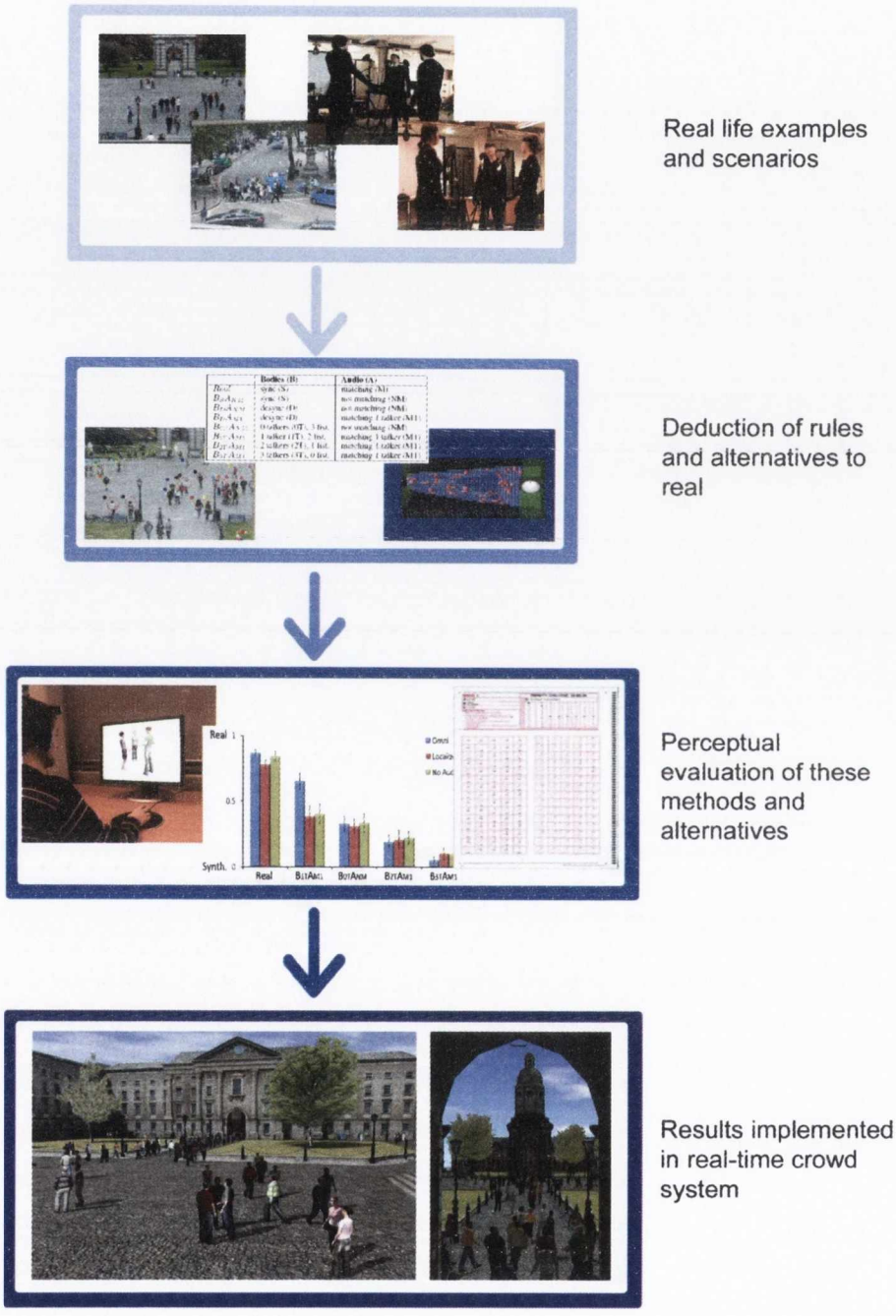



Figure 1.1: Overview of the work carried out for this thesis; outlining our methodology of using real world data against which to compare our rules for evaluation purposes.

2. Validation of these rules across a number of different viewpoints to ensure that they are robust for a navigable real-time crowd system. We also identified camera viewpoints where less detail is needed for behaviour of the crowd, and where more randomness is acceptable [Publication 1].

3. Appropriate methodologies for adding groups of pedestrians in the environment. Metrics for plausible combinations of these groups were found through psychophysical experiments and the results were then implemented in our crowd system [Publications 3 and 5].

4. Motion capture and audio recording methods for 3 person conversations, along with a database of natural free-flowing conversations between 3 people; with both male and female actor groups [Publications 2 and 4].

5. Generation of new conversations for virtual groups using this data-set. This has been developed for conversations using only body motion, i.e., without audio, and also with both omni-directional audio and 3D localised audio [Publications 2 and 4].

6. New results from the psychophysical experiments that were carried out for each of these contributions, which served to validate our crowd formation methods and conversation generation methods and provide insights into the role audio plays in these conversations. These results are not only useful to real-time crowd developers, but also to neuroscientists, as we have added to the body of research on social conversation and humans' ability to localise audio.

## 1.5 Summary of Chapters

The remainder of this thesis has been structured as follows:

- **Chapter 2** provides an overview of previous and related work in the field of behaviour of real and virtual humans.

- **Chapter 3** looks specifically at previous work on group behaviour; human groups and computer graphics applications that employ groups of virtual characters.

- **Chapter 4** gives an overview of our real-time crowd system, Metropolis, and the different areas that are currently under research on the project. Here, we also present our motion capture system, particularly for our multi-person sessions, and outline the pipeline involved in processing the captured conversational data into virtual

representations in our crowd system.

- **Chapter 5** outlines the methodologies used and a series of experiments carried out to evaluate plausible configurations of our real-time crowd. We consider methods to position and orientate agents plausibly in different scenes, investigate the robustness of these methodologies across different camera angles, and investigate the role of groups in a dynamic scene.
- **Chapter 6** focusses on conversing groups, and outlines methods to create new conversations from a limited data set of body motions and conversational audio. We also conduct a series of experiments to evaluate these new conversations, and investigate the role audio plays in multisensory perception of conversations.
- **Chapter 7** summarizes our contributions to the field and also discusses plans for future work.

# Chapter 2

# Background and Related Work

## 2.1 Introduction

One of our main goals is to populate an urban pedestrian environment with a realistic crowd, who look, walk and behave in a realistic manner. More specifically, the work described in this thesis is related to simulating the behaviour of the crowd. Crowd behaviour is a widely researched area, both with respect to humans and animals, for many years. Many different approaches have been taken for different purposes such as video games, where the focus is on making sure the crowd looks aesthetically pleasing to the user, and health and safety crowd simulations, where it is more important to ensure that the movement of the crowd is as accurate to real life as possible in order to identify possible hazards to the safety of the crowd. Other factors will affect the approach taken, such as the size of the crowd, or how detailed the characters and environment will need to be. In this chapter, we will discuss properties of human behaviours, focussing on the behaviour of pedestrians and crowds, and outline different approaches for modelling aspects of human and crowd behaviour. We will also discuss the role of human perception in aiding the field of computer graphics in creating and evaluating methods for realistic behavioural simulations.

## 2.2 Flocking and Herding

Reynolds [Rey87] implemented a distributed model to replicate the motion of a flock of birds or a herd of animals on land. In this system the agents, or "boids", are simulated as

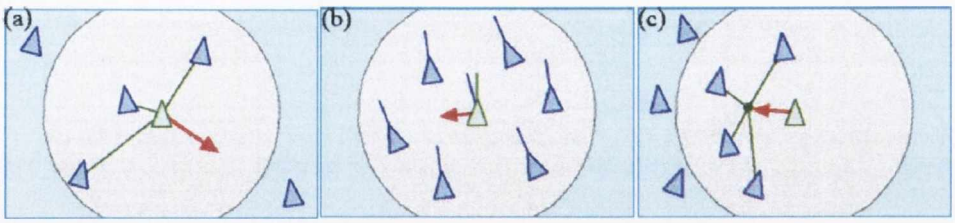



Figure 2.1: "Boids" steering behaviours: (a) separation, (b) alignment and (c) cohesion [Rey99].

an elaborate particle system. They maintain a flocking behaviour by moving based on the positions and velocities of their neighbours. Along with collision avoidance of static and dynamic objects, the boids perform three types of steering behaviours: *separation* to avoid crowding, *alignment* to steer towards the flock's average heading direction and *cohesion* to move towards the groups' average position (see Figure 2.1). Each boid has direct access to the scene database, but only reacts to flock-mates within a certain volume (the agent's perceptual field of view), which is defined by the boid's *distance* from the centre of the flock and its *angle* from the direction of travel. The steering behaviour results in the appearance of the flock moving as a single entity, but with variation within the flock. The repulsive force for each boid for all nearby flock-mates is computed first, then summed together to produce the overall steering force (separation). Next, the average position of all nearby boids is computed, then a steering force is applied towards this position (alignment). Finally, the average unit forward vector is found for all nearby flock-mates. The difference between this vector and the boid's current vector is applied as a steering force (cohesion). The sum of these three steering force vectors is the combined steering vector. Better control is achieved by normalising and controlling 3 different parameters for each steering behaviour (weighting, distance and angle). The boids behaviour model was also extended to include new behaviours such as Seek, Flee, Pursuit, Evasion and Arrival.

Tu and Terzopoulos [TT94] developed a framework for the behaviour and motion of virtual fish. The movements of the characters were based on a physics based fish model using a mass spring system. This simulated muscle control mechanism was designed to replicate that of fins. The behaviour of the characters was based on their perception of the environment, their mental states and habits. The user defines the preferences for the fish character at start-up for likes and dislikes of factors such as brightness and cold. A mental state is then defined by three variables: Hunger, Libido and Fear. The sensory perception of the fish is a 300 degree field of view, which can select only useful information about the world such as object colours, distances and sizes. Based on these three inputs,

an intention generator chooses and executes a behaviour routine, which in turn runs the appropriate motor controllers for realistic motion of the fish (see Figure 2.2 for an example of an intention generator and final images for a predator fish).

Other flocking behaviours have been implemented by using point mass systems, robots and cyclists [BH97, BMH98]. The algorithm for grouping behaviours implemented consists of two parts: a perception model to determine the neighbours of the agent and an algorithm to find the desired position based on the velocities and positions of the other agents. A lower level control algorithm determines the velocity necessary to achieve that position and tries to apply that velocity to the agent. Each agent in a group can "see" the locations and velocities of the set of all other agents in the group and uses this to compute a desired position relative to them. This position maintains a desired separation from the other agents. Then the set of desired positions is weighted by the distance between two agents and averaged to compute a desired group position. When taking obstacles into consideration, the effect of an obstacle on the movement of the group plus an offset is added to the desired position, thus updating it to avoid the obstacle. This algorithm was then amended to be applied to three different dynamic group types: point mass system, one legged robots and cyclists.

Bayazit et al. [BLA02] used similar methods to perform local level flocking behaviours, and introduced the idea of global roadmaps for new grouping behaviours. The use of roadmaps allows the representation of global information for navigation, using adaptive roadmaps to enable communication between agents. Rules are then associated with roadmap nodes and edges to allow specification of local behavioural rules. Using the rule-based roadmaps, additional behaviours are added to the group: *homing* (reaching a pre-defined goal), *goal searching*, defining how agents position themselves relative to other flock members when going through a narrow passage and *shepherding* (where an external agent influences the behaviour of the flock).

## 2.3 Human Crowds

The aim of this research is to create behaviours for virtual humans and crowds in a virtual urban environment. The most important element of this goal is the perceived realism of the crowd and therefore it is critical for us to investigate pedestrian behaviours and tendencies in real life scenarios. Much previous research on this topic has been carried out in the fields of urban planning and civil engineering.

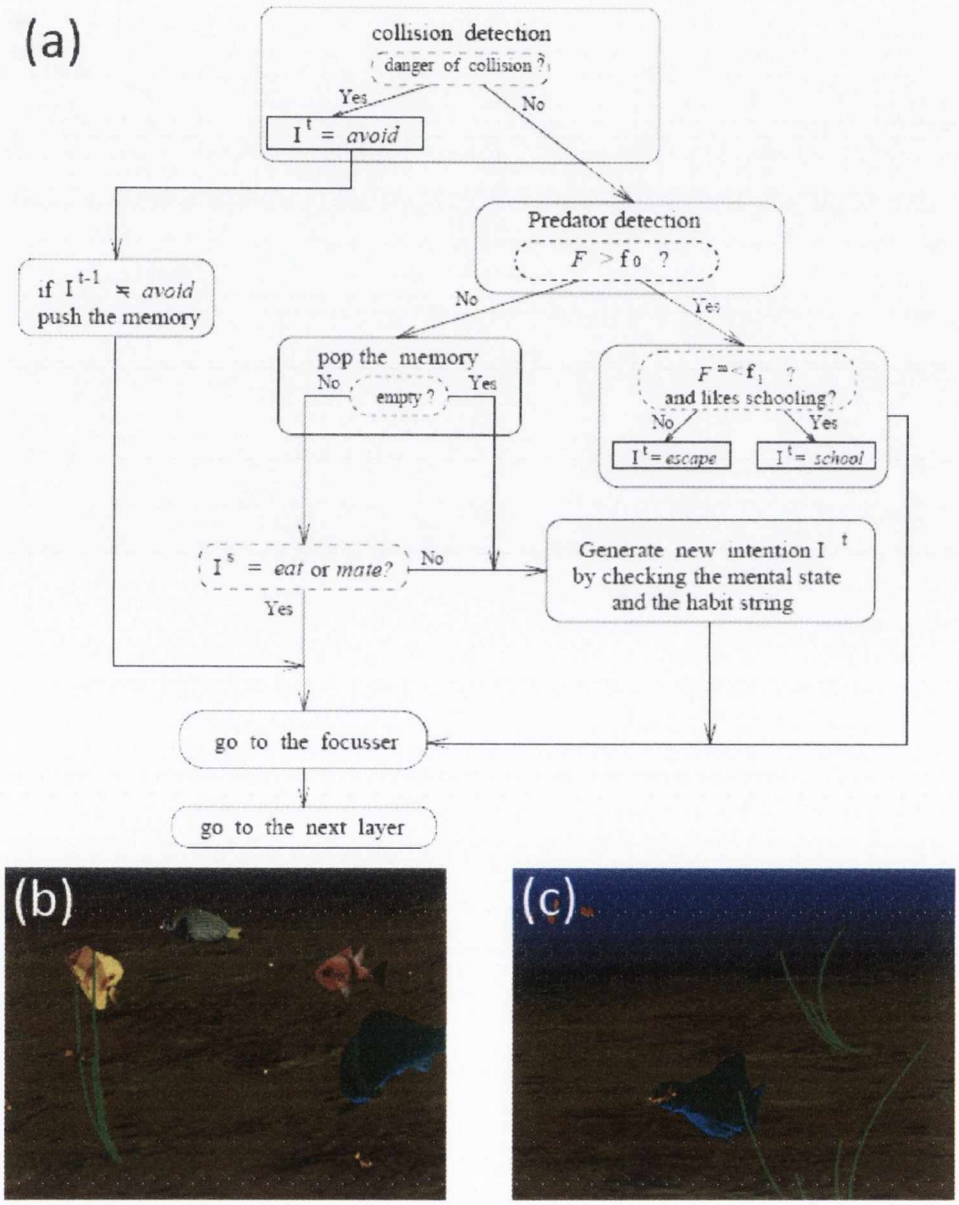



Figure 2.2: (a) Intention generator for predator fish and (b,c) resulting behaviours of stalking and eating prey [TT94].

### 2.3.1 Urban Planning Research

Zacharias [Zac01] discusses assumptions about how pedestrians perceive their environment, how they plan their walking itineraries based on this perception and how urban planners make use of these assumptions. From a town planning perspective, planned walking environments are becoming more and more important since the increased level of motorised transport in our cities and with shifts in culture leaning towards higher levels of consumerism. Of particular interest to urban planners is density and direction of movement and how pedestrians' perception of their environment affects the choices they make when it comes to planning their journeys. They are also interested in particle models of pedestrian movement, which allow for those random behaviours of pedestrians that are inherent to the pedestrians themselves as opposed to being caused by properties of the environment around them. This is especially of interest in the area of traffic engineering. Engineers also study the volume flow of pedestrians in open spaces. Since pedestrians do not move in evenly distributed patterns, realistic designs need to be established to ensure that levels of facilities in different areas are appropriate to the pedestrian traffic volume [Khi94].

As pedestrians become more familiar with the spatial layout of an area, they will use this knowledge to minimise path lengths from one destination to another. For urban planners, this is important as business areas can be designed around these minimised distance paths that are used more often [Khi90]. From the perspective of creating a behaviour system, it may be an important factor for a realistic system that agents appear to learn from the environment around them and use minimised distance paths. However, path minimisation may not be used by all pedestrians, for example tourists will be more likely to explore an area rather than trying to reach a final destination as quickly as possible. Also, when people are exploring an area, direction changes tend to be small to avoid getting lost. Therefore, a mixture of tourists and business characters in an urban environment could be defined by parameters such as walking speed, direction changes and path minimisation decisions.

Environmental aspects such as aesthetics, accessibility, climate and lighting affect pedestrian traffic volumes and walking speed in different areas [Geh03]. Urban planners focus on the image of an area, its location in relation to obtaining a desirable climate, ensuring that it is easily accessible and well lit to ensure maximum pedestrian traffic. For a virtual environment, it may be possible to use these real life tendencies to define weights for different paths for a virtual crowd.

Ambient sound is also a factor that affects pedestrian behaviour [KG80]. It has been shown that when the level of input from the environment is increased, helpfulness of pedestrians decreases. Korte and Grant studied this effect and found that when traffic noise is increased, pedestrians' peripheral attention is decreased and they tend to increase their walking speed. Since our crowd system is also focussing on creating realistic sounds, with appropriate sound levels throughout the city, this type of information could be extremely important in creating realistic behaviour patterns.

Traffic engineering research on pedestrian behaviour has identified four main areas of pedestrian behaviour [DH03]. These areas are:

1. **Use of Space and Walking Speed**: Walking speeds of pedestrians are affected by *personal characteristics* (e.g., age, gender, size etc), *journey characteristics* (e.g., purpose, length) *infrastructure properties* (e.g., type, shelter), *environment characteristics* (e.g., ambient sound, weather conditions) and *pedestrian density*. Pedestrians' lateral and longitudinal spatial use are also important to consider, since pedestrians rarely travel in a straight line. Deviations will increase with faster walking speeds and the relationship between longitudinal space needed and speed of the pedestrian is shown below, where V is walking speed, $V_f$ is average free walking speed (approx. $1.34m/s$), $A$ is the required area (comprised of longitudinal and lateral space) and $A_{jam}$ is the largest area for which walking is possible (approx $0.19m^2$) [DH03]:

   $$A(V) = A_{jam} - 0.52ln(1 - \frac{V}{V_f})$$

   (Original reference: [Wei93])

2. **Dynamic Lane Formation:** The phenomenon of the emergence of bi-directional flows of pedestrians in crowds is known as dynamic lane formation [Old68]. These lanes are usually formed on the right hand side, for both left-hand and right-hand traffic regulated countries, and apply to corridors, pathways and crossings.

3. **Intersection Stripe Formation:** This refers to pedestrian behaviour at an intersection, where two pedestrian flows can intersect with each other, and form small lanes within this intersection. This means that neither group needs to stop to allow the other to pass. For an illustration, see Figure 2.3.

4. **Interaction and Walking behaviours:**

**Scanning**

Goffman [Gof71] described how pedestrians observe the environment around them using a subconscious scanning process. The scanning process only extends to the pedestrians immediately surrounding a person and is usually elliptical rather than circular, narrow either side and longest in front of the person. This ellipse changes depending on the density of the surrounding pedestrians. The idea of using scanning to avoid collisions, as with all interactive walking behaviours, will only work as long as there is *co-operation* between pedestrians. There are two moments critical to avoiding collisions when walking among pedestrians. The first, *emission of critical signs,* communicates to others what a pedestrian is going to do, e.g., a small movement of the shoulder to indicate a turning direction. The second, *establishing moment,* refers to the recognition by both pedestrians that critical signs have been exchanged. Once these two steps have occurred, changes in the pedestrians' courses are carried out.

**Side-Step**

This was described by Wolff [Wol73] as the *step-and-slide movement*, occurring mostly between members of the same gender. Pedestrians that are interacting will not necessarily aim to completely avoid contact with others. They will angle their bodies slightly, turn shoulders and take a tiny side-step rather than change their course completely. This will likely result in a slight brush with the other pedestrian, but will only avoid collisions if there is *co-operation* between interacting pedestrians.

**Granting Space**

Pedestrians will also grant a certain amount of space to others [DS75]. It has been shown that they tend to give a wider berth to groups than to individual pedestrians. Dabbs and Stokes also suggest that pedestrians grant more space to men than women. However, Sobel and Lillith [SL75] observed that women are granted more space than men. Willis et al. suggest that gallantry plays a role in deciding who gives way to avoid collisions e.g., younger groups give way to older groups.

While the literature is consistent on this point, there are conflicts with other research regarding whether more space is given to different genders, ages and ethnicities.

Figure 2.3: Demonstration of stripe formation in an intersection of two dynamic groups [HJ09]

#### Bottlenecks

Dampen and Hoogendoorn [DH03] use controlled experimental approaches to derive human pedestrian behaviour. They found that at low density conditions, pedestrians will walk through the middle of a bottleneck, maximizing the space between them and the walls. During higher density conditions, they found that pedestrians tend to walk diagonally behind each other maximizing the space for others. They also noted the tendency of pedestrians to spread out in a wider formation at either end of a bottleneck.

## 2.4 Models for virtual human behaviour

Once high-level patterns in human pedestrian behaviour have been identified, it is then important to consider different methods for implementing this behaviour for virtual humans. This section looks at different models implemented to simulate aspects of human behaviour in virtual human characters.

### 2.4.1 Macroscopic models for virtual human behaviour

There have been many models for replicating the movement of humans, many of which have been investigated by Helbing et al. [Hel91, HFV00, HJM+06, Hel92, HM95, HBJW05]. A mathematical model was implemented considering velocity, attracting and repulsing forces, unexpected detours, queuing behaviours and the formation of groups [Hel91]. Another model used fluid dynamics to recreate the collective movement of pedestrians, giving interesting results such as those mentioned in Section 2.3, specifically in relation to bottleneck streams and the development of flow lanes (see Figure 2.4) [Hel92]. Internal motivations, or social forces, have been investigated, looking at acceleration towards a desired velocity, maintenance of personal space from other pedestrians and modelling attraction

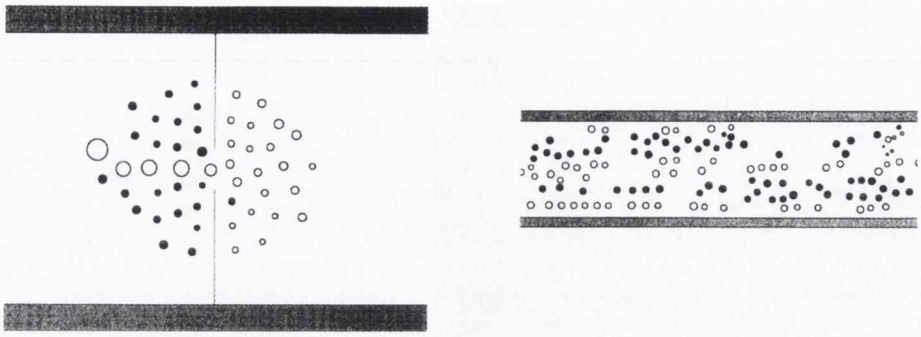
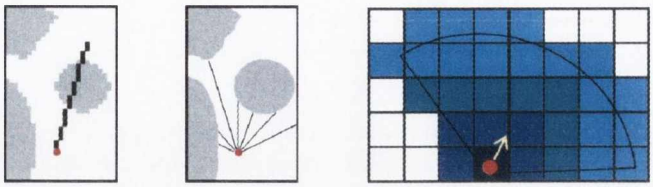


Figure 2.4: Results from Helbing and Molnar's social force model for pedestrian dynamics showing: (left) bottleneck streams and (right) flow lanes. Black circles are travelling from left to right and white circles are travelling right to left. The size of the circle indicates velocity [HM95].

effects towards other pedestrians or objects in the environment [HM95]. Mathematical representations for these social forces are given and a factor is added to allow for *fluctuations* to represent the element of randomness that is present in human behaviour. Particle models have been applied to represent continuous bottleneck flows and the preconditions and consequential behaviours of panic, such as interactions between individuals becoming more physical, jamming and clogging at exits, fallen people turning into obstacles [HJM+06, HBJW05].

Antonini et al. [ABW06] presented a discrete choice model for short term walking behaviour. This model is made up of: a *choice list* of a discrete number of possible speeds and directions, a list of *attributes* describing the choices, *socio-economic* characteristics to describe the decision maker and a term to allow for randomness.

Shao and Terzopoulos [ST05] integrated motor, perceptual, behavioural and cognitive components for individual characters in a virtual train station. The environment itself is represented hierarchically with a topological map representing links between different regions, a perception map for local information about stationary objects within a region and path-maps with quad-tree mapping for long and short range path planning with the A* algorithm. The A* algorithm computes the shortest path between nodes from an initial position to a goal position. It takes into account not only the distance from the current position to the next node, but the overall distance from the initial position. The pedestrians in this environment perceive and analyse the environment around them, using it to make decisions about their behaviour. They do this by making perceptual queries about ground height and stationary and dynamic obstacles (other pedestrians) of the perceptual



(a) (b) (c)

Figure 2.5: Sensing objects: (a) examining map cells along the rasterized eye ray, (b) fan of eye rays used for broader view and (c) perceiving dynamic objects with a tiered sensing fan [ST05].

environment description level in different ways (see Figure 2.5). Six reactive behaviours are defined to connect actions to these perceptions depending on whether the obstacle is dynamic or static, or how close it is. The use of these six behaviours allows for the emergence of real human behaviours such as flow lanes and queuing. Higher level motivational and navigational behaviours are applied to enable the pedestrians to navigate the environment autonomously. The highest level of behaviour is cognitive control, incorporating global path planning and a goal-stack mechanism connecting cognitive and behavioural controls. This was simulated with 1400 characters in real-time.

Blumberg and Galyean [BG95] used the idea of creating a sensory system to allow the character to perceive the environment around them. They also integrated a level of directability of the character to allow the user to interact with and direct the virtual character at different levels. This directability is added as a motivational control, where the user can change these internal variables of the character that represent its goals. The user can also access different parts of the behaviour of the character during run-time and execute specific demands. Furthermore, the user can add imaginary sensory input to the characters' perception which may trigger certain behaviours. This was implemented in the ALIVE project [MDBP95] for a dog character, as opposed to a humanoid character, but it still provides insight into different methods of applying and controlling the behaviour of a virtual character.

Niederberger and Gross [NG05] developed a level-of-detail concept for behavioural modelling for reactive and proactive agents. They vary the time spent on behavioural modelling per rendered frame depending on the visibility of the agent. Agents that are visible are assigned more behavioural simulation time, resulting in more sophisticated proactive behaviours than for agents in the distance or out of view. Figure 2.6 gives an overview of their system. A level-of-detail *classification* is assigned to each agent (a), while at the same time

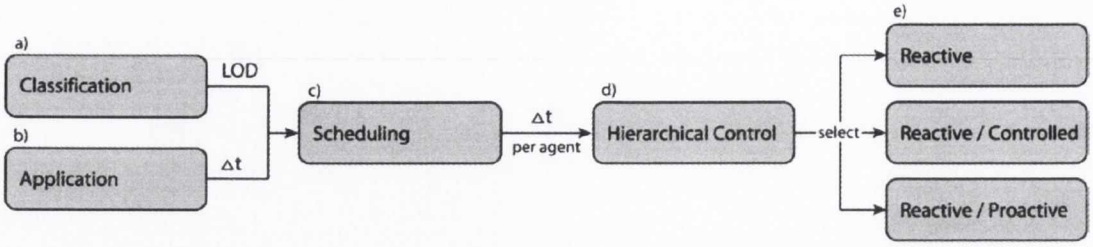



Figure 2.6: Overview of the level-of-detail method for reactive and proactive real-time behaviour described in [NG05].

the *application* receives an allotted amount of time to invest in computing the behaviour (b). A *scheduling* algorithm decides which agents should be active and assigns behaviour computation time to each active agent (c). A *hierarchical control* decides whether that agent is independent or whether it will follow an agent higher in the hierarchy (d). Then, either a *Reactive* (e.g., avoid static obstacle), *Reactive / Controlled* (e.g., stay within a group) or *Reactive / Proactive* (e.g., perform high level goal) behaviour will be selected, based on the previous 4 steps.

### 2.4.2 Path planning and Navigation

One of the most important factors in implementing behaviour for virtual humans is enabling the characters to navigate the environment they are in. For data-driven approaches, such as ours, path following can be restricted by the range of motion-captured data available for animating the character for different movements. Kamphuis et al. [KPOL05] presented a method of path planning for a single character taking animation constraints into account. A control space for angular and linear velocity is constructed from the motion capture data consisting of a set of points. Each point corresponds to one motion capture clip of a walk cycle. Then, given an input vector with the desired locomotion, the three nearest control space motion points are selected. A barycentric interpolation of those selected points is then carried out to determine the output animation. The input vector can be changed continuously, thus allowing real-time animation for the character.

Pettré et al. [PSL02] have also addressed the problem of planning virtual human motion considering animation restrictions. The navigation of the character through the environment is a composition of Bezier curves using control points. Given initial and final positions and orientations, the control points are calculated with respect to the distance between the initial and final positions. Once the path is planned, a sub-controller syn-

thesises a walk animation for the computed path using characterised motion capture data files for each frame.

Kuffner [Kuf98] split the problem of navigation into two parts for a dynamic environment with an even terrain: goal directed path planning and following. The algorithm consists of four computation phases: *initialization* refers to the preprocessing of motion capture data for a single linear walk cycle, with other inputs being a 3D description of the environment and a goal; the *projection* phase involves projecting a 2D representation of the 3D environment; the *path search* phase finds a collision free path from the initial position to the end goal using Dijkstra's algorithm or a modified A* method; in the final phase, *path following*, a PD controller for position and velocity along with motion capture data is used to synthesise the final motion of the character for the determined path.

Brogan and Johnson [BJ03] developed a behavioural model for realistic path planning using observed paths of human pedestrians in two controlled environments. The motivation for using human walk paths was to allow higher level control of the generation of animations (such as seek and avoid), along with an implied increase of realism of these animations. They use three evaluation metrics to validate their paths compared to observed paths: distance error, area error and speed error. When compared to the commonly used A* algorithm, they found that this new model performed better for both distance and area errors.

The methods discussed so far only allow for path planning on flat terrains. Recently, methods have been developed to allow for ceiling and floor constraints for real-time path planning for virtual humans [Lam09]. The topology of unstructured 3D triangular meshes is extracted by analysing them using a prismatic spatial subdivision based on Delaunay triangulation [BY98] and allowing for human constraints (e.g., ceilings, steps). At runtime, optimised path generation is performed as well as footprint placement algorithms for the animation. Path generation for a given topological map is done by localising the initial and final positions inside the topological map, then connecting them via waypoints generated by the segments of the map from their cells using A*.

## 2.5 Virtual Crowds

There are a number of challenges when it comes to simulating a large scale crowd system such as Metropolis. The environment is complex and there is a large number of pedestrians to consider. This does not only lead to challenges for behaviour, but there are many

challenges for animation and rendering also. This section will look at other crowd systems and outline the different challenges faced and solutions proposed to generate human behaviour models on a large scale for virtual crowds.

There are a variety of approaches to simulating a scalable crowd system. Sung et al. [SKG05] investigated the problem of environment navigation for a large number of pedestrians. They considered the demands on animation of the characters for scenarios where the agents need specific positions, orientations and body poses at different times. They do this by representing the space of possible motions on a motion graph allowing seamless connection of different clips. The path planning for the agents is done using probabilistic roadmaps to navigate the complex environment and produce approximate motions to satisfy environment constraints. A search algorithm is then used to find a motion that conforms to the approximate motion. Their algorithm also allows for continuous adjustment of the agents position, orientation and speed, so as to allow for animations outside the restrictions of the motion graph. This algorithm is scalable up to 300 characters in real-time. However, Metropolis does not currently have the constraints of needing agents in a particular place at specific times so this may not be applicable to our project.

Stylianou et al. [SFC04] adopted a top-down approach for simulating a large number of pedestrians. The movement of the agents in a virtual city is computed at a global level to maintain flux and density. At a lower level, the city is made up of a network of nodes describing the walkable areas and the flow of the agents is modeled as a random walk, keeping track of the numbers of agents entering and leaving nodes. The aim here is to control the densities inside each node, so if an agent cannot yet leave a node, they will wander or stop to look at attractions until they can leave. This method is appropriate when it is not necessary to have close interaction between the characters in a virtual environment but would not be appropriate for Metropolis as the agents do not have higher level goals or path planning, and do not interact with each other in any way. It also yielded slow frame rates for 7000 agents due to the low level collision avoidance expense.

### 2.5.1 Hierarchical models of agents for crowd simulations

Thalmann [Tha01] describes a bottom-up approach to simulate the dynamics of human environments by defining agents in a crowd as autonomous and independent, thus allowing them to interact in an evolving environment in real-time. The agent architecture divides the agent simulation into low-level simulation factors, such as locomotion or object interaction, and higher level behaviours consisting of goals, beliefs, plans and internal states.

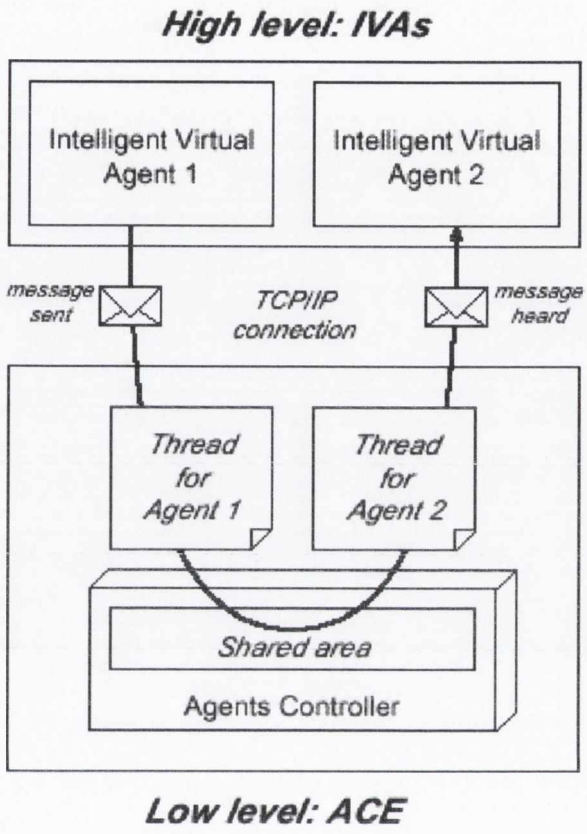



Figure 2.7: Overview of architecture to represent agents in a crowd with high level behaviours and low-level physical simulators [Tha01].

The latter are based on Rao and Georgeff's Beliefs, Desires and Intentions (BDI) architecture [RG98]. Agents share controllers for the low-level simulation, but not the high level. When two agents communicate, the message goes from the high level, or Intelligent Virtual Agent (IVA), to the lower level or Agents' Common Environment (ACE), where it is placed in the shared area (see Figure 2.7). A second agent can then retrieve the message and select the correct action to perform to achieve its internal goals. The internal state of the agent can evolve over time so that the agent can react dynamically.

Hierarchical models are presented by Musse et al. [MGT99, Mus00, MT01]. They implemented a crowd system, called ViCrowd [MGT99, Mus00], to model and generate crowds where the autonomous nature of the individuals in the crowd could be of more or less sophistication. Different types of control were applied to the agents depending on the objective of the simulation: *programmed* (pre-defined behaviours), *autonomous* (rule-based

behaviours) and *guided* (interactively controlled). In a guided crowd, the user could control the crowd via a series of commands, communicated through a Textual User Interface, whereas the autonomous crowds followed rules for inherent behaviours such as goal seeking and collision avoidance. The idea of a hierarchical model was then applied to the crowd itself, not just the behaviours [MT01]. The structure of the hierarchy was *crowds*, *groups* and then *individuals*. The groups were the most complex of the three structures and within the groups were the different levels of autonomy referred to above. This allowed the crowd simulation to be adapted for different purposes, with simple behaviours implemented to simulate large crowds, or more complex behaviours implemented for increased realism. The complex behaviours included goal changing, attraction to objects in the environment, repulsion from other agents or objects, internal beliefs and goals and emotional status allowing variation in animation.

### 2.5.2 Navigation and path planning for crowds

Arikan et al. [ACF01] present a path planning algorithm for large numbers of agents in a game-type virtual environment. The application also employs culling techniques by using a *proxy* simulator to conduct path planning and collision avoiding computations for areas of the environment that are invisible to the user. Pre-computed data structures are used to speed up path planning for all fixed obstacles in the world when an order is given by the user to go to a particular destination. The path chosen is then altered to allow for collision avoidance with dynamic obstacles at run-time. Then, when the agent is moving, collision detection is computed at each frame and if a collision is detected, the path is altered. The proxy simulator keeps track of each agent's approximate location and the environment is divided into cells. The user's viewpoint is an overhead rectangular view, so only objects in the cells on this rectangle are visible. When an object leaves a visible cell, it is switched to the proxy simulator and vice versa. The agents using the proxy simulator can perform 4 events: stop (go from dynamic to static), replan (compute another path plan if not at destination), entry (enter into visible cell) and reinsert (at time intervals, possible interactions are recomputed, due to delays in the path finding for agents in the visible cells).

Sud et al. [SAC+07, SGA+07] have investigated real-time path planning for crowds of agents in dynamic environments. They use *Adaptive Elastic ROadmaps* (AERO), global roadmaps that continuously update to take dynamic obstacles and inter-agent interaction forces into account and compute collision-free paths in complex environments. In order to

make the navigation more efficient, they use *link bands* to alter the local dynamics and resolve multi-agent collisions. AERO defines a set containing each agent in the environment with a position, radius and goal position, and also a set of static and dynamic obstacles. They restrict the motions of the agents to the free space in the environment and, based on the states and goals of the agents, define a sequence of states for all agents in the free space. In order to avoid the problems that occur when using local methods of navigation, they use a global method. The road-map is time varied using milestones and links to make the global navigation more efficient for each agent. Each milestone is the position of an agent, and a link connects two milestones along a path in a curve. Then, a graph search algorithm is used to compute a path between two configurations.

Paris et al. [PDB06, PPD07] have also investigated the topic of path planning for crowds of agents. This approach involves topological extraction of the environment. The environment is divided into three different layers representing cells, groups of cells and zones of groups for three different types of nodes (corridor, dead-end and crossing). The path planning consists of 3 steps: first, plan from current zone to destination, then plan within the first two zones, then plan within the first two groups. This allows efficient path planning as the highest level is computed, and the lower levels can be computed as the agent moves through the environment. Reactive path planning is also implemented as necessary if the environment changes.

Ondřej et al. [OPOD10] present a local collision avoidance steering algorithm for crowds based on synthetic vision for the agents. They use a *bearing angle* (angle between the agent and obstacle) to determine whether a collision will occur, which happens when the derivative of this angle is zero. They also calculate the time-to-interaction (*tti*) which is the time remaining before the minimum distance between an agent and the obstacle is reached. Based on possible collisions "seen" by the agents (from bearing angle), and the level of danger associated with the collision (tti), they will change their direction and slow down to avoid collisions within the environment. Simulations of this method show self organising patterns such as lane formation from two opposing groups of agents in a corridor type scenario.

