Aaron S.W. Wong, Stephan K. Chalup, Shashank Bhatia, Arash Jalalian, Jason Kulk, Steven Nicklin & Michael J. Ostwald 'Visual gaze analysis of robotic pedestrians moving in urban space', Architectural Science Review, 55(3), 2012

Available from: http://dx.doi.org/10.1080/00038628.2012.688013

Visual Gaze Analysis of Robotic Pedestrians Moving in Urban Space

Aaron S. W. Wong[1], Stephan K. Chalup[1], Shashank Bhatia[1], Arash Jalalian[1], Jason Kulk[1], Steven Nicklin[1] and Michael J. Ostwald[2]

[1]School of Electrical Engineering and Computer Science, The University of Newcastle, Australia

[2]School of Architecture and Built Environment, The University of Newcastle, Australia

Corresponding author: Tel: +61 2 49218628; Fax: + +61 2 4921 6993; Email: aaron.wong@uon.edu.au

ABSTRACT: This study is founded on the idea that an analysis of the visual gaze dynamics of pedestrians can increase our understanding of how important architectural features in urban environments are perceived by pedestrians. The results of such an analysis can lead to improvements in urban design. However, a technical challenge arises when trying to determine the gaze direction of pedestrians recorded on video. High "noise" levels and the subtlety of human gaze dynamics hamper precise calculations. However, as robots can be programmed and analysed more efficiently than humans this study employs them for developing and training a gaze analysis system with the aim to later apply it to human video data using the machine learning technique of manifold alignment. For the present study a laboratory was set up to become a model street scene in which autonomous humanoid robots of approximately 55*cm* in height simulate the behaviour of human pedestrians. The experiments compare the inputs from several cameras as the robot walks down the model street and changes its behaviour upon encountering "visually attractive objects". Overhead recordings and the robot's internal joint signals are analysed after filtering to provide "true" data against which the recorded data can be compared for accuracy testing. A central component of the research is the calculation of a torus-like manifold that represents all different 3D head directions of a robot head and which allows for ordering extracted 3D gaze vectors obtained from video sequences. We briefly describe how the obtained multidimensional trajectory data can be analysed by using a temporal behaviour analysis technique based on support vector machines that was developed separately.

Keywords: Gaze Analysis, Localisation, Manifold Learning, Pedestrians, Robots, Saliency, Tracking

INTRODUCTION

In the late 1970s William H. Whyte famously analysed the behaviour of pedestrians in complex urban environments and demonstrated that pedestrians' actions were governed by a combination of strategic and opportunistic decisions. Strategic decisions typically included the desire to move to particular points in space such as bus stops and train stations. Opportunistic actions were typically governed by things seen along the strategic path, for example shops, seats, friends, or obstructions (Whyte, 1980). At the present time pedestrian simulation software is being used to assist the design of major buildings and infrastructure, typically sports and transport centres as well as urban spaces. However this existing software is almost entirely based on simulations of strategic movement. For any pedestrian simulation software to be useful for more general urban or architectural analysis, it must be able to simulate opportunistic decision-making. This implies that it must have some capacity to predict how people will react in environments where there are multiple different, conflicting visual cues – often called "visual attractors" or "visually attractive objects" -– competing for a pedestrian's attention. However, this remains a complex problem to the present day because the relationship between human gaze and human reactions has never been adequately extracted (from video recordings) and modelled (in software). Accurately modelling human pedestrian behaviour relates to opportunistic decisions, which are based on judgements made about current gaze direction. These are notoriously difficult to identify due to the high noise levels in video recordings and the subtlety of gaze dynamics. This interdisciplinary project aims to begin the process of solving this task by using techniques from robotics, machine vision and architectural space analysis.

The question of how humans interact with the built environment has been investigated from many perspectives in different disciplines. A large number of studies in architecture, cognitive science and environmental research investigated how the built environment can influence lifestyle (Auchincloss & Diez Roux, 2008; Frank, Engelke, & Schmidt, 2003; Gao & Gu, 2009; Saarloos, Kim, & Timmermans, 2009). However, their results were often inconclusive, in part because it is difficult to determine whether features of the environment or genetic factors may be dominant in controlling human behaviour. The way in which the aesthetics of the environment shapes the behaviour of human pedestrians remains heavily theorised, but poorly understood. The impact of factors such as path widths, connection to open space and the presence of obstacles or attractive objects is similarly unknown. In order to overcome these design problems associated with the way pedestrians move in space, some past studies used agent-based models or computer simulations (Auchincloss & Diez Roux, 2008; Gao & Gu, 2009) while others dealt with observations of humans (Frank, Engelke, & Schmidt, 2003; Saarloos, Kim, & Timmermans, 2009).

