# Metropolis: multisensory simulation of a populated city

Carol O'Sullivan and Cathy Ennis
Trinity College Dublin
Email: Carol.OSullivan@tcd.ie

Fig. 1. Close-up of two conversing characters in Metropolis

***Abstract*—Creating realistic populated virtual environments is a challenge that many graphics and VR researchers are currently tackling. There are many interesting problems to solve, such as rendering and animating large and varied crowds efficiently and realistically in believable surroundings, and creating plausible behaviours and sounds for the individual inhabitants and their environment. This is the challenge that we are addressing in the Metropolis project, where our aim is to create the sights and sounds of a convincing crowd of humans and traffic in a complex cityscape. Exploring the perception of virtual humans and crowds is also integral to our approach, through psychophysical experiments with human participants.**

## I. Introduction

Metropolis is an interdisciplinary research project involving graphics, sound engineering and neuroscience. The main motivation is to create a convincing crowd that inhabits a multisensory city environment – recreating all the sights and sounds of a busy city full of pedestrians. The main focus of the project is the real-time visual, motion and audio simulation of crowds, groups and individual human inhabitants of a virtual city. Rather than trying to simulate all the physically precise properties of a real-life crowd, our aim is to create a simulation that appears plausible and compelling to the viewer. To this end, we harness 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 considering the simulation of real-time crowds for applications such as games, there are a number of trade-offs that need to be made in order to optimise quality while maintaining high and constant frame-rates. Much computational and memory resources are needed to execute rendering, animation, audio and behaviour processes to simulate a large crowd, so sometimes accuracy and detail must be sacrificed. As a result, perceptible artifacts can appear, such as low level of detail rendering of the crowd models, degradation of the animation of the humans' motions, and other limitations such as low variety in the appearance and motions of the crowd, or low believability of crowd behaviour due to a lack of human characteristics such as emotional expressiveness. Our aim is therefore to simulate a large-scale heterogenous crowd in real-time while optimising the resources available to us, using principles of human multisensory perception to ensure that the experience is believable and immersive for the user.

The brain integrates information in order to create a final percept across all the senses, which are highly interactive. A stimulus perceived in one sense can often have a direct effect on another, so our aim is to take this into account when simulating a crowd of virtual humans. If we are looking at a crowded scene, will hearing the sounds associated with such an environment enhance our feeling of presence? Can the properties of the audio change the perceived interpretation of a scene and its inhabitants, or could multisensory information be used to distract from and disguise anomalies within the scene? The main research activities in the Metropolis project involve the visual simulation of the city and crowd, the creation of a realistic soundscape for the environment, and perceptual research focussed on exploring people's expectations of real and virtual human crowds. These combined efforts target a common objective of exploiting knowledge of human multisensory perception to identify and enhance the features that increase the plausibility of the final simulation, while ameliorating perceptible artifacts that reduce realism and presence.

978-0-7695-4419-9/11 $26.00 © 2011 IEEE
DOI 10.1109/VS-GAMES.2011.9

Fig. 2. Metropolis scene depicting both conversational groups and pedestrians

## II. Background

The inspiration and the backdrop for Metropolis came from an earlier project, called Virtual Dublin [1], which was a 3D virtual simulation of Dublin city, populated with virtual humans (See Figure 3). This crowd system was based on the novel idea of *Geopostors* [2], 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 threshold pixel-to-texel ratio, which was validated through perceptual experiments [3]. The results were minimal popping artifacts and a large scale crowd rendered in real-time.

While large numbers of characters could be displayed in real-time in this system, other aspects limited the plausibility achieved. Firstly, the agents in the crowd exhibited only basic behaviours, 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. In the Metropolis project, artists generated realistic building models and a small number of highly detailed virtual human models with very realistic appearance, in order to support the overall research goal of the project to make 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.

## III. System Overview

The Metropolis system is written in C++ code, with several sub-systems allowing for individual research teams to implement their own changes individually, with subsequent 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. Figures 1, 2 and 4 show the Metropolis system in its current form.

The different sub-systems for Metropolis handle:

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

We will discuss these systems in further detail in Section IV. There is also an experiment system, which allows members of the Metropolis team to conduct perceptual experiments using the real-time crowd system. This useful tool provides an alternative to the usual practice of presenting video clips to participants, as it allows for automatic random ordering of the stimuli and also incorporates many functionalities of the Metropolis system, thereby providing the experimenter with easy access to control specific variables. Commonly used experimental paradigms are supported, such as different types of staircase procedures, method of constant stimuli, and others. 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, thereby allowing for variables to be modified on the fly.