Loscos et al. [LMM03] have also implemented local laws for collision avoidance, and focus on generating realistic motion for densely populated environments. They represent the environment using a binary image with black for the buildings and white for the ground. Then a convolution filter is used to define an area around the buildings as pavement. Goals are then placed at the corners of pavements and pedestrians are placed next to them at run-time and assigned a new goal to reach adjacent to their goal. Collision

detection is performed by comparing the direction trajectory and velocity of each agent and the distance between each agent. Depending on these factors, an agent will deviate by an appropriate angle, slow down or stop. They also create lane flows by having agents follow the trajectory of agents recently in their cell when there are high densities. Tecchia et al. [TLCC01, TC00] have proposed a fast algorithm for collision detection also and implemented this in an Agent Behaviour Simulator (ABS), which is a platform for urban behaviour development. The platform consists of four layers: *inter-collision detection layer* (collisions between agents), *collision detection layer* (collision with obstacles), *behaviour layer* ($2^{32}$ behaviours can be coded on a colour map and an agent can choose the correct behaviour for the cell it occupies) and a *callback layer* (more complex behaviours for certain cells such as boarding a bus).

Lerner et al. [LCL07] set out to generate crowds that displayed varied behaviours individually without defining an explicit behaviour model. They used a data-driven example based approach to achieve this, allowing the agents to learn from real-world examples. A database is constructed from input video of real world pedestrian behaviours by manually tracking pedestrians in the video to generate a set of trajectories, which are stored as examples in the database. At runtime, the database is queried for similar examples that match those of the simulated pedestrian and the closest matching example is selected as the resulting trajectory. Lee et al. [LCHL07] use a data-driven approach to simulate virtual human crowds imitating real crowd behaviour. They recorded crowd videos in a controlled environment from an aerial view. Users must manually annotate video frames with static environment features and can semi-automatically track multiple individuals in order to provide their trajectories. This data informs an agent movement model used to simulate a crowd that behaves similarly to those observed in the video.

### 2.5.3 Variety and roles in crowd behaviour

In any large crowd system, one of the big challenges to improve realism is to create an impression of variety. This has been researched in terms of animation and rendering [MLD+08, TOY+07] as well as for agent behaviour. Durupinar et al. [DAPB08] map behaviour parameters to personality traits, thus allowing the user to control the behaviour without needing intricate knowledge of the crowd system parameters and creating variety in the behaviours of the virtual crowd. This was carried out in a Hi-Density Autonomous Crowd Simulator (Hi-DAC) which has also assigned leader and follower roles to agents in a crowd evacuation simulator [PB06, PAB07]. Here, the agents behave in different ways

depending on whether they are leaders or followers. To navigate the environment, agents will communicate paths to each other or, depending on their role, will either find another exit or follow a leader's movements (see Figure 2.8).

The use of roles to control the behaviour of agents has also been used as an alternative to scripting behaviours by hand [MDCO02] and in conjunction with psychological models as a means to communicate information about the environment between agents [POSB05]. Other psychology based models assign vision and attention capabilities to agents thereby building a mental model of the world and the surrounding agents [RD05b, RD05a]. After time, if an agent has not "seen" a certain agent, that agent will disappear from its mental model. Collision avoidance is carried out for the neighbouring agents, but is not always accurate, e.g., when agents are behind a wall. Accuracy is also lost for a larger number of agents, as each agent in view is monitored less frequently as the number of visible agents increases.

Ju et al. [JCP+10] present a crowd model that can morph between different crowd formation behaviours for a range of crowd sizes. This model takes both example data from processed crowd videos, and results of a simple rule based crowd simulator as inputs. A formation model then determines the group formation from the local position of a number of individuals relative to their neighbours, which construct a formation distribution. This is done for a number of sample formations, so each distribution has a number of samples. A trajectory graph is then used to tell the model how each individual should move. The trajectory graphs contain a short sequence of 2D points that represents the movements of randomly sampled crowd data. The morphable model then generates new simulations for an arbitrary crowd size based on this source data. One simulation generates a static formation, which can serve as an initial crowd configuration. Another provides a dynamic crowd, travelling along the selected trajectory.

## 2.6 Informed Environment

To guide characters' behaviour in a virtual environment, information and behaviours can be embedded in the environment itself, so that different objects are associated with different actions. This method has the advantage that, for environments that do not change, it is known what each objects' function will be. It also allows for a certain amount of pre-processing, which can lighten the load for behaviour computation at run-time. This can be a great advantage for real-time systems.

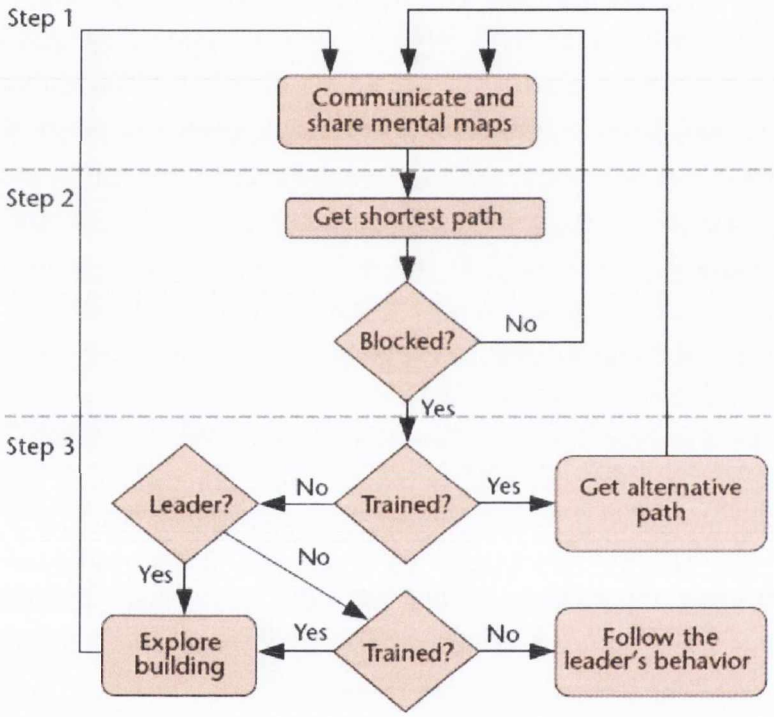



Figure 2.8: Overview of evacuation navigation algorithm in Hi-DAC [PB06].

Thomas and Donikian [TD00] presented a model for simulating virtual environments that contain sufficient information to allow the characters within the environment to carry out different appropriate behaviours (driving and pedestrian). The environment is represented by two structures: *hierarchical* and *topological*. The hierarchical structure contains the breakdown and relation of the composition of the town, i.e., a town is composed of quarters, which in turn are composed of road networks and so on. This continues down to the level of buildings and free spaces for pedestrians, and lanes and intersections for driver characters (see Figure 2.9). The topological structure contains graphical information to allow agents to navigate the world. The world contains three different types of regions: constrained (e.g., lane, corridor or sidewalk), intersection (where characters can change routes) and free regions (no circulation restrictions). Each region type is associated with a different algorithm that specifies the locomotion of the character appropriate for that region. This idea was extended by Badawi and Donikian [BD07] and applied to agents' interaction with individual objects within an environment. Here, Interactive Surfaces (IS) were used to describe surfaces of interest on objects and the space affected by interaction with the object (for example, a door would have the handle as a surface of interest, and the arc it follows when opened as the space affected). A set of basic actions are defined and each object describes its individual interaction process through a combination of these actions.

Farenc et al. [FBT99] also used the idea of a hierarchical environment to represent a virtual city and apply semantic behaviours to different regions. They designed the environment in a similar way as described above, but also used lures and smart objects to provide information about behaviours. Lures could be either bounding box information for collision avoidance, or tagged with action knowledge. For example, seating in a U-shape would contain a bounding box which also covers the area inside the U-shape, preventing characters from using this area. Defining the seating as a group of lures allows the pedestrians to enter the space between the seating, since it is not defined as part of the area for seating. Action knowledge contains information to help generate motions. For example, a stairs would contain number of steps and the height of the steps, to help parameterise a climbing motion. Kallmann [Kal01] proposed another feature of smart objects: interaction features, which allow a smart object to describe not only interaction surfaces, but their functions such as movements and purposes. Interaction features are *intrinsic object properties* (physical properties like weight, moving properties and a text description of the general purpose of the object), *interaction information* (information for actors on how to interact e.g., interaction parts and actor positioning), *object behaviour* (describes the reaction of an object to different interactions) and *expected actor behaviour* (e.g., expected position for characters to be in before interaction). Smart objects were employed by Pe-

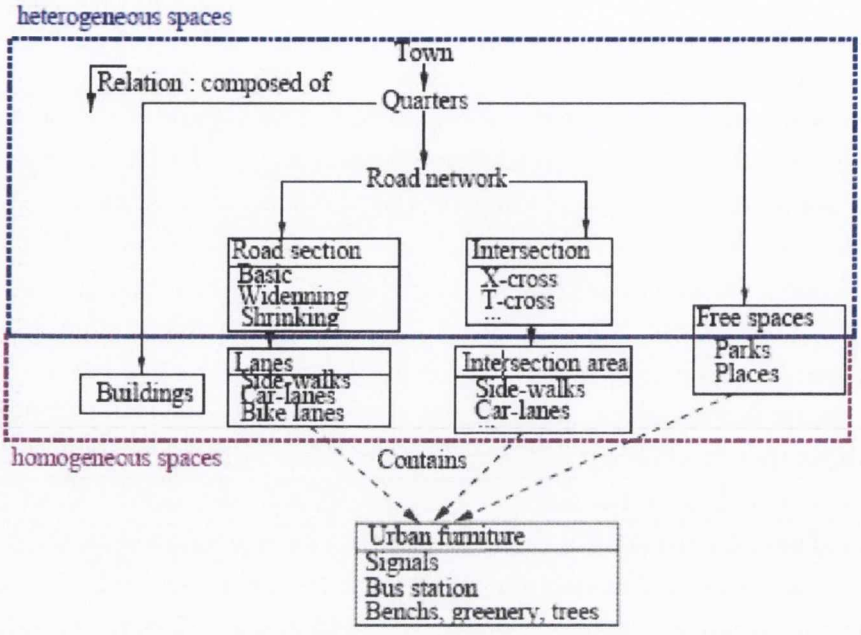



Figure 2.9: Hierarchical structure of a town according to Thomas and Donikian [TD00].

ters et al. [PDMO03] as the basis for a framework for modelling interactions, with the addition of information for directing gaze behaviours for the agents for both top-down and bottom-up attention capture. They also allowed for multiple agents interacting with the object, with instructions on how to interact with each other simultaneously.

## 2.7 Perception

A vast amount of knowledge exists about the perception of how humans behave, how they walk and interact. The work described in this thesis builds on the steadily growing body of research that combines these two fields and applies knowledge about human perception of virtual environments and characters to computer graphics simulations. This chapter will give a brief outline of research into human perception in computer graphics and will present different applications where it can be applied. We will also discuss perceptual studies specifically related to virtual humans and crowds and their relevance, particularly for real-time applications.

### 2.7.1 Perceptually Guided Graphics

Many different types of computer graphics applications can use perceptual evaluation as a guide to improve their simulation. Knowledge of how a user views a virtual scene can be used to draw attention to objects the developer wants the user to focus on, or detract attention from objects or anomalies in the scene the developer wants to hide from the user. Special tasks can be set to determine thresholds for many different applications e.g., how many different characters are needed for a large crowd to appear varied, or how many characters need to depict an emotion before the crowd as a whole appears to convey that emotion.

These types of studies are also very useful to create an increased sense of presence in virtual environments. In these environments, the user typically uses a Head Mounted Display (HMD) to view a virtual world, blocking any visual information from the real world around them. A realistic virtual environment viewed correctly through a HMD can give the user a real sense that they occupy the virtual space around them, especially when their movements in the real world can be tracked and applied in the virtual environment. These applications have many uses; both for entertainment purposes and more serious aims, for example to help rehabilitate patients with brain injuries [TSN+05], or to motivate spinal injury patients to aid recovery [BR00]. Virtual environments have also been known to help people with serious phobias such as a fear of heights [RHK+95], or a fear of flying [RHW+96]. However, creating a sense of "presence" can be quite difficult, as there are a few well known issues in delivering a virtual environment with an adequate level of realism.

There are some issues with using HMDs for these virtual displays, in that there is a known problem with depth and distance perception; people tend to underestimate depth when using these displays [JIS+08, WRMW95]. There can also be problems with time lagging between the movement of the person in the real world and the virtual representation [MMTS10]. These and other factors can affect a user's sense of presence and can hinder the effectiveness of the environment. As a result, many perceptual studies have been conducted to measure the effect of these kinds of anomalies on how involved the user feels in the environment or how well they are able to perform a given task, thereby providing insights into how to improve the experience for the user.

For complex and highly detailed scenes, perceptual methods can be a useful tool to help reduce computational complexity, while retaining the important detail in a scene. One way to do this is to use subtle perceptual methods to guide the user's gaze in a scene. McNamara et al. [MBG08] use an eye-tracker to show how gaze behaviour can be influenced

using a method called subtle gaze direction. Another way to help reduce complexity is to use eye-tracking to find out where a user looks in the scene and reduce the complexity of the regions that are not focussed on too often. Sundstedt et al. [SCCD04] investigate a user's perception in a visual task to enable certain degradations in quality and reduced-frame-times.

Perceptual studies are also used in the visualization of complex data to show how certain approaches can affect a user's understanding of a data-set. Healey et al. [HTER04] show how artistic techniques can positively influence user understanding of complex multi-dimensional weather and agricultural data-sets. Results from perceptual studies investigating other stylisation techniques have been used to show that certain stylised rendering methods can effect user task performance in visual search tasks, particularly for complex scenes [RD09].

Of particular interest and relevance to the work presented in this thesis is the work on perception of audio desynchronisation by Carter et al. [CST+10]. In the movie industry, temporal desynchronisations between voice and picture can happen during the dubbing process or during a broadcast. Carter et al. conducted perceptual studies investigating the effect this desynchronisation has on viewer enjoyment and found that there was an effect on the perceived emotion in the scenes and consequently the viewer's enjoyment.

Looking at virtual characters, there have been many perceptual studies investigating how to create and animate these characters to make them appear more plausible. We investigate perception of body motions of characters so it is important to look at results from similar studies on character animation. The first major study of human motion of relevance to us was by Johansson [Joh73], where he found that 12 moving light points were sufficient to represent the movement of a human. This is known as biological motion and has been often used since in experiments of human motion. The fact that we are so attuned to human motion and can identify specific properties (such as different walks, gender) from little information means that it can be quite difficult to fool people when it comes to animating characters. Reitsma and Pollard [RP03] investigate user sensitivity to errors in ballistic human motion. They found that as they added errors to a jumping motion (see Figure 2.10), participants were sensitive to the changes in the motion, particularly for added acceleration rather than added deceleration. Jörg et al. [JHO10] found that people were also sensitive to synchronisation errors in some types of finger motion for virtual characters. They also found that desynchronisation in finger motions of characters can change the interpretation of the virtual movie the user is watching.

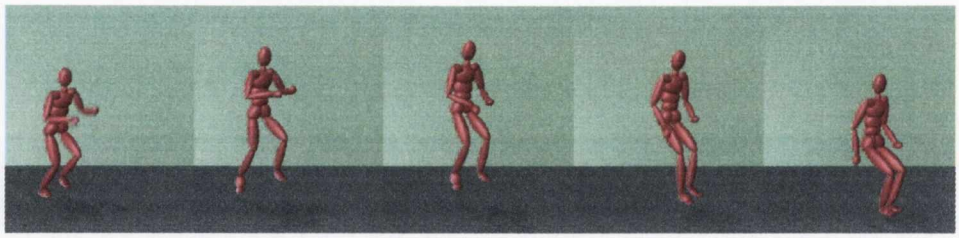


Figure 2.10: Example motion used in [RP03]

### 2.7.2 Perceptually Guided Crowds

For virtual crowds, in particular for applications that run in real-time, exploiting knowledge about how a user views a crowded scene can be a very useful way of maintaining a sense of realism for the user, while keeping the computation to a minimum, which is important to keep the frame rate of the simulation high. When running a simulation of a virtual crowd, there are a number of properties that come into play. The behaviour for the agents needs to be calculated according to one of the methods mentioned in Section 2.5. These characters, and the environment they inhabit, then need to be displayed, or rendered, to the screen. Characters then need to be animated as they move around the environment. Each of these processes can be very costly, so as the number of agents in a crowd grow, the computational overheads need to be reduced.

One method of doing this is by using a Level-of-detail approach; where the objects and characters closest to the camera viewing the scene are of the highest detail. However, objects that are further away from the camera can switch to a less detailed version at some point, while objects that are out of view only get rendered or animated when they come into view. Several perceptual experiments have been motivated by the need to find the best distance from the camera to switch to these lower levels of detail, so that the experience and sense of realism for the user is undisturbed. Hamill et al. [HMDO05] investigated how effective one of these lower level representations (called impostors) is for virtual humans. They displayed a model of a virtual character at a distance from the camera, then moved the character towards the camera at a constant speed, switching from an imposter to a fully detailed character at some point. Participants were asked to say if they noticed a change. They found that imposters can be successfully used at a distance where the pixel to texel ratio is 1:1 (where a pixel to texel ratio describes the number of pixels on screen occupied by an image).

Another problem with a large crowd of agents is that to keep computational expense down, only a limited number of different characters can be used. However, this results in a large number of the same characters, or clones, occupying the virtual environment, which will reduce the overall level of realism of the scene. Recent studies have examined ways to create variety in these characters by way of colour or texture modulation of their clothes, variety in the motions of the characters, changing some facial properties like adding makeup or beards. McDonnell et al. [MLD+08] found that adding variation to the appearance of the character was more important than adding variation to the motion of the character. In a subsequent study [MLH+09], they found that the illusion of variety can be created by selective variation, i.e., only the important regions of the character need to be varied, which can be as effective as varying the appearance of the entire character. In particular, they found that the addition of accessories to the head (e.g., hats, glasses), changing the texture of the top of the clothes (e.g., t-shirt design changes) and facial texture changes (e.g., makeup or facial hair) were equally effective, but that changes to the lower part of the body, or changing the geometry of the face were not effective. Results from these studies have been used to create a varied crowd using a small number of character models in our Metropolis system.

We aim to build on this specific field of research to find ways to populate a virtual environment with a crowd in a meaningful way. We also want to create variety in interactive agents, where we implement new conversations from a limited data-set of motions and audio. For both of these processes, we conduct perceptual experiments to deduce which methodologies are appropriate to achieve a plausible result for the user.

## 2.8 Discussion

In this chapter we have discussed many properties of human behaviour in urban environments. We have also outlined many existing methods and approaches to modelling these behaviours for many computer graphics different applications. The behaviour we focus on in this thesis relies heavily on user feedback through a series of perceptual studies, so we describe how these types of studies have been used for graphics applications in the past, and particularly how they can be of use for real-time crowd applications, whose realism can be limited by the need to optimise computational resources. We also focus on conversational agents. While we did not address the topics of conversational and group behaviours here, we concentrate solely on them in Chapter 3. We also present an overview of existing models for conversational behaviour for computer graphics applications.

# Chapter 3

# Overview of Group Behaviour

## 3.1 Introduction

In a crowd simulation such as Metropolis, where the setting is a urban environment with a number of large pedestrian areas, there will be a need to fill areas with agents who are not only individually navigating their way around the environment, but also engaging in more meaningful and varied activities. As the objective of the Metropolis project is to create as realistic a crowd as possible, the addition of agents interacting with the environment, for example, sitting on benches, and each other, is a key goal. We would like our agents to stand or walk around in groups, appearing to communicate with each other in a meaningful way. The agents should also be able to queue in an orderly fashion, wait at traffic lights, or watch a street performer. In order to implement these kinds of group activities, we first want to discover how groups form in human crowds, and identify important properties of the behaviour of these groups. This chapter presents an overview of the behaviour of smaller groups, including communicative behaviours, and discusses some applications of these agents in graphics applications.

## 3.2 Dynamic Groups

The first kind of groups we will discuss are dynamic groups. By dynamic groups, we refer to people, or agents in a simulation, who navigate around an environment in groups of two or more. Of particular interest to us is how individuals in this group interact with each other, how the group navigates as a whole, different group formations and how they

avoid collisions with each other.

### 3.2.1 Pedestrian Dynamic Groups

It has been shown that the majority of pedestrians do not travel individually, but will walk in a group of two or more. Some studies have shown that the number of pedestrians in a crowd travelling in a group is as high as 70% [MPG+10]. James [Jam53] found in studies of pedestrian areas in the 1950s that most groups were of size 2. Groups of size 3 and 4 were less common but still more frequent than groups of 5 and larger.

Moussaïd et al. [MPG+10] have investigated the effects of these social groups on the dynamic behaviour of the crowd as a whole and documented some interesting findings. They observed the movements of groups of 2, 3 and 4 in a low and medium density crowd and the effect of the organization of these groups on the behaviour of the crowd. They then developed a model based on their observations. They found that the speed of the group as a whole tended to be smaller as the group size was larger, regardless of the crowd density. They also found that at different densities, the formation pedestrians walked in changed; when in a low density crowd, group members tend to walk in a horizontal formation, perpendicular to the walking direction. However, when the pedestrians had less space around them as crowd density got higher, they tended to walk in a "V" shape with the group members on the outside of the group slightly in front of the member(s) in the middle. The authors felt that this formation was to aid communication within the group as much as possible, while avoiding collisions within and outside the group.

Based on their findings, Moussaïd et al. created a model to represent these social interactions of a moving group by calculating a social interaction term for the response of each pedestrian to other group members. This term is composed of an acceleration term for the vision field of each agent (for communication), acceleration term for attraction to the centre of mass of the group and a repulsion term to avoid collisions within the group. They found that results for their simulation closely matched their observed angle distributions throughout the groups (see Figure 3.1) and also showed a similar linear decrease in speed as group size increases.

### 3.2.2 Dynamic Groups In Graphics Applications

Musse and Thalmann [MT97, Mus00] created a model for human crowds and investigated sociological group behaviours. Their model assigns parameters for agents that are part of

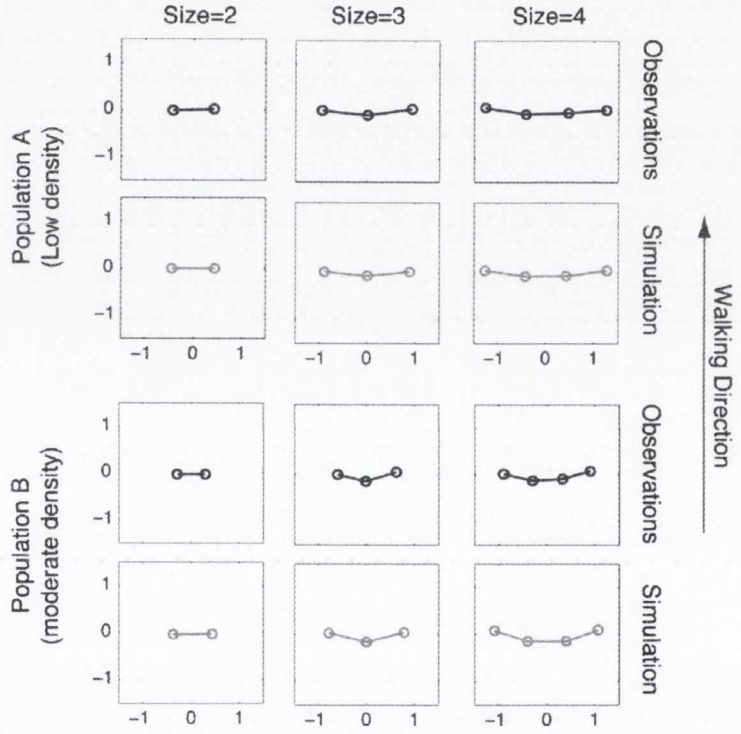



Figure 3.1: Comparison between observed and simulated average group formations for different group sizes as found by Moussaïd et al. [MPG+10]

a group: *list of goals*, *list of interests*, *emotional status*, *relationship with other groups* and *domination value*. Based on these parameters, agents can leave or join groups, become the leader of a group and share emotions depending on their domination parameter. The group behaviour is composed of *goal seeking* (agents follow the groups direction) and *flocking* (ability to walk together). Behaviour at an individual level involves walking, collision avoidance and interaction with other agents. *Splitting*, *adding* and *space adaptability* were added to this system at a later stage to add more realism.

Groups and their emergent behaviours were also investigated in the work of Kankri et al. [Kan05, LMM03]. Group leaders were determined before others joined the group using sociological parameters such as leadership and sociability. When the conditions are satisfied and the agent becomes a leader, then an emotional parameter is increased, making the agent *happier*. In order for an agent to join a group, an *interaction quality* must be high enough, then the agent adds the intention to join the group to its internal goals. After an interaction, the agent can join the group if its leadership parameters are acceptable. The new agent will then take on the goals of the group, i.e., the group leader's goals. A cohesion force is applied to keep the agent near the group position. This model also allows for three different types of group formations and modifications to the spacing between agents.

Hostetler and Kearney [HK02, Hos02] described a method to represent three dimensional curved pathways as ribbons in space to allow for navigation of small groups of pedestrians in urban environments. These walkways define the geometry of the surface, taking into account obstacles (such as fire hydrants, bins) for two preferred directions of travel. This does not restrict the behaviour of single pedestrians, just those travelling in groups. The ribbons also describe spatial relations for groups by defining any obstacles in front of or beside the ribbon structure. The 3D curve is defined by a spine with a surface normal defined at each point to allow the ribbon to twist around the spine. Points on this coordinate system are then defined by distance along and offset from the spine. The pedestrian walks are defined by an action space of three turning parameters and three speed parameters. A *constraint proxy* votes on each cell in the action space at each time step to apply positive or negative weights to all the turning and speed parameters. Pedestrians query a series of pursuit points along the path and calculate a pursuit direction. When two or more pedestrians are navigating the path, their individual position is calculated from the group average position and their orientation is the same. A *slip* metric is used to measure local distance between members of a group and a maximum slip denotes how far ahead or behind members of a group are. Pedestrians also maintain a minimum separation or

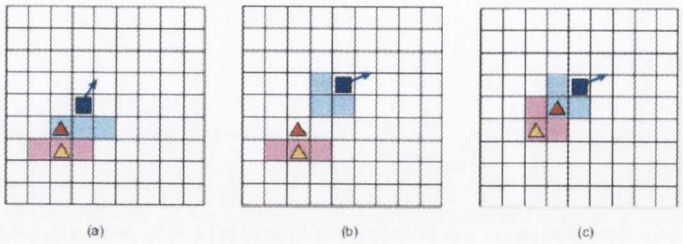



Figure 3.2: Group motion using advised tiles [Mar02].

personal space from others and avoid obstacles by shifting the pursuit point to the edge of the obstacle. This results in the entire group avoiding larger obstacles, while allowing individuals within the group to steer in order to avoid smaller obstacles.

Marchal [Mar02] describe a method of simulating group motion in a cell type grid, also using leaders and members of a group. Figure 3.3 outlines the process. The leader (square) advises three tiles for group members to occupy behind him (a) as does the first member (red triangle). When the leader moves, the advised tiles move too (b). The two members (red and yellow triangles) move to keep the connectivity of the group (c). For each agent, one of its neighbouring cells must contain another agent from the same group, otherwise it has become disconnected from the group. The resulting movement is not particularly useful for urban pedestrians as the members tend to follow each other rather than travel beside each other, however the grid mechanism could possibly be adapted for such purposes.

## 3.3 Conversational Characters and Groups

Another important element for our crowd system is the concept of "conversational groups". As Metropolis is based in Dublin, and centered in Trinity College, there will be many different sized groups of tourists, students, staff and others within this environment. These groups will not only be navigating around the environment but, as in real life, there will be some who are "idle", i.e., conversing with each other, waiting on others, or perhaps watching a street performer. While the characters are idle, they will still need to interact and converse in a meaningful way. This activity therefore needs to be visually and aurally plausible to the viewer.

### 3.3.1 Non-Dynamic Groups

There are a number of different kinds of groups that can be found in a crowd scene, which tend to exhibit different types of behaviours. A small group of people would usually be quite sociable and intimate. The members of these type of groups will generally tend to at least be acquainted with one another, therefore communicating with each other with each individual providing feedback within the group [DDN95]. Such groups can also affect the behaviour and movement patterns of the crowd around them [Sti69], so they are an important property of the crowd to investigate. As the number of people in a group grows and the time allowed for each person to equally participate is reduced (5-10%), the function of the group tends to be more of a one way communication, or smaller cliques within the group can be formed [DDN95]. In an outdoor scenario, these types of groups could be small groups of sightseers following a tour-guide, or watching a street perfromer. The communication tends to be one-way, with any feedback, such as questions, being asked in a formal manner. Other large groups gather without specific purpose to communicate with one another, for example members of a crowd stopping to watch a street performer.

Large groups can also be composed of individuals and small groups attending a concert or more emotional events such as protests or sports celebrations. Aveni [Ave77] interviewed members of a large crowd leaving a college football game. They found that for a social event like a football game, 74% percent of people attended the game in groups, rather than alone. Of these groups, the majority were in couples (54%), with much a smaller proportion attending in groups of three (18%), four (16%) or larger (12%). There is a similar pattern here to the ratios in dynamic pedestrian groups described earlier. There is a slight difference for crowds gathering at these events in that the number of groups with more than three members is slightly higher than was noted for purely pedestrian groups. There is a possibility that for a college environment, such as the one Metropolis is set in, that there will be differences in group sizes also, since there can be large groups leaving classes together, or socialising with each other at different break times.

For the scope of this thesis, we focus solely on smaller groups. We wish to create conversational groups for our crowd system, and while we do plan to add medium and large sized groups as found at tours or demonstration, they are not the focus of the thesis. Much work has been carried out regarding communication properties of individuals and within small groups, so we will present in this section an overview of this research.

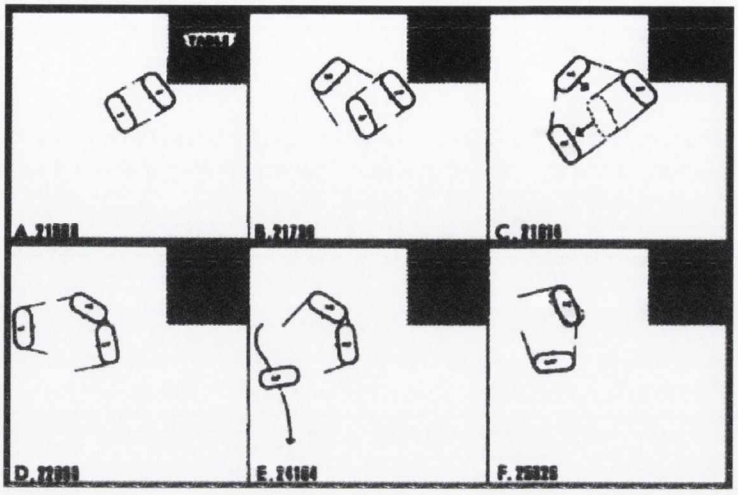


Figure 3.3: Group formation changes for individual joining the group (a-c) and leaving the group (d-f) [Ken90].

### 3.3.2 Communication And Gestures

There have been many studies outlining general properties of the formation of conversational groups. Dunbar et al. [DDN95] suggest that there is a limit on the number of people who can interact in a spontaneous conversation. They conducted a study looking at the size of conversational groups, and the number of speakers in these groups, in a variety of settings over a 12 month period. Their results suggest that, due to the limitations imposed by speech production and detection, the upper limit on a conversational group appears to be four people.

Kendon [Ken90] presented a system to describe the structure and properties of a conversing group in relation to spatial positioning and orientation. He describes the space in between a group of conversers as an *o-space* and describes the formation surrounding that as an *F-Formation.* The idea that each member of the group works to maintain this o-space with the positioning of their bodies is the *F-Formation system.* The F-Formation is described mainly by the location and orientation of the lower bodies of the members, i.e., by feet placement. He describes a number of different arrangements of groups that are considered F-Formations:

Groups of 2:

- **Vis-a-vis:** where each member is facing one another. This is more commonly found on pathways.

- **L-arrangement:** where each member is aligned as the two arms of an "L". This arrangement is more often found in open spaces.
- **Side-by-side:** where each member stands beside each other, facing the same direction. This is most commonly found at edges of an area, e.g., facing out from walls or looking onto a football pitch.

Groups of 3 or more can be found in various formations; circular, semi-circular, linear and rectangular. The role an individual plays within a group can be dependent on his/her position in the arrangement [Ken73]. Circular arrangements with groups of 3 or more tend to allow for equal contributions to the conversations, whereas other arrangements, such as a rectangular arrangement where there is a "head" of the group, will tend to have one individual talking more than the other members of the group.

Kendon also discusses how when members of the group join or leave, how others in the group will reposition themselves to ensure that the o-space is maintained exclusively for the group, and to occupy the position most comfortable for communication. Jan and Traum [JT07] define a social force model for this movement, based on four main forces. The first force is an attraction force of a listener towards the speaker. The second force is a repellant towards outside noise. A third force reflects the individual's desire to not stand too close to other agents. Finally, there is a force towards a circular formation of all conversation participants.

If a group of people are conversing, an important property of this conversation is the information they convey. This comes from the words they say, how they say it, and how they gesticulate while they are saying it. All of these cues signal to the person they are communicating with how they should react. Non-verbal communication is an important element in conversation. People can use gestures to get help illustrate a point they are making. Mostly, however, we gesticulate spontaneously, without a great deal of deliberate planning. It has been well documented that speech and these non-verbal communication methods are very heavily linked [GM05, Ken94]. Ekman and Friesen [EF69] describe a number of categories of non-verbal communication:

- **Emblems:** These are non-verbal acts that have a direct verbal translation. Generally, these are the only kind of gestures that are deliberate. They are usually used when verbal communication is prohibited by noise or distance e.g., asking a friend do they want another drink in a busy bar, or asking for the time by pointing at a watch.

- **Illustrators:** These are gestures that are directly tied to speech and are used to illustrate what is being said. There are six different types of such gestures: *batons* which are used to emphasise a word as a 'beat' to the word; *ideographs* are used to sketch a train of thought, like using your fingers to illustrate a number of points; *deictic movements* are pointing gestures to an object; *spatial movements* illustrate a distance or spatial relationship; *kinetographs* illustrate a bodily action, such as running, or crying; *pictographs* are used when the speaker draws a picture of what they are talking about.

- **Affect Displays:** These are movements that are triggered by basic emotions. They are extremely prevalent in facial movements, like smiling when someone is depicting happiness, frowning when depicting sadness or disagreement, or eyebrow raises to indicate surprise. These affect displays vary greatly across culture and are socially learned.

- **Regulators:** These are gestures that ensure that the conversation between two or more people is dynamic. Examples include eye gaze directed at the talker to show attention or interest in what they are talking about. Other gestures can be used to show disinterest, or to ask the speaker to repeat or elaborate.

- **Adaptors:** These are interactions with ourselves, others or objects during conversation to perform a physical function, which are rarely used for communication. There are three types of adaptors. *Self adaptors* are when a person touches a part of their body, for example, putting a hand to their face, or the licking their lips. *Alter-adaptors* are movements that are aimed at another person, for example, giving or taking objects to or from a person. *Object-adaptors* refer to an interaction with an object, such as using a tool.

### 3.3.3 Studies on Human Communicative Behaviour

With so much information available on how conversational groups form, and how we can use gesture to communicate in many different ways, it follows that much research has been carried out on the perception of these communications. We present in this section a selection of studies looking at human communication.

Clarke et al. [CBF+05] investigated whether people were able to identify the emotional content of a conversation between two people based on biological motion alone. They found that people were quite well able to correctly identify most of the six emotions used (anger,

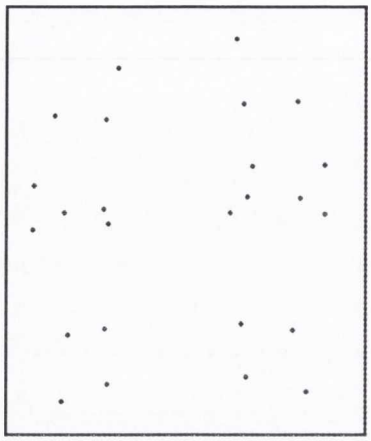


Figure 3.4: Example of the stimuli presented to participants in Rose and Clarke [RC09]

disgust, fear, joy, sadness and romantic love). They also examined the importance of the interaction between the people conducting the conversation to conveying the emotion of these conversations. Here they found that when the movement of only one person in the conversation was displayed, people were no longer able to recognise joy or love. They also used a mirror image of the actor and found that when viewing these conversations, participants were no longer able to recognise sadness, joy or fear. This study shows that biological motion alone provides sufficient cues for the perception of emotion. However, it also highlights our reliance on interactions between the conversers in order to make these judgements, since it can become unclear, when one of the cues is removed.

Rose and Clarke [RC09] then went on to examine how good people are at detecting talker/listener roles using biological motion alone (Figure 3.4). They presented participants with a clip depicting the biological motion of two actors taking part in a scripted conversation while exhibiting a variation of the six Ekman basic emotions [Ekm92]. They found that for conversations depicting anger, joy, sadness and love, people were able to identify the speaker correctly. So for emotional conversations, people are quite accurate at identifying speaker roles using very little information. However, Rose and Clarke's study was carried out for emotional conversations, and does not address the issue of sensitivity to gestures and roles for non-emotional, natural conversations. This reiterates previous research that shows that people are acutely sensitive to gesturing and its role in interpersonal communication [McN96].

There has also been some research into the perception of male and female conversational behaviour. Briton and Hall [BH95] examined perceived differences in non-verbal behaviour

of hypothetical men and women. They asked participants to read a list of non-verbal behaviours on a scale of one to ten depending how likely a man or a woman is to carry out these behaviours. They found that participants believed women to use more expressive gestures than men, and that they were better at communicating in this way. They found that participants believed men to be louder, more interruptive, nervous and less fluid than women. They also found that women tended to rate women's gestures more believable than men, but no such effect with men.

Hannah and Murachver [HM99] investigated the effect of speech-style and gender on perceived conversational behaviour. They asked male and female confederates to conduct a conversation with participants in a facilitative or non-facilitative manner. A confederate is an actor participating in an experiment pretending to be a subject, but actually working with the researcher. Facilitative conversation styles consisted of the confederate interrupting and looking away from the participant seldomly, while giving minimal responses frequently (nods and feedback responses such as "mmmm" or "uh-huh"). Non-facilitative conversation styles looked away and interrupted frequently and did not give frequent minimal responses. Females tend to use a facilitative conversational style and males tend to be less facilitative. So they wanted to find out if it was gender or conversational style that affected participants' perception of the conversation. They did this by measuring how much the participants spoke in the conversations with confederates. They found that conversation style did affect participants' speech behaviour, rather than the gender of the confederate. They also found subtle differences in how male and female participants responded to conversational style. They found that female participants maintained their conversation style regardless of the gender or conversation style of the confederate, while male participants tended to use a more facilitative conversation style when conversing with a facilitative female confederate.

### 3.3.4 The Role Of Audio In Conversations

Krahmer and Swerts [KS07] conducted a set of experiments to investigate the effect of visual beats on the prominence of speech (how much a word stood out in a phrase). In a first experiment they examined the production of a visual beat affected the production of speech. They found that when participants were asked to produce a visual beat when saying a target word in a short sentence, this had a small effect on the duration, energy and frequency of the stressed word. They then conducted two more experiments to find if these differences were perceivable in conversation.

In the second experiment, three independent labellers listened to the audio clips and rated them on a scale of 0 to 2 (no accent, minor accent and clear accent) depending on whether they noticed a change in pitch for each target word under the various conditions. They found that when a target word was accompanied with a visual beat, the perceived prominence of that word increased, regardless of whether the speaker had added a pitch accent or not. In their final experiment they examined the perception of seeing a visual beat on the perceived prominence of a word. Participants listened to recordings of different conditions and rated the prominence of each target word on a 10 point scale. They had both audio only (no video) and audio visual conditions. They found for this experiment that when participants saw a visual beat, the word was perceived as more prominent than when audio alone is used. They also found that a beat gesture was more effective in this way than eyebrow movements, and that speakers gesturing style also had subtle effects. So it is clear that when we converse, we use both audio and visual information to infer importance of stressed words in a sentence.