What is striking in the simulation studies of human or agent behaviour is that most rely on recording the trajectories and resting locations of essentially point-like agents (that

is, the pedestrian is represented as a single point in space with no other physical characteristics). Some agent models appear to be more complex, displaying a full 3-dimensional pedestrian body with animated skeleton. But in most cases this is superficial and only for graphical purposes, while the underlying body dynamics are still very basic (Bandini, Manzoni, & Vizzari, 2009; Magnenat-Thalmann & Thalmann, 2005; Magnenat-Thalmann, Jain, & Ichalkaranje, 2008). All of this means that the vast majority of past research into the modelling and prediction of pedestrian movement, regardless of whether it has used observations of humans or computer agents, grossly simplifies the pedestrian to a point a space and vector or movement.

The present project expands the scope of this type of investigation in two ways.

I.) First, the paper highlights an aspect of human visual behaviour that is encoded in the dynamics of the visual gaze vector. Our hypothesis is that the gaze vector is an important feature of human visual behaviour and its dynamics reflects interaction with the visual environment. This approach was previously proposed and addressed in pilot simulation experiments (Jalalian, Chalup, & Ostwald, 2011). There are some precedents to this way of thinking about pedestrians including a virtual crowd model that employs graphically sophisticated 3-dimensional pedestrians that can display “look-at” behaviour if they pass close to a window (Maim, Haegler, Yersin, Mueller, Thalmann, & Gool, 2007). This behaviour follows a rule that lets a wandering agent slow down and look through the window until it reaches a certain distance thereafter it resumes a faster walk. Other more sophisticated agent models can avoid collisions (Molnár & Starke, 2001), sense the type of surface they walk on, and can incorporate a number of simple artificial intelligence features (Aschwanden, Haegler, Halatsch, Jeker, Schmitt, & Gool, 2009). However, even if these pedestrians may look graphically sophisticated, their visual behaviour is still both repetitive and simplistic when compared to that of humans. Obviously there is a trade-off between the complexity of the agent model and what is possible in autonomous agent simulations (Musse, Kallmann, & Thalmann, 1999). The present study differs from previous work by putting an emphasis on sophisticated modelling and analysis of visual gaze dynamics. It is known that there are differences in the visual behaviour of different groups of people. For example, children are more likely to focus their attention to task-irrelevant objects compared with adults (Egan, Willis, & Wincenciak, 2009). This can include, for example, focusing on an advertisement for ice cream while crossing a street, or looking at trees and buildings rather than watching the path while walking. Signal noise and individual subject differences may therefore obscure some of the signals associated with visual gaze behaviour, making them only statistically detectable.

II.) The second new aspect is that this study investigates robotic pedestrians as an intermediate stage between simulated pedestrians and real human pedestrians. These three stages in a larger study, of which the present paper describes the second one, are as follows:

Stage 1 (Fully artificial): Pedestrians and their environments are simulated. The artificial worlds of the simulated agents are modelled using plans of real urban spaces. Pilot experiments for stage 1 were previously reported by (Jalalian, Chalup, & Ostwald, 2011).

Stage 2 (Semi artificial): Robotic pedestrians are programmed to operate in a human-like manner in a laboratory environment that reflects features of a real urban space.

Stage 3 (Real world): Human pedestrians are analysed by using video recordings of their movements in selected urban areas. Stage 3 involves several computer vision challenges and is planned as a future component of this project.

Pedestrian studies using simulations (as in stage 1) or video analysis of real humans (as in stage 3) are relatively common (Ono, Okabe, & Sato, 2006; Ren, Rahman, Kehtarnavaz, & Estevez, 2010). It is a newer idea to employ robotic pedestrians to increase our understanding of the relationship between fully simulated pedestrian agents and real world human pedestrians. The use of real robots involves a number of challenges associated with gaze direction detection that will be addressed in the present paper.

Visual gaze direction can be derived from a person's head position and eye centre location. It has been shown that within controlled environments where video with sufficient accuracy is feasible, it is possible to detect the eyes and the head direction of pedestrians and to use this information to calculate a visual gaze direction (Valenti, Staiano, Sebe, & Gevers, 2009). A number of tools for eye centre or eye corner location exist (Valenti, Staiano, Sebe, & Gevers, 2009). However, these typically require that a medium-to-high resolution image of the subject can be obtained. Such studies are also typically conducted in circumstances where a camera can directly face the subjects, such as where the subject sits in front of a computer or in a car. The present study aims to contribute to the eventual development of a system that enables the recording of pedestrian behaviour and the calculation of the 3D gaze direction from a distance and where, in general, only a low resolution image of the head can be obtained and eye centres are typically not detectable.