## IV. Research Areas

### A. Rendering

The rendering sub-system handles the display of the environment and crowd on the selected display device. The environment consists of the buildings and other props contained in the city, such as lights, trees and litter bins. The real-time rendering of the crowd can be achieved using a variety of methods, including point-based rendering, impostors, level of detail geometry and others. Research is ongoing into methods

Fig. 3. Scenes from the Virtual Dublin system, the precursor to Metropolis

Fig. 4. Large crowd of pedestrians in Metropolis

for reducing computation while introducing variety, thereby creating the impression of detail using a limited amount of resources [4], [5] (see Figure 5) and developing a rendering system capable of creating and displaying large numbers of fully animated characters using commodity hardware [6].

We use an open-source rendering engine called Ogre (Object-Oriented Graphics Rendering Engine) to render the Metropolis environment [7]. Ogre is a multi-platform scene-oriented 3D graphics engine, which 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.

### *B. Animation*

The movement of each agent is simulated in the animation sub-system. An in-house animation system has been developed 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 [8] (See Figure 6). The animation system then applies the appropriate motion for each agent. The animations we use for Metropolis locomotion are motion-captured walk loops with a variety of different speeds and turning rates. Our approach is based on a *parametric space*, in which we automatically derive motion parameters from the captured motion clips (e.g., speed/curvature of a walk cycle). A 2D triangulation of this space then allows us to use linear methods to synthesize a motion with arbitrary parameters. This allows for fast and flexible animation of mid-level of detail locomoting characters.

Fig. 5. Variety is achieved efficiently through fast lookup of different colour outfits, facial and clothing textures, and accessories

We have recently perceptually evaluated the realism of this type of processing, and discovered that slowing down fast motions is much more acceptable than speeding up slower motion clips [9].

We have another layer of functionality in the animation subsystem that controls the motions for more complex behaviours, including the conversational characters. For these characters, we use Natural Motion's Morpheme [10] solution. This is an animation engine, which provide the user with an interface to test animation transitions, and create finite state machines (FSM) to transition between a set of animations. The system takes as input processed animations, which in our case have been motion captured and post-processed in a modelling package, and outputs a Finite State Machine (FSM). These FSMs and the corresponding sets of animations are then used in Metropolis to create more complex behaviours, such as conversing groups.

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 about the appearance of the actor being recorded. 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 motion capture 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 [11]. 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 three degrees of freedom for each marker. Rotational information for the markers is calculated relative to the orientation of three 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.

For a motion capture session, the actor or actors whose

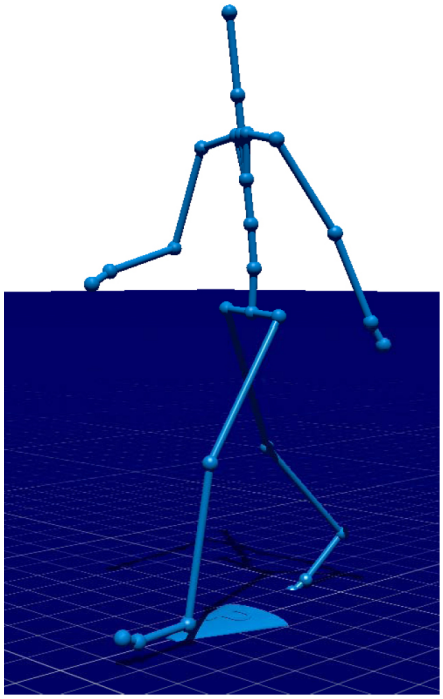

Fig. 6. Animated skeleton, with proxy bone shown on ground plane

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 Metropolis, we used 53 markers per body (Figure 7), with no facial, or finger motion captured (current research on the project is extending this work to include both faces and hands [12], [13]). 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.