While Krahmer and Swerts have shown that people are sensitive to temporal alignment of video and audio signals for facial motion, Giorgolo and Verstraten [GV08] investigated the perception of speech and gesture integration for body motion in the absence of facial motion. They conducted an experiment to examine the effect of synchronisation between audio and visual signals on how pleasing participants found video clips of a person talking. They recorded an unscripted dyadic conversation, where one person is explaining the plot of a movie to another. Short clips of these videos were presented to participants, with the face of the speaker blurred (see Figure 3.5). Delays from -1000ms to 1000ms, in intervals of 250ms, between audio and video were induced and participants were asked to say whether the clips were synchronised or not. They found that people were very sensitive to these temporal delays and only accepted delays of $\pm 250ms$ or no delay. This implies that both visual and aural signals are closely integrated in the brain when viewing human conversations.

Of interest to us, for our real-time crowd system where the audio between the characters can be localised at the positions to increase reality, is whether humans have the ability to localise audio in 3d space correctly. We did not find any information about people's ability to do so in a 3D environment, but previous work has shown that on a horizontal plane, people were able to localise binaural (i.e., in both ears) tone audio signals, but only for tones of a higher frequency. Musicant and Butler [MB84] investigated localisation abilities for tones only, rather than any linguistic audio. Muller and Bovet [MB02] have also investigated the ability to localise audio, but used linguistic stimuli. They examined

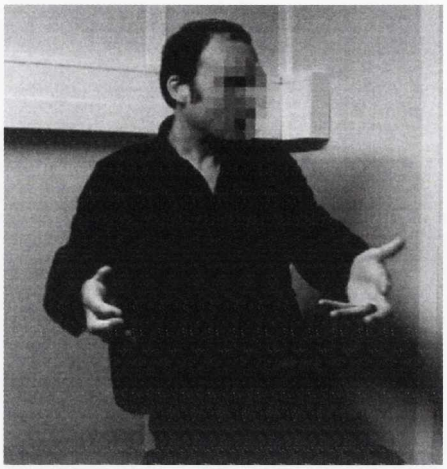


Figure 3.5: Example of the stimuli presented to participants by Giorgolo and Verstraten [GV08]

performance and reaction times in a task where participants had to localise, in the horizontal plane, monaural (i.e., in one ear) audio signals containing participants' first names. They found that participants were equally able to localise their own name compared with the names of others, but that there was a difference in reaction times. Their results imply that people use not only acoustic cues, but some sort of semantics when processing audio localisation.

### 3.3.5 Conversational Agents in Virtual Applications

In this section, we will discuss several applications of conversing agents in computer graphics. Many of these approaches are in the general field of Human Computer Interaction (HCI). Cassell et al. [CPB+94] have looked at the use of gestures and conversation in HCI. They describe a system which automatically generates conversation text, and intonation, using a discourse analyser. This text and intonation is then used to generate conversational behaviours using appropriate speech, facial expressions and hand gesturing for a virtual agent. The discourse analyser runs through the text of the script and identifies cues based on the semantics of content for one of 4 types of gestures: *iconics* (representing something concrete being talked about), *metaphorics* (representing something abstract being talked about), *deictics* (representing a point in space) and *beats* (hand movements for emphasising words). Facial animation is generated based on the intonation of the voice, or by hand. There are four rules for controlling the gaze of the character: *planning* (e.g.,

looking away to organise thoughts), *comment* (e.g., nodding head to agree), *control* (e.g., looking at listener to denote their turn to speak) and *feedback* (e.g., looking at listener during pauses to determine his/her reaction).

Cassell et al. [CVB01] then developed this system further to allow the user to input the text they wish to see spoken by the virtual agent. Here, they present a dialogue analyser that tags the input text with one of five language tags: clause (a full phrase such a sentence), theme and rheme (the topic of conversation and new information being presented about that topic), word newness (whether the word has been said before), contrast (words that may be opposites or identical meanings of each other) and objects and actions (for example, "dog" is an object and "walking" is an action). A number of behaviour generators then use these tags to generate appropriate behaviours. The first is a default gesture that is used if no information is available for a more specific gesture, called a beat gesture generator. There is also a surprising feature iconic gesture generator, used if one of the objects identified by the tagger has an unusual feature. Next there was an action iconic gesture generator, which determines actions for those tagged in the rheme. A contrast gesture generator identifies objects that are in contrast with other objects, usually conjoined in the text with an "or", e.g., "would you like an apple or an orange?". An eyebrow flash generator is used when new material has been identified. A gaze generator was used for appropriate gaze for the agent throughout the conversation. Finally, an intonation generator controls a text-to-speech engine. The system then uses a behaviour selection module to resolve conflicts between different gesture types and prioritises the behaviours, which are subsequently scheduled to be in synchrony with the input text and the character is animated appropriately (Figure 3.6).

Vilhjálmsson and Cassell [VC98] presented a method to automatically generate body gestures and interactive conversational behaviours such as turn-taking and feedback. They describe a multi-user system where users communicate with each other using avatars. When a user inputs text, his/her avatar generates appropriate body gestures, facial expressions and simple body functions. Perceptual evaluation results showed that users found the autonomous behaviours generated by the model more expressive than avatars with no behaviours, manually controlled behaviours and a combination of autonomous and manual behaviours [CV99]. Participants also felt that they had more control over the conversation when the behaviour of the avatar was autonomous. Cassell and Vilhjálmsson suggest that this is due to the importance of the textual contribution for multi-party conversations for users, and how focussing on manual control of avatar behaviours can distract them from the focus of the conversation. A later architecture combines this model with

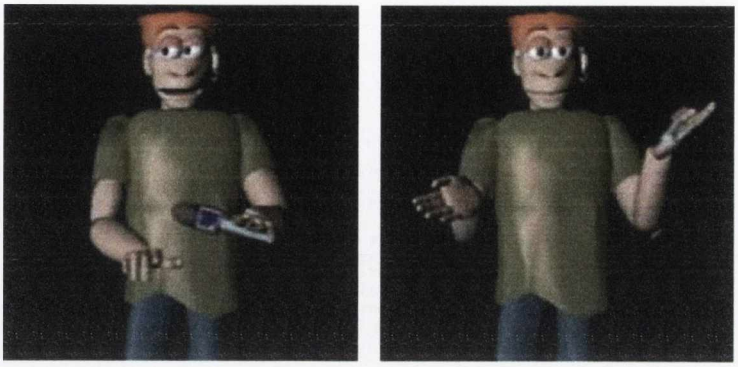


Figure 3.6: Example of the output from the BEAT system [CVB01]

the earlier discourse analyser [CVB01] to generate appropriate voice synthesization and communicative behaviours for avatars in a multi-user environment [Vil04].

This work on communication between virtual agents was further developed through a behaviour mark up language that unifies the multi-modal behaviour generation processs for Embodied Conversational Agents (ECAs) [VCC+07, KKM+06]. Focussing solely on the behaviour planning component of conversational agent construction, this model splits the process into three stages; intent planning, behaviour planning and behaviour realisation. The markup language mediates between behaviour planning and behaviour realisation. The XML-based language allows for coordination between speech, gesture, gaze, head and body movement. Each behaviour has 6 animation phases, each of which has a sync-point that is used to synchronise many different behaviour types at the same time. It also allows for feedback between the behaviour planner and behaviour realiser modules. The latter can send messages informing the former of different events, i.e., if the behaviour has been realised successfully; the behaviour realiser is unable to carry out the planned behaviour to its full extent; or a behaviour, or series of behaviours has to be cancelled for some reason.

A gesture generation system presented by Neff et al. [NKAS08] can recreate a specific speaker's gesturing style. Video is used as an input into the system, where the speech and motions of the speaker are annotated to produce a statistical model of his/her gesturing style. The speech annotation focusses on the theme, rheme and focus of the utterance. The focus refers to the main subject of the sentence, introducing context, and allows alternative themes/rhemes to be disregarded. The motion is annotated according to phases, phrases and units. Phase refers to the gesture phase; either preparation, stroke or hold. At the phrase level, a number of phrases are defined into a gesture. A unit encompasses an entire gesture phrase beginning from and returning to a rest position. The dimensions for hand

position are recorded for each gesture and an animation lexicon produced containing the following information: hand orientation, posture information and data for after-strokes. A gesture profile is built using all this information to estimate probabilities for word and gesture use for the speaker. A graph structure is used for both the speech and gestures for the new speaker input text. This in done in four stages. Many candidate gestures are first produced using the statistical model. The next stage selects candidates, and specifies properties of the gestures (such as handedness and hand-shape). Gesture units are created and planned in the third stage by merging gestures that occur within 1.5 seconds of each other. Head and body rotations are added to the character and gestures are then animated, by extending the gesture data provided to a string of sentences, using a combination of kinematic and dynamically simulated animation.

Levine et al. [LTK09] presented a system that generates gestural body animations automatically using speech, rather than text input. Their system uses a training phase that correlates gesture and prosody, or intonation, of a database of motion-capture data and audio recordings, to build a probabilistic model of gestures. Gestures are extracted from that motion data, while syllables are extracted from the audio data, along with pitch and intensity values, which are used to compute a prosody descriptor for each audio segment. This information is used to calculate parameters of a hidden Markov model (HMM), which is then used to select an appropriate motion capture segment for each segment of a live audio stream. Results from a user study showed that the model used was preferable to random synthesis of gestures for the same videos.

Pelachaud [Pel09] conducted a number of perceptual studies investigating gesture expressivity and quality for virtual agents. Results from a study where participants were presented with facial animations of agents showing singular emotions (happiness, sadness) and masked emotions (such as masking disappointment with happiness). They found that people were able to identify the more complex masked emotions for a number of different methods. This implies that humans are sensitive to subtleties of human emotion, even for virtual agents, and that the detail of these agents are very important.

### Conversational Groups In Virtual Applications

Gesturing behaviour and interactions between the user and an embodied communicative agent is very important to learn about the important properties that have been implemented for virtual characters. Of particular interest to us for the Metropolis system is the idea of group interactions. We will discuss applications where these kinds of static or

conversing groups have been implemented.

Shared virtual environments are increasing in popularity. Massively Multi-player Online (MMO) Games such as Second life, where the user is given an avatar and navigates the environment to interact with other users' avatars, have become particularly popular. Such avatars demonstrate another application area where social behaviours are an important factor in increasing the sense of presence in the environment for the user, especially in group conversations. Pedica and Vilhjálmsson [PV08, PV09] present a platform that produces realistic reactive behaviours for individuals participating in group conversations. Unconscious reactions tend to be rare in such applications, because explicit user control is always needed for navigating the avatar around the virtual world and does not allow for low level reactive behaviours. So, Pedica and Vilhjálmsson use a steering layer for their avatars that can be combined with locomotion behaviours. They base their platform on Kendon's F-Formation system, discussed in Section 3.3.2. The steering layer contains prioritised and weighted behaviours for keeping personal distance, social equality, group cohesion, common attention and domain awareness. Each of these behaviours, which are only activated when needed, influence the re-positioning and re-orienting of the avatar in the conversational group by producing attraction or repulsion forces as necessary. A perceptual evaluation of this technology implies that the addition of territorial behaviours can increase believability of virtual conversers. Participants completed a web survey comparing videos where territorial behaviours are present and absent for four different scenarios (joining a conversation, avoiding a conversation, moving within a conversation and passing close to a conversation taking place). They found that the presence of these behaviours are present increase the overall believability of some of these scenarios, such as joining a conversation, or avoiding a conversation. However, some territorial motions displayed within conversations were found to be no more appealing than when there are no behaviours present.

Hostetler [Hos02] also addresses the problem of positioning and orienting of agents in a conversational group. He proposes that when a number of agents are standing in a group, the attention of the group members will tend to converge on a focal point. He introduces six "proxies" that determine the behaviour of an agent in such a group. The first "track focal point" proxy orientates the pedestrian towards the focal point of the group, while "maintain stationary formation" and centering proxies ensure that the agent maintains a certain distance from the focal point, depending on the number of other individuals in the group, while "avoid pedestrians" proxy allows group members to maintain a personal space within the group. The "avoid obstacles" ensures that an agent will maintain a distance

Figure 3.7: Sample scene utilising role-passing LOD [OCV+02]

from obstacles in the environment. For example, if a speaker is positioned near a wall, the other members of the group will position themselves to leave a gap between themselves and the wall. Finally, the "inertia" proxy is used to avoid fluctuations in behaviour by remembering the last winning position and orientation and giving this position and orientation a vote as a preference. An "election" is held at each time step, where each proxy is viewed as a voter that votes independently of the others. Each proxy is given a weight in the election and once the voter puts forward their weighted vote, an electioneer selects the most suitable action. Hostetler also proposes a method for combining dynamic groups and these static groups by introducing stopping and starting intermediate behaviours and using these to blend between two separate dynamic and static behaviours.

O'Sullivan et al. [OCV+02] employed a behaviour generation toolkit based on Cassell et al.'s [CVB01] gesture generation system to generate levels of detail for conversational behaviour of groups of agents. They first apply the concept of conversational levels of detail, which uses the BEAT framework described in Section 3.3.5. Filter rules remove generated behaviours based on a group's visual saliency within the scene. Secondly, level of detail Artificial Intelligence (LODAI) is facilitated by a process of role-passing, where agents are given the ability to take on different roles, depending on the situation they are in. For example, an agent in a work meeting will behave differently than an agent socialising in a bar (Figure 3.7). A role is layered on top of a very basic agent behaviour at appropriate times within a simulation. The roles specify when an agent should interact with others in the environment or how the agent should interact with objects around him/her and defines a specific set of required animations. This means that a virtual

world can be populated with basic agents who can behave appropriately when they enter different areas of the world. It also allows for level of detail because the roles can be assumed and discarded when needed, without the need to consider behaviours not specific to the situation. They propose that this is further utilised by varying levels of detail depending on an agent's saliency in the scene.

## 3.4 Discussion

In this chapter, we have presented an overview of group anthropology, properties of non-verbal communication and some applications for computer graphics. The prevalence of group behaviour in everyday life highlights the importance of the presence of interactive groups in pedestrian crowd simulations that aim to closely reflect real-life scenarios. Such groups are still largely missing in video games and other real-time crowd simulations, although they are omnipresent in everyday life. While much work has been carried out examining the properties of the non-verbal behaviour of individuals, and interactions between two or more human conversers, very little is known about how we expect a group of virtual conversers to behave. These kinds of groups are becoming more prominent with the advancement of shared virtual environments and the work described here highlights the need for studies on how such virtual groups are perceived, and therefore how to deliver a high sense of realism to the user in crowd simulations.

# Chapter 4

# System and Methods

## 4.1 Introduction

In this chapter we will give an overview of the Metropolis crowd system that forms the framework for the work described in this thesis. We will discuss the motivations behind the project, the different research areas being pursued and the software solutions used. We will also provide an overview of our motion capture pipeline, mentioned in relation to capturing our conversational groups.

## 4.2 Metropolis Crowd System

### 4.2.1 Objectives

The motivation behind the Metropolis project is to create a convincing crowd that inhabits a multisensory environment. The project addresses the issues involved in real-time visualisation, motion and audio simulation of a large multisensory human crowd. Rather than simulate physically realistic properties of a real-life crowd, our aim is to create a simulation that is plausible to the user. In order to do this, we use existing software solutions wherever possible, and develop new algorithms, models and metrics, guided by the multisensory perception of the user, to achieve a highly realistic simulation depicting the sights and sounds of a busy metropolis: in our case the city of Dublin.

When dealing with real-time crowds, there are a number of trade-offs that need to be made, which can lead to limitations in these kinds of simulations. Many of the computational

resources can be expended trying to render, compute behaviour and animation for a large scale crowd, and this inevitably leads to resources for other areas being reduced. As a result, many real-time simulations have artifacts in the rendering of the crowd, some animation degradation and other limitations such as low variety in the appearance and motions of the crowd, or a poor feeling of immersion of the user in the crowd due to a lack of human characteristics such as emotions. We aim to simulate a large-scale crowd in real-time while utilising our limited resources to give the illusion of a large heterogenous crowd, that behaves as the user expects.

As discussed in Chapter 3, the brain integrates information in such a way that it creates a final percept across all of our senses. If our senses are highly interactive, and something we perceive in one sense can often have a direct effect on another, we want to find out what the consequences are for a virtual crowd. If we are looking at a crowded scene, will we feel more involved if we can hear the sounds we associate with such scenes? Can we distract the user using audio to make them see a scene differently, or can we provide multisensory information to disguise anomalies within a scene? This project is particularly focussed on questions about multisensory perception, with various groups working on different aspects: a Computer Science group working on the simulation of the crowd; an Engineering group working on creating a realistic audio backdrop to the environment; and finally, a Neuroscience group dedicated to finding out what is expected from these crowds, and they compare to human crowds. These groups all work together as a team to combine their efforts with the objective of exploiting knowledge of human multisensory perception to identify and enhance the features that increase plausibility and ameliorate perceptible artifacts that reduce the realism of the final simulation.

### 4.2.2 Background

The inspiration and the backdrop for Metropolis came from an earlier project, called Virtual Dublin [HO03], which was a 3D virtual simulation of Dublin city, populated with virtual humans. This crowd system spawned the novel idea of *Geopostors* [DHOO05], where the crowd was rendered using a hybrid system of geometry based characters and impostors. The system imperceptibly switched between the two rendering styles at a distance from the camera that was found through perceptual experiments [HMDO05], which resulted in minimal popping artifacts and a large scale crowd rendered in real-time.

While large numbers of characters could be displayed in real-time, other aspects limited the plausibility of the system. Firstly, the agents in the crowd exhibited only basic behaviours,

Figure 4.1: Screenshots of Virtual Dublin, the precursor to Metropolis.

where they did not interact with each other or the environment. Secondly, the appearance of the buildings and the agents were not very high resolution, or as visually realistic as desired. There was also limited variety in the crowd. For Metropolis, artists have been recruited to generate realistic building models and a small number of high-res agents with very realistic appearance, while the research goals of the project involve making the crowd as realistic as possible. Virtual Dublin also lacked any audio information, or any psychological input to the behaviour of the agents. Metropolis was therefore a natural continuation of this project, focussing on the agents that make up the crowd, to improve the overall perceptual fidelity of the simulation.

### 4.2.3 System Overview

The Metropolis system is written in C++ code, using Microsoft's Visual Studio. The code itself has been split into a number of different systems to allow for individual research teams to implement their own changes individually, but then allows for easy integration into the Metropolis system. A scripting system has also been integrated to allow for changes to be made to the system without the need for recompiling. Below are some screenshots to give an idea of the current status of the Metropolis system (Figures 4.2, 4.3 and 4.4).

Figure 4.2: Screenshot of Metropolis. This scene is in the front square of the Trinity College Dublin campus and in the crowd, conversational groups as well as dynamic agents can be seen.

Figure 4.3: Screenshot of Metropolis. This is a close-up of one of our conversational groups.

Figure 4.4: Screenshot depicting the capacity of our system to render a large crowd in real-time.

The different sub-systems for Metropolis handle:

- Rendering
- Animation
- Behaviour
- Audio
- Experiments
- Scripting

Our rendering, animation, behaviour and audio sub-systems allow the various groups to efficiently conduct their research and implement their solutions. We will discuss these systems in further detail in Section 4.2.4. There is also an experiment system, which allows members of the Metropolis team to conduct perceptual experiments using the real-time crowd system. It is extremely useful, and preferable to using video clips for presenting stimuli to participants, as it allows for automatic random ordering of the stimuli and also incorporates many functionalities of the Metropolis system, which gives the experimenter easy access to control specific variables. Finally, there is a scripting manager, which

provides a way for developers working on the system to add additional functionality to their code. Using this manager, we can make any function in the system scriptable, allowing for the changing of variables on the fly.

### 4.2.4 Research Areas

#### Rendering

This system organises the rendering of the environment and crowd to the display device. The environment consists of the buildings and other props contained in the city, such as lights, trees and litter bins. The rendering of the crowd is a complex procedure, with different methods being utilised to render large crowds in real-time. Research is ongoing into this aspect such as methods for reducing computation while introducing variety, creating the impression of detail using a limited amount of resources and developing a rendering system capable of creating and displaying this amount of detail on fully animated characters using commodity hardware [MLH+09, MLD+08, LPDO10].

We use an open-source rendering engine called Ogre (Object-Oriented Graphics Rendering Engine) to render Metropolis [MR04]. Any research conducted uses the functionality of this engine. Ogre is a multi-platform scene-oriented 3D graphics engine, that provides developers with an interface between high level coding classes and objects and rendering system libraries like OpenGL. The main advantages of using a system like this is to allow for level-of-detail rendering, scripting functionality and exporter compatibility with modelling software such as 3D Studio Max. The interface also displays simulation information, such as frame rate and triangle count to the user.

#### Animation

This system deals with how the movement of each agent is computed. We have an in-house animation system for the locomotion of the agents in the virtual world, which receives information from the behaviour system about where each agent should be at each frame (position and velocity values) and calculates a proxy position for that desired position. The animation system will then apply the appropriate motion for each agent. The animations we use for Metropolis locomotion are a motion-captured walk loop with different speeds and turning rates.

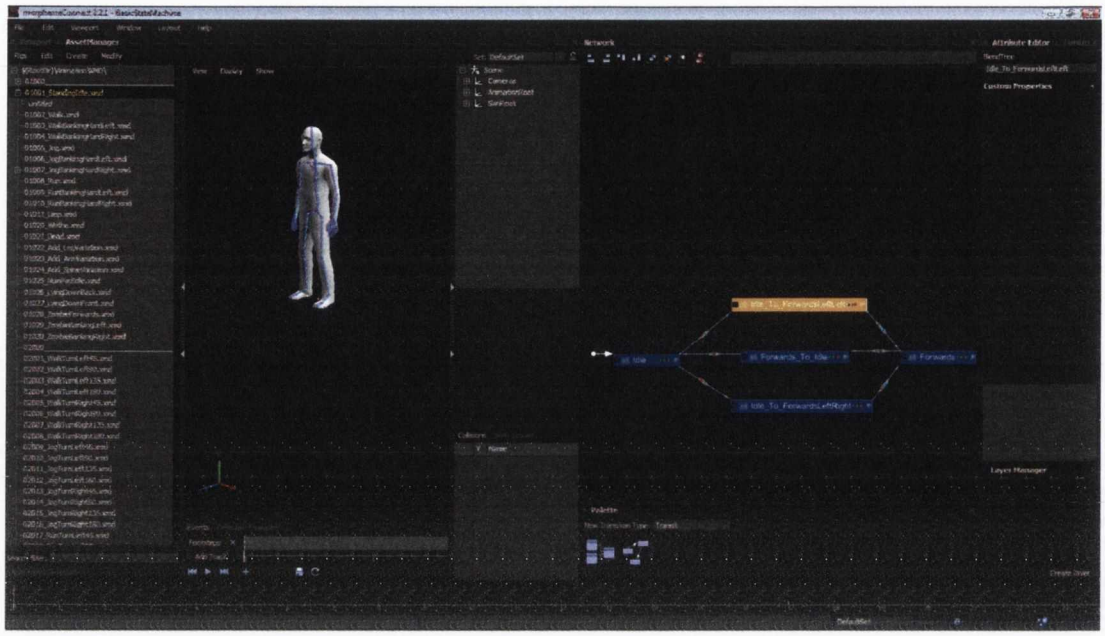


Figure 4.5: Screenshot of Morpheme animation engine, showing animation selection area, real-time animation viewer and FSM editor.

We have another level in the animation sub-system that controls the motions for the conversational characters. For these characters, we use Natural Motion's Morpheme [Mor09]. This is an animation engine, which gives the user an interface to test animation transitions, and create finite state machines (FSM) to transition between a set of animations (Figure 4.5). The system takes as input animations that are exported from a modelling package such as 3DS Max (which, in our case, are created using motion capture) and exports as output an FSM and the corresponding set of animations, which can then be used in Metropolis.

### Behaviour

The behaviour sub-system in Metropolis controls the behaviour of the agents who navigate the environment. As this is the main focus of this thesis, we will describe in chapter 6 how we implement the behaviour of our conversational agents. The path-finding and steering behaviours for our agents is based on the method described by Paris et al. [PDB06]. The behaviour manager works from the assumption that human behaviours are situated within a space and time in the environment around them. Therefore, the system tackles the behaviours of the agents in two different ways:

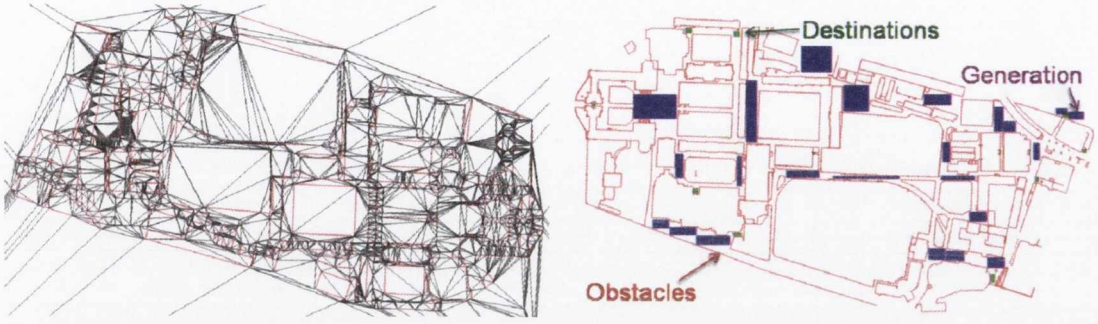



Figure 4.6: Part of our environment (Trinity College) subdivided according to CDT (left), and the informed environment showing obstacles, destination and generation locations (right).

- Describing the environment to identify behavioural components.
- Managing the movement of the agents within this environment.

**Environment Description** The size of the virtual environment we wish to include in our simulation raises a problem for describing the environment in a way that can be used to generate behaviour for our crowd, as a grid-based solution is not appropriate. Instead a method is employed based on Constrained Delaunay Triangulation (CDT) [Che87]. This involves editing the environment into a 2D ground representation, with a small elevation. An in-house tool uses CDT to produce a representation of the environment that describes areas which are either obstacles, generation locations for agents to enter the environment and destination locations to which agents navigate throughout the simulation (Figure 4.6).

**Agent Navigation** Navigation of the agents within this informed environment takes place over two steps. First, for the *global navigation*, the path for each agent is planned from its current position to its destination at each time step. A predictive geometric model is then used for *local navigation*. By local navigation, we refer to the efforts of the agent to follow its computed path while reactively avoiding collisions (Figure 4.7).

**Traffic Simulation** Since Metropolis is situated in an urban environment, it is important to include traffic simulation in the system. Therefore, we have also implemented behaviour for cars and buses to navigate around the environment, where they stop at traffic lights, and perform other functions such as rational over-taking (Figure 4.8). The vehicles in Metropolis are based on a method that involves layering a road network within

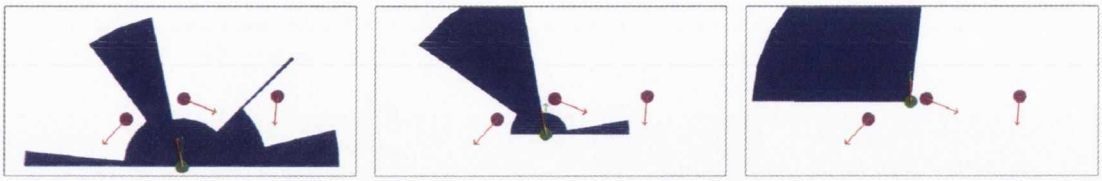


Figure 4.7: Predictive geometric collision avoidance for an agent for three time steps, showing the allowed space the agent can use at each time step in blue.

Figure 4.8: Screenshots of traffic simulation in Metropolis, showing navigable demarcations rendered (left) and a congested scene outside Trinity College (right).

the virtual environment [GR09]. A navigable map is then automatically generated to allow the vehicles find their way around the city. Reactive driving behaviours, such as collision avoidance, are implemented using Fuzzy Logic [GR06].

### Audio

The first problem being tackled by the audio team is how to create ambient noise that matches the sounds heard walking through different parts of a city. Trinity college is a highly pedestrianised area, so the main ambient noise will consist of human-made noise and other nature noises. However, outside the college, there is a lot of traffic and the audio will therefore need to adapt to reflect the user's position in the world. This is achieved using noise-maps, to calculate where audio sources should be placed around the virtual world [MRPD08] (Figure 4.9). Challenges include finding the best way to play back long repetitive audio files, such as traffic audio, setting up a surround sound experience to ensure the user gets the best sense of presence with the directionality and levels of the audio in the system, and synchronising the sounds of footsteps with the animation.

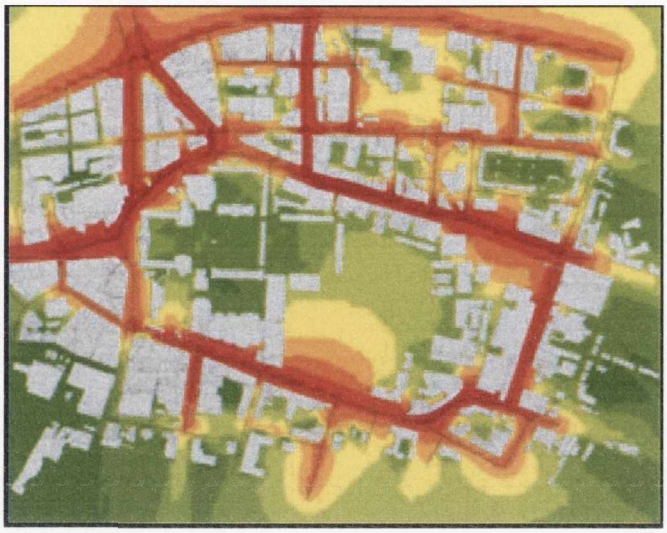


Figure 4.9: Noisemap showing audio levels at different locations around Trinity College.

**Neuroscience**

Neuroscience research consists mainly of psychophysical experiments to compare and contrast the differences between how real and virtual crowds and scenes are perceived. Topics that are currently under investigation are depth perception in a virtual environment, along with horizontal distance perception. We also examine the role audio plays in multisensory scenes for a search task. Finally, we have investigated the role of emotion in real and virtual crowds [MJM+09]. The Metropolis system facilitates running such experiments in a practical way. Experiments are also conducted in the real world, and using a Head Mounted Display (HMD).

## 4.3 Motion Capture Method

We animate our characters using motions that have been captured from real humans. Motion capture involves recording the movements of the joints of an actor at a high frequency. Only the movements alone are captured, represented by a set of moving points, with no information of the appearance of the actor. Performance capture refers to a capture session that includes extra information such as facial expressions, finger motions, voices or eye movements.

We use an optical system, i.e., image sensors triangulate the 3D position of our actor(s) from a number of light-emitting cameras with markers placed on the body of the actor [Men99]. While different kinds of markers can be used for this, we use *passive markers.* These markers are small plastic balls coated in a retro-reflective material that reflects light back to the capture cameras. This provides each camera with a 2D image of where the marker is in space. Using a number of cameras (we use 13), this allows for tracking of enough markers for one or more actors, with 3 degrees of freedom for each marker. Rotational information for the markers is calculated relative to the orientation of 3 or more markers.

For our conversational agents, we conducted a number of motion capture sessions on a group of three actors simultaneously, both with and without audio recording. In this section, we will describe the process of how these motions processed and applied to a virtual character.

### 4.3.1 Motion and Audio Capture Set-up

For a motion capture session, the actor or actors whose movements are being captured typically wear a motion capture suit, which, in our case, are full body suits. Markers are then attached to the suit using velcro at specified points around the body. For the studies described in this thesis, we used 53 markers per body (Figure 4.10), with no facial, or finger motion captured (current research on the project is extending this work to include both faces and hands [JHO10, McD10]). We attach the markers to joints, such as the elbows, knees or pelvis and also use reference markers to help the system deduce the orientation of the body.

Our motion capture system consists of 13 cameras positioned around a rectangular space in our research lab (Figure 4.11). The majority of the cameras are affixed at a height near the ceiling, but we can also place them lower down on tripods, to allow for concentration on the markers nearer the floor, such of those on the feet, as these can often cause problems in processing. In our case, we positioned four cameras in this way.

When recording our conversations with three actors, we had to use a small capture space. Because of increased complications of capturing three people simultaneously, we ensured that the cameras were more focussed on a smaller area to maximise their capturing ability with the extra occlusions introduced by multiple bodies in the space. For the sessions where we recorded audio along with motions, we also had four microphone stands; one in front of each actor capturing their individual contribution to the conversation and another

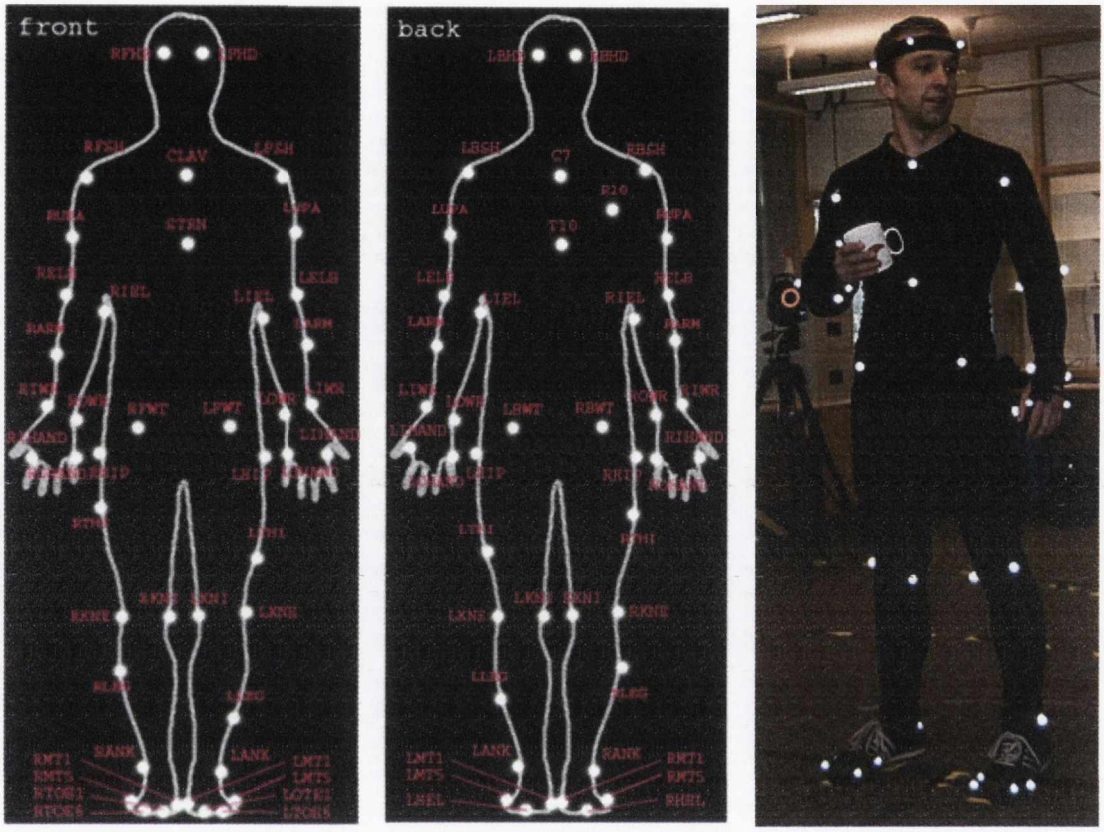



Figure 4.10: Showing marker placement for our motion capture skeleton (left and centre) and a photograph of an actor in one of our motion capture suits, with a motion capture camera in the background (right).

camera equidistant from each actor, recording the conversation as a whole. This introduced further occlusions and added to the complexity of processing the captured data.

The solution we use for our Motion Capture is from Vicon [Vic10], who provide both the cameras and related hardware, and the software, Vicon IQ, as an interface (see Figure 4.12 for a screenshot). At the start of a capture session, the entire system needs to be calibrated. This is done by using an object with markers attached at known positions to calibrate the cameras and determine the origin of the capture space. This also provides the camera positions and lens distortion.

We then select the skeleton that is to be used for the capture session that matches the marker placement on the actor and perform a calibration capture trial. This is a short capture session where the actor moves each joint in the body so that the Vicon system can

Figure 4.11: Photograph of our research lab, showing four of our Vicon cameras.

calculate the range of motion and length of the bones of the actor. This session typically starts and ends with a T-pose, where the actor stands straight with his/her feet forward, arms out and palms down. This is done separately for each person if the session involves multiple actors. A skeleton is fitted to this data and it is used as the reference skeleton for the remainder of the capture session (Figure 4.13). We will describe the pipeline in Vicon, 3DS Max and Morpheme in the next section.

### 4.3.2 Pipeline from Motion Capture to Metropolis

In this section, we will discuss the pipeline we used for capturing the motions of our conversational agents. We will start with the Vicon skeleton we mentioned in the previous section, and describe how we process movement data from the cameras in Vicon IQ, and how we prepare the data to be used in Morpheme using 3DS Max.

Because of the length of the conversations captured, and the fact that we captured three actors simultaneously, we processed the movements of each actor individually from start to finish, and then reconstructed the conversational groups in Metropolis. However, as all three actors were captured simultaneously, they were also fully synchronised in the system.

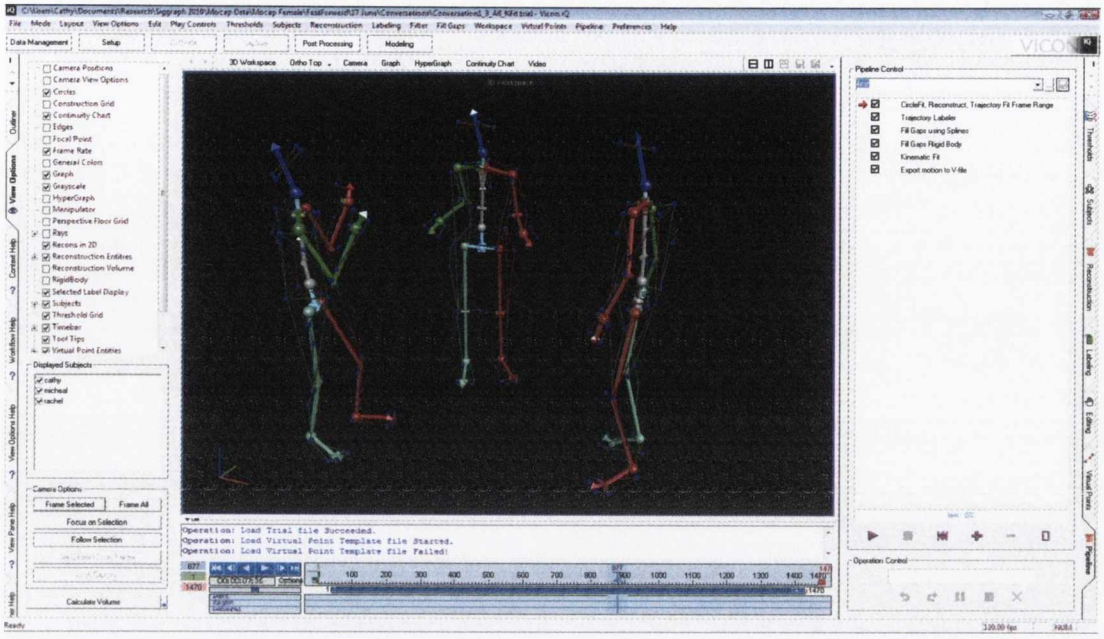


Figure 4.12: Interface of Vicon IQ, our motion capture software showing data for a 3 person conversation.