Thus, the remainder of this paper is structured as follows: (a) the robots involved in the study are described; (b) the robotic pedestrian experiments and their results supported by overhead tracking are reported; (c) it is discussed how robots can localise and detect salient objects when they leave the lab and act in a real environment; (d) techniques and experiments for visual gaze analysis without overhead tracking are described and lastly; (e) a discussion and conclusion is presented. The present paper develops and expands on the methods and results recorded in previous research by the authors (Wong, Chalup, Bhatia, Jalian, Kulk, & Ostwald, 2011).

ROBOTIC PEDESTRIANS

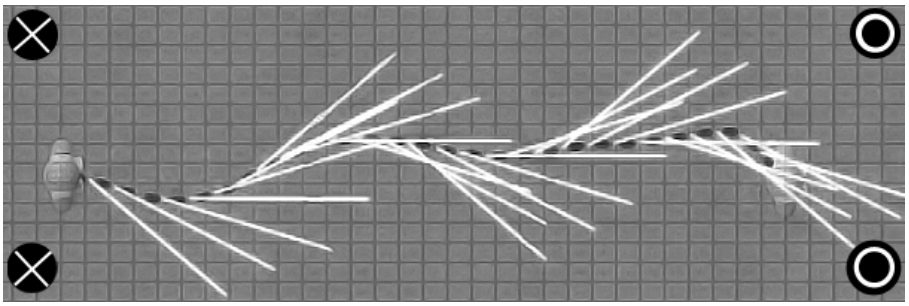
The robotic platform used for this paper is a humanoid robot (Gouaillier, et al., 2009). The robot stands 55cm tall, weighs 5kg and is equipped with 21 degrees of freedom. The robot uses an internal x86 processor to enable autonomous operation. The on-board processor performs all object recognition and cognitive functions required to walk through the environment. There are two degrees of freedom in the head, allowing the head pitch to range from −38.5º to 29.5º, and the head yaw to pan between ±119.5º. The robot has a forward facing camera in the head with a 45º by 34.35º field of view. The camera produces images at 30Hz with a resolution of 640 by 480 pixels. In the robot's chest there are ultrasonic distance sensors that are used to prevent the robot from running into walls or other obstacles.

A. Omnidirectional Walk of the Robots

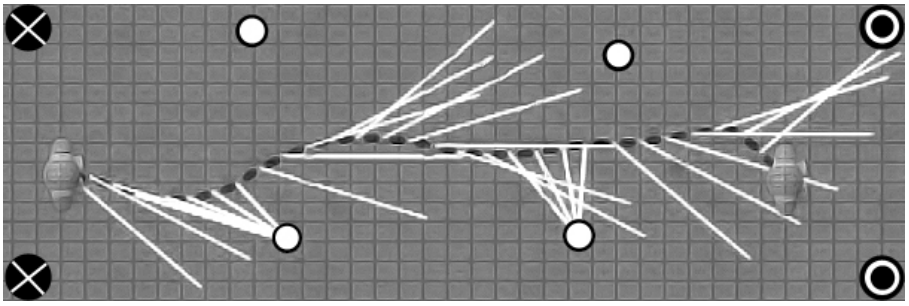
The robot is equipped with omni-directional walk engine. Walk patterns are generated online using inverse kinematics from a simple Zero Moment Point (ZMP) trajectory that is calculated from user specified step parameters (Gouaillier, et al., 2009). The particular parameters used in the present project have been selected to provide a small improvement in stability over the default settings, and a significant improvement in speed. The walk engine makes use of ZMP feedback to stabilise the robot. The robot is capable of walking at a forward speed of approximately 16 cm/s, a sideward speed of approximately 12 cm/s, and a turning rate of 0.5 rad/s. The walk is configured so that the translational velocity is directed towards the target, and the rotational velocity is directed such that the target is brought in front of the robot. However, given that the maximum translational speed is significantly higher than the rotational speed, the robot is frequently walking with a non–forward translation velocity. Consequently, the walk velocity vector is a good indication of the robot's target.

B. Implementation of Gaze Behaviour

The robot is capable of searching for objects by panning its head, and is capable of tracking an object. The searching head movements consist of slow and smooth motions calculated to scan over a sector in front of the robot as explained by the schematic drawing in Figure 1 (a). The head panning speed is varied such that the speed of the ground in the robot's image is at a constant regardless of its distance from the robot. Consequently, the searching movements are reasonably fast when scanning areas close to the robot, and slow when scanning along the horizon. An urban attractor, like a park bench, bus stop or shop window (in this study represented by an orange sphere [small white circle in Figure 1 (b)] with a diameter of about 9 cm), is tracked by maintaining the