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

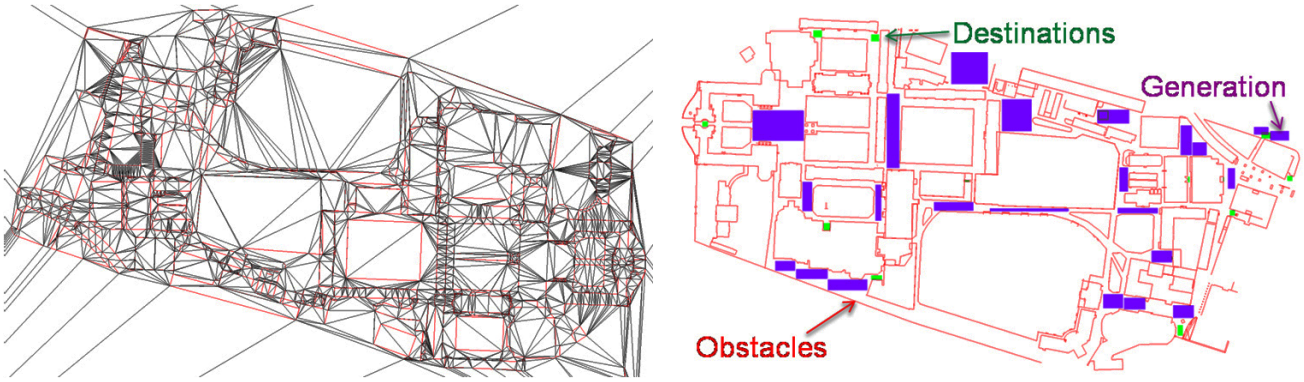


Fig. 8. Part of our environment (Trinity College) subdivided according to Constrained Delaunay Triangulation (CDT) (left), and the informed environment showing obstacles, destination and generation locations (right).

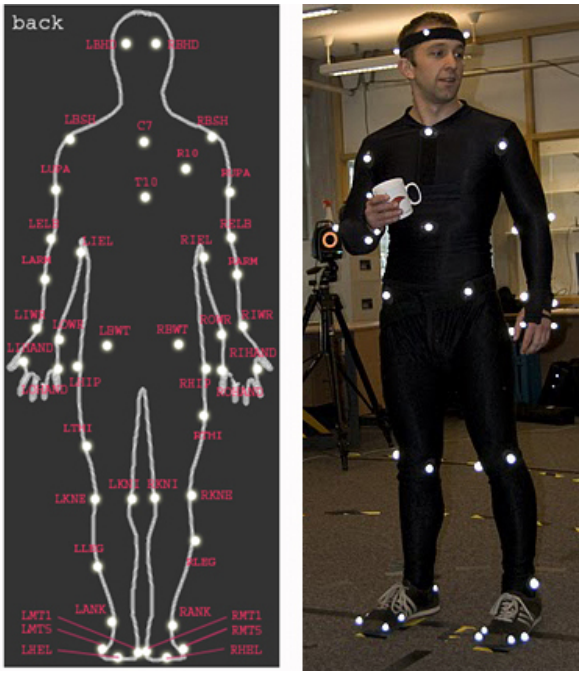


Fig. 7. An actor in a motion capture suit, with marker labels on the left

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 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.

### C. Behaviour

The behaviour sub-system in Metropolis controls the the agents who navigate the environment. The path-finding and steering behaviours for our agents are based on the method described by Paris et al. [14]. 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:

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

*a) 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 – a grid-based solution is not appropriate due to complexity of the environment. Our method is based on Constrained Delaunay Triangulation (CDT) [15], which involves editing the environment as a 2D ground representation, with a small elevation. An in-house tool uses CDT to produce a representation of the environment that annotates areas as being either obstacles, generation locations for agents to enter the environment, or destination locations to which agents navigate throughout the simulation (Figure 8).

*b) Agent Navigation:* Navigation of the agents within this informed environment occurs in two phases. 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 9).

*c) 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 10). The vehicles in Metropolis are based on a method that involves layering a road network within the virtual environment [16]. 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.

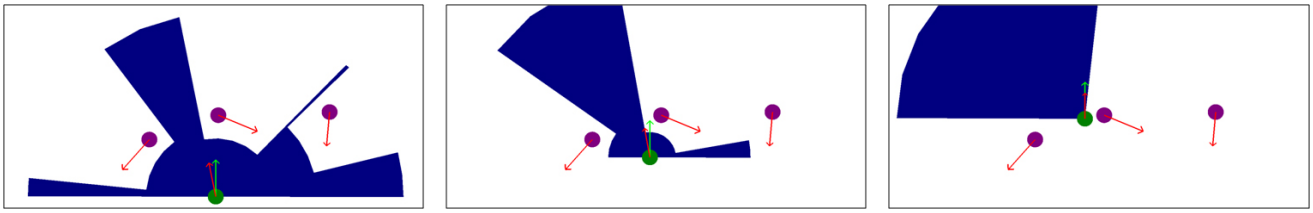

Fig. 9. Predictive geometric collision avoidance for an agent over three time steps, showing the allowed space the agent can use at each time step in blue.