### Vicon IQ

**Circle Fit:** Once we have completed a capture session, we are left with *raw capture data*, which will be saved as a file in Vicon. If this is opened in the 3D workspace of IQ, the scene will appear to be empty. The first step to viewing the data is to apply a circle fit to the data to show the markers that were captured by the camera. This operation takes the data that has been captured by the cameras and displays them as they were captured in the 3D workspace. The marker data for each subject in the capture session is shown (Figure 4.14(a)), so the data needs to be separated out for each subject in the next step.

**Trajectory Labelling:** We now need to fit the calibrated skeletons to the captured markers. This enables the system to recognise each actor, and allows for each person's movement to be processed separately. We now have a skeleton fit for the actor we wish to work on. We first select the skeletons captured from the calibration session for each actor, and add them to the scene. We then run a *trajectory labeler* operation, that will automatically fit each skeleton to a set of markers (Figure 4.14(b)). We can then remove two actors from view and continue processing them one at a time.

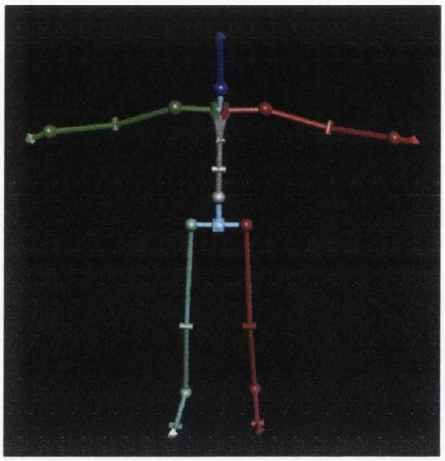


Figure 4.13: Example of a Vicon skeleton in T-Pose.

In some cases, particularly for a capture session like ours, with three actors and extra occlusions, the trajectory labeller will miss some markers, or label some incorrectly. Therefore, a manual trajectory label is conducted after running the operation in the pipeline. This ensures that each marker is labelled correctly for each actor, making the process easier down the line.

**Fill Gaps:** Even after manual labelling, it is likely for a complicated session with three people to have markers missing for some frames. A marker may even be missing for an entire capture session due to occlusion. So, Vicon IQ offers two methods to fill any gaps in the time-line for each marker automatically:

- Fill gaps rigid body: This method of filling gaps for a marker uses a calculation of where a marker should go, based on where other markers on the body are. This is the preferred method for large gaps of missing data.
- Fill gaps using Splines: This extrapolates the position of the marker using the last known position of the marker and the next reappearing position. This method is more suited to short durations of missing information.

**Kinematic Fit:** Once the actor's markers have been labelled accurately for the duration of the capture session, the next step is to calculate the bone animation data for each person. This is done by running the *kinematic fit*. This will affix the skeleton to each actor in the workspace (Figure 4.14(d)).

**Motion Export:** The calculated motion can then be exported in V-file format, which is

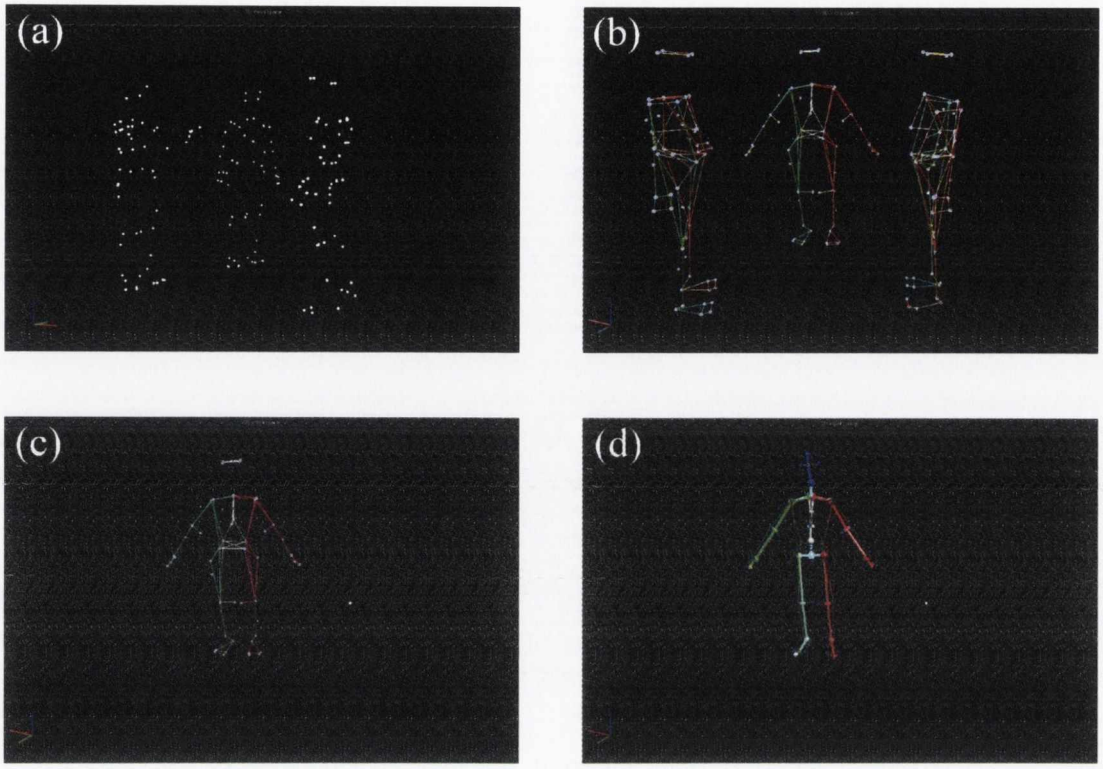



Figure 4.14: Illustration of data processing pipeline in Vicon: (a) shows a circle fit of each subject, (b) shows a skeleton fit and trajectory label operation performed for each subject, (c) shows 2 subjects removed to manually label one at a time and (d) shows a kinematic fit applied to the body after labelling and filling gaps.

a binary motion data format that contains the marker data and global and local rotation data, so that it can be used with other software.

### 3D Studio Max

We use an in-house exporter to transform V-files into a BVH, a format which can be used in 3DS Max. Once we have the BVH, we apply the animation to a biped (a generic two legged character rig that can be adjusted for different characters) and save. Each animation is then checked for artifacts, and they are manually fixed using key-frame animation. We then note each actor's position and orientation from the origin at the first frame of the conversation, and reposition the character at the origin in XY space, facing forward (-90° in Max). This is then exported as a Morpheme readable format (XMD), for the next step of the process.

### Morpheme

Once we have our animations in XMD format, we need to set up a Finite State Machine (FSM) for each conversation for our agents in the Metropolis system. This can be done very easily using Natural Motion's Morpheme software. We select a number of starting positions within the conversation where we want our short animation clips to start. Once we label these events, an FSM can be created in the viewport by selecting an Idle base state, and providing links between this and each of the events we want to use as a starting point for our conversational agents. Once we have built the state machine, it is then exported for each character for each conversation. This is then used, along with the XMD files in Metropolis, where we set up each conversation separately for our agents. As the position of the characters in the XMD files are front facing, we offset each character from the origin by the offset amount we recorded in 3DS Max.

Figure 4.15 shows the Asset Manager workspace, where the events are selected, and the viewport, where the user can see in real-time the animation on the virtual character. Figure 4.16 shows the Network section of the workspace, where the FSM is created. Here, we can see the simple FSMs for one of our conversations. Within the FSM, we have 6 links from our Idle state (State 1) to our conversation state (State 2), reflecting the 6 events we had selected in the Asset Manager workspace, and one link back from the conversational states to Idle state.

## 4.4 Discussion

In this chapter, we have outlined the motivations behind Metropolis, and given an overview of the different research questions being addressed by the various groups involved. These groups collaborate closely to achieve the most immersive, perceptually plausible experience possible for the users of Metropolis. We also provided an overview of our motion capture pipeline and described what happens at each stage of our motion capture process and the subsequent export of the animations into the real-time Metropolis system.

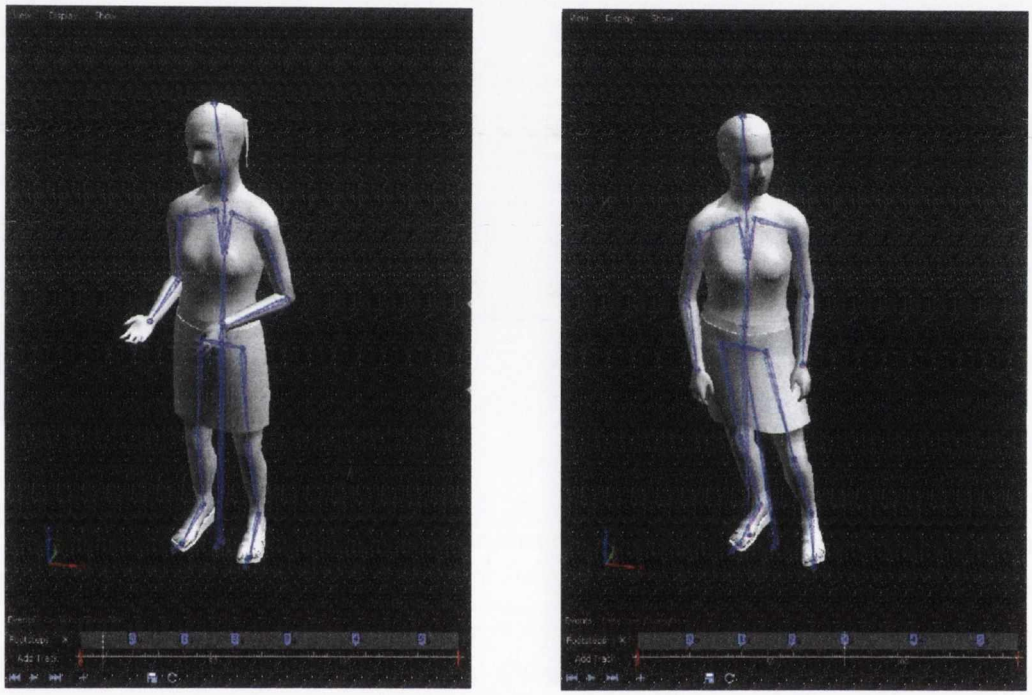


Figure 4.15: Screenshots of Asset Manager workspace, where events (shown at bottom of image) for the FSM are chosen: (left) shows a conversation clip of the talking motions for the character and (right) shows a event with a listening motion.

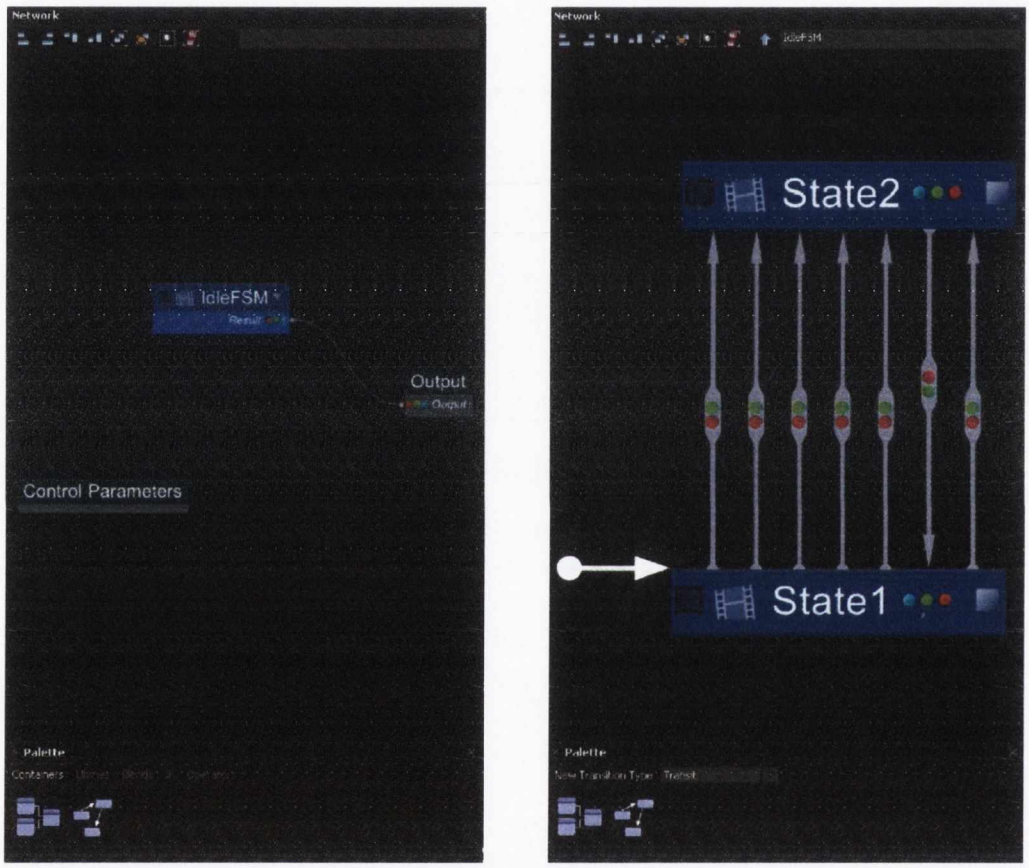



Figure 4.16: Screenshots of Network workspace: (left) shows the FSM structure for a character conversation and (right) shows inside the IdleFSM, with links between idle and conversational states for 6 events.

# Chapter 5

# Plausible Configurations

## 5.1 Introduction

Since studies focusing on how viewers perceive and evaluate the plausibility of synthesised behaviours remain scant, initially we started to work on the first step in an evaluation methodology that was developed to allow us to better elucidate this complex issue. The work described in this chapter consists of the construction of experiments to evaluate viewer's general impressions of static scenes, accounting for factors such as the density, orientation and positioning of pedestrians. We also considered the viewpoint being used. Rather than considering pedestrian properties in isolation, a central theme in our methodology was to also consider the *context* in which it appeared. There can be many different aspects of context, which we define for our purposes as being of two general types: firstly, nearby pedestrians and objects that may affect an individual, and secondly, the type of walking zone that an individual inhabits. As humans, our tendency to adhere to the context of our surrounding environment leads to very specific behaviour patterns as follows:

- Pedestrians tend to walk on paths.
- Pedestrians are often directed towards exit or goal positions when mobile.
- Pedestrians do not walk into obstacles or other individuals.
- Members of the crowd will travel in groups, usually in proximity to each other, and may have other properties in common.
- In restricted areas, pedestrians tend to follow invisible flows.

In this chapter, we will outline some methodologies we have used for crowd formations. We will also discuss a number of perceptual studies conducted as an evaluation of these methodologies and their results. All of the work described in this chapter was carried out in collaboration with Christopher Peters, who was the research fellow working on the behaviour for the Metropolis project during this time. Christopher developed the tool (Metroped) we used for generating some of the stimuli for our experiments. We designed the perceptual experiments together, but I created the stimuli, performed the statistical analysis, and reported the results for inclusion in the papers that we wrote together on the topic.

## 5.2 Static Crowd Formation Experiments

Two separate studies were carried out to investigate the effective methods for populating a crowd with virtual characters. The first step in this research consisted of a short experiment exploring the effects of the orientation of characters in a crowded scene on the perceived realism of that scene. Our second study consisted of a follow on experiment with three blocks, looking at position, orientation and position and orientation combined. For these studies, we created a set of rules for generating plausible positions and orientations for characters in a crowded scene. These rules were developed, in part, based on well known properties of crowd dynamics e.g. tendencies of crowds to form lanes when walking, especially in constrained areas. Other rules were developed from observing real crowds in pedestrian environments, and noting repeated patterns in their behaviour e.g., directions people are likely to walk in an open space relative to the entrances and exits of that area.

Our methodology for these experiments consists of four phases: for *data collection*, data *annotation*, *reconstruction* and *modification*. The methodology was the same across both crowd formation experiments for data collection, annotation and reconstruction phases, but differed slightly for our modification phase.

### 5.2.1 Data Collection Phase

For the data collection phase of our two experiments, a number of videos were taken of two different locations, each representing an archetypical pedestrian movement zone. We refer to these as a constrained or *corridor* location and an unconstrained or open location. An open location represents a relatively large space where pedestrians can to be seen crossing

in many varying directions due to the presence of many possible exits and entrances. In contrast, a corridor location is more constrained, usually with a single entrance/exit at either end and therefore tends to enforce bi-directional movement.

A number of still images were extracted from each video, to be used as a basis from which to create reconstructions of the scenes depicting the real positions and orientations of individuals. These will be referred to here as the *real* category of scenes.

### 5.2.2 Annotation Phase

Each still image was annotated manually to highlight individuals' positions and orientations and their groupings for the stimuli for each experiment (see Figure 5.1(b)). Groups were deemed to consist of one or more individuals, according to their localisation in space and aided by a visual inspection of the video clip surrounding the still image being annotated. Each group was designated by an ellipse, which covered all members of the group and was colour-coded according to whether the corresponding group was static (*black*) or mobile (*yellow*). The orientations of individuals were classified as belonging to one of the following 8 rotations specifying cardinal directions: 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°. Each direction was associated with a unique colour code, to aid visual recognition of the general characteristics of the scene, such as the number of groups containing one, two or three individuals.

### 5.2.3 Reconstruction Phase

Once a photograph had been annotated, the scene was reconstructed by using it as a view-port background in *3D Studio Max* and fitting our 3D model by manually tweaking the virtual camera parameters to ensure an acceptable fit between the photograph and the model. Next, the positions of virtual characters were manually matched up with their real-life counterparts from the photograph, providing a good approximation to the composition of the original scene (see Figure 5.1(c)). Orientations were classified as belonging to one of the 8 cardinal directions and all pedestrians are orientated to match the nearest corresponding direction from this list. This was for both the first experiment, and the orientation block of the second experiment.

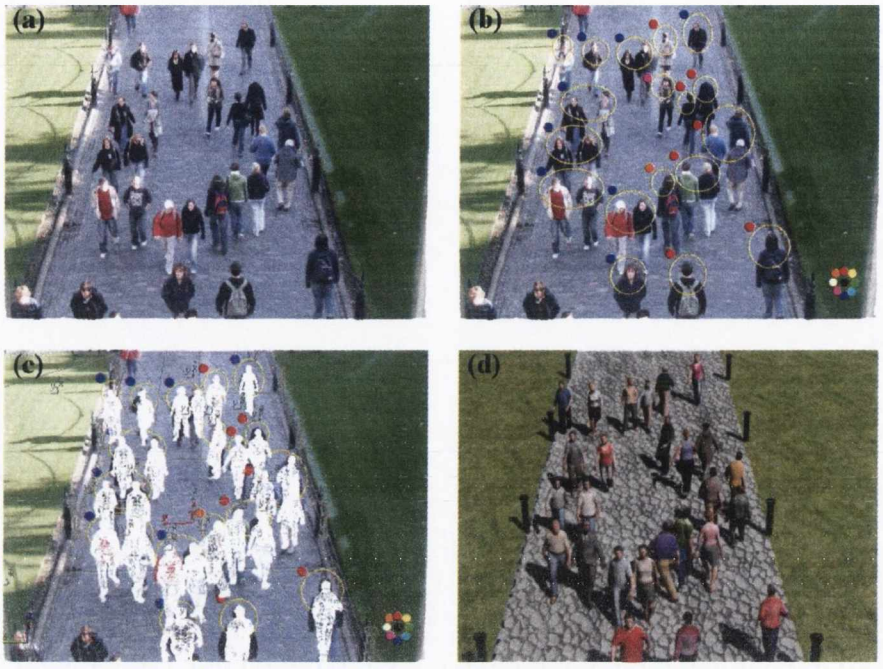



Figure 5.1: Reconstructing the scene. From (a) initial photograph to (d) a virtual scene.

### 5.2.4 Orientation Experiment Modification

Since an important focus for us is to consider the *context* of individuals within the scene, rather than looking at the pedestrian characteristics in isolation, a number of context rules were defined for positioning and orientating characters. There can be many different aspects relating to context, which belong to three general types: firstly, nearby pedestrians, objects and obstacles that may affect an individual; secondly, the type of walking area that an individual inhabits, e.g., in order to specify the general direction of *flow* in that area. Finally, groups may play a large role in people's perceptions of crowds and pedestrians e.g., group size and the number of groups in a scene. Pedestrians were represented in the scenes with either a posed human figure or else a direction-less pawn figure, the latter of which was used in the position studies.

For our first experiment, we examined the role of scene context for characters' orientations only. We identified a number of basic rules for modifying pedestrian orientation and accounting for context. The orientations of all individuals in the scene is manually set to one or more of the 8 cardinal directions, according to one of the four following basic orientation rules:

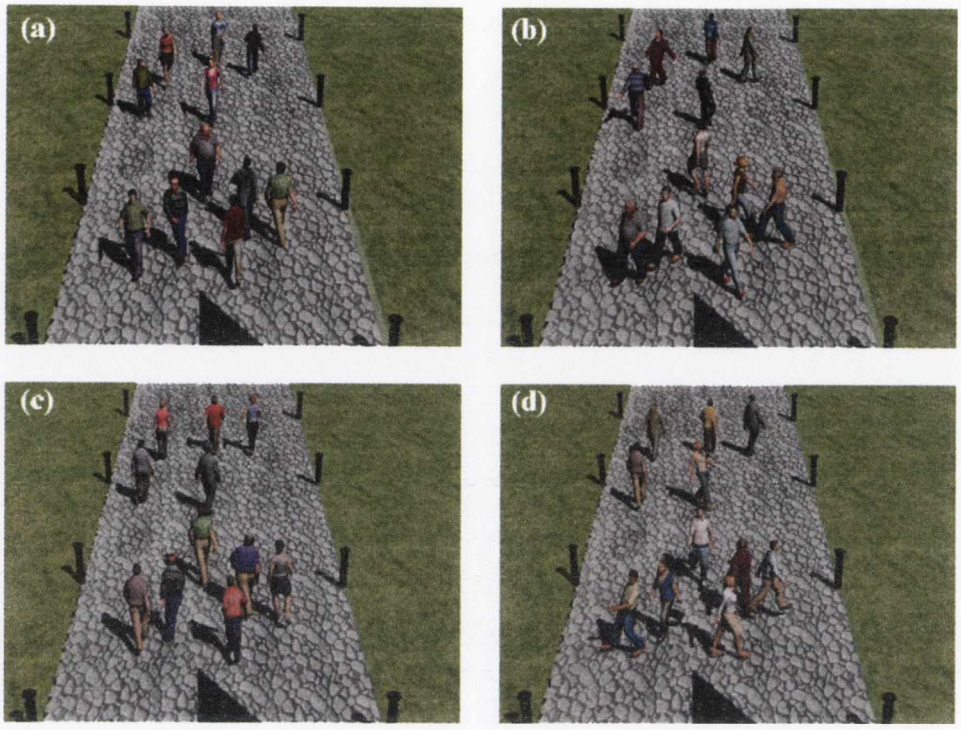



Figure 5.2: Example of the basic orientation rules. Reconstruction (a) of the original photograph, (b) random rule, (c) uniform rule and (d) even rule applied to all pedestrians.

1. Original - the orientation of each individual is aligned to match the annotated orientation from the original scene (Figure 5.2(a)).
2. Random - each individual is assigned an orientation chosen at random from one of the 8 cardinal directions (Figure 5.2(b)).
3. Uniform - a single orientation is chosen at random from one of the 8 directions and all characters are aligned to match it (Figure 5.2(c)).
4. Even - individuals are chosen at random and assigned an orientation from one of the 8 cardinal directions so as to fit into an overall even distribution (Figure 5.2(d)).

### Orientation Context Rules

In addition to these four basic orientation rules, the methodology includes three context sensitive rules for altering orientations. These take into account other individuals and the environment (and are applied in the following order):

1. Flow Sensitive - the orientation of each individual is chosen randomly from a subset of the 8 cardinal directions. This subset is created from the allowable flow directions for the position of the character, based on a ground flow-tile representation.

2. Adjacency Sensitive - the orientation of each individual is chosen randomly from the 8 cardinal directions, but any direction leading to inappropriate facings is disqualified e.g., a character walking into a lamp post would be considered an inappropriate facing.

3. Group Sensitive - Each individual is assigned orientations according to their group membership, rather than on an individual basis, if they were part of a group of 2 or more.

When combining rules, basic rules are applied first in 3DS Max and then contextual rules are manually applied afterwards in the order listed above, overriding the basic rules. In the case of flow sensitivity for example, the result is that the basic rule is only applied in ground areas that have not been assigned flow directions.

### 5.2.5 Position and Orientation Experiments Modification

For our second experiment, we investigated pedestrian formations looking at the following properties: only individuals' positions, only their orientations and both their positions and orientations. For this experiment, our scenes were created with a combination of 3DS Max and the aid of an in-house tool, Metroped (see Figures 5.3 and 5.4).

Metroped was created by Christopher Peters [PO] as a tool to support and generate AI behaviours for virtual pedestrians. It allows the user to annotate a virtual environment with zone and path information to highlight waypoints for character navigation. Along with this annotation tool, it imports a scene mesh to visualise the environment and can export data in xml format to be used in other applications. Of interest to us, is the application of this tool for research purposes. The user can annotate an environment using a grid, which then allows him/her to create positions and orientations for crowd formations (as seen in Figure 5.3). For those rules modifying the original positions or orientations for this experiment, a number of steps were carried out in MetroPed to allow for the automatic generation of data. A grid was created in order to fit the area that was to appear in the final scene. Each cell in the grid was then manually assigned with attributes, such as walkability and flow direction(s), if any. In order to ensure that characters did not face inappropriate directions at the borders of the scenes, where there are usually obstacles,

a flow direction was added to these borders. These were a width of one cell grid, and allowed characters close to borders only to be orientated facing away from the obstacle (when random rules were not employed). As this grid is created manually, an element of intuition needs to be employed by the user. For our open and corridor locations, the placement of flow cells was heavily guided by videos of pedestrians in the actual location. Once this grid has been created, the user can then choose parameters such as; number of characters to be positioned, number and sizes of groups and whether random placements or context placements should be employed. Metroped then uses either random functions or our set of context rules to either give positions and/or orientations for a number of characters within the scene. Once a scene was generated with the tool, it was exported to 3D Studio Max as a set of dummy nodes, each of which contained the transformation for a particular pedestrian.

In this experiment, the density of pedestrians in the two locations (corridor and open) was kept as equal as possible, to minimise the variation in responses from participants. Based on the area of the location visible to the viewer, it was calculated that 30 pedestrians in the open zone would roughly correspond in crowd density terms to 12 in the corridor zone, and therefore still images with these numbers were selected from the extracted video stills for the respective location types.

### Position Context Rules

For the position formations, the characters are displayed as *pawn* characters with no discernible orientation. The images used for the position block of the experiment are shown in Figure 5.5. Here, the only information available is the position of the characters. The images for this block of the experiment contained Real (positions of characters were the same as those in the still image), Random (characters assigned random positions within the grid) and Context; where individuals were assigned a position according to the following context rules (applied in the following order):

1. Bounds Sensitive - An individual can only be assigned a position that is part of a designated walkable area. In these experiments, grass was regarded as being *out-of-bounds* during the application of contextual rules.

2. Group Sensitive - Individuals will be assigned a position to maintain the number and size of groups in the original still image. In other words, in the case of a group of pedestrians, the position of the group only will change, not the position of the individuals making up the group.

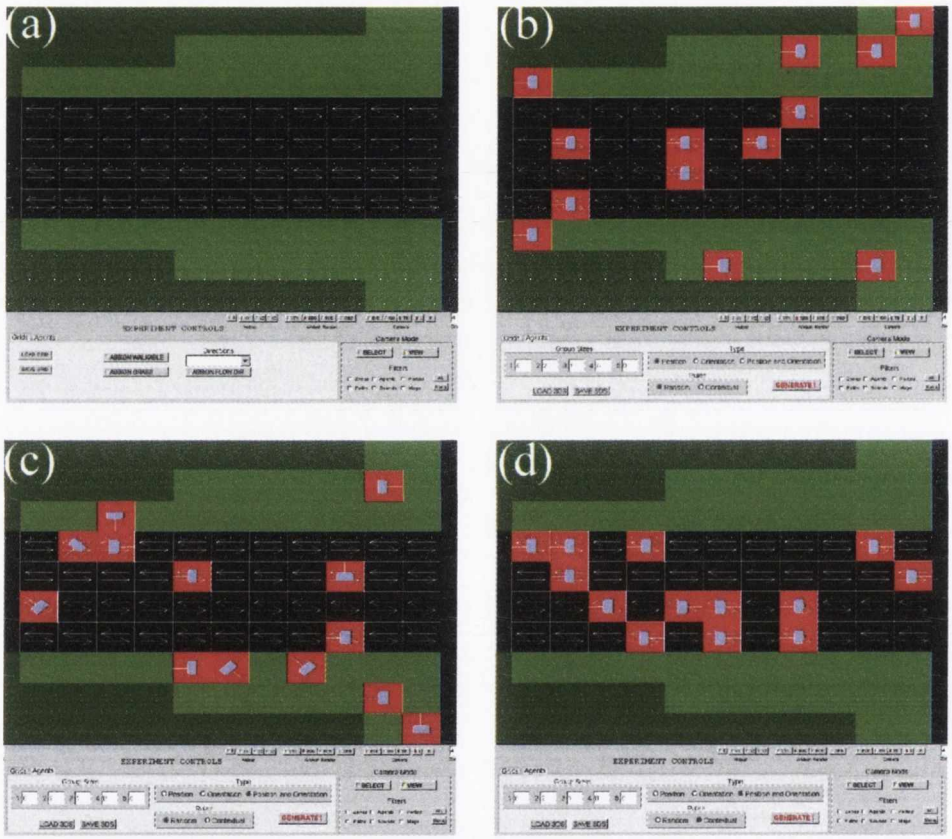



Figure 5.3: View of MetroPed for the corridor location: (a) design of the grid with out-of-bounds and flow directions; (b) examples of output for random position; (c) random position and orientation and (d) context modified position and orientation rules.

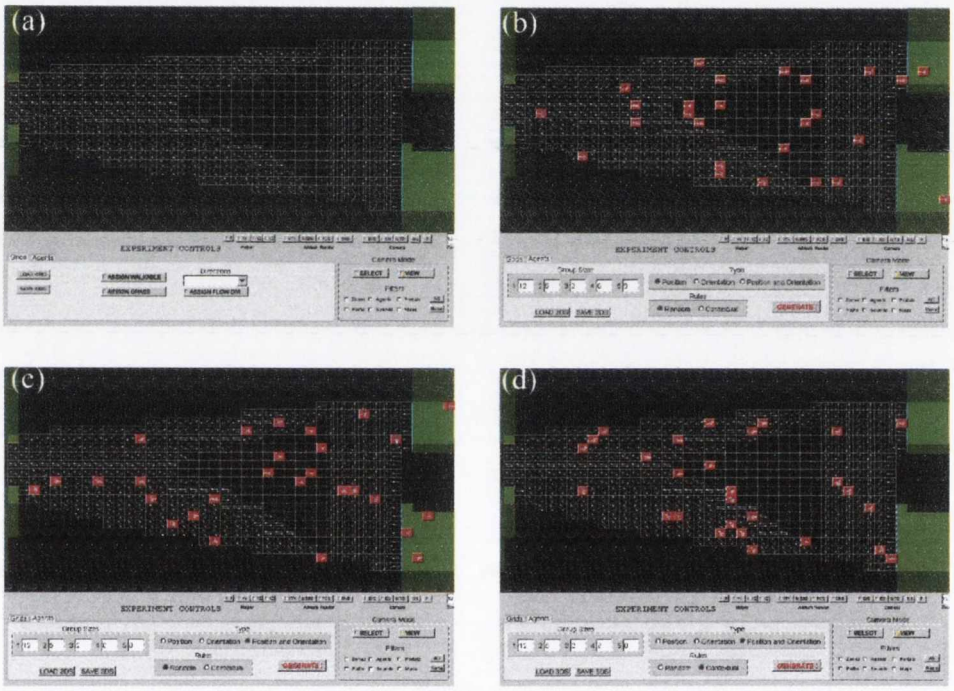



Figure 5.4: Generation of pedestrian transformations for the open location: (a) design of the grid with out-of-bounds and flow directions; (b) examples of output for random position (c) random position and orientation and (d) context modified position and orientation rules.

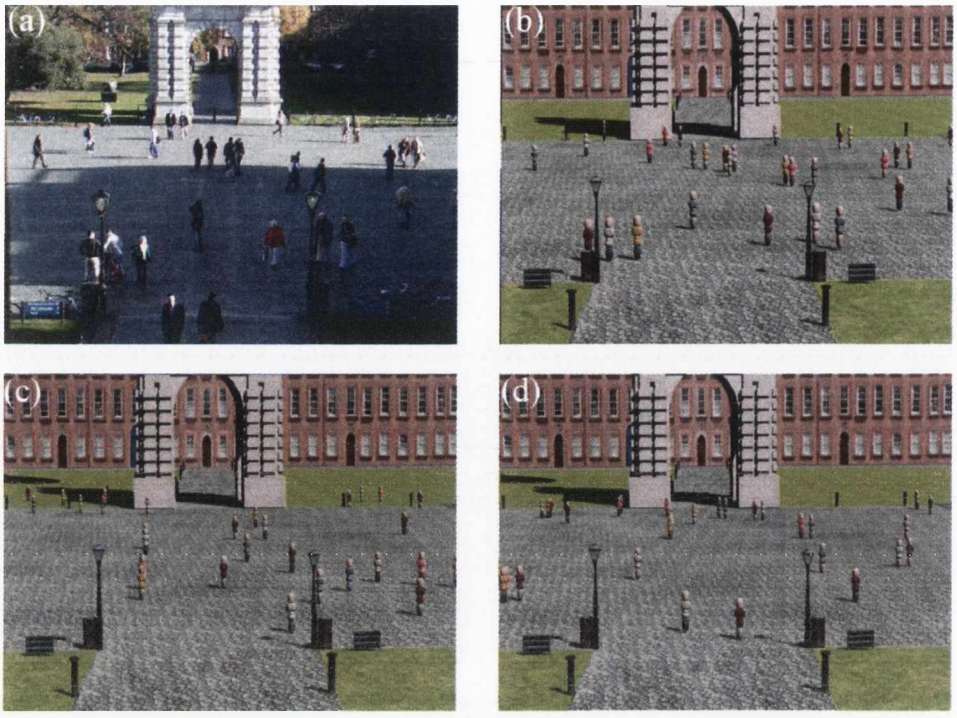



Figure 5.5: Position images for the open location: (a) the original still image; (b) virtual Real image; (c) virtual Random image and (d) virtual Context image.

**Orientation Rules**

The modified images used for the orientation block contained standard human character models with discernible orientations (see Figure 5.6). The character positions remained the same as those from the original still image; only their orientations were changed. Again we had Real, Random (each individual assigned one of the 8 cardinal orientations at random) and Context; according to the context rules: flow sensitive, adjacency sensitive and group sensitive (as described in Section 5.2.4).

## 5.3 Orientation Experiment

For the first of our experiments looking at formations of pedestrian crowds, we looked at the role of context on character's orientations. We defined a number of *master scenes*, each of which was a reconstruction based on an original, annotated photograph of the scene (see Section 5.2.3). In these scenes, characters' positions were matched as closely

Figure 5.6: Orientation images for the corridor location: (a) the original still image; (b) virtual Real image; (c) virtual Random image and (d) virtual Context image.

as possible to those in the original photograph, and orientations were matched according to the closest of the 8 cardinal directions defined. No attempt was made to match the actual models or poses of characters with the original scenes; however, we ensured that the model and poses assigned to a character differed between each image. Fabricated scenes were created manually by applying our orientation rules (see Section 5.2.4) to the master scenes in order to change the orientations of the individuals. There were therefore three general categories of imagery used in our study: master scenes corresponding to the real scenes, images derived from the master scenes where orientation was altered as defined by the non-contextual orientation rules, and images derived from the master scenes where orientations were defined according to the context sensitive rules. For the context sensitive rules, flow lanes were created manually for the scene from an inspection of our annotated corpus of pedestrian movement. Adjacency rules were not applied, as the crowd density in our corpus did not create orientation conflicts e.g., pedestrians appearing to walk into each other.

The goals of this perceptual study were to test if participants would find the reconstructed real scenes more plausible than those where the orientations were synthetically generated, and for the latter, to compare responses between those that were created with and without the use of the context sensitive rules. Our first hypothesis was that participants would preferentially rate those original images reconstructed from actual scenes (the master scenes) above those created manually according to the application of our rules. Secondly, we hypothesised that for the images where the orientations were synthetically generated, those that considered the contextual rules detailed in Section 5.2.4 would receive more favourable ratings, than those not considering any form of context.

### 5.3.1 Method

Twenty five participants (9F, 16M) ages 18 to 30, sat in front of a projected display in a research lab. Divided into two groups, they were shown a different randomised ordering of images and given an instruction sheet: two photographs of the corridor and open location were shown and they were told that the images they were about to see were derived from real photographs, but in some the character orientations were real, while in others they were synthetically generated. For each image displayed, participants were asked if they thought the orientations of the characters were real or synthetically generated. Between each trial, a blank screen was displayed for 5 seconds after which a noise alerted participants for the next trial. In this, and our other experiments on crowd formations,

we drew participants' attention to either the position or orientation of the characters, and asked them to focus on this when making their decisions. We did this to ensure that participants' responses about how realistic they found the scenes to be were not due to unintentional artifacts, such as exaggerated character poses or outfit colour schemes, as we did not want anomolies of these properties to affect our results.

We had seven conditions, so each image seen by participants could thus be categorised as belonging to one of the following seven different types depending on the rules applied to the orientations: Original, Random no context (RandomNC), Random with context (RandomC), Even no context (EvenNC), Even with context (EvenC), Uniform no context (UniformNC), and Uniform with context (UniformC). Those listed as *With Context* employ a basic rule modified by the contextual rules (Section 5.2.4), while those listed as *No Context* employ only a basic rule. A total of 72 trials were presented to participants, broken down as follows: 2 realism levels (real and synthetic orientations counter-balanced) $\times$ 2 locations (open and corridor) $\times$ 6 orientation rules (RandomNC, RandomC, EvenNC, EvenC, UniformNC and UniformC) $\times$ 3 repetitions.

### 5.3.2 Results

We averaged responses over each of the three repetitions for each orientation type, for both open and corridor locations. A two factor ANalysis Of VAriance (ANOVA) with repeated measures showed a main effect of orientation type ($F_{6,144} = 18.54, p < 0.001$). Post-hoc analysis was then performed using a standard Newman-Keuls test for pairwise comparisons among means.

Figure 5.7 shows that, overall the original master scenes, featuring real orientations, were judged as *real* significantly more times than any of the other orientation types ($p < 0.03$ in all cases), implying that people are able to distinguish the real cases from the synthetic ones based primarily on differences in orientation. In comparing the synthetic scenes, overall RandomC, EvenC, UniformC and UniformNC were perceived as the most realistic, while RandomNC and EvenNC were perceived as the least realistic of all cases ($p < 0.001$). This suggests that the addition of our context sensitive rules improves the perceived plausibility of the scenes. Looking more closely at the application of context to each of the basic rules in isolation, it can be seen that for the Random and Even cases, the application of contextual rules results in significantly higher perceived realism for the orientations.

For the Uniform case, the use of contextual rules did not have a significant effect on plausibility. One possible explanation for the higher than expected performance of UniformNC

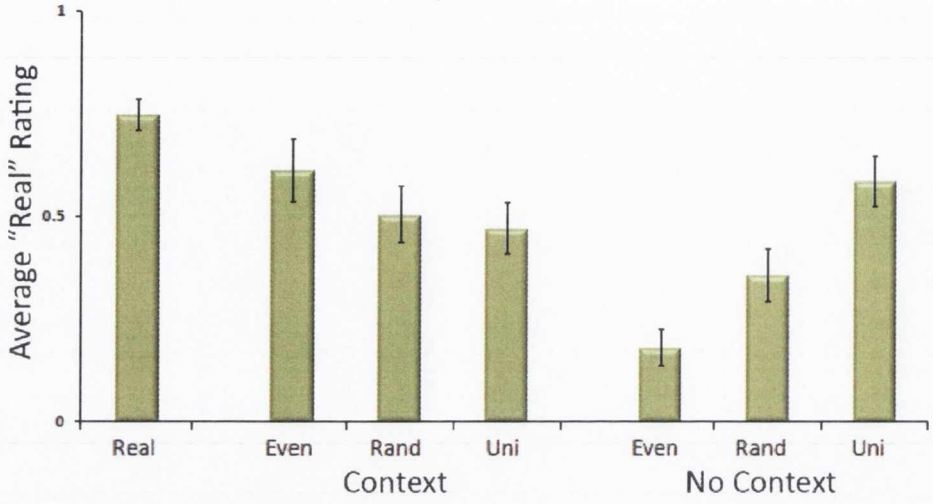



Figure 5.7: Main effect of orientation. On the Y axis of the graph, 1 corresponds to a 'real' response and 0 corresponds to 'synthetic'.

could be that the orientations of all individuals were judged to be in a plausible flow direction, particularly for the constrained location case (see Figure 5.2(c)). In the open location UniformNC case, close inspection reveals that a number of pedestrians in the background appear to be walking towards grass or obstacles. It is possible that in the amount of time available, participants only viewed those characters nearer the camera when making their assessment or could not properly see those in the background. Further studies employing eye-tracking equipment may provide more clues to resolve this, as it is of relevance for employing level of detail schemes.

This relates to another important consideration in our work, that is, the effect of the type of walking zone on perception. There was an interaction between the location and orientation type ($F_{6,144} = 22.77, p < 0.001$). Figure 5.8 shows that the ratings for RandomNC and EvenNC were significantly different for the open location than for the corridor location; these rules were considered significantly more realistic in the open location ($p < 0.001$ in all cases). The opposite effect occurred for RandomC and UniformC, where the open location was considered significantly less natural than the corridor ($p < 0.001$). This seems to suggest that in an open location, viewers are more tolerant to differing orientations than when they are viewed in the context of a constrained location, due to the restricted number of possible directions of movement. For corridor scenes specifically, RandomC, EvenC and UniformC were perceived to be equally natural to the Real orientation and therefore any of these orientation types could be used in a similar walking zone type. From our data,

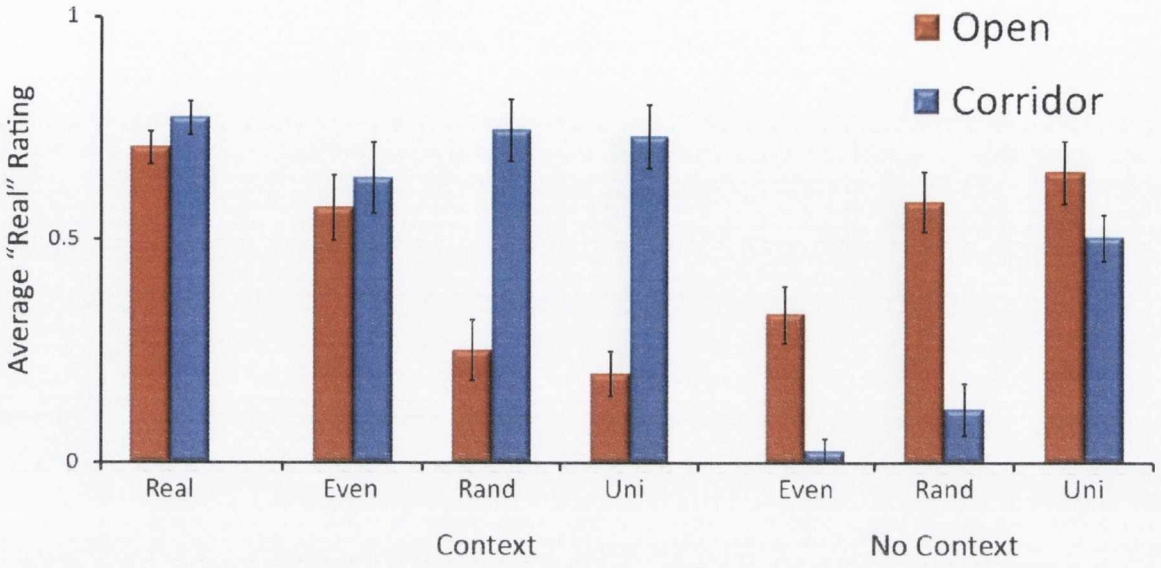



Figure 5.8: Interaction between orientation and location. On the Y axis of the graph, 1 corresponds to a 'real' response.

it can be seen that RandomNC, EvenC and UniformNC were perceived similarly to the Real orientation type for the open location and therefore could be suitable for use in open location types.

Overall, it can be seen that the success of the application of the contextual rules is contingent to different degrees on the type of location that the pedestrians inhabit. In the constrained location, the contextual rules (i.e., RandomC, EvenC and UniformC) are more significant in contributing towards the perception of plausibility. However, for the less constrained open location, the use of contextual rules does not appear to be so important (i.e., RandomNC and UniformNC, as well as EvenC performed well in these situations) looking at orientation alone for these scenes.

## 5.4 Position and Orientation Experiment

Our next experiment in this series concentrated on position along with orientation. We reduced the number of synthetic conditions and used only Real, Random and Context for each block of this experiment to allow us to focus on position as well as orientation properties. We considered position alone, orientation alone (using the new conditions) and position and orientation in combination.

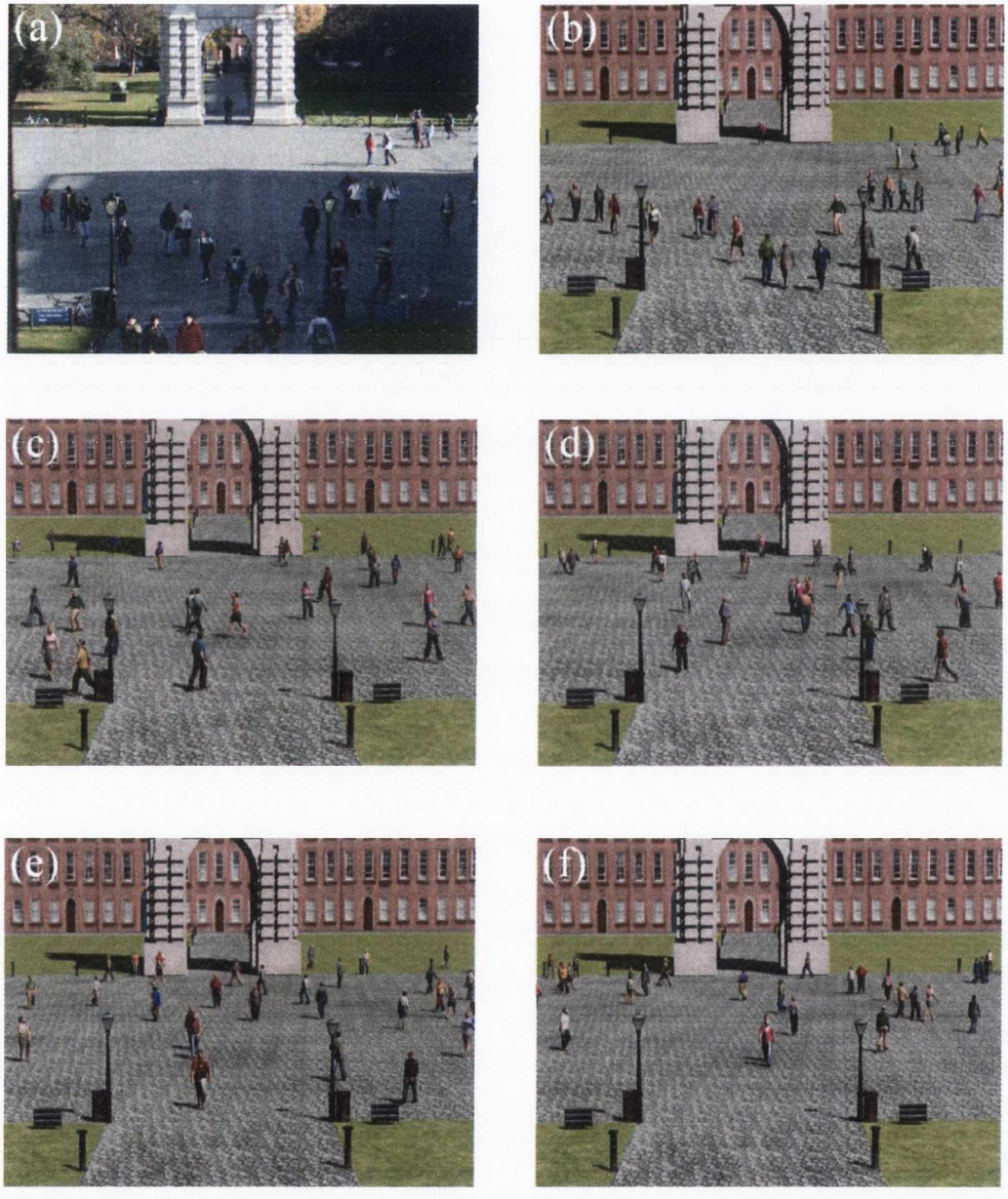



Figure 5.9: Position and orientation images for the open scene: (a) Photograph; (b) Real; (c) Random; (d) Position Context; (e) Orientation Context and (f) Both Context.

### 5.4.1 Method

Thirty two participants (12F, 20M) age 18 to 30, were seated in front of a computer screen. They were told that the experiment consists of three blocks and were given an instruction sheet: two photographs of the corridor and open locations were shown and participants were told that the images they were about to see were derived from real photographs. However, in some the character formations were real, while in others they were synthetically generated. For the first block of the experiment the participants were told to focus only on the positions of the characters. For each image displayed, participants were asked if they thought the positions of the pawn figures were real or synthetically generated. For the second block, participants were asked to look at the orientations of the characters only and judge if they were real or synthetically generated. For the final block of the experiment, participants were asked to take both position and orientation of the characters into account and to judge whether the scenes were real or synthetically generated. The reason that we presented the blocks in this order was to avoid biasing participants. If the pawn figures were viewed after the humanoid characters, this could have caused them to perceive the scenes as less realistic due to the reduced realism of the characters, which was not the effect being tested. Furthermore, the scenes with position and orientation combined were presented during the final block, to prevent participants from taking position into consideration when conducting the orientation only trial. Between each trial, a blank screen was displayed for 5 seconds, after which the number of the next trial was displayed alerting participants.

For the first two blocks of the experiment, an equal number of Real and Virtual scenes were presented to the participants in 2 different random orders. The virtual scenes were divided into either Random or Context. So, participants viewed 24 trials in total for each block: 2 Locations (Corridor/Open) $\times$ 2 Realism Levels (Real/Synthetic) $\times$ 2 Formation Types (Random/Context) $\times$ 3 Repetitions.

For the final block of the experiment, an image could be categorised as belonging to one of the following five different types: Original, Random, Position Context (context rules applied for position only), Orientation Context (context rules applied for orientation only) and Both Context (see Figure 5.9 for examples). This experiment consisted of 48 trials, broken down as follows: 2 Locations (Corridor/Open) $\times$ 2 Realism Levels (Real/Synthetic) $\times$ 4 Formation types (Random, Position Context, Orientation Context, Both Context) $\times$ 3 Repetitions.

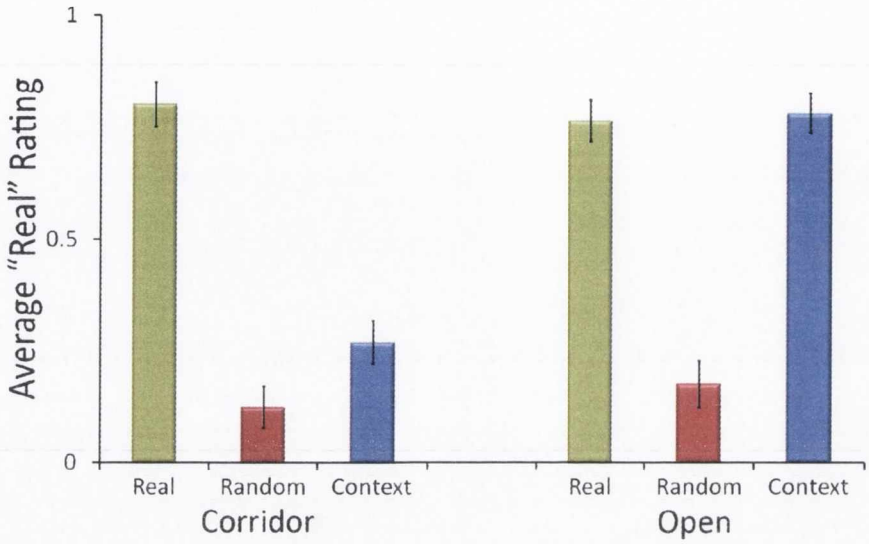



Figure 5.10: Graphs showing position results for Corridor and Open locations.

### 5.4.2 Results

#### Position Results

We averaged responses over each of the three repetitions for each position type. A two factor ANOVA with repeated measures showed a main effect of scene ($F_{1,31} = 17.895, p < 0.01$), in that the open scene was perceived to be more realistic than the corridor scene, and position type ($F_{2,62} = 70.077, p < 0.01$) where the real positions were perceived to be more real than virtual positions. There was also interaction between the two ($F_{2,62} = 23.476, p < 0.01$), where the real positions were perceived as real more often for the corridor location than the open location but the virtual positions were both perceived as real more often for the open location. Post-hoc analysis was then performed using a standard Newman-Keuls test for pairwise comparisons among means.

We found that the original corridor scenes with real positions were judged as *real* significantly more times than either the context or random virtual corridor scenes ($p < 0.01$ in all cases), implying that participants are able to distinguish the real cases from the synthetic ones based primarily on differences in position for constrained locations. However, participants perceived the scenes with context ruled positions to be more real than those with random positions ($p < 0.05$).

Looking at the open location, while participants perceived the real positions more *real* than

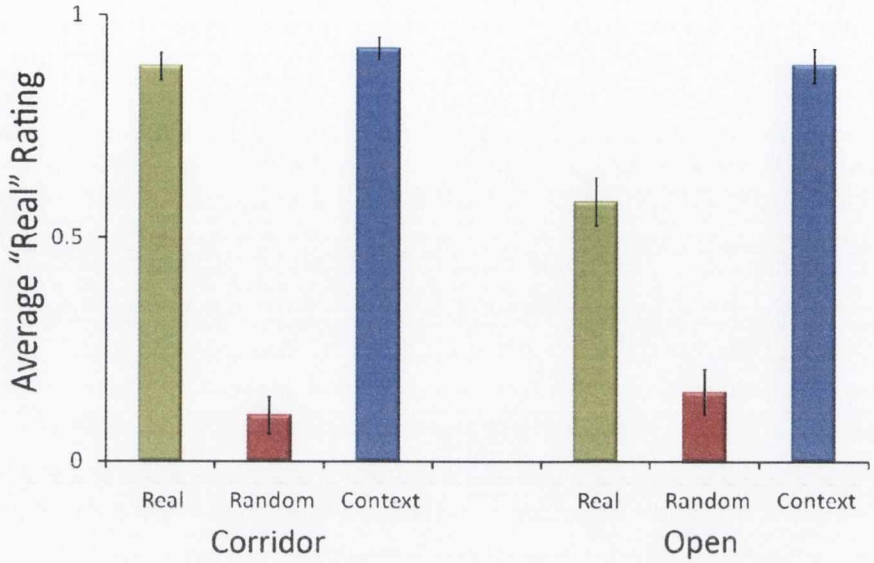



Figure 5.11: Graphs showing orientation results for Corridor and Open locations.

the random positions ($p < 0.01$), they judged the scenes with context-based positions to be as realistic as those with the real positions. Figure 5.10 shows these findings on a scale of 0 to 1, where 0 means they were perceived as synthetically generated and 1 means they were perceived as real.

As can be seen from Figure 5.10, it can be seen that, while the context rules applied did improve the perceived realism of the scene, the participants could still distinguish the real positions from the synthetically generated ones. For for the open location, the context rules applied had a greater effect on participants' perception of the scene. The fact that participants judged the scenes with our context rules to be as real as those scenes with real positions indicates that these rules would be an adequate way to populate scenes when positioning characters in an open or unconstrained location. While the rules do not have such a strong effect on the realism of corridor locations, it has been shown here that they would be a suitable alternative to random positioning for characters, while not necessitating manual placement of individuals.

### Orientation Results

We averaged responses over each of the three repetitions for each orientation type. A two factor ANOVA with repeated measures showed a main effect of location ($F_{1,31} =$

$11.508, p < 0.01$), where the corridor location was perceived to be more real for this experiment, and orientation type ($F_{2,62} = 162.04, p < 0.01$) where context orientations were perceived to be the most real, followed by the real orientations, with the random orientations being judged the least real. There was also interaction between location and orientation ($F_{2,62} = 12.040, p < 0.01$), where participants judged the real and context scenes as real more often for the corridor location but judged the random scene as real more often for the open location. Post-hoc analysis was then performed using a standard Newman-Keuls test for pairwise comparisons among means.

We found that the original corridor scenes with real orientations were judged as *real* significantly more times than the random virtual corridor scenes ($p < 0.01$), but the scenes with context rules were judged to be as real as the original scenes, implying that participants are no longer able to distinguish the real cases from orientations generated using our context rules for constrained locations.

Looking at the open location (see Figure 5.11), while participants perceived both the real and context-based orientations to be more *real* than the random orientations ($p < 0.01$), they judged the scenes with context-based orientations to be *more* realistic than those with the real orientations. One explanation for this interesting result is that the unconstrained nature of the open location evokes a tolerance when viewing character formations in these location types than in more constrained types, where viewers are more adept at spotting peculiarities.

**Position & Orientation Results**

We averaged responses over each of the three repetitions for each type of formation. A two factor ANOVA with repeated measures showed a main effect of location ($F_{1,31} = 15.754, p < 0.01$), where the open location was perceived to be more real than the corridor location, and formation type ($F_{4,124} = 54.093, p < 0.01$), where scenes with both position and orientation context rules were judged to be almost as *real* as the real scenes. There was also interaction between the two ($F_{4,124} = 16.615, p < 0.01$) where participants perceived the real formations to be more real for the corridor location, but judged each virtual formation to be more real for the open location. Post-hoc analysis was then performed using a standard Newman-Keuls test for pairwise comparisons among means.

We found that for the corridor locations, the real formations were judged as *real* significantly more times than any of the four virtual formations ($p < 0.05$ in all cases). This implies that, given position and orientation information for a constrained location, par-

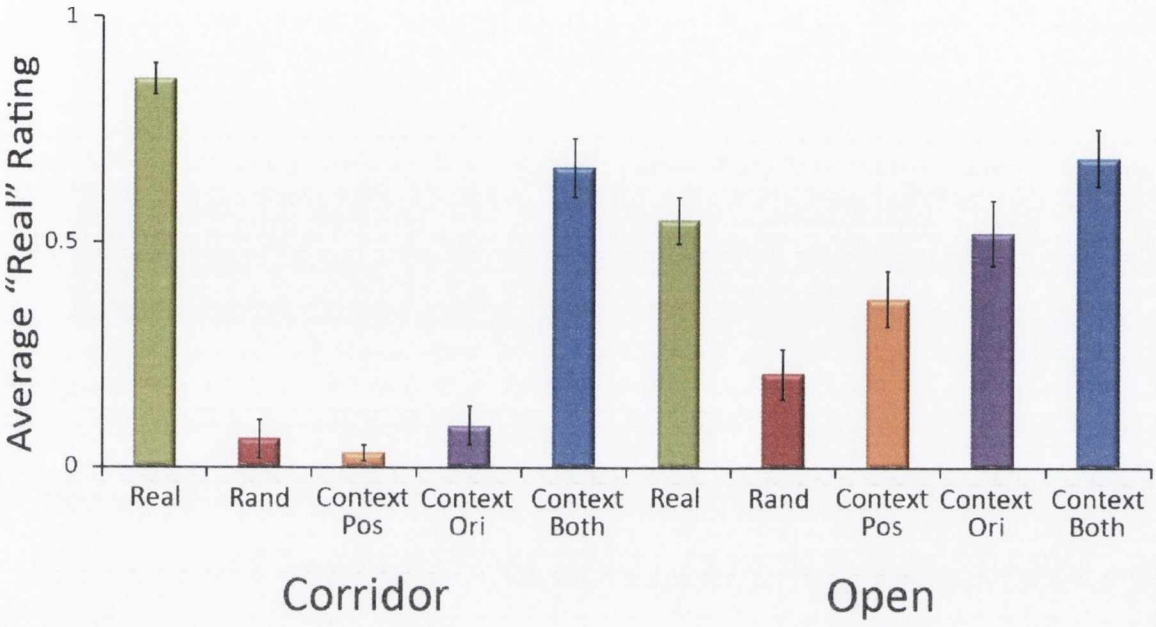



Figure 5.12: Graphs showing position and orientation results for Corridor and Open locations.

ticipants are able to tell real scenes from virtual scenes. Having said that, the scenes where both position and orientation context rules had been applied were perceived as real significantly more often than either the random, position context or orientation context scenes ($p < 0.01$ in all cases).

In the open location, the results were, again, more complicated, with the real scenes being judged as real more often than the random formations and scenes with position context information only ($p < 0.01$ in all cases). For the scenes with orientation context and both position and orientation context, the participants judged them to be as real as the real scenes. This reiterates the possibility that participants find it more difficult to distinguish between real and synthetic formations for less constrained locations.

From Figure 5.12, it can be seen that, while participants can still differentiate between real and synthetic corridor scenes, they cannot differentiate between these for open scenes. Interestingly, for the corridor location, the use of either position or orientation context on its own was not effective, as participants perceived these to be as synthetically generated as the random scenes. However, for the open location, participants perceived the scenes with random positioning and orientation context rules to be as real as both the real scenes and the scenes with both position and orientation context rules applied. This would imply

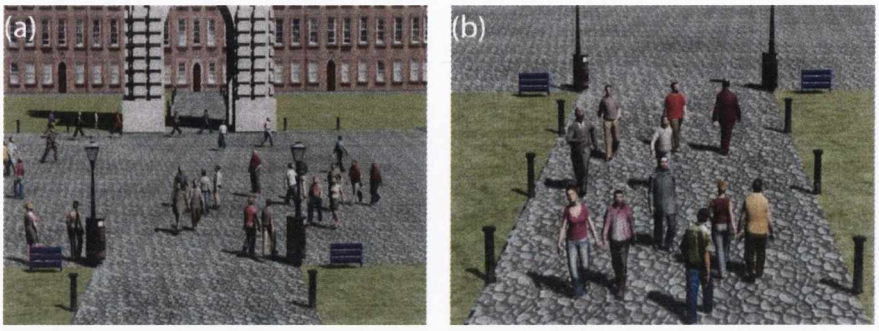



Figure 5.13: Showing example of real scenes used in Experiment 2 for (a) the open location and (b) the corridor location.

that, in an unconstrained location with a large number of people, orientation seems to be of greater importance when it comes to plausibility.

Overall, it can be seen that participants found the context formations less plausible in the more restricted, or corridor locations. This could be due to highly repeated formations, with little variation. Slight extensions to our context rules could add an element of variety to these rules, however, and potentially improve the plausibility of the scenes. In these locations, people can often walk slightly diagonally, or side-step, to avoid collisions with on coming individuals or groups. We did not allow for this in our context rules, we disregarded collisions by ensuring characters were facing the same direction, rather than facing into each other. An addition of side-stepping ability or allowing for slightly diagonal directions to avoid collisions could potentially improve the plausibility of context formations in more constricted locations.

## 5.5 Real Scene Variation Experiment

For each block of our position and orientation experiment, different real formations were used, with participants perceiving some of these as plausible a relatively low number of times. We conducted a second experiment in order to investigate whether any of these results could have been caused by any specific real formation appearing unrealistic.

### 5.5.1 Method

Using the same method as outlined in Section 5.2.3, we reconstructed four virtual replicas of different real scenes for both the open location and the corridor location. We kept the number of characters in each scene constant, to ensure no effect of population density. There were three different variations of each scene, where we placed different character models in the same positions and orientations. This was done to avoid participants becoming familiar with a specific character model thus influencing their judgements. An example of one of the four different real scenes used for this experiment for each location can be found in Figure 5.13.

In addition to the real scenes, we created 6 scenes using the context rules as outlined in Section 5.2.5 for both locations. We also generated 6 scenes for each location using random placement. Unlike our previous experiments, where random placement included out-of-bounds regions such as grass areas, we only allowed random placement within walkable areas. This was due to the difference in the ratio of walkable areas to out-of-bounds regions across the two different locations. For the open location, there is only a small area out-of-bounds towards the back of the scene, whereas in the corridor location, almost half of the scene is designated as out-of-bounds. This difference could have had an effect on participants' perception of random behaviours, since an inappropriate position in the constrained location could be more easily noticed than in the unconstrained location. An example of a context scene for the open location and a random scene for the corridor location can be seen in Figure 5.14. For each random and context scene, we kept the number of characters constant, with 30 characters in the open location scenes and 11 in the corridor location scenes. For the context scenes, we kept the number and sizes of groups the same as in the real scenes.

Ten participants (1F, 9M) aged 18 to 30, were seated in front of a computer screen and were given an instruction sheet. For each image displayed, participants were asked if they thought the pedestrian formations were from a real scene, or whether they were synthetically generated. A total of 48 trials were displayed for 4 seconds each as follows: 24 Real Formation Scenes (2 Locations $\times$ 4 Variations $\times$ 3 Repetitions) and 24 Synthetic Formation Scenes (2 Formation Type $\times$ 2 Locations $\times$ 6 Variations). Between each trial, a blank screen was displayed for 5 seconds, after which the number of the next trial was displayed to alert participants.

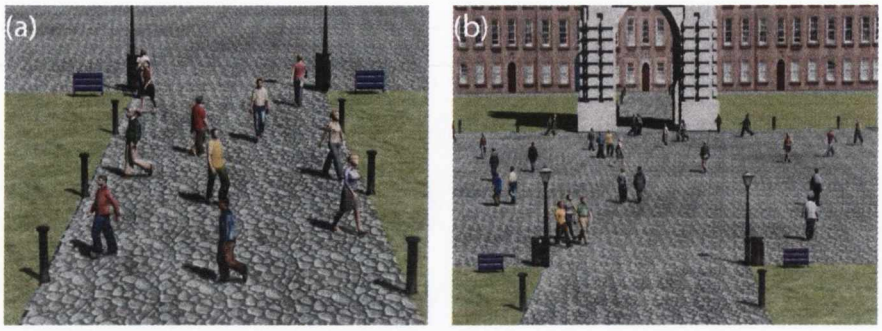



Figure 5.14: Showing (a) random and (b) context random scenes for the constrained and unconstrained locations respectively.

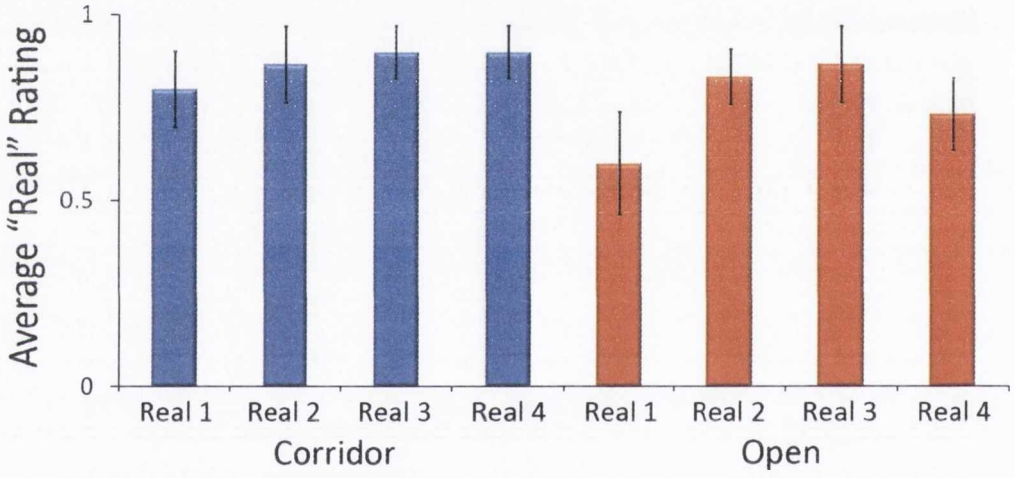



Figure 5.15: Graphs showing results for corridor and open locations for 4 real scenes.

### 5.5.2 Results

The images were presented to participants in two different randomised orders. We found no effects of ordering between the two groups.

We averaged responses over each of the three repetitions for each of the 4 real scenes. There appears to be some slight variation in the ratings for the real scenes as can be seen Figure 5.15). In particular, Real 1 appears to have received a lower real rating than the other real scenes for the open location. Looking at this scene, it was the only real scene that contained no groups of three characters. Each other scene contained a group with three characters. The slight difference in clustering of characters in this scene could be

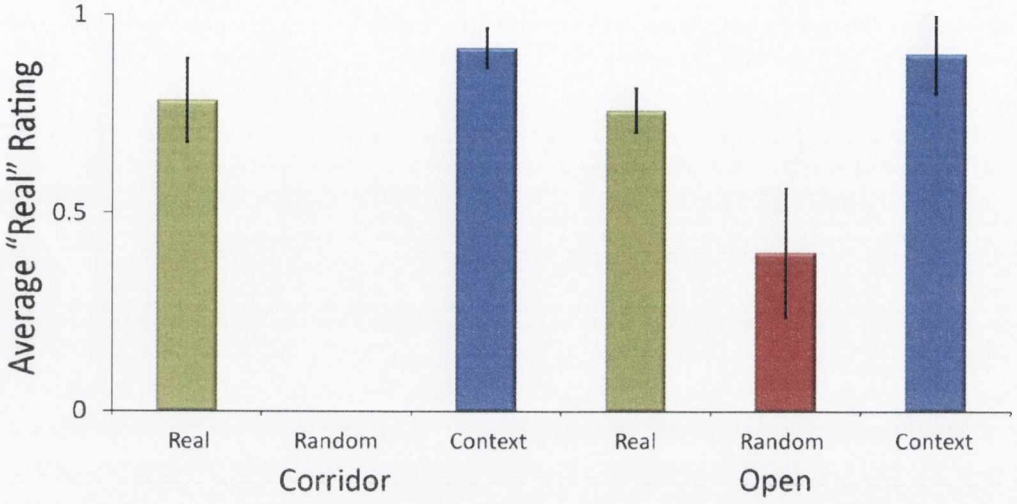



Figure 5.16: Graphs showing all results for corridor and open locations (Note: for the corridor location, random scenes received a zero rating averaging across all participants).

one possible reason for the slightly lower real rating for this image. However, an ANalysis Of VAriance (ANOVA) showed there were no statistically significant differences between the four real scenes for either the corridor or open locations. This implies that the low ratings for some of the real scenes in our previous experiments were not a result of a particular scene appearing any more or less realistic than the others, but rather as a result of the effect of the context and random rules on the participants' perception of the scene. Because of this, we were able to average across all of the real scenes for further analysis.

We averaged responses over each of the three repetitions (where the characters were varied for each repetition, but not the positions and orientations) for each of the 3 formation types (real, random and context). A one factor ANalysis Of VAriance (ANOVA) with repeated measures showed a main effect of formation type ($F_{2,18} = 45.044, p < 0.0001$). Post-hoc analysis was then performed using a standard Newman-Keuls test for pairwise comparisons among means. We found that participants perceived the random formation to be real less often than the context or real formations but that participants perceived the context scenes to be real as often as the real scenes (see Figure 5.16). This result was expected for the open location, based on the results from our previous experiments, where participants found context scenes to be as realistic as real scenes in many situations. There was a slightly different result for the corridor location from our previous experiment however. Before, participants could distinguish between the real scenes and our context scenes, whereas in this experiment they could not.

We did not find an effect of location type, meaning that the participants' perception of the scenes did not vary between the open and corridor locations. This differs from our results found previously, where we found an effect of scene in each block. A possible reason for this result is the alteration of the random placement rules to exclude placement of characters in out-of-bounds regions of the scene. It is likely that the large difference in size of the out-of-bounds areas in the two different locations had an effect on the participants' perception of the scene. However, since they were still able to correctly identify the random placement scenes, it would imply that this was not the only cue used to identify such behaviours. Other contextual information is still important, such as grouping and orientation.

Another factor that could possibly have an effect on a viewer's perception of a scene is the viewpoint through which they are viewing the scene and the distance from the camera to the characters. In these experiments, the distance from the characters varies across the locations and this may have an effect on the perception of the formations within the scenes. Because of the location used to obtain the videos for the real scenes, there is also a difference in the viewpoint between the corridor and open locations. We conducted a further experiment to examine whether the camera viewpoint has an effect on perceived realism in pedestrian crowd scenes.

## 5.6 Camera Viewpoint

A fourth study conducted looked at the effect of camera viewpoint on a viewer's ability to identify random behaviours in a pedestrian crowd scene. The purpose of this experiment was to take a first step into investigating how factors other than the behaviour of the crowd itself can affect the plausibility of crowd scenes and, from there, finding out how generalizable the results already found were.

### 5.6.1 Method

In order to create our images for our experiment stimuli, we used MetroPed to position and orientate the characters in the scene as in Section 5.2. For this experiment, we only included virtual formations, creating scenes with random positioning and according to our full context rules (both position and orientation rules combined). Once the formations had been created in MetroPed, they were then exported and loaded into the Metropolis Fat Client system. This was then used to generate images for the experiment.

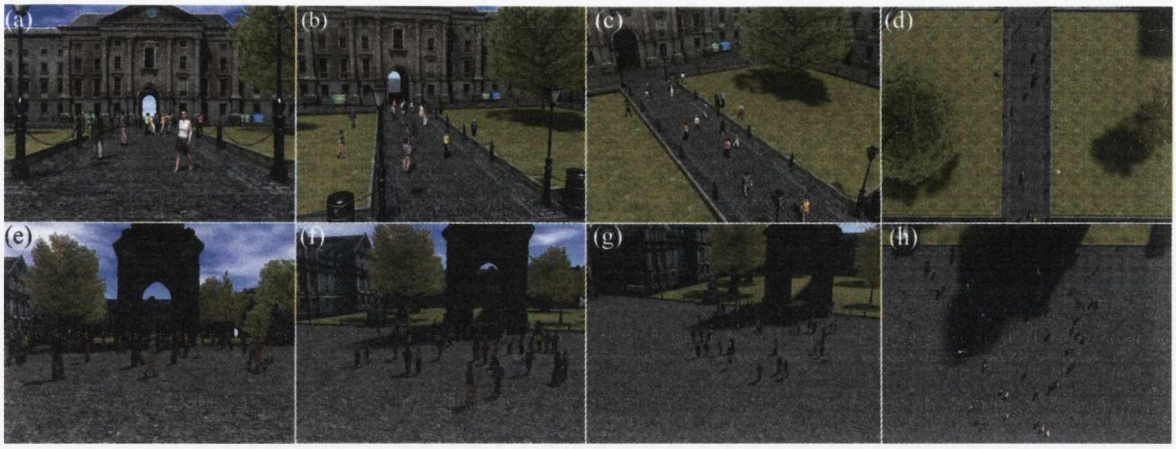



Figure 5.17: Showing each viewpoint (eye-level (1), canonical (2), isometric (3) and top-down(4)) for the corridor location (a) to (d), and open location (e) to (h).

We chose four different camera viewpoints, which can be seen in figure 5.17. The first viewpoint we chose was *eye-level* (viewpoint 1). This was due to the prevalence of first person video games, where the player sees everything through the viewpoint of the main character in the game. The second viewpoint we used was a *canonical* viewpoint (viewpoint 2). The idea of a canonical viewpoint was introduced by Palmer et al. [PRC81] and has been shown to aid the memory of virtual objects [GS08]. A canonical viewpoint of an object can be regarded as the viewpoint first imagined visually, or the viewpoint that is selected as the best angle at which to take a photograph. In free exploration tasks, the canonical viewpoint will often be inspected for the longest period of time [BTBV99]. Object recognition has been shown to be viewpoint dependent and in recognition experiments, canonical viewpoints tend to have the lowest response time and error rate. A canonical viewpoint of an object is often a rotation of 10 degrees about each axis [Tar95]. We approximated this viewpoint in our scene by using a slightly elevated camera angle with a small rotation giving us an offset from the X and Y axes. This was done manually, so as to obtain a camera position that resulted in the most visual information for the scene, maintaining an equal distance from the nearest possible character position for each location. The *isometric* view (viewpoint 3) was chosen as a mid-point between our canonical and top-down viewpoints, as there was a very large difference between these two angles. We positioned the camera higher up and more angled to the scene than our canonical view, but maintained a perspective view of the scene. Finally, we used a *top-down* view (viewpoint 4), which can be an important viewpoint to consider when larger crowds and more areas of a scene need to be within view e.g., in movies with large-scale scenes.

### 5.6.2 Experiment

Twenty three participants (4F, 19M) aged 18 to 30, were seated in front of a computer screen and were given an instruction sheet. For each image displayed, participants were asked if they thought the scenes were realistic or not. A total of 48 images were displayed for 4 seconds each, which consisted of the following: 2 Locations (Open/Corridor) $\times$ 2 Formation Types (Random/Full Context) $\times$ 4 Viewpoints $\times$ 3 Repetitions. Between each trial, a blank screen was displayed for 5 seconds, after which the number of the next trial was displayed to alert participants.

### 5.6.3 Results

We averaged responses over each of the three repetitions for each type of formation. A two factor ANOVA with repeated measures showed a main effect of formation type ($F_{1,23} = 25.887, p < 0.01$), where scenes with context rules were judged to be more *real* than the random scenes. There was also a main effect of viewpoint ($F_{3,69} = 4.2014, p < 0.01$), where participants judged the canonical view of the scene to be less real than any of the other three views. There were also interactions between location and formation ($F_{1,23} = 7.4572, p < 0.05$), where context formations were more real in the corridor scene, as opposed to the open scene, where random formations were found to be more real in the open scene. This was similar to effects found in our previous study, where random formations were more plausible in an open scene than in a corridor scene. The final interaction was between formation and viewpoint ($F_{3,69} = 3.1268, p < 0.05$), where the top down viewpoint was perceived to be most realistic for context scenes, whereas the eye-level view was perceived to be the most plausible viewpoint for random formations.

Breaking the analysis down across formation types, and by analysing the participants responses to context and random scenes separately, we found that for the context scenes, there was no effect of either location or viewpoint. This result indicates that participants judge the context rules applied to a scene to be equally realistic, regardless of whether it is a corridor or open scene and which viewpoint is used. This is an interesting result for us as it implies that our context rules will work effectively using any viewpoint, and therefore transitioning between the viewpoints when traversing the virtual city in our crowd system will not result in loss of perceived realism.