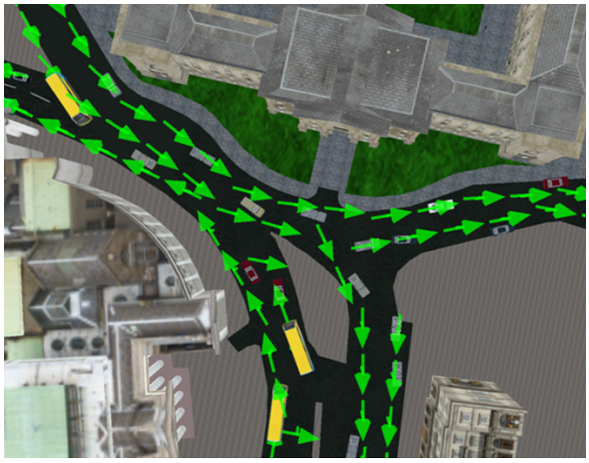

Fig. 10. Traffic simulation in Metropolis using navigable demarcations

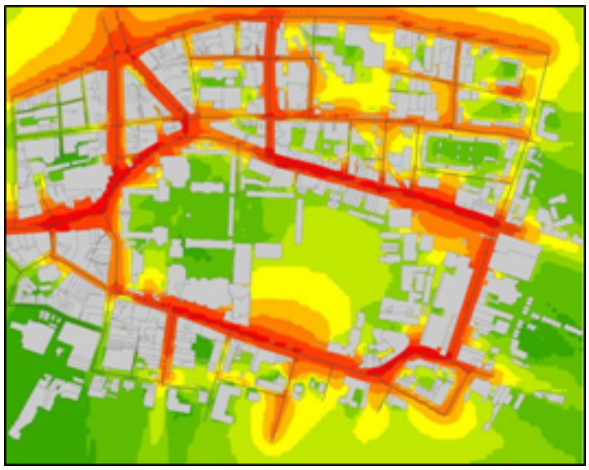

Fig. 11. Noisemap showing audio levels at different locations around Trinity College Dublin

### D. Audio

The first problem being tackled by the audio team is how to create ambient noise that matches the sounds heard while walking through different parts of a city. Trinity College, which is at the centre of Dublin City, is a highly pedestrianised area, so the main ambient noise should consist of human-made and other natural sounds. 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 [17] (Figure 11). 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.

Fig. 12. Experiments were conducted to determine the plausibility of different crowd formations in a variety of contexts, from differing viewpoints

### E. Multisensory Perception

Neuroscience research consists mainly of psychophysical experiments to explore how real and virtual crowds and scenes are perceived. Topics that are currently under investigation include depth perception in a virtual environment and the role of audio when performing a visual search task in a crowded virtual environment [18]. The role of emotion in real and virtual crowds is also being explored [19], [20]. We have also conducted experiments into the effects of scene context and camera viewpoints (see Figure 12) on the perceived realism of crowd formations [21] and explored the effects of desynchronised body motions in conversational groups [22] and the role of audio in such scenes [23]. Finger and facial motion are also important factors in the creation of believable humans, and we are investigating the level of realism that needs to be achieved in order to create plausible groups and crowds of characters in real-time [12], [24]. 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).

## V. Conclusion

The Metropolis system provides a flexible framework for conducting research on different aspects of crowd and city simulation. For example, we are using the platform to integrate procedural city modelling methods [25] and natural transitions between different types of environment representations [26]. Future work involves the creation of more complex and believable individual and group behaviours, group dynamics and physical interactions between crowd characters, global illumination for more realistic environmental effects, and the simulation of complex interactions between people and vehicles. We also plan to use the Metropolis engine to simulate historical reconstructions of Dublin through the ages, and as a training and treatment framework for people at risk of social exclusion, both ideal application areas for serious game technologies.