When looking at the responses for the random formations alone, there was a main effect of location ($F_{1,23} = 7.9331, p < 0.01$), where the open scene was seen as more real than

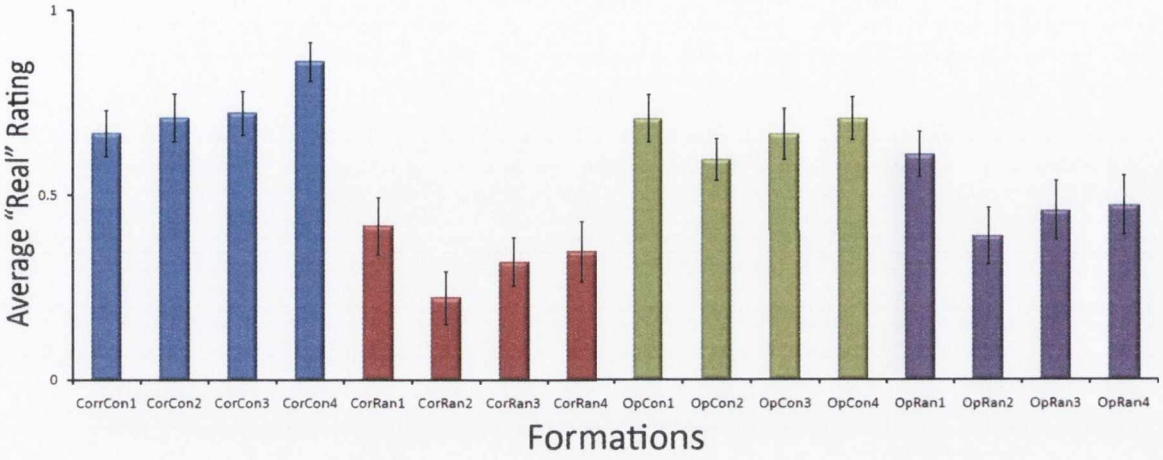



Figure 5.18: Shows responses for both locations (Cor/Op), both formations (Con/Ran) and each viewpoint (1-4).

the corridor scene. There was also a main effect of viewpoint ($F_{3,69} = 4.5257, p < 0.01$), where random formations viewed from the canonical viewpoint was perceived to be the least realistic out of the four viewpoints. Looking at Figure 5.18, it can be seen that when the camera is at eye-level, random formations are more plausible than when viewed from other viewpoints. This result is important for us in terms of using computational savings when our system is using this camera viewpoint only, since viewers are less inclined to notice randomness of the characters' positions and orientations when viewing a scene from this level. Post hoc analysis using Newman-Keuls comparison between means showed that viewpoint 1 (eye-level) was rated as real significantly more often than viewpoint 2 (canonical) but there were no differences between either viewpoint 3 (isometric) or viewpoint 4 (top-down). This suggests that, in general, using an eye-level view will result in a more realistic representation of pedestrian crowd formations to a viewer. Figure 5.19 shows participants' responses for each viewpoint across both open and corridor locations and random and context formations.

## 5.7 Dynamic Groups

The idea of investigating the context of a particular scene and the role it plays in the plausibility of the positions and orientations of a character in a scene was quite an important factor for us to consider. Through the development and perceptual evaluation of context rules for positions and orientations, we had developed a sensible way to populate 2 differ-

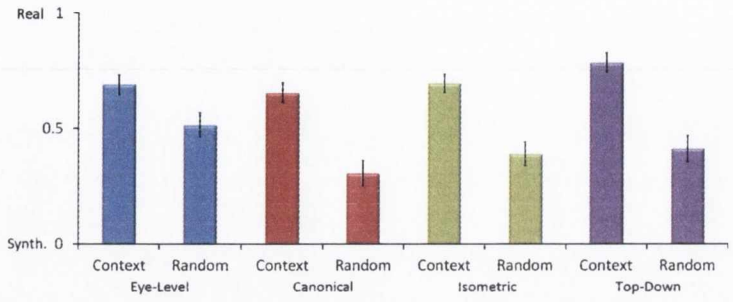



Figure 5.19: Shows participant responses for each viewpoint for context and random formations, across each location.

ent types of scenes, of which there are many throughout the virtual model of Dublin that Metropolis is based on. This was an important first step towards creating an initialization procedure to eventually be part of a realistic behaviour system.

The next topic we wanted to investigate was the idea of groups in a virtual environment. As mentioned before, a large percentage of pedestrians in an urban setting travel in groups of two or more. This implied that in order for our system to appear realistic to the user, there was a need for group behaviour in our system. However, very little was known about the level of importance the presence of groups has on the plausibility of a scene; will a user automatically perceive a scene to be more realistic if there are groups present, or can certain group combinations actually detract from the overall realism of a scene? In order to investigate this, we conducted a study on the effects of dynamic groups on the realism of a pedestrian crowd. We presented participants with a variety of short animations, each containing pedestrians organised into different ratios of individuals, pairs and groups of 3, and investigated how realistic they were perceived to be.

### 5.7.1 Corpus Analysis

In order to obtain some data about the presence of groups in pedestrian scenarios, we selected a random selection of 20 frames of five and a half minutes of video footage of an open scene. We obtained the footage at a time of day when pedestrian traffic density tended to be low or medium, with 25 to 42 pedestrians in each scene. Higher density traffic is possible in this area at other times (for example, lunchtime). However, we did not include recordings for those periods because a large number of occlusions would make annotation significantly more complicated and error prone.

From what we observed, approximately 37% of pedestrians travelled singly, 46% in pairs,

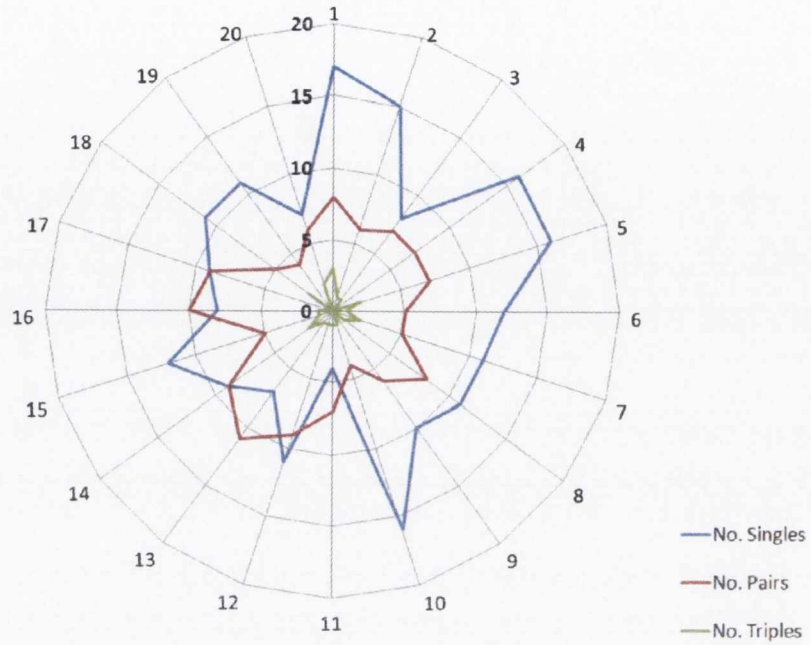



Figure 5.20: Showing frequencies of numbers of single pedestrians in a scene for 20 random frames of video, numbers of pairs and numbers of triples.

11% in a group of three, and just under 6% in larger groups. Figure 5.20 shows that the number of single pedestrians was the highest in most of the frames, with an average of 11.05 per scene. Pairs were the next most frequent occurrence, with the average being 6.9. Finally, trios were the least frequent, with the average being 1.15. This result concurs with empirical observations that singles or pairs are most frequent in pedestrian simulations, whereas groups of three or more are less frequent. This result informed our hypotheses for conducting our perceptual study on dynamic groups.

### 5.7.2 Method

We generated a variety of short, three second animations in the Metropolis crowd system, MetroPed (see Figure 5.21 for screenshots from the animations shown to participants). In each animation, the virtual pedestrians simply walked forward at constant speed ($0.95m/s$): their starting positions were set to ensure that no collision avoidance or turning maneuvers were necessary. This ensured that local factors, such as turn velocity, or avoidance distance, could not affect viewer ratings. Scenes were created according to two general categories: *groups* and *no groups*. For the *groups* category, each animation

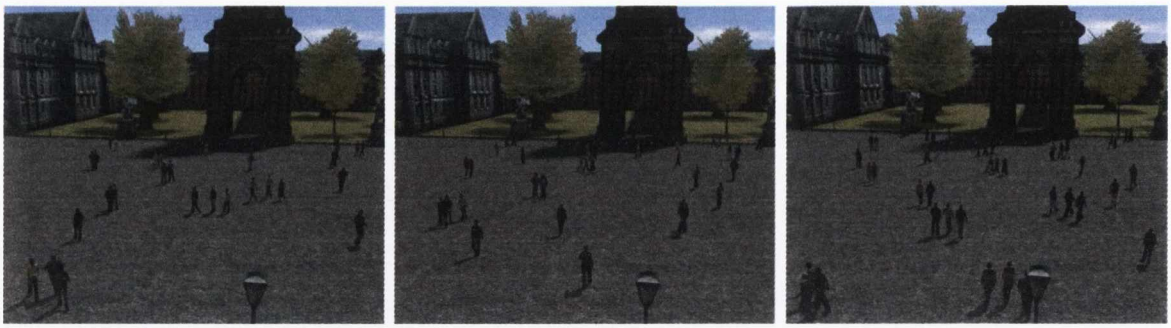


Figure 5.21: Example scenes from a sample of animations that were rated by participants in the perceptual study. These images corresponding to group ratios 1:1:1 (left), 2:1:1 (center) and 1:1:2 (right).

| Case | Ratio | Groupings | Total Characters | Hypothesis |
|---|---|---|---|---|
| 1. | 1:1:1 | 6S, 6P, 6T | 6+12+18= 36 | Plausible |
| 2. | 1:1:2 | 6S, 6P, 12T | 6+12+36= 54 | Implausible |
| 3. | 1:2:1 | 6S, 12P, 6T | 6+24+18= 48 | Plausible |
| 4. | 1:2:2 | 6S, 12P, 12T | 6+24+36= 66 | Implausible |
| 5. | 2:1:1 | 12S, 6P, 6T | 12+12+18= 42 | Plausible |
| 6. | 2:1:2 | 12S, 6P, 12T | 12+12+36= 60 | Implausible |
| 7. | 2:2:1 | 12S, 12P, 6T | 12+24+18= 54 | Plausible |

Table 5.1: Group ratios used for the *groups* category of animations. The third column shows the numbers of singles (S), pairs (P) and triples (T) when 6 is chosen as the base value, giving the resulting total number of characters column 4. In each case, we hypothesised that the ratios would be Plausible or Implausible

consisted of a varying number of groups and individuals, specified by a group ratio $S{:}P{:}T$, with $S$ as the proportion of single individuals in the scene, $P$ the proportion of pairs, i.e., groups of size 2, and $T$ the proportion of triples, i.e., groups of size 3. We enumerated seven group ratio combinations (see Table 5.1). We found 6 pedestrians to be an appropriate choice as a base value for generating pedestrian numbers from the ratios. Thus, for a ratio of 1:1:1 (see Case 1. in Table 5.1), we obtain 6 singles, 6 pairs and 6 triples (6S, 6P, 6T), giving a total of 36 pedestrians.

The *no groups* category of animations contained solely individuals. The total number of individuals in these animations were matched with the maximum, average and minimum number of pedestrians over all the scenes in the *groups* categories and an equal number of *groups* and *no groups* scenes were created.

### 5.7.3 Experiment

Thirty three participants (7F, 26M) age 18 to 30, were shown the short animations in a randomised order and were asked to judge whether each scene was plausible or not. Participants were shown a total of 42 animations. Of these, 21 contained animations of three different densities (36, 51, and 66 characters), but no groups. The other 21 animations were divided among the 7 different group combination ratios (Table 5.1) with 3 repetitions of each. Between each trial, a blank-screen was displayed for 5 seconds, after which the number of the next trial was displayed alerting participants.

We hypothesised that some of the group ratios in Table 5.1 would be more plausible for viewers than others. Specifically, we hypothesised that cases biased towards higher ratios of individuals and pairs would be more plausible than those biased towards more pairs and triples. Therefore, in the final analysis, we conducted comparisons between three conditions: *plausible groups*, *implausible groups* and *no groups*.

### 5.7.4 Results

A single factor ANalysis Of VAriance *ANOVA* with repeated measures for each of the three pedestrian population levels was conducted for the *no groups* condition. There was no significant difference in the responses of the participants to changes in pedestrian population alone, suggesting that any variance in participant responses to the group stimuli would likely not be influenced by the change in number of pedestrians in each scene (see column 4 of Table 5.1). This result allowed us to consider the three *no groups* cases as a single condition.

There was also no significant difference when performing a two factor analysis across the *groups* conditions and *no groups* conditions. This was expected, since the *groups* condition was composed of both hypothesised plausible and hypothesised implausible group ratios.

We also conducted a single factor ANOVA analysis, with repeated measures design, for all of the cases (i.e., all seven group ratios and the additional no groups condition). This highlighted a main effect of the grouping ratio adopted i.e., the 8 cases of no groups, 1:1:1, 1:2:2, 2:2:1, and so on ($F_{7,224} = 3.6349, p < 0.01$). From Figure 5.22, it can be seen that the ratios 1:2:1 and 2:1:1 were rated as real more often than any other group ratio, followed by 1:1:1 and 2:2:1. This was followed by the no groups condition, and the implausible groups ratios, 1:1:2, 2:1:2 and 1:2:2. Post-hoc analysis was then performed using a standard Duncans test for pairwise comparisons among means. This test concluded

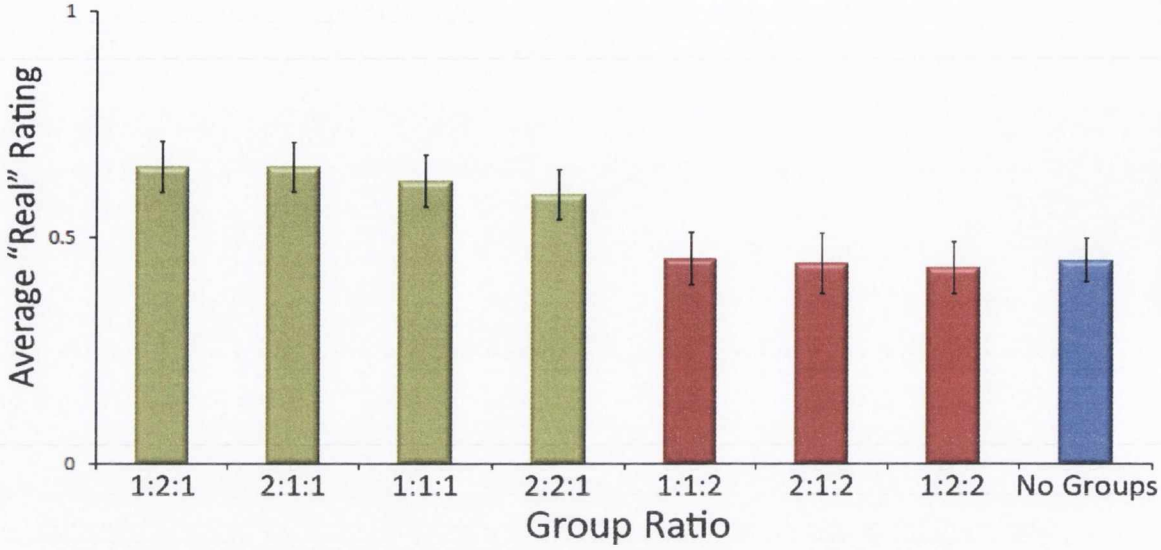



Figure 5.22: Graph showing 'real' ratings for each group ratio (singles:pairs:triples), with 1 representing a rating of 'Real' and 0 representing a rating of 'Synthetic'.

that participants perceived the group ratios 1:2:1, 2:1:1 and 1:1:1 as real more often than group ratios 1:1:2, 2:1:2, 1:2:2 and the No Group condition ($p < 0.05$ in all cases).

We then averaged across our plausible, implausible and No Groups conditions and performed an ANOVA with repeated measures design. The overall results of our comparison between the three primary conditions, plausible groups, implausible groups and no groups, can be seen in Figure 5.23. The results indicate that those animations in the plausible groups category were rated as real more often than those in the implausible groups and no groups categories ($F_{2,64} = 8.6173, p < 0.05$), suggesting that the addition of groups to crowds can increase the plausibility of the scene, provided that care is taken in terms of the ratios of individuals and groups added.

## 5.8 Discussion

The studies carried out in this chapter look at the importance of evaluating the perceptual effects of virtual scenes and how results from these evaluations can guide the population of such scenes. From these studies, we now know that in order to create perceptually adequate crowd formations, there are a simple set of rules that can be followed, relating the characters to the environment around them, that can result in realistic formations

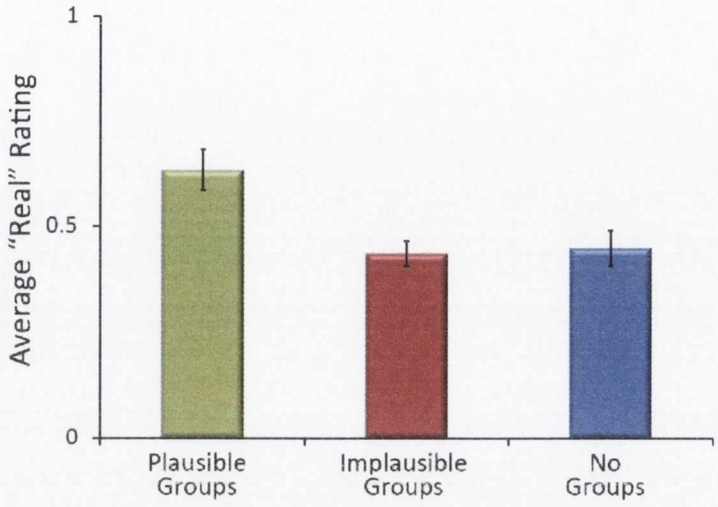



Figure 5.23: Graph showing 'real' ratings for the three main conditions, with 1 representing a rating of 'Real' and 0 representing a rating of 'Synthetic'.

suitable for many different purposes.

We also now know that different camera viewpoints can be used to mask random behaviour in a crowd. If a low level of detail is being used to simulate the behaviour of a crowd, then it is best to keep the camera at an eye-level, rather than at angles that provide the user with more information, thereby permitting them to notice anomalies in the behaviour of the crowd. However, if the crowd formation is perceptually accurate to the user, then it does not matter what camera angle is used, as it will seem equally plausible in all.

Finally, when it comes to the idea of an urban pedestrian crowd, we now know that the presence of realistic groups is an important factor for added realism for the viewer. However, these groups must be added in a plausible way. Our work only looked at dynamic groups, who were navigating through the environment. We follow on this work looking at methods to populate our crowd scenes with other types of groups consisting of conversational agents.

# Chapter 6

# Conversational Groups

## 6.1 Introduction

The results from our previous study, described in Section 5.7, demonstrated that the addition of groups to a system such as Metropolis could improve how realistic the behaviour of the characters appear to the viewer. However, the addition of groups alone did not improve the plausibility of the scenes tested, as the group combinations needed to be plausible. In a crowd system such as the one used for the Metropolis system, i.e., a multisensory, urban and highly pedestrianised environment, groups in transit will not be the only groups needed to help create a sense of presence. There will be a high percentage of the crowd standing in groups, interacting and conversing, in particular in the open areas. In order to further investigate how to populate our virtual environment with such conversational characters, and to determine how they should interact with each other in a realistic way, we conducted a number of experiments investigating viewers' perception of virtual conversing groups.

There are limitations on the creation of realistic conversing groups in a real-time crowd system; problems can arise with the volume of data exhausting available memory resources, so there is a need to ensure that motion captured data is as reusable as possible. There are also issues with replaying real conversations. In a dynamic scene, characters will be leaving and joining groups at different times, and it will not always be possible to synchronise character animations. Also, when the camera is moving through the crowd and a conversing group becomes salient, the characters will not necessarily be synchronised with each other due to the complexity of finding suitable blending animations for looping

or transitioning animations. So, in order to account for these issues, it is unlikely that it will always be possible to replay our conversations as they were captured. For this reason, we wanted to investigate methods to mix and match unsynchronised, body motions for characters from different conversations to create new plausible conversations. We also need to find out how best to match these body motions to the audio of the conversations, so we conducted a series of experiments to investigate these challenges. These new conversations will help create a sense of variety in our crowd, which is an important factor to make the scene appear real.

Results from these experiments could also be relevant to video games, where cut-scenes have been around for at least two decades [Rou98], and more recently the use of performance capture (where both motion capture and audio are recorded together) has been growing in popularity as it produces the most alive and realistic characters. However, it is not always possible due to the cost of hiring A-list actors or stars (such as Roger Federer or Rafael Nadal in Topspin 3$^{TM}$) to perform motion and voice, or because of the location constraints necessary for high quality audio [Edg10]. In these cases, audio and motion are often captured separately. However, little is known about what effects, if any, possible desynchronisation has on the perception of the final sequence.

We first investigated participants' sensitivity to temporal desynchronisation (where each of the characters' body motions came from different conversation clips) for body motion alone. We then carried on this work with a second set of experiments, where we looked at multisensory conversing characters, with different types of audio signals. For these experiments, body motion desynchronisation was as before, with the added complexity of whether or not the characters' body motions matched the voices heard. For each experiment discussed in this chapter, participants were over 18 years of age, naïve to the purpose of the experiment and from a range of disciplines. Ethical approval was granted for all experiments, and participants were recruited via email. They were given a book voucher to compensate them for their time.

This work was carried out in collaboration with Rachel McDonnell, who is the research fellow working on the animation system for Metropolis. We designed the perceptual experiments and created the stimuli together, and jointly performed the statistical analysis and reported the results for inclusion in the papers that we wrote together on the topic. In the experiments described in Section 6.3 Rachel McDonnell took the lead. For the experiments described in Sections 6.4, 6.6 and 6.7.2 I did the majority of the analysis and writing.

## 6.2 Body Motion Desynchronisation Experiments

The first set of experiments was designed to investigate participants' sensitivity to body motion desynchronisation for conversations with visual information only, i.e., in the absence of audio information. We conducted two experiments, with different stimuli and a slightly altered task, to indicate the first steps in how to go about creating new conversations from a limited data set. In order to do this we combined unsynchronised body motions from each character, picking them from different conversations.

### 6.2.1 Experiment Design

Three male actors participated in a motion capture session which was conducted on a 13 camera Vicon optical system, with 53 markers per actor. We chose groups of three, for each of our experiments containing conversational characters, as a representative number for small groups. Complications with motion and audio capture for multiple individuals simultaneously meant that it was desirable to keep the group size as small as possible in order to ensure high quality data capture. Larger groups would have induced more occlusion in the motion capture data, and overlap in the capture of each individual's audio. However, while keeping the group size small was important, we also wanted to ensure that our conversations would contain certain group dynamics, such as turn taking, interruptions and gaze shifting. In order to have these more complex properties, it was necessary to capture these conversations with more than two individuals. The actors were instructed to stand at the corners of a triangle in the capture space of the motion capture system for the start of each recording. During the recording, they were free to move around close to their starting point. Since we were not recording finger motion, we placed fabric around their fingers to restrict movement in this area. Before the session, a list of topics was given to the actors and they indicated which topics they were most comfortable discussing. All actors were non-professionals, familiar with each other, the environment, and the motion capture setup. Conversation was free-flowing and unscripted, which resulted in natural dialog. Recordings ranged in length from one to three minutes. (Note: please see section 6.5.1 for the stimuli creation process for our single stimulus body desynchronisation experiment.)

Two different styles of conversation were recorded. The first type had a dominant speaker, where one person from the group spoke while the others politely listened. Each actor took a turn at being the dominant speaker, resulting in three recordings of this type of

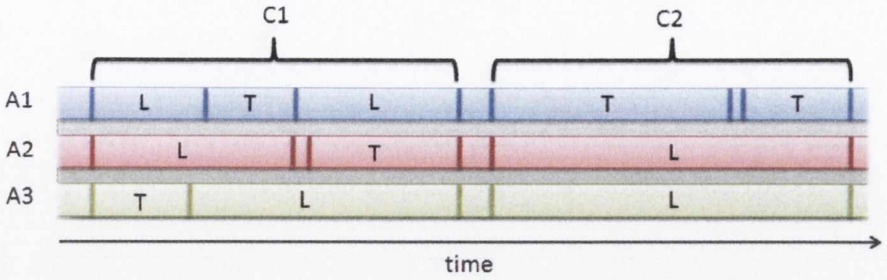



Figure 6.1: Example of talker/listener (T/L) annotation of two conversation recordings (C1 and C2) for three actors (A1, A2, A3). Each T/L clip is greater than 10 seconds in length.

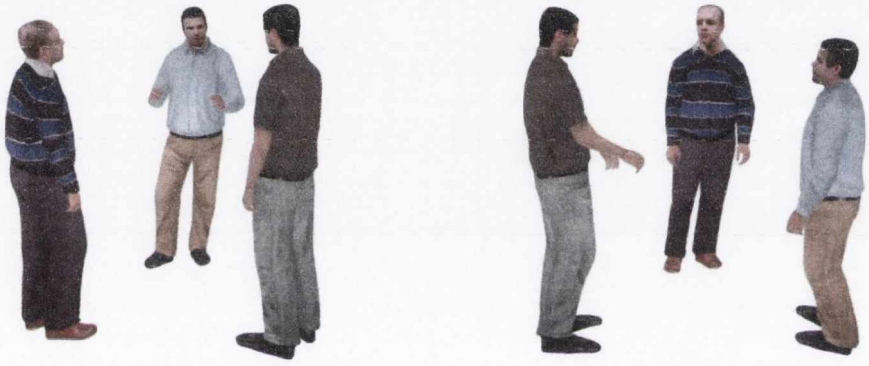


Figure 6.2: Examples of synchronised dominant-speaker conversations used in the experiment.

conversation. The second style was a debate, where each actor had an affirmative or negative position on the topic being discussed. This style of conversation allowed for interruptions while discussing different opinions. We recorded four debates in total.

Using video footage as reference, each recording of the dominant speaker conversations was annotated for each actor to indicate at which frames of animation they were talking or listening for longer than 10 seconds (Figure 6.1).

For the virtual scene, we chose three male virtual characters to represent our three actors. We ensured that the characters were the same relative height as the actors, in order to minimise retargeting errors. A white background was used in order to provide good contrast. The camera was placed so that one character was centrally focused and the gestures of the other two were clearly visible (Figures 6.2 and 6.3). For each trial, the camera was randomly focused on one of the three characters, in order to avoid any effect of character preference.

Figure 6.3: Examples of synchronised debates used in the experiment.

## 6.3 Exp 1: 2 Alternative Forced Choice Experiment

In our first experiment looking at body motion desynchronisation, twenty-eight volunteers (17M, 11F) took part. A 2 Alternative Forced Choice (2AFC) paradigm was used, where the following were presented in random order: a 10-second clip of a real (synchronised) conversation (Figures 6.2 and 6.3), and a 10-second clip of an altered (desynchronised) conversation. Participants were asked to indicate which conversation was the real one, using a left mouse click to indicate the first conversation, or a right click to indicate the second conversation. We chose to limit the trial durations to 10 seconds, rather than ending the trial upon participant response. This was to ensure that participants neither rushed their responses or became fatigued throughout the experiments. This 10 second trial duration also reduced the variability in the content participants viewed, as we could ensure to the best of our ability that there were no animation artifacts in the clips presented to them. Participants were questioned after the experiment and reported that the 10 second trials were of an adequate length to allow them to make their decisions.

The experimental system was developed using an open-source renderer that used DirectX 9.0 and a commercially available animation engine. This allowed us to randomly seed the experiment for each participant to allow for much variation in the stimuli (i.e., repetitions were randomly selected from the recordings). This avoided any effects that may have occurred due to the repetition of particular animation sequences. The experiment was run on a workstation with 2GB of RAM, an 8-series GeForce graphics card on a wide-screen 24-inch LCD monitor.

| Block | Term | Bodies (B) Meaning |
|---|---|---|
| Both | Real $B_S A_0$ | sync (S) |
| Debates | $B_D A_0$ | desync (D) |
| Dom. Sp. | $B_{0T} A_0$ | 0 talkers (0T) |
| | $B_{1T} A_0$ | 1 talker (1T) |
| | $B_{2T} A_0$ | 2 talkers (2T) |
| | $B_{3T} A_0$ | 3 talkers (3T) |

Table 6.1: Notation used for describing experimental conditions. $B_x$ denotes body motion condition and $A_0$ denotes that no audio was used for these experiments.

Our notations for synthetic conversations are in "$B_x A_y$" format, where $B$ refers to Body motions and $A$ refers to Audio (see Table 6.1 for an overview of notational conventions). In our body motion experiments we used visual only stimuli, so the body motion conditions are suffixed by "$A_0$".

Participants viewed two counterbalanced blocks, where block 1 showed synchronised and desynchronised dominant speaker conversations and block 2 showed synchronised and desynchronised debates. Throughout this study, we will use the terms *synchronised / desynchronised* to refer to the body motion of our characters. In the dominant speaker block, we tested the number of talkers condition, in order to determine if this had an effect on the realism of the desynchronised conversations. Against the synchronised condition (real) we tested: 3 talkers ($B_{3T} A_0$), 2 talkers ($B_{2T} A_0$), and 1 talker ($B_{1T} A_0$), where the others in the group were listeners. We also tested a 0 talker condition ($B_{0T} A_0$) where no character was talking, to represent the situation of a break in conversation. Ten-second motion clips were chosen at random from the three dominant speaker recordings to match these conditions (Figure 6.4). We ensured that the virtual characters were consistently matched to the same actors (e.g., the 3 talker case had a clip from each of the actors rather than 3 from one actor). Therefore, in both the real and altered conversations, the virtual characters had the same personality traits. For the real conversations in this block, 10-second synchronised motion clips were randomly chosen for each character from the 3 dominant speaker recordings. Twelve pairs were shown in total (3 repetitions of each number of talkers).

In the debates block, we tested the synchronised (real) against desynchronised conversations ($B_D A_0$). We also tested the topic context condition, in order to determine if desynchronised clips from the same recording would be more acceptable than those from different recordings. As before, we ensured that each virtual character was consistently

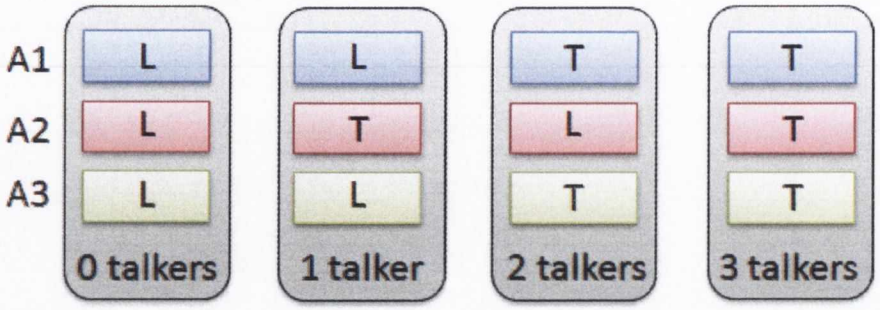



Figure 6.4: Examples of desynchronised dominant speaker conditions, where 10-second talker or listener (T/L) clips are chosen at random from the annotated recordings.

matched to an actor. For the real debate conversations, 10-second synchronised motion clips were randomly chosen for each character from the 4 debate recordings. Twelve pairs were shown in total (6 repetitions of each topic context).

We hypothesised that desynchronised debates would be less noticeable than desynchronised dominant speaker conversations, due to the fact that breaking turn-taking etiquette may be more noticeable for polite conversations since there should only be one talker at a time. Conversely, debates by nature are more chaotic and tend not to follow turn-taking rules, which may result in them appearing more realistic when desynchronised. We also hypothesised that participants would find 0 or 3 talkers particularly unrealistic, and that desynchronised conversations from out-of-context debates may appear less realistic, due to the level of argumentation not matching.

### 6.3.1 Results

For each participant, we recorded their ability to identify synchronised and desynchronised animation sequences for both debates and dominant speaker conversations. We averaged the results over all repetitions for each participant. Since corresponding conditions were not present for dominant speaker and debates, we averaged over the number of talkers for the dominant speaker and the topic context for debates. A repeated measures ANalysis Of VAriance (ANOVA) was then conducted to determine if there was an effect of conversation type. A main effect was found ($F_{1,26} = 5.2, p < 0.05$), which was due to the desynchronised dominant speaker conversations being easier to detect (70% detection) than the desynchronised debates (61% detection) (Figure 6.5).

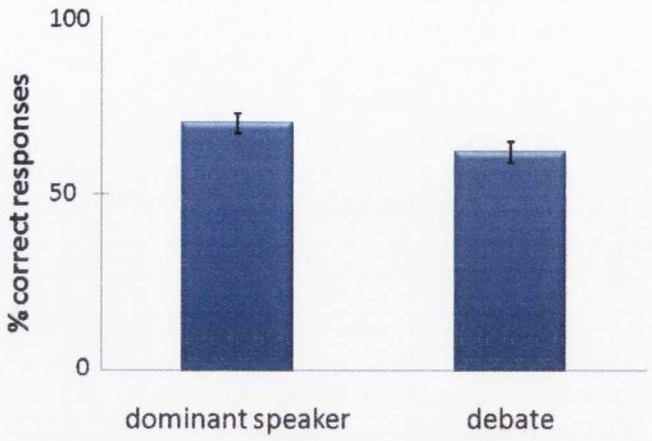



Figure 6.5: Main effect of conversation type. A correct response is when a user correctly identifies the real conversation.

In order to determine if there was an effect of the number of speakers in the desynchronised dominant speaker conversations, a repeated measures ANOVA was conducted. Surprisingly, we found no effect of speaker number, which implies that desynchronised conversations using 0 or 3 talkers did not stand out as any more unrealistic than the others (Figure 6.6). We feel that this could be due to the fact that humans put their own interpretations on the conversations, making them more tolerant of different heated debates or breaks in conversation.

We also found no effect of topic context for the debates, where participants were able to detect anomalies equally for both the same and different debate topics. Again, this could be due to people associating their own story or scenario with the conversations that they viewed, thus making desynchronised conversations appear plausible at times.

Fifty percent of our participants (9M, 5F) were informed less about the task than the others (8M, 6F). Half were told only that the desynchronised conversations were "altered in some way," while the other half were specifically told that they were desynchronised. A between-groups ANOVA on the results showed that there was no difference between the detection abilities of the participants in the first group and those in the second. This implies that even when participants knew what they were looking for in terms of anomalies, this did not change their detection performance.

Differences have been reported between male and female interpretations of conversations (e.g., [HM99]). Therefore, we used a between-groups ANOVA to determine if there were differences between the detection rates for our male and female participants. We found

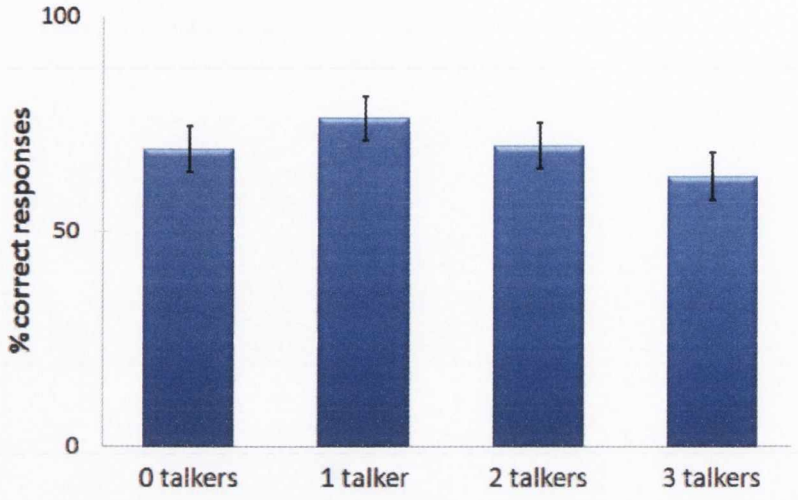



Figure 6.6: No main effect of number of speakers in desynchronised dominant speaker conversations was found. A correct response is when a user correctly identifies the real conversation.

no effect, which implies that conversation interpretations were consistent regardless of the sex of the participant.

## 6.4 Exp 2: Single Stimulus Experiment

In our second body motion desynchronisation experiment, we also explored participants' sensitivity to body motion desynchronisation in the absence of audio. We conducted a similar experiment as before. However, in this instance, we used a richer motion capture data set, with both male and female actor groups (stimuli creation discussed in Section 6.5.1), and participants were given a more intuitive task. Instead of a 2AFC paradigm, we asked them to make their judgement while viewing one stimulus at a time, in order to determine how realistic they found both the real and synthetic conditions to be in isolation, rather than making a comparison between the two.

Twelve volunteers (10M, 2F) took part in this experiment. As in our previous experiment, it was conducted in two blocks (*debates* and *dominant speaker* conversations) shown in random order. There were four synthetic conditions for the dominant speaker block and one for the debates block as outlined below. As in all succeeding experiments, participants were first shown an example of a real and synthetic conversation (from a conversation clip not used in the experiments). They were told that the real conversations depicted the body

motions played back exactly as captured, while the synthetic ones were altered in some way. We did not influence their decisions by explicitly informing them that conversations were desynchronised.

As before, it was conducted in two blocks (*debates* and *dominant speaker* conversations) shown in random order. In the dominant speaker block, we again tested the number of talkers condition: 3 talkers ($B_{3T}A_0$), 2 talkers ($B_{2T}A_0$), 1 talker ($B_{1T}A_0$), and 0 talkers ($B_{0T}A_0$), where the others in the group were listeners (Table 6.1). We had three repetitions for each actor group for each condition, and counterbalanced the synthetic conditions with real conversations. So in total we had (2 $\times$ real/synthetic $\times$ 4 conditions $\times$ 2 actor groups $\times$ 3 repetitions) 48 trials.

In the debates block of this experiment, we did not test the topic context condition as before. We simply tested the condition of whether the body motions were synchronised (real) or randomly selected from different conversations ($B_DA_0$). As for the dominant speaker conversations, we had three repetitions for each actor group, counterbalancing for real and desynchronised conditions, giving us a total of (2 $\times$ real/synthetic $\times$ 2 actor groups $\times$ 3 repetitions) 12 trials.

Also as before, we matched each actor to a virtual character consistently throughout the experiment. This experiment was run on a workstation with 4GB of RAM, a Creative SB Audigy 2ZS soundcard and an 8-series G-Force graphics card. The stimuli were displayed on a 24-inch widescreen monitor. Participants viewed each trial for 10 seconds and were asked to indicate using a mouse click whether they thought that the conversation they viewed was *real* or *synthetic*. We found through interviewing participants that 10 seconds was an adequate time for them to make their decisions. We randomly associated the right or left mouse button with the *real* response so as to avoid any bias towards a particular button-press. After a participant gave his/her response, a cross was displayed to focus attention on the centre of the screen.

We hypothesised that debate style conversations would be more plausible than dominant speaker conversations. We also hypothesised from our previous experiment that there would be no effect regarding number of talkers in dominant speaker conversations. Based on previous research [BH95], we hypothesised that there might be a difference in participant responses for male and female actor groups.

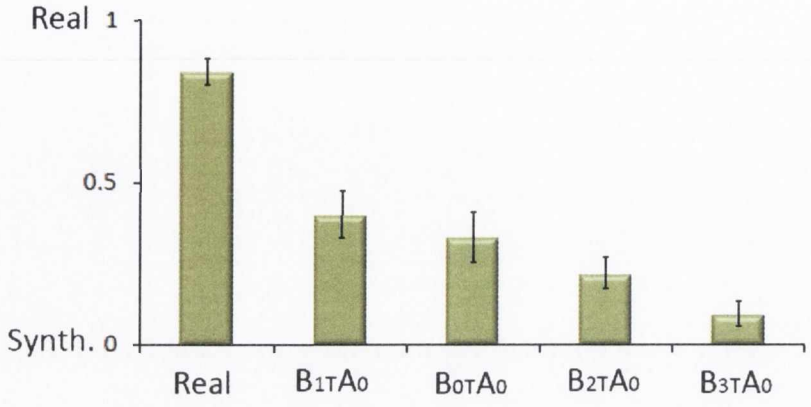



Figure 6.7: Results for our single stimulus body motion experiment for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition.

### 6.4.1 Results

Participant responses were averaged over repetitions for each condition. For our analysis of this experiments, we conducted a 3-way repeated measures ANalysis Of VAriance (ANOVA) with within-subjects factors of *AV condition*, *actor group* and *conversation type*. *AV condition* refers to the various audio/visual desynchronisation combinations, for both real and synthetic conversations (which were counterbalanced). *Actor group* refers to the male and female motion captured groups, while *Conversation type* refers to dominant speaker conversations and debates (which were separated into two blocks). Post-hoc analysis was conducted using Newman-Keuls tests for comparison of means and only significant results at the 95% level are reported.

The results for this experiment can be seen in Table 6.2, which contain the 'real' rating means and standard deviations, and Figures 6.7 and 6.8, which graphs the mean 'real' ratings and their standard error.

For the dominant speaker block, we found a main effect of AV condition ($F_{4,44} = 27.193, p < 0.00005$), where the real conversations were found to be more plausible than any other condition. For the most part, our results replicated the effects of our previous experiment. However, post-hoc analysis showed that participants found our $B_{3T}A_0$ to be less plausible than our $B_{1T}A_0$ and $B_{0T}A_0$ conditions. This contradicts our results from our previous experiment, where we did not find that conversations where three characters were animated with talker body motions were particularly unrealistic for the dominant speaker conversations. These different results could be due to the fact that we conducted this

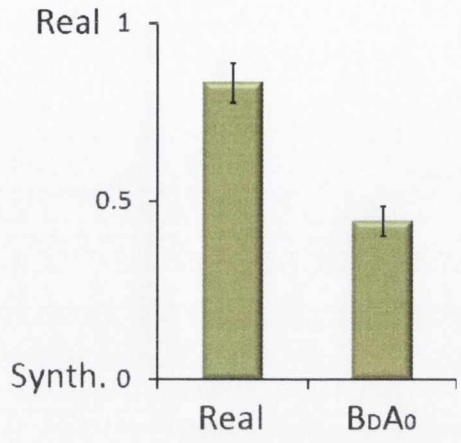



Figure 6.8: Results for our single stimulus body motion experiment for debates, showing mean 'real' ratings and standard error bars for each condition.

experiment with a more intuitive task that yielded more sensitivity to desynchronisation.

For the debates block, we also found a main effect of AV condition ($F_{1,11} = 20.731, p < 0.0005$). Again, participants found our real condition more plausible than our desynchronised condition. We found an overall main effect of conversation type ($F_{1,11} = 10.066, p < 0.05$), where participants found the debates more plausible than the dominant speaker conversations. There was no effect of actor group, implying that participants found it equally difficult to determine real conversations for both male and female motions. We also found a main effect of conversation type ($F_{1,11} = 10.066, p < 0.05$). As before, the debates appeared more plausible overall than the dominant speaker conversations.

| **Block** | **Condition** | **Mean** | **St. Dev.** |
|---|---|---|---|
| Debates | Real | 0.83 | 0.14 |
| | $B_D A_0$ | 0.44 | 0.19 |
| Dom. Sp. | Real | 0.84 | 0.14 |
| | $B_{0T} A_0$ | 0.33 | 0.27 |
| | $B_{1T} A_0$ | 0.40 | 0.25 |
| | $B_{2T} A_0$ | 0.22 | 0.16 |
| | $B_{3T} A_0$ | 0.10 | 0.13 |

Table 6.2: Mean 'real' ratings and standard deviations for No Audio experiment.

Our results show that people can detect real synchronised conversations above chance using body motion alone. We also showed that desynchronisation of debate or argumentative-style conversations is less noticeable than for dominant speaker (polite) conversations when

looking at body motion alone. Therefore, using debate-style conversations played back in different orders could be a good solution to create variety for idle groups in virtual crowd scenes for scenes where no audio is present. Our next step following these experiments was to investigate the addition of audio to these virtual conversers and to identify the implications for viewers' perception of body motion desynchronisation.

## 6.5 Multisensory Desynchronisation Experiments

Following on from our body motion desynchronisation experiments, we then wanted to investigate the importance of matching audio to body motion for virtual conversing characters. We wished to find out if conversations appeared plausible if only the talker in the conversation was matched to the audio. Of interest to us also was whether it would be possible to use unmatched audio for the motions being displayed in a conversation, once the motions are synchronised as they were captured. If people are sensitive to desynchronisation of audio and visual information, we wanted to further investigate which modality people would rely on when judging realism, and how implausible a conversation would appear if both audio is unmatched and motions are desynchronised. We conducted a set of experiments in order to answer these questions.

### 6.5.1 Experiment Design

For these experiments, we created a number of real and synthetic conversational scenarios depicting a group of three individuals. The real conversations involved replaying the body motions and audio as captured. For this study, we will use the terms *synchronised / desynchronised* to refer to the body motion of our characters, as before, and *matched / unmatched* when referring to audio (as we do not desynchronise the voices from each other, we simply mis-match the entire audio clip from the body motions) for the desynchronised conditions. Please see Table 6.3 for the notation used in our multisensory experiments.

The first synthetic condition with audio we investigated was **synchronised bodies not matched to audio ($B_S A_{NM}$).** All three body animations were chosen from the same point in time of a conversation, but a different audio clip was chosen at random. For this and all following random conditions, we ensured that the audio never matched the body motions by chance. In the case of body motion desynchronisation, the motions were never from the same clip. The rationale for this condition was that, if a high sensitivity

| | **Bodies (B)** | **Audio (A)** |
|---|---|---|
| *Real* | sync (S) | matching (M) |
| $B_S A_{NM}$ | sync (S) | not matching (NM) |
| $B_D A_{NM}$ | desync (D) | not matching (NM) |
| $B_D A_{M1}$ | desync (D) | matching 1 talker (M1) |
| $B_{0T} A_{NM}$ | 0 talkers (0T), 3 list. | not matching (NM) |
| $B_{1T} A_{M1}$ | 1 talker (1T), 2 list. | matching 1 talker (M1) |
| $B_{2T} A_{M1}$ | 2 talkers (2T), 1 list. | matching 1 talker (M1) |
| $B_{3T} A_{M1}$ | 3 talkers (3T), 0 list. | matching 1 talker (M1) |

Table 6.3: Notation used for describing experimental conditions.

threshold was found, this would allow us to add variety to a scene by playing a particular conversation animation with random audio.

Our second synthetic condition was **desynchronised bodies not matched to audio** ($B_D A_{NM}$), where each of the three body animations were chosen at random with a randomly chosen audio clip. If this condition was found by participants to be plausible, it would mean that random body motions and audio clips could be played together, giving a large range of variety. This would allow us to resolve blending problems and would easily facilitate dynamic group formation.

Our next three conditions were **1, 2 and 3 talkers, with audio matched to 1 talker** ($B_{1T} A_{M1}$, $B_{2T} A_{M1}$, $B_{3T} A_{M1}$). Here, either 1, 2 or 3 talker body animations were chosen from different conversations at random (the other characters used random listener body animations) with audio randomly matched to one of the talker motions. We used these conditions in order to determine whether any random animation could be chosen for other characters in a group, once one talker was matched to the audio. This would be an easy way to populate a crowd scene with conversing groups.

For completeness, our final synthetic condition was chosen to investigate the effect of conversational audio being played with no character displaying a talker body motion: **0 talkers, audio not matched** ($B_{0T} A_{NM}$). This contained 3 listener body animations, chosen at random, with a random audio clip playing.

Our motion capture session for these experiments differed from that for our first body motion experiment (we used this set up for our single stimulus body motion experiment (6.4)). Two sets of actors participated in recording sessions this time, where we recorded both their voices and body movements. The first set contained three males aged between 25 and 31. The second had three females aged between 22 and 28 (see Figure 6.9). We

Figure 6.9: Our audio and motion capture setup with three female actors.

chose two groups of actors in order to test if our results were generalizable. Briton and Hall [BH95] found that participants perceive females to be more expressive with gestures and that female participants rated female gestures higher than male gestures, which were found to be less fluid and more interruptive. All actors were non-professionals and were accustomed to the motion capture setup and environment. We also ensured that actors knew each other and were informed about the topics that they would be discussing in advance, in order to ensure natural, realistic conversations.

Motion capture was conducted using a 13 camera Vicon optical system, with 53 markers per actor. The markers were placed on the major joints and at regular intervals on the body, in order to capture accurate body motion. We did not capture finger or face motion as it was conversational body motion that was of most interest to us in this study. An AKG C-414 omni-directional microphone was placed on a tripod in the centre of the actor triangle to record their voices from all directions while they were being motion captured. Also, we placed a Behringer C-2 studio condenser microphone in front of each actor to record only their part of the conversations, using a MOTU-896HD external soundcard. We wanted to collect audio for all actors simultaneously and each actor individually so that we could play audio from the centre of the conversation (Section 6.6), and also position individual audio tracks corresponding to 3D positions of characters on screen (Section 6.7.2). A clapboard with motion capture markers attached was used to indicate the start of a conversation clip. Actors were instructed to place their feet in pre-specified positions

on the edges of a triangle at the start of each capture in order to prevent significant changes in their positions. In advance of each recording, we adjusted the preamp gain for each actor to ensure that no audio distortion occurred due to microphone proximity. Thereby, we minimised the audio recording constraints and the actors were informed that they could move around freely within the motion capture zone once capture had begun.

As before, we captured and recorded two different conversation types: *debates* and *dominant speaker*. Debate conversations were free-flowing in nature, where each actor expressed a strong opinion on the topic being discussed, and interruptions were common. Dominant speaker conversations allowed only one speaker to talk at a time, while the others politely listened and were not allowed to interrupt. In total, we recorded 30 dominant speaker conversations (5 per each of the 6 actors) and 10 debates (5 with the female actor group and 5 with the male). Dominant speaker conversations lasted approximately 30 seconds, while debates lasted between 2 and 3 minutes.

We chose six virtual characters to represent each actor in the experiments (Figure 6.10). The characters were chosen to approximately match the actors in age, weight and height, to minimise re-targeting errors. Throughout the experiments, we matched the motions of each actors to their virtual character. Also, as before we used a white background to provide good contrast, and randomly placed the camera focussing on one character, while keeping the other two clearly visible.

Figure 6.10: Examples of stimuli used in our experiments: (L) a real female debate and (R) a real male dominant speaker conversation.

For each dominant speaker conversation, we chose two different temporal offsets from the start of the conversation to begin a 10 second conversation clip. For each debate conversation, we chose six different offsets. For the dominant speaker conversations, we annotated clips to tag each character as either a speaker or listener. For the virtual

representations of the real conversations, we played the correct conversational audio and motion capture clips simultaneously. The synthetic conversations were made up of the conditions displayed in Table 6.4 and described in Section 6.5.1.

In order to discover what role audio plays in viewers' sensitivity to desynchronisation, we used this stimuli in a set of experiments that examined how audio and body motion affect the perception of virtual conversations. The first explored omni-directional audio, where we played the same mono audio track in both headphones. We wanted to determine whether the addition of audio had any effect on participants' sensitivity to body desynchronisation. The second experiment considered localised audio, where we played a separate audio track from each character's 3D position on screen through the headphones. We conducted this experiment to determine whether the addition of richer, localised audio would have any effect on participants' ability to distinguish real from synthetic conversations.

The real-time experimental system was developed using a commercially available animation system and an open-source renderer. This allowed us to seed each experiment randomly for each participant. In order that participants did not always associate a voice with a character, we colour modulated the characters at every trial to disguise them. Also, we placed the camera so that one of the characters was centrally focussed, but randomly chose which character to focus on at each trial. See Figure 6.10 for examples of stimuli.

The experiments were run on a workstation with 4GB of RAM, a Creative SB Audigy 2ZS soundcard and an 8-series G-Force graphics card. The stimuli were displayed on a 24-inch widescreen monitor and participants used Sennheiser HD 202 headphones to listen to the audio (Figure 6.11). Participants viewed each trial for 10 seconds and were asked to indicate using a mouse click whether they thought that the conversation they viewed was *real* or *synthetic*. The audio was played at a pre-determined volume that we found to be high enough to determine clearly the content of the conversation, but not too high as to cause discomfort. This level was maintained for all participants in all experiments with audio. We found through interviewing participants that 10 seconds was an adequate time for them to make their decisions. We randomly associated the right or left mouse button with the *real* response so as to avoid any bias towards a particular button-press. Similarly, 50% of our stimuli were real, in order to avoid any bias. After a participant gave his/her response, a cross was displayed to focus attention on the centre of the screen.

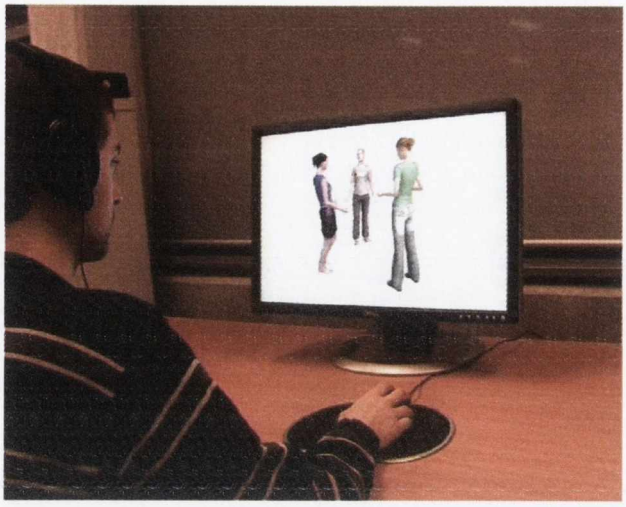


Figure 6.11: Participant in Localised Audio experiment.

## 6.6 Exp 3: Omni-Directional Audio Experiment

In this experiment, we tested participants' sensitivity to audio that is matched or unmatched to body motion. Does a talker's body motion in a conversation need to match the audio? For the dominant speaker conversations, we hypothesised that when the audio did not match the body motion of the speaker (or when more talking bodies were present than the number of voices heard) participants would find these conversations most unrealistic. Will it look plausible to use audio from a different conversation as long as the body motions are synchronised? How unrealistic do the conversations appear when both bodies are desynchronised and audio is unmatched? Since in the debates block, the conversations contained more complex dynamics and we had previously found that desynchronising debate body motions appeared quite realistic, we postulated that unmatched audio would have a similar effect regardless of the synchronisation of body motion.

To investigate these issues, nineteen new volunteers (11M, 8F) took part in this experiment. As before, it was conducted in two blocks shown in random order. The conditions we tested for both blocks were *actor group* and *AV condition*. A breakdown of the AV conditions for this experiment can be found in Table 6.4.

| Block | Factor | AV Condition | Total Trials |
|---|---|---|---|
| Debates | Real | Real ($B_S A_M$) | 18 |
| | Synth. | $B_D A_{NM}$ | 6 |
| | | $B_S A_{NM}$ | 6 |
| | | $B_D A_{M1}$ | 6 |
| Dom. Sp. | Real | Real ($B_S A_M$) | 36 |
| | Synth. | $B_S A_{NM}$ | 6 |
| | | $B_D A_{NM}$ | 6 |
| | | $B_{0T} A_{NM}$ | 6 |
| | | $B_{1T} A_{M1}$ | 6 |
| | | $B_{2T} A_{M1}$ | 6 |
| | | $B_{3T} A_{M1}$ | 6 |

Table 6.4: Experimental design for both Omni-Directional and localised Audio experiments, showing total number of trials (50% male actors, 50% female).

### 6.6.1 Results

As in our previous experiments, we averaged participants responses over repetitions for each condition. We again conducted a 3-way repeated measures ANalysis Of VAriance (ANOVA) with within-subjects factors of *AV condition* and *actor group*. For the omni-audio experiment, there was a between-subjects factor of *sex of participant*. As before, post-hoc analysis was conducted using Newman-Keuls tests for comparison of means.

Our statistical analysis showed no effect of *actor group*, which demonstrated that sensitivity to synthetic conversations was independent of the actors used. Also, in contrast to results found by Briton and Hall [BH95], there was no effect of the sex of the participant or any interactions, implying that both males and females perceived the non-verbal behaviours similarly for both male and female virtual characters.

For both blocks, we again found that participants were able to distinguish between the real condition and synthetic conditions, as the real were found to be more plausible than any other condition. Our results from the dominant speaker block of this experiment reveal some interesting results for the effect of audio on participants' perception of synthetic conversations (Figure 6.12). We found a main effect of AV condition ($F_{6,102} = 40.63, p < 0.00005$ in all cases). Post-hoc analysis showed that when there was one talker body motion ($B_{1T} A_{M1}$ and $B_S A_{NM}$), participants perceived these synthetic conversations to be equally real, regardless of whether the gestures of the talkers matched the audio or not. We also found that, as the number of talkers (depicted by body motions) grew, conversations were

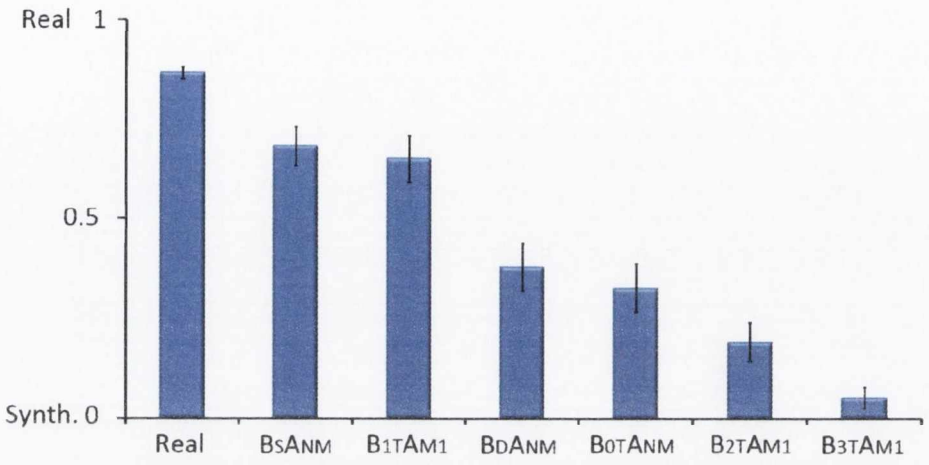



Figure 6.12: Results for our Omni-directional Audio experiment for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition.

found to be progressively less realistic, since the audio only contained one talker. Zero talkers and desynchronised body motions were also found to be unrealistic, but less so than conditions containing more than one talker (Table 6.5).

For the debates, we again found a main effect of AV condition ($F_{3,51} = 33.63, p < 0.00005$). For these conversations, we found that desynchronisation can be masked to some degree by ensuring that one character's motion matches the audio, even if the other two characters are given random motions $B_D A_{M1}$, as this was the second most plausible condition, after real (Figure 6.13). We also found that when the audio was not matched to any character, synthetic conversations were considered to be equally the least realistic, regardless of whether the body motions themselves were synchronised ($B_S A_{NM}$) or desynchronised ($B_D A_{NM}$).

We also found an effect of conversation type ($F_{1,18} = 11.546, p < 0.005$), where the dominant speaker conversations were found to be more plausible overall. This result was the opposite of our findings when looking at conversations using body motion alone. This implies that when participants are presented with more information, as in an audio-visual conversation, that they try to match the two signals together in a plausible way. The debates will be less matched than the dominant speaker conversations as the speaker voice is very likely to change at a different time than the talker role in the body motions. For the dominant speaker conversations, there is no changing of roles of talkers, so the same voice is heard throughout each trial, making it more plausible when one speaker body motion

| Block | Cond. | Mean | St. Dev. |
|---|---|---|---|
| Debates | Real | 0.85 | 0.14 |
| | $B_D A_{NM}$ | 0.32 | 0.27 |
| | $B_S A_{NM}$ | 0.37 | 0.23 |
| | $B_D A_{M1}$ | 0.52 | 0.20 |
| Dom. Sp. | Real | 0.87 | 0.096 |
| | $B_{0T} A_{NM}$ | 0.33 | 0.27 |
| | $B_{1T} A_{M1}$ | 0.65 | 0.25 |
| | $B_{2T} A_{M1}$ | 0.19 | 0.21 |
| | $B_{3T} A_{M1}$ | 0.05 | 0.11 |
| | $B_S A_{NM}$ | 0.68 | 0.21 |
| | $B_D A_{NM}$ | 0.38 | 0.25 |

Table 6.5: Mean 'real' ratings and standard deviations for Omni-directional Audio experiment.

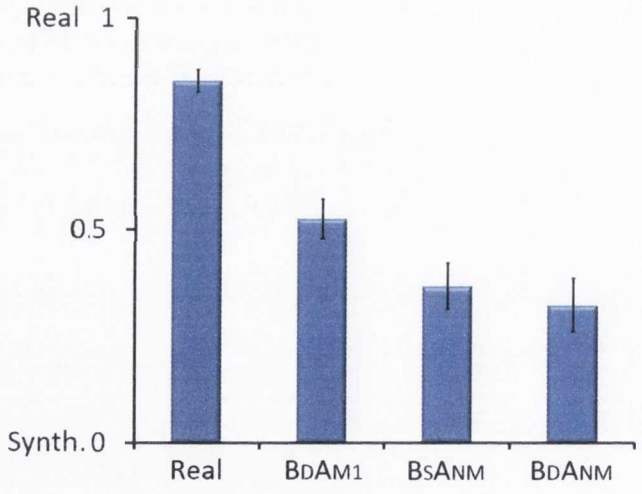



Figure 6.13: Results for our Omni-directional Audio experiment for debates, showing mean 'real' ratings and standard error bars for each condition.

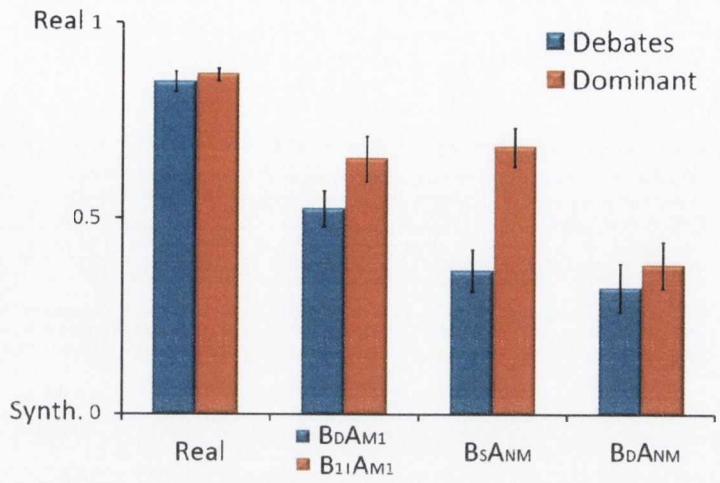



Figure 6.14: Comparison for matching conditions across our Debates (blue series) and Dominant Speaker conversations (orange series).

is displayed.

Finally, there was an interaction between conversation type and AV condition ($F_{3,54} = 6.6032, p < 0.005$), where some of the debates conditions were found to be less plausible than their matching dominant speaker conditions. Post-hoc analysis showed that participants found conversations where one talker matched the audio ($B_{1T}A_{M1}$ and $B_DA_{M1}$) more plausible for more polite conversations than debates. This was also true for our synchronised body motions not matching audio ($B_SA_{NM}$) condition. This implies that desynchronisation affects these two different conversation types in different ways, especially for the more plausible synthetic conditions. In some cases, it would be preferable to use dominant speaker conversations to ensure adequate plausibility.

## 6.7 Exp 4: Localised Audio Experiment

### 6.7.1 Baseline

We next decided to investigate the effect of 3D audio, and first tested how accurate people were at localising conversational audio. We postulated that people would correctly identify the 3D location of directional audio for virtual conversations. Will different pitches affect the ability of participants to localise audio? Previous research has shown that higher pitched tones are easier to localise than lower pitched ones [MB84]. Our six different actors

provided a reasonable range of pitch, and we hypothesised that the ability to accurately localise male and female voices would be affected by this pitch difference.

To prepare the stimuli for this experiment, we used the audio recordings from the three Behringer C-2 studio condenser microphones, which recorded the speech of each individual separately. We used only dominant speaker conversations for this experiment, so only one speaker would be heard during any given trial. Each character was assigned a static audio source at their position from where the 3D audio was played. Each static source had position, direction, orientation and fall-off properties, which could be matched to the position and orientation of the characters in each trial.

For this experiment, each trial had three characters on screen. However, this time the characters displayed no animation and appeared in a standing pose. There was one condition of *actor voice*, in order to determine whether there were differences in participants' ability in localising the voices of our six actors. In each trial, one actor voice was played from a random character's position on screen, and there were nine repetitions of each actor's voice.

Fifteen new volunteers took part in this experiment (9M, 6F). They were asked to listen to each conversation and decide from which character the voice was coming. They responded by pressing either the 1, 2 or 3 key to indicate the position of the chosen character on screen.

**Results**

We conducted a repeated measures ANOVA, where the within-subjects factor was *actor voice* and the between-subjects factor was *sex of participant*. We found no main effect of actor voice or sex of participant, which means that both male and female participants could localise audio for a range of actor voices. An independent-samples t-test was conducted to compare the accuracy of participants and we found that participants could correctly localise the audio for each actor voice well above chance ($p < 0.0005$ in all cases) as can be seen in Figure 6.15.

Our results in this baseline found that not only were participants well able to identify the 3D location of directional audio, *they were equally able to localise the voices of each actor, regardless of the sex of the actor.* This is an interesting result, as there was a conflict based on previous literature that the difference in the pitch of the audio could result in difficulties in localisation [MB84, MB02]. However, our results show participants were

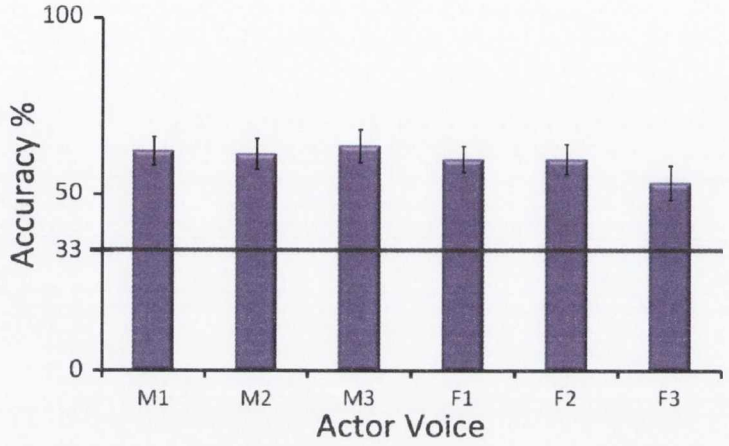



Figure 6.15: Results for our Audio localisation Baseline experiment showing participant accuracy in localising conversational audio well above chance (33%) for each of our three male and three female actor voices.

equally able to localise conversational audio for our six actor voices, implying that pitch does not affect their ability to do so. There was no effect of *sex of participant*, meaning that both males and females could localise sound equally as well.

### 6.7.2 Localised Audio Experiment

Our baseline experiment showed that people can accurately localise conversational audio. However, does the addition of a localised audio signal affect participants' ability to recognize synthetic conversations? We hypothesised that the addition of a reliable localised sound source would increase a participant's dependency on audio as a factor when making their decisions, resulting in increased recognition of synthetic conversations in most cases. However, since we found from our Omni-Directional Audio experiment that conversations where there was one speaker matching in body motion and audio ($B_D A_{M1}$ for the debates and $B_{1T} A_{M1}$ for dominant speaker conversations) were more realistic, the addition of a more reliable auditory cue might also make these conditions even more realistic.

Thirteen volunteers took part in this experiment (8M, 5F). We prepared the audio for this experiment in the same manner as in our baseline and the experiment procedure was as outlined in our Omni-Directional Audio experiment. Stimuli were presented to participants in one block each for *debate* and *dominant speaker* conversation types.

The conditions for the dominant speaker block were similar to those for the Omni-Directional Audio experiment (Table 6.4), except that in this experiment, audio localisation was also added. All but the following two conditions contained audio localised correctly for the character speaking. The desynchronised body motion, not matched to audio condition ($B_D A_{NM}$), had audio localised at a random character position. The zero talker body motions, audio not matched condition ($B_{0T} A_{NM}$) now congruently contained no audio.

The conditions for the debates block matched those for the Omni-Directional Audio experiment. The audio localisation for the debates was achieved by positioning a sound source at each character's location. For the real conversations, the three sound sources were localised at the correct characters. For the condition when one character's body motions were matched to the audio ($B_D A_{M1}$), the sound source was localised at the matching character, while the other two sound sources were randomly positioned at the two remaining characters. For the remaining conditions ($B_S A_{NM}$ and $B_D A_{NM}$), the sound sources were randomly positioned at each character.

### 6.7.3 Results

Our statistical analysis for this experiment was similar to that for our Omni-directional Audio experiment with within-subject factors of *AV condition* and *actor group* and between-subjects factor of *sex of participant*. As before, post-hoc analysis was conducted using Newman-Keuls tests for comparison of means.

Our results were similar to those with the omni-directional audio, which suggests that the addition of localised audio was not any better than omni-directional audio at helping participants to better distinguish real from synthetic conversations (Figures 6.16 and 6.17 and Table 6.6).

As before, we found the real condition to be more plausible than any synthetic condition for both debates and dominant speakers ($F_{3,33} = 28.94, p < 0.00005$ and $F_{6,66} = 14.193, p < 0.00005$ respectively). For the dominant speaker block we found the most plausible synthetic condition to be $B_S A_{NM}$, followed by $B_{1T} A_{M1}$ and $B_{0T} A_{NM}$ (Figure 6.16). Our least plausible conditions were $B_{2T} A_{M1}$ and $B_{3T} A_{M1}$ ($p < 0.00005$ in all cases). We found that for the one talker body motion matched to audio condition ($B_{1T} A_{M1}$), that there was a difference in participants' responses, where they found the male characters more realistic than the females. This could be due to the level of expressivity of the talker

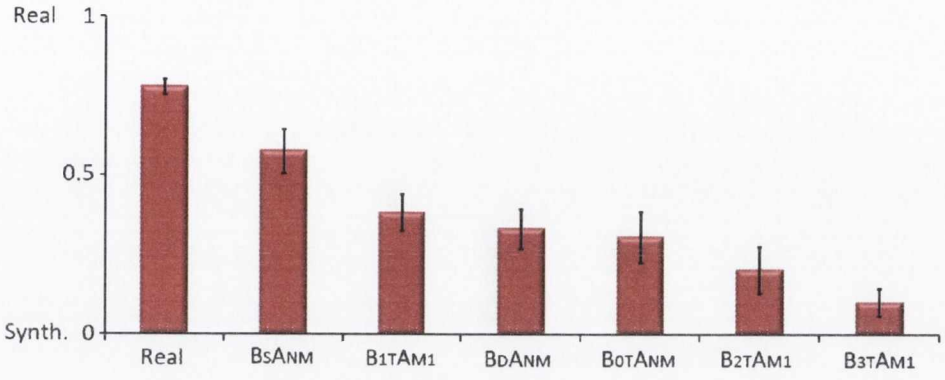



Figure 6.16: Results for our localised Audio experiment for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition.

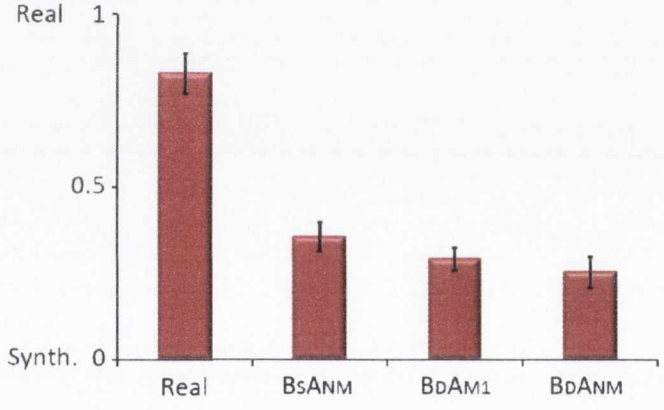



Figure 6.17: Results for our localised Audio experiment for debates, showing mean 'real' ratings and standard error bars for each condition.

gestures of the male characters, but warrants further investigation. As expected, we found no effect of sex of participant.

For the debates, we found that, when using localised audio information, participants found each of our synthetic conditions to be equally as plausible (Figure 6.17). So for these conversations, matching the audio to one of the characters body motions ($B_D A_{M1}$) is no longer an adequate way to mask desynchronisation. This will be discussed in more detail in our next section.

| Block | Cond. | Mean | St. Dev. |
|---|---|---|---|
| Debates | Real | 0.83 | 0.17 |
| | $B_D A_{NM}$ | 0.26 | 0.16 |
| | $B_S A_{NM}$ | 0.30 | 0.15 |
| | $B_D A_{M1}$ | 0.36 | 0.21 |
| Dom. Sp. | Real | 0.78 | 0.13 |
| | $B_{0T} A_{NM}$ | 0.31 | 0.29 |
| | $B_{1T} A_{M1}$ | 0.38 | 0.21 |
| | $B_{2T} A_{M1}$ | 0.21 | 0.26 |
| | $B_{3T} A_{M1}$ | 0.10 | 0.16 |
| | $B_S A_{NM}$ | 0.58 | 0.25 |
| | $B_D A_{NM}$ | 0.33 | 0.23 |

Table 6.6: Mean 'real' ratings and standard deviations for localised Audio experiment.

## 6.8 Audio Experiments Cross-Analysis

In order to find the effects of the addition of localised audio, compared to omni-directional audio on desynchronisation, we compared our conditions for these two experiments. We found slightly different results across these two experiments. For the debates, when using omni-directional audio, we found that our desynchronised condition where one speaker matched the audio $B_D A_{M1}$ was more plausible than our other synthetic conditions ($B_S A_{NM}$ and $B_D A_{NM}$). However, when a localised audio signal was used, participants no longer found this condition to be anymore plausible than the other synthetic conditions (Figure 6.18). We found a similar effect with the dominant speaker conversations; our one talker body motion condition, $B_{1T} A_{M1}$, was equally as plausible as our synchronised bodies with unmatched audio condition ($B_S A_{NM}$) for the Omni-directional Audio experiment. However, when the audio signal contained localisation information in our final experiment, our $B_{1T} A_{M1}$ condition was now less plausible than $B_S A_{NM}$ (Figure 6.19). We feel that the differences in these trends may be caused by the localised audio accelerating identification of the speaker in the trial, leaving participants to pay further attention to the other characters on screen and noticing the body motion desynchronisation.

For our statistical analysis to compare these two experiments, we conducted a 2-way ANOVA with within-subjects factor of *AV condition* and between-subjects factor of *audio signal level*. Looking at the dominant speaker block, we found an effect of AV condition ($F_{6,180} = 48.873, p < 0.00005$). Most conditions were equally as plausible across both Omni-directional Audio and localised Audio experiments. Post-hoc analysis showed that the condition with one talker body motion matched to the audio ($B_{1T} A_{M1}$) was

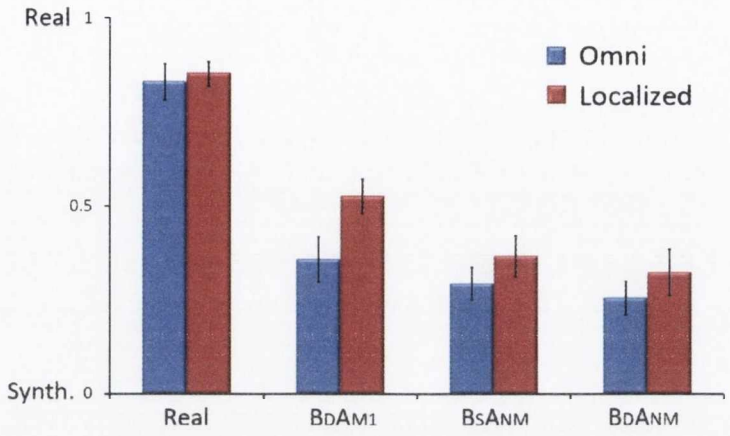



Figure 6.18: Comparison of results for Omni-directional (blue) and localised Audio (red) experiments for debates, showing mean 'real' ratings and standard error bars for each condition.

less plausible with localised audio than with omni-directional audio. For the debates, there was no effect of audio signal level, meaning that the use of localised audio does not significantly effect masking of desynchronisation. We found an effect of AV condition ($F_{3,90} = 62.333, p < 0.00005$), but post-hoc analysis showed no significant differences across the two different experiments for matching conditions.

## 6.9 Three Experiments Cross-Analysis

We compared the results found for our single stimulus body motion experiment (No Audio), Omni-Directional Audio and localised Audio experiments for matching conditions. We wished to determine the overall effect the addition of an audio signal had on participants' sensitivity to desynchronised talking bodies.

For the dominant speaker blocks, we cross analysed matching conditions (Figure 6.20). We found that the addition of audio (omni or localised) did not affect the perception of desynchronisation of the talking bodies. The one talker body motion condition was perceived to be more realistic with omni-directional audio than with localised or no audio. We had hypothesised that the localised audio would improve the realism of this condition, but participants actually found it to be synthetic more often. This unexpected result could be due to the localised audio accelerating identification of the speaker in the trial, thereby allowing attention to be shifted to the desynchronised listener motions. Perhaps

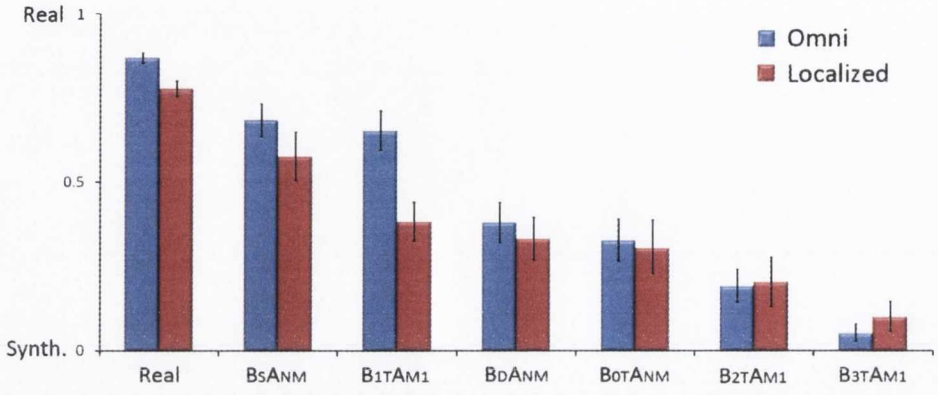



Figure 6.19: Comparison of results for Omni-directional (blue) and localised Audio (red) experiments for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition.

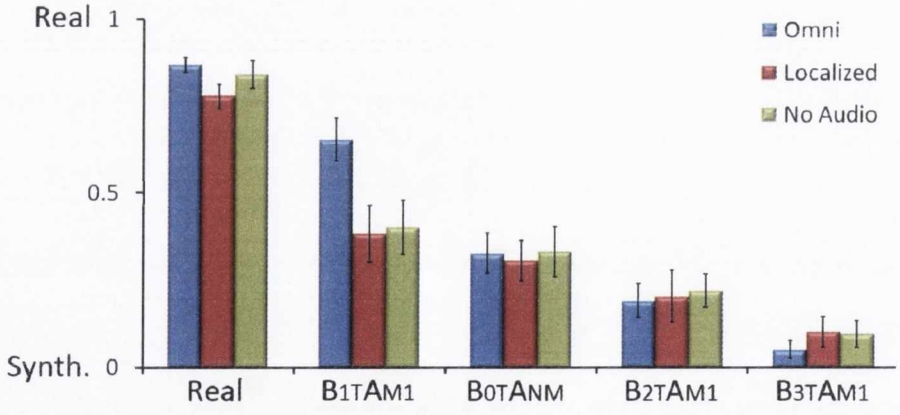



Figure 6.20: Results across 3 experiments for dominant speaker conversations, showing mean 'real' ratings and standard error bars for each condition. Note: for No Audio (green) ignore $A_x$ label, all are $A_0$.

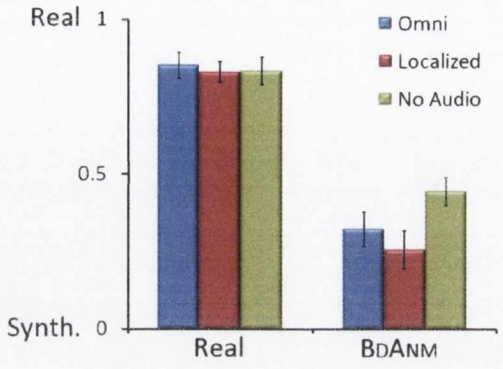



Figure 6.21: Results across 3 experiments for debates, showing mean 'real' ratings and standard error bars for each condition. Note: for No Audio (green) ignore $A_x$ label, all are $A_0$.

the addition of head-look modifications (i.e., ensuring that the listeners attend to the speaker) would make this condition more realistic.