## Acknowledgment

We would like to thank the Metropolis team in Trinity College Dublin, and Science Foundation Ireland who provide the funding for this project, grant number S.F.I.-06IN.1I96.

## References

[1] J. Hamill and C. O'Sullivan, "Virtual dublin - a framework for real-time urban simulation," *Journal of WSCG*, vol. 11, pp. 221–225, Feb. 2003.

[2] 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*, 2005, pp. 95–102.

[3] 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)*, vol. 24, no. 3, pp. 623–633, 2005.

[4] R. McDonnell, M. Larkin, B. Hernandez, I. Rudomin, and C. O'Sullivan, "Eye-catching crowds: Saliency based selective variation," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 55:1–55:10, 2009.

[5] R. McDonnell, M. Larkin, S. Dobbyn, S. Collins, and C. O'Sullivan, "Clone Attack! Perception of Crowd Variety," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 26:1–26:8, 2008.

[6] 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*, 2010, pp. 1–1.

[7] I. Milne and G. Rowe, "Ogre: Three-dimensional program visualization for novice programmers," *Education and Information Technologies*, vol. 9, no. 3, pp. 219–237, 2004.

[8] M. Prazak, L. Kavan, R. McDonnell, S. Dobbyn, and C. O'Sullivan, "Moving crowds: A linear animation system for crowd simulation," in *I3D '10: Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games Posters Session*, 2010, pp. 1–1.

[9] M. Prazak, R. McDonnell, and C. O'Sullivan, "Perceptual evaluation of human animation timewarping," in *ACM SIGGRAPH Asia Sketches and Applications (SIGGRAPH Asia'10)*, 2010 (In Press).

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

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

[12] S. Jörg, J. 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*, 2010, pp. 129–133.

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

[14] S. Paris, S. Donikian, and N. Bonvalet, "Environmental abstraction and path planning techniques for realistic crowd simulation: Research articles," *Computer Animation and Virtual Worlds*, vol. 17, no. 3-4, pp. 325–335, 2006.

[15] L. Chew, "Constrained delaunay triangulations," in *SCG '87: Proceedings of the third annual symposium on Computational geometry*, 1987, pp. 215–222.

[16] 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*, ser. Lecture Notes in Computer Science, 2009, vol. 5506, pp. 1079–1086.

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

[18] J. S. Chan, C. Maguinness, S. Dobbyn, P. McDonald, H. J. Rice, C. O'Sullivan, and F. N. Newell, "Aurally aided visual search in depth using 'virtual' crowds of people," *Journal of Vision*, vol. 10, no. 7, p. 886, 2010.

[19] 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*, vol. 6, no. 3, pp. 1–11, 2009.

[20] "Perceiving emotion in crowds: the role of dynamic body postures on the perception of emotion in crowded scenes," *Experimental Brain Research*, vol. 204, no. 3, pp. 361–72, 2010.

[21] C. Ennis, C. Peters, and C. OSullivan, "Perceptual effects of scene context and viewpoint for virtual pedestrian crowds," *ACM Transactions on Applied Perception (Impact Factor: 1.447)*, vol. 8, no. 2, p. In Press, 2011.

[22] R. McDonnell, C. Ennis, S. Dobbyn, and C. O'Sullivan, "Talking bodies: Sensitivity to desynchronization of conversations," *ACM Transactions on Applied Perception (Impact Factor: 1.447)*, vol. 6, no. 4, 2009.

[23] C. Ennis, R. McDonnell, and C. O'Sullivan, "Seeing is believing: Body motion dominates in multisensory conversations," *ACM Transactions on Graphics (Impact Factor: 3.619)*, vol. 29, no. 3, 2010.

[24] R. McDonnell and M. Breidt, "Face reality: investigating the uncanny valley for virtual faces," in *ACM SIGGRAPH ASIA 2010 Sketches*, 2010, pp. 41:1–41:2.

[25] B. Cullen and C. O'Sullivan, "A caching approach to real-time procedural generation of cities from gis data," *Journal of WSCG*, vol. 19, no. 3, pp. 119 – 126, 2011.

[26] Y. Morvan and C. O'Sullivan, "Handling occluders in transitions from panoramic images: A perceptual study," *ACM Transactions on Applied Perception (Impact Factor: 1.447)*, vol. 6, no. 4, 2009.