Similarly, for the debates we cross analysed matching conditions (Figure 6.21). Whether the participants viewed the stimuli in the presence or absence of audio (both omni-directional and localised), did not significantly affect their responses. From this, we deduced that audio did not influence the results, even in these more complex dynamic conversations.

## 6.10 Discussion

Overall, we found similar trends across all experiments; with or without audio information, participants were able to recognise real conversations with high levels of accuracy. Therefore, when populating a crowd with conversing agents, for the most salient groups it would be preferable to use synchronised body motions matched to audio where possible. However, when this is not feasible (e.g., when more variety is needed from a limited database), we provide guidelines in Section 7.2 to help mask desynchronisation and maintain realism for such scenes.

For situations that call for increased variety in groups based on a finite data set, or when animation issues mean that characters cannot always be synchronised, then a method that will result in an acceptable level of realism is to match one character to the audio track. For scenes without audio, assigning one talker to a group gives a better sense of realism.

It is important to note that the stimuli for this experiment were focussed on by participants for ten seconds each, with no audio or visual distractors. Taking this into account, many of our conditions produced promising results, in particular the one talker matched to audio conditions ($B_D A_{M1}$ for debates and $B_{1T} A_{M1}$ for dominant speaker conversations). When bodies were synchronised with each other, but not matched to the audio ($B_S A_{NM}$), this was also perceived to be quite realistic. Our results could also apply to larger groups, especially for dominant speaker conversations, since attention would remain focussed on the single speaker, regardless of group size. Perhaps adding more characters for debates would only serve as distractors, due to the chaotic nature of the conversations.

### 6.10.1 Alternative Approaches

We chose to create new conversations by desynchronising body motions of the characters from the voices, maintaining the audio as it was recorded for the purpose of these experiments. We did this because it would be possible to fully desynchronise the characters body motions from each other, however, desynchronising the actors voices from the recordings could produce some strange effects due to voice cross overs, especially for the more dynamic debate style conversations. Audio processing can be more challenging as it is more difficult to separate actors voices if there is any overlap during the conversation. Also, processing audio signals can result in audible artifacts, where the voice can sound artificial or robotic. However, a similar evaluation could potentially be carried out with high quality audio processing and separating the actor voices, or a state of the art speech synthesiser. Alternatively, since participants are more tolerant of desynchronisation for the dominant speaker conversations, different actor voices could be recorded and applied to character motions to give many different conversations, as this would be much more straightforward than a fully audio and body motion capture session.

Alternative animation methods could also be used to investigate desynchronisation between audio and motion. Procedural or physics-based animation methods can be partially synchronised to the audio by selecting synchronisation points in the audio and applying an appropriate gesture to the motion of the character. This method would allow for both full synchronisation if each word in a sequence is synchronised to an appropriate gesture, and partial synchronisation, if only a selected portion of the sequence is synchronised. Finding out if partial synchronisation is enough to make the conversation appear plausible could greatly reduce the workload for animators. It could also be of use for motion captured data, as new conversations could be created by modifying parts of the motions

Figure 6.22: Screenshot of our real-time crowd system with dynamic agents and conversing groups.

to important sync points in the audio.

### 6.10.2 System Implementation

We have implemented three of the experimental conditions in our real-time crowd system (*Real*, $B_D A_{M1}$, and $B_D A_{NM}$ debates). Ten groups of conversers were placed in an open scene, amongst two hundred pedestrian characters (see Figure 6.22). A sound source was located at each of the characters in each of the groups using the OpenAL audio library (as in the Localised Audio experiment). The addition of conversing groups enhanced the overall realism of the simulation, especially for fixed camera viewpoints. However, with the camera in walk-through or fly-through mode (as shown in the supplemental video), we observed that setting plausible parameters for audio was non-trivial. Finding the correct levels of attenuation, directionality and gain in order to create a plausible simulation was challenging. In particular, there was a mismatch between the 3D audio effects of a large out-door scene and the small screen display. Furthermore, with the large amount of visual and auditory distractors when panning through the scene, the desynchronisation in even the worst case ($B_D A_{NM}$) seemed more plausible. The effects of desynchronisation and audio parametrization of conversing groups when viewed in different scenarios will be explored in future work. The effects of facial animation will also be investigated.

# Chapter 7

# Conclusions and Future Work

In this chapter, we outline the contributions of this thesis, and discuss potential directions for future work.

## 7.1 Summary of Contributions

### 7.1.1 New Methodologies

New methods have been developed for populating static scenes with virtual characters. Based on properties of the scene, a simple set of rules have been developed to position and orientate characters in a sensible way. Our rules take factors like walkable areas, obstacles, neighbouring pedestrians and groups into account. These methodologies have been shown to be a suitable method for semi-automatically populating virtual scenes, and are a preferable alternative to random placement, while not needing the attention and time taken for manual placement.

We have also presented methods for generating new conversations for virtual agents using a limited data-set of body motions and audio recordings. This has been done for conversational agents using body motion alone, omni-directional audio and also localised audio. Evaluation of the two methodologies developed here has enabled us to deduce guidelines for developers who wish to populate a virtual scene or generate groups of conversational agents with restricted resources.

### 7.1.2 New Experimental Results

A number of perceptual studies were conducted, to evaluate the plausibility of the methods we developed for our crowds. From these experiments, we have a number of results that provide novel insights into human perception of virtual crowds. This is relevant to computer graphics, as these results helped in the deduction of guidelines to aid developers working with real-time crowds. The results are also relevant to the field of psychology, as our results provide information about how virtual crowds are perceived, and the most salient interactions in a virtual conversation.

#### Pedestrian Formations

Results from our position and orientation experiments show that a viewer's ability to distinguish between real and artificial scenes depends heavily on the context of the scene and how characters adhere to this context (e.g., characters walking in bi-directional lanes in a constrained location). We found that contextual factors are vital when considering the perceived realism of pedestrian formations. From a modelling standpoint, the results from these experiments, most significantly those taking both position and orientation into account, imply that the contextual rules presented here form an effective general starting point from which to populate urban environments. The fact that most of the participants were familiar with the areas used for both scenes, in particular the open location, implies that they were aware of the general directions of flows of pedestrians and this is possibly reflected in their judgements of realism throughout the experiments.

We also investigated an unexpected result of low realism ratings for some of our real scenes, and found that there was no difference in how participants perceived the realism of 4 different real scenes. This implies that our results obtained for our static formations were not due to any peculiarities in any one individual real image, but rather to the effects of all formation types.

When examining the validity of our context rules for a number of camera viewpoints, we found that people are less able to identify random formations of characters when the camera is at eye-level, but that they were better able to identify these random formations as being unrealistic when viewing the scene at an angle that gives more information. However, we found that our context rules were found to be equally as realistic across all viewpoints. These results are useful in two ways: they suggest that the context rules are effective in both corridor and open locations no matter what angle they are viewed

at, indicating that they are appropriate for initializing a crowd system; and when the camera is at eye-level, the accuracy of the positions and orientations of the crowd is not as perceptible to the viewer. This opens opportunities for reducing the simulation burden when certain behaviours of our agents do not need to be fully accurate, and could possibly apply to other aspects such as rendering and animation of the characters also.

### Dynamic Groups

Findings from an experiment investigating the importance of groups found that the presence of groups can add realism to a dynamic crowd scene, if the group formations are plausible. The results found here suggest that further investigation of group behaviour is an important factor in creating a plausible crowd.

### Conversational Groups

Results from experiments looking at conversational groups yielded some interesting findings about what we expect from virtual conversers. Firstly, it was found that when people look at a group of conversers when they have access to visual and audio information for the characters, they tend to rely more on the body motion of the characters than what the characters are saying. Similar trends were found across matching conditions, regardless of whether the conversations were body motion alone, with omni-directional or localised audio. So, the addition of audio did not help highlight or mask conditions where body motions were desynchronised any more than when viewing the conversations without audio.

For polite, structured conversations, the number of talkers is more important. Our results show that for these conversations, when there was omni-directional audio present, it did not matter whether the audio matched the gestures of the characters, once the number of voices heard matched the number of talker body motions seen. So, when looking at these conversations, people look for the correct talker and listener roles to match the number of voices they hear, but it is not as important whether the listeners are paying attention to the speaker, or whether the speaker's gestures match what he/she is saying.

Other interesting results we found were that desynchronisation in conversations affects different conversations in different ways. Overall, our findings were that dominant speaker conversations were more plausible when desynchronised, for matching conditions. So, it would be preferable to use these conversations if desynchronisation is needed. Also, it was

shown that people are able to localise conversational audio for a three person conversation to an accurate degree. This had not previously been known and, while localisation does not help to mask desynchronisation, it could add heightened realism to these conversations if people can localise the speaker's position. Finally, there were mixed views in the literature regards differences in male and female conversational behaviour, but no differences were found either between male and female actors or participant responses.

### 7.1.3 Motion Capture Database

To carry out the experiments on conversational agents, numerous motion capture sessions were carried out, where the audio and body motions of three actors were captured simultaneously, for natural conversations. This has resulted in a large database of conversational animations covering a variety of topics, for both males and females, which can be used for more perceptual experiments and can be implemented in our crowd system.

## 7.2 Guidelines

From the results we have obtained from our perceptual studies, we have deduced some recommendations for developers who wish to apply our methodologies to virtual crowds.

When populating a pedestrian scene:

- The context of the location is an important consideration when creating pedestrian formations in a crowd scene.
- The context rules presented here are more important when the location is of a more constrained nature. When the scene is less constrained, less attention may be required for plausible positions and orientations due to the more open nature of the location and possibility for more behaviours.
- When the camera is at eye-level, less processing may be required to generate plausible positions and orientations than when the camera is at an angle that provides a more visible view of the individuals in the scene.

For situations that call for increased variety in groups based on a finite data set, or when animation issues mean that characters cannot always be synchronised, we provide the following guidelines:

- Localisation of audio does not increase realism of conversing groups, so may not be worth additional implementation effort.
- Audio can be plausibly assigned on-the-fly to dominant speaker conversations by ensuring appropriate talker/listener roles, regardless of audio matching or body desynchronisation.
- Debates will be more difficult to implement, as they will only appear sufficiently plausible if at least one talker in the group is matched to audio.
- No special considerations need to be taken into account when using male or female characters in these circumstances.

## 7.3 Future Work

### 7.3.1 Conversational Model

Future plans for Metropolis include the development of a conversational model for the idle agents, using the methodologies and results described in this thesis. We want to build on the work carried out to date and add some other features to the model:

**Facial motion:** Firstly, it would be important generate some basic facial animation for our conversers, by animating the mouth of the talker characters. Although this does not seem to be important for the identification of speaker roles within a conversation, since our participants were very accurate at identifying speakers using body motion alone, it likely have an effect on how realistic our conversations appear. The addition of facial animation could potentially have an effect on how plausible participants found our desynchronised conversations to be, so this will also have to be investigated.

**Complex scenes:** It is important to note that the stimuli for our conversation experiments were focussed on by participants for ten seconds each, with no audio or visual distractors. Taking this into account, many of the desynchronised conditions produced promising results. However, conducting a similar experiment in a crowd scene might yield more positive results for some of the desynchronised conversations that did not appear plausible. This would mean increased options for variety.

**Additional parameters:** It is also necessary to explore other parameters to ensure the most varied crowds possible. When implementing the groups in Metropolis, we found that estimating the correct levels of attenuation, directionality and gain in order to create a

plausible simulation was challenging. So it would be useful to investigate this further to ensure optimal levels with respect to matching an outdoor scene to a screen display. It is likely that our results could also apply to larger groups, especially for dominant speaker conversations, since attention would remain focussed on the single speaker, regardless of group size, so we will examine the effects of desynchronisation on groups of different sizes, as well as at different distances. It is also very possible that ensuring that the listeners attend to the talker will further increase the realism of synthetic conversations.

**Eye-tracking:** Another way in which we could find out more about how we perceive virtual conversing agents is from eye-tracker data. Using an eye-tracking device, we could determine what participants attend to in a virtual conversation. For example, we might find that people pay more attention to head orientation or gestures. This information would allow us to improve the believability of the desynchronised conversations, by altering the salient areas.

Using the methodologies and results already collected, along with findings from these future plans, we hope to build a model to generate plausible conversers based on motion captured gestures. Our aim is to create dynamic scenarios where agents can join and leave conversational groups in a convincing manner. Groups will need to be of different sizes, agent positions within a group will vary and plausible ways to transition between conversations will be needed. By ensuring that the listeners attend to the talker, we should be able to further increase the realism of synthetic conversations.

### 7.3.2 Dynamic Groups

Once we have a model for our conversational agents, the next step will be to improve the behaviour of the dynamic groups for Metropolis. Ideally, these groups will interact with each other and exhibit reactive behaviours to other individuals and groups and the environment around them. Once the groups have been integrated into our navigation behaviour system, we will then set up a method for our agents to change between conversational and dynamic roles. Agents should be able to leave and join groups, and groups should be able to go from conversational to dynamic, and vice versa. This will include formation changes, and likely animation changes, as plausible dynamic group members should still interact with one another.

### 7.3.3 Awareness of Environment

While we do not necessarily want our agents to have high level cognition, we do want them to appear to be aware of their surroundings. This is an important factor that could potentially increase the plausibility of their behaviour. For instance, we will want our dynamic agents to interact with the traffic in the environment, i.e., crossing the road at appropriate times and only at designated locations. This will involve the extension of the virtual world to include more traffic areas and the annotation of the environment with areas where pedestrians can cross. We will then increase the functionality of the crowd behaviours to allow them to only cross at appropriate times, and link these behaviours up with our traffic simulator.

Another way for agents to give the impression of intelligence, is for them to be aware of other agents around them. This is done at a low level with our collision detection and avoidance for our dynamic agents, but could be extended for conversational agents and implemented at a higher level. For instance, if a group are conversing, and a dynamic agent passes nearby, with a view to joining the group, it would be appropriate for the conversing agents to be aware of his presence and for the group formation to change once he has joined the group. Having this behavioural property could also be important if a situation arises where the user navigates the environment with an avatar; if the agents in the environment react to the presence of the user's avatar, it could increase their sense of presence in the simulation. However, these behaviours are only needed for characters that are highly salient to the user, and are not needed for agents in the background. As our agents appear more clever, there will be a greater need for Level-Of-Detail methods to maintain real-time frame rates, and challenges will be presented regarding where to transition to higher level behaviours, with no loss of plausibility to the user.

# Bibliography


[ABW06] G. Antonini, M. Bierlaire, and M. Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.

[ACF01] O. Arikan, S. Chenney, and D. A. Forsyth. Efficient multi-agent path planning. In *Proceedings of the Eurographics workshop on Computer animation and simulation*, pages 151–162, 2001.

[Ave77] A. F. Aveni. The not-so-lonely crowd: Friendship groups in collective behavior. *Sociometry*, 40(1):96–99, 1977.

[BD07] M. Badawi and S. Donikian. The generic description and management of interaction between autonomous agents and objects in an informed virtual environment. *Computer Animation and Virtual Worlds*, 18(4-5):559–569, 2007.

[BG95] B.M. Blumberg and T.A. Galyean. Multi-level direction of autonomous creatures for real-time virtual environments. In *SIGGRAPH 1995: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 47–54, 1995.

[BH95] N. J. Briton and J. A. Hall. Beliefs about female and male nonverbal communication. *Sex Roles: Journal of Research*, 32(1-2):79–90, 1995.

[BH97] D. Brogan and J.K. Hodgins. Group behaviors for systems with significant dynamics. *Autonomous Robots*, 4:137–153, 1997.

[BJ03] D. Brogan and N.L. Johnson. Realistic human walking paths. In *Proceedings of the 16th International Conference on Computer Animation and Social Agents (CASA 2003)*, pages 94–101, 2003.

[BLA02] O.B. Bayazit, J-M. Lien, and N.M. Amato. Better group behaviors in complex environments using global roadmaps. In *Proceedings of the eighth international conference on Artificial life*, pages 362–370, 2002.

[BMH98] D. Brogan, R. Metoyer, and J.K. Hodgins. Dynamically simulated characters in virtual environments. *IEEE Computer Graphics and Applications*, 18(5):58–69, 1998.

[BR00] T.L. Bennett and G. Riva. Virtual reality in rehabilitation of spinal cord injuries: A case report. *Rehabilitation Psychology*, 45(1):81–88, 2000.

[BTBV99] V. Blanz, M.J. Tarr, H.H. Bulthoff, and T. Vetter. What object attributes determine canonical views? *Perception*, 28:575–600, 1999.

[BTM+08] S. Bernard, J. Therien, C. Malone, S. Beeson, A. Gubman, and R. Pardo. Taming the Mob: Creating believable crowds in Assassin's Creed. Presented at Game Developers Conference (San Francisco, CA, Feb 18–22), 2008.

[BY98] J. Boissonnat and M. Yvinec. *Algorithmic geometry*. Cambridge University Press, New York, NY, USA, 1998.

[CBF+05] T.J. Clarke, M.F. Bradshaw, D.T. Field, S.E. Hampson, and D. Rose. The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34(10):1171–1180, 2005.

[Che87] L.P. Chew. Constrained delaunay triangulations. In *SCG '87: Proceedings of the third annual symposium on Computational geometry*, pages 215–222, 1987.

[CPB+94] J. Cassell, C. Pelachaud, N.I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *SIGGRAPH 1994: Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420, 1994.

[CST+10] E.J. Carter, L. Sharan, L. Trutoiu, I. Matthews, and J.K. Hodgins. Perceptually motivated guidelines for voice synchronization in film. *ACM Transactions on Applied Perception*, 7(4):1–12, 2010.

[CV99] J. Cassell and H. Vilhjálmsson. Fully embodied conversational avatars: Mak-

ing communicative behaviors autonomous. *Autonomous Agents and Multi-Agent Systems*, 2(1):45–64, 1999.

[CVB01] J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *ACM SIGGRAPH 2001: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486, 2001.

[DAPB08] F. Durupinar, J. Allbeck, N. Pelechano, and N.I. Badler. Creating crowd variation with the OCEAN personality model. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, pages 1217–1220, 2008.

[DDN95] R. Dunbar, N. Duncan, and D. Nettle. Size and structure of freely forming conversational groups. *Human Nature*, 6:67–78, 1995.

[DH03] W. Daamen and S. Hoogendoorn. Controlled experiments to derive walking behavior. *Journal of Transport and Infrastructure Research*, 3(1):39–59, 2003.

[DHOO05] S. Dobbyn, J. Hamill, K. O'Conor, and C. O'Sullivan. Geopostors: a real-time geometry / impostor crowd rendering system. In *SI3D '05: Proceedings of the 2005 symposium on Interactive 3D graphics and games*, pages 95–102, 2005.

[DS75] J.M. Dabbs and N.A. Stokes. Beauty is power: The use of space on the sidewalk. *Sociometry*, 38(4):551–557, 1975.

[Edg10] Edge online: Performance art. Available at: http://www.next-gen.biz/features/performance-art, January 2010. Accessed: 27 September 2010.

[EF69] P. Ekman and W.V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98, 1969.

[Ekm92] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200, 1992.

[FBT99] N. Farenc, R. Boulic, and D. Thalmann. An informed environment dedicated to the simulation of virtual humans in urban context. *Computer Graphics Forum*, 18(3):309–318, 1999.

[Geh03] J. Gehl. *Life Between buildings.* The Danish Architectural Press, 2003.

[GM05] S. Goldin-Meadow. *Hearing gesture: How our hands help us think.* Belknap Press, 2005.

[Gof71] E. Goffman. *Relations in Public: Microstudies of the Public Order.* Basic books, 1971.

[GR06] A. Gerdelan and N. Reyes. Synthesizing adaptive navigational robot behaviours using a hybrid fuzzy A* approach. In *Computational Intelligence, Theory and Applications*, pages 699–710. 2006.

[GR09] A. Gerdelan and N. Reyes. Towards a generalised hybrid path-planning and motion control system with auto-calibration for animated characters in 3d environments. In *Advances in Neuro-Information Processing*, volume 5506 of *Lecture Notes in Computer Science*, pages 1079–1086. 2009.

[GS08] P. Gomez and J. Shutter. Memory for objects in canonical and non canonical viewpoints. *Psychonomic Bulletin and Review*, 15(5):990–944, 2008.

[GV08] G. Giorgolo and F.A.J. Verstraten. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, pages 31–36, 2008.

[HBJW05] D. Helbing, L. Buzna, A. Johansson, and T. Werner. Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transportation Science*, 39(1):1–24, 2005.

[Hel91] D. Helbing. A mathematical model for the behavior of pedestrians. *Behavioral Science*, 36(4):298–310, 1991.

[Hel92] D. Helbing. A fluid dynamic model for the movement of pedestrians. *Complex Systems*, 6:391–415, 1992.

[HFV00] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407(6803):487–490, 2000.

[HJ09] D. Helbing and A. Johansson. Pedestrian, crowd and evacuation dynamics. *Encyclopedia of Complexity and Systems Science*, 16, 2009.

[HJM+06] D. Helbing, A. Johansson, J. Mathiesen, M.H. Jensen, and A. Hansen. Analytical approach to continuous and intermittent bottleneck flows. *Physical Review Letters*, 97(16):168001–1 – 168001–4, 2006.

[HK02] T. Hostetler and J. Kearney. Strolling down the avenue with a few close

friends. In *Proceedings of Eurographics Ireland Workshop on Computer Graphics*, pages 7–14, 2002.

[HM95] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995.

[HM99] A. Hannah and T. Murachver. Gender and conversational style as predictors of conversational behavior. *Journal of Language and Social Psychology*, 8(2):153–174, 1999.

[HMDO05] J. Hamill, R. McDonnell, S. Dobbyn, and C. O'Sullivan. Perceptual evaluation of impostor representations for virtual humans and buildings. *Computer Graphics Forum (Eurographics 2005)*, 24(3):623–633, 2005.

[HO03] J. Hamill and C. O'Sullivan. Virtual dublin - a framework for real-time urban simulation. *Journal of WSCG*, 11:221–225, February 2003.

[Hos02] T. Hostetler. *Controlling steering behavior for small groups of pedestrians in virtual urban environments.* PhD thesis, 2002.

[HTER04] C.G. Healey, L. Tateosian, J.T. Enns, and M. Remple. Perceptually based brush strokes for nonphotorealistic visualization. *ACM Transactions on Graphics*, 23(1):64–96, 2004.

[Jam53] J. James. The distribution of free-forming small group size. *American Sociological Review*, 18(5):pp. 569–570, 1953.

[JCP+10] E. Ju, M.G. Choi, M. Park, J. Lee, K.H. Lee, and S. Takahashi. Morphable crowds. *To appear in ACM Transactions on Graphics (SIGGRAPH ASIA 2010)*, 2010.

[JHO10] S. Jörg, J.K. Hodgins, and C. O'Sullivan. The perception of finger motions. In *APGV '10: Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pages 129–133, 2010.

[JIS+08] J.A. Jones, J.E. Swan II, G. Singh, E. Kolstad, and S.R. Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *APGV '08: Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 9–14, 2008.

[Joh73] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1973.

[JT07] D. Jan and D.R. Traum. Dynamic movement and positioning of embodied agents in multiparty conversations. In *AAMAS '07: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–3, 2007.

[Kal01] M. Kallmann. Object interaction in real-time virtual environments. *PhD Thesis, Dept. of CS, École Polytechnique Fédérale de Lausanne*, 2001.

[Kan05] S. Kankri. Virtual human crowds with sociological behaviours. Interim report, Universite Pierre et Marie Curie Paris, 2005.

[Ken73] A. Kendon. *The role of visible behaviour in the organization of social interaction*, chapter 2, pages 29–74. Academic Press, 1973.

[Ken90] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, 1990.

[Ken94] A. Kendon. Do gestures communicate? A review. *Research on language and social interaction*, 27(3):175–200, 1994.

[KG80] C. Korte and R. Grant. Traffic noise, environmental awareness and pedestrian behavior. *Journal of Environment and Behavior*, 12(3):408–420, 1980.

[Khi90] C.J. Khisty. Non-euclidean metrics in nonmotorized transportation. *Transportation Research Record*, 1281, 1990.

[Khi94] C.J. Khisty. Evaluation of pedestrian facilities: Beyond the level-of-service concept. *Transportation Research Record*, 1438, 1994.

[KKM+06] S. Kopp, B. Krenn, S. Marsella, A.N. Marshall, C. Pelachaud, H. Pirker, K.R. Thórisson, and H. Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *International Conference On Intelligent Virtual Agents*, pages 21–23, 2006.

[KPOL05] A. Kamphuis, J. Pettré, M. Overmars, and J-P. Laumond. Path finding for the animation of walking characters. In *Poster Proceedings of Eurographics/ACM Siggraph Symposium on Computer Animiation*, 2005.

[KS07] E. Krahmer and M. Swerts. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414, 2007.

[Kuf98] J.J. Kuffner Jr. Goal-directed navigation for animated characters using real-time path planning and control. In *Proceedings of the International Workshop on Modelling and Motion Capture Techniques for Virtual Environments*, volume 1537 of *Lecture Notes In Computer Science*, pages 171–186, 1998.

[Lam09] F. Lamarche. Topoplan: a topological path planner for real time human navigation under floor and ceiling constraints. *Computer Graphics Forum*, 28(2):649–658, 2009.

[LCHL07] K.H. Lee, M.G. Choi, Q. Hong, and J. Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 109–118, 2007.

[LCL07] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3):655–664, 2007.

[LMM03] C. Loscos, D. Marchal, and A. Meyer. Intuitive crowd behaviour in dense urban environments using local laws. In *Proceedings of the Theory and Practice of Computer Graphics*, pages 122–129, 2003.

[LPDO10] M. Larkin, S. Paris, S. Dobbyn, and C. O'Sullivan. Every last detail: density based level of detail control for crowd rendering. In *I3D '10: Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games Posters Session*, pages 1–1, 2010.

[LTK09] S. Levine, C. Theobalt, and V. Koltun. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics*, 28(5):1–10, 2009.

[Mar02] D. Marchal. Simulating pedestrian crowd behaviour in virtual cities. *PhD Thesis, École Polytechnique Fédérale de Lausanne*, 2002.

[MB84] A.D. Musicant and R.A. Butler. The influence of pinnae-based spectral cues on sound localization. *The Journal of the Acoustical Society of America*, 75(4):1195–1200, 1984.

[MB02] B.S. Muller and P. Bovet. Performance and reaction times in monaural localization of first names in the horizontal plane. *Brain and Language*, 82(1):1 – 9, 2002.

[MBG08] A. McNamara, R. Bailey, and C. Grimm. Improving search task performance

using subtle gaze direction. In *APGV '08: Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 51–56, 2008.

[McD10] R. McDonnell. Facial animation for real-time conversing groups. In *To appear in Proceedings of The ACM / SSPNET 2nd International Symposium on Facial Analysis and Animation*, 2010.

[McN96] D. McNeill. *Hand and mind: What gestures reveal about thought.* University of Chicago Press, 1996.

[MDBP95] P. Maes, T. Darrell, B. Blumberg, and A. Pentland. The ALIVE system: full-body interaction with autonomous agents. In *Computer Animation '95., Proceedings*, pages 11–18, 1995.

[MDCO02] B. Macnamee, S. Dobbyn, P. Cunningham, and C. O'Sullivan. Men behaving appropriately: Integrating the role passing technique into the aloha system. In *Proceedings of AISB '02 Symposium on Adaptive Agents and Multi-Agent Systems*, pages 59–62, 2002.

[Men99] A. Menache. *Understanding Motion Capture for Computer Animation and Video Games.* Morgan Kaufmann Publishers Inc., 1999.

[MGT99] S.R. Musse, F. Garat, and D. Thalmann. Guiding and interacting with virtual crowds in real-time. In *Proceedings of Eurographics Workshop on Animation and Simulation*, pages 23–34, 1999.

[MJM+09] R. McDonnell, S. Jörg, J. McHugh, F. Newell, and C. O'Sullivan. Investigating the role of body shape on the perception of emotion. *ACM Transactions on Applied Perception*, 6(3):1–11, 2009.

[MLD+08] R. McDonnell, M. Larkin, S. Dobbyn, S. Collins, and C. O'Sullivan. Clone Attack! Perception of Crowd Variety. *ACM Transactions on Graphics*, 27(3):26:1–26:8, 2008.

[MLH+09] R. McDonnell, M. Larkin, B. Hernandez, I. Rudomin, and C. O'Sullivan. Eye-catching crowds: Saliency based selective variation. *ACM Transactions on Graphics*, 27(3):55:1–55:10, 2009.

[MMTS10] J.D. Moss, E.R. Muth, R.A. Tyrrell, and B.R. Stephens. Perceptual thresholds for display lag in a real visual environment are not affected by field of view or psychophysical technique. *Displays*, 31(3):143 - 149, 2010.

[Mor09] Natural motion morpheme. Available at: http://www.naturalmotion.com/morpheme.htm, September 2009. Accessed: 27 September 2010.

[MPG+10] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5(4):e10047, 2010.

[MR04] I. Milne and G. Rowe. Ogre: Three-dimensional program visualization for novice programmers. *Education and Information Technologies*, 9(3):219–237, 2004.

[MRPD08] P. McDonald, H. Rice, F. Pilla, and S. Dobbyn. Communicating environmental noise data using virtual soundscaping. In *Proceedings of Internoise 2008*, 2008.

[MT97] S.R. Musse and D. Thalmann. A model of human crowd behavior: group inter-relationship and collision detection analysis. In *Proc. Workshop of Computer Animation and Simulation of Eurographics*, pages 39–51, 1997.

[MT01] S.R. Musse and D. Thalmann. Hierarchical model for real time simulation of virtual human crowds. *IEEE Transactions on Visualization and Computer Graphics*, 7(2):152–164, 2001.

[Mus00] S.R. Musse. Human crowd modelling with various levels of behaviour control. *PhD Thesis, École Polytechnique Fédérale de Lausanne*, 2000.

[NG05] C. Niederberger and M. Gross. Level-of-detail for cognitive real-time characters. *The Visual Computer: International journal of Computer Graphics*, 21(3):188–202, 2005.

[NKAS08] M. Neff, M. Kipp, I. Albrecht, and H.P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):1–24, 2008.

[OCV+02] C. O'Sullivan, J. Cassell, H. Vilhjálmsson, J. Dingliana, S. Dobbyn, B. McNamee, C. Peters, and T. Giang. Levels of detail for crowds and groups. *Computer Graphics Forum*, 21(4):733–741, 2002.

[Old68] S.J. Older. Movement of pedestrians on footways in shopping streets. *Traffic Engineering and Control*, 10(4):160–163, 1968.

[OPOD10] J. Ondřej, J. Pettré, A-H. Olivier, and S. Donikian. A synthetic-vision based

steering approach for crowd simulation. *ACM Transactions on Graphics*, 29(4):1–9, 2010.

[PAB07] N. Pelechano, J.M. Allbeck, and N.I. Badler. Controlling individual agents in high-density crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 99–108, 2007.

[PB06] N. Pelechano and N.I. Badler. Modeling crowd and trained leader behavior during building evacuation. *IEEE Computer Graphics and Applications*, 26(6):80–86, 2006.

[PDB06] S. Paris, S. Donikian, and N. Bonvalet. Environmental abstraction and path planning techniques for realistic crowd simulation: Research articles. *Computer Animation and Virtual Worlds*, 17(3-4):325–335, 2006.

[PDMO03] C. Peters, S. Dobbyn, B. Macnamee, and C. O'Sullivan. Smart objects for attentive agents. In *Proceedings of 11th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media*, 2003.

[Pel09] C. Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630 – 639, 2009.

[PO] C. Peters and C. O'Sullivan. Metroped: A tool for supporting crowds of pedestrian ai's in urban environments. In *Proceedings of AISB Symposium on AI and Games Short Papers*, pages 64–69.

[POSB05] N. Pelechano, K. O'Brien, B. Silverman, and N.I. Badler. Crowd simulation incorporating agent psychological models, roles and communication. In *Proceeedings of First International Workshop on Crowd Simulation*, pages 21–30, 2005.

[PPD07] S. Paris, J. Pettré, and S. Donikian. Pedestrian reactive navigation for crowd simulation: a predictive approach. *Computer Graphics Forum*, 26(3):665–674, 2007.

[PRC81] S. Palmer, E. Rosch, and P. Chase. *Canonical Perspective and the Perception of Objects*, pages 135–151. Hillsdale, 1981.

[PSL02] J. Pettré, T. Simeon, and J.P. Laumond. Planning human walk in virtual

environments. In *IEEE/RSJ International Conference on Intelligent Robots and System*, volume 3, pages 3048–3053, 2002.

[PV08] C. Pedica and H. Vilhjálmsson. Social perception and steering for online avatars. In *Intelligent Virtual Agents*, volume 5208 of *Lecture Notes in Computer Science*, pages 104–116. 2008.

[PV09] C. Pedica and H. Vilhjálmsson. Spontaneous avatar behavior for human territoriality. In *Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 344–357. 2009.

[RC09] D. Rose and T.J. Clarke. Look who's talking: Visual detection of speech from whole-body biological motion cues during emotive interpersonal conversation. *Perception*, 38(1):153–156, 2009.

[RD05a] S.J. Rymill and N.A. Dodgson. A psychologically-based simulation of human behaviour. In *Proceedings of EG UK Conference on Theory and Practice of Computer Graphics*, pages 35–42, 2005.

[RD05b] S.J. Rymill and N.A. Dodgson. Psychologically-based vision and attention for the simulation of human behaviour. In *Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, pages 229–236, 2005.

[RD09] N. Redmond and J. Dingliana. Investigating the effect of real-time stylisation techniques on user task performance. In *APGV '09: Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, pages 121–124, 2009.

[Rey87] C.W. Reynolds. Flocks, herds, and schools: A distributed behavioral model. In *Computer Graphics (Proceedings of SIGGRAPH '87)*, pages 25–34, 1987.

[Rey99] C.W. Reynolds. Steering behaviors for autonomous characters. In *Proceedings of Game Developers Conference*, pages 763–782, 1999.

[RG98] A.S. Rao and M.P. Georgeff. *Modeling rational agents with a BDI-architecture*, pages 317–328. 1998.

[RHK+95] B.O. Rothbaum, L.F. Hodges, R. Kooper, D. Opdyke, J.S. Williford, and M. North. Effectiveness of computer-generated (virtual reality) graded exposure in the treatment of acrophobia. *American Journal of Psychiatry*, 152(4):626, 1995.

[RHW+96] B.O. Rothbaum, L. Hodges, B.A. Watson, G.D. Kessler, and D. Opdyke. Virtual reality exposure therapy in the treatment of fear of flying: A case report. *Behaviour Research and Therapy*, 34(5-6):477–481, 1996.

[Rou98] R. Rouse. Embrace your limitations – cut-scenes in computer games. *ACM SIGGRAPH 1998: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 32(4):7–9, 1998.

[RP03] P. Reitsma and N. Pollard. Perceptual metrics for character animation: Sensitivity to errors in ballistic motion. *ACM Transactions on Graphics*, 22(3):537–542, 2003.

[SAC+07] A. Sud, E. Andersen, S. Curtis, M. Lin, and D. Manocha. Real-time path planning for virtual agents in dynamic environments. In *Proceedings of IEEE Virtual Reality Conference*, pages 91–98, 2007.

[SCCD04] V. Sundstedt, A. Chalmers, K. Cater, and K. Debattista. Top-down visual attention for efficient rendering of task related scenes. In *VMV '04: Proceedings of the 9th annual workshop on Vision, Modeling and Visualization*, pages 209–216, 2004.

[SFC04] S. Stylianou, M. Fyrillas, and Y. Chrysanthou. Scalable pedestrian simulation for virtual cities. In *Proceedings of the 2004 ACM symposium on Virtual reality software and technology*, pages 65–72, 2004.

[SGA+07] A. Sud, R. Gayle, E. Andersen, S. Guy, M. Lin, and D. Manocha. Real-time navigation of independent agents using adaptive roadmaps. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, pages 99–106, 2007.

[SKG05] M. Sung, L. Kovar, and M. Gleicher. Fast and accurate goal-directed motion synthesis for crowds. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 291–300, 2005.

[SL75] R.S. Sobel and N. Lillith. Determinant of nonstationary personal space invasion. *Journal Of Social Psychology*, 97:39–45, 1975.

[ST05] W. Shao and D. Terzopoulos. Autonomous pedestrians. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 19–28, 2005.

[Sti69] I.B. Stilitz. The role of static pedestrian groups in crowded spaces. *Ergonomics*, 12(6):821–839, 1969.

[Tar95] M.J. Tarr. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, 2(1):5582, 1995.

[TC00] F. Tecchia and Y. Chrysanthou. Real time rendering of densely populated urban environments. In *Proceedings of the Eurographics Workshop on Rendering Techniques '00*, pages 45–56, 2000.

[TD00] G. Thomas and S. Donikian. Modelling virtual cities dedicated to behavioural animation. *Computer Graphics Forum*, 19(3):71–80, 2000.

[Tha01] D. Thalmann. The foundations to build a virtual human society. In *Proceedings of the Third International Workshop on Intelligent Virtual Agents, Lectures in Computer Science, Vol. 2190*, pages 1–14, 2001.

[TLCC01] F. Tecchia, C. Loscos, R. Conroy, and Y. Chrysanthou. Agent behaviour simulator (ABS): A platform for urban behaviour development. In *Proceedings of Games Technology Conference 2001*, pages 17–21, 2001.

[TOY+07] D. Thalmann, C. O'Sullivan, B. Yersin, J. Maïm, and R. McDonnell. Populating virtual environments with crowds. *Eurographics Tutorials*, 1:25–124, 2007.

[TSN+05] T. Tanaka, S. Sugihara, H. Nara, S. Ino, and T. Ifukube. A preliminary study of clinical assessment of left unilateral spatial neglect using a head mounted display system (HMD) in rehabilitation engineering technology. *Journal of NeuroEngineering and Rehabilitation*, 2(1):31, 2005.

[TT94] X. Tu and D. Terzopoulos. Artificial fishes: physics, locomotion, perception, behavior. In *ACM SIGGRAPH 1994: Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 43–50, 1994.

[VC98] H. Vilhjálmsson and J. Cassell. Bodychat: autonomous communicative behaviors in avatars. In *Proceedings of the second international conference on Autonomous agents*, pages 269–276, 1998.

[VCC+07] H. Vilhjálmsson, N. Cantelmo, J. Cassell, N.E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. Marshall, C. Pelachaud, Z. Ruttkay,

K. Thórisson, H. van Welbergen, and R. van der Werf. The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science*, pages 99–111. 2007.

[Vic10] Vicon software. Available at: http://www.vicon.com/, September 2010. Accessed: 27 September 2010.

[Vil04] H. Vilhjálmsson. Animating conversation in online games. In *Entertainment Computing ICEC 2004*, volume 3166 of *Lecture Notes in Computer Science*, pages 139–150. 2004.

[Wei93] U. Weidmann. Transporttechnik der fussgänger - transporttechnische eigenschaften des fussgängerverkehrs (literaturstudie). Literature research, Institut füer Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau IVT an der ETH Zürich, 1993.

[Wol73] M. Wolff. Notes on the behaviour of pedestrians. In *In People in Places: The Sociology of the Familiar*, chapter 3, pages 35–48. 1973.

[WRMW95] J.P. Wann, S. Rushton, and M. Mon-Williams. Natural problems for stereoscopic depth perception in virtual environments. *Vision Research*, 35(19):2731 – 2736, 1995.

[Zac01] J. Zacharias. Pedestrian behavior and perception in urban walking environments. *Journal of Planning Literature*, 16(1):3–18, 2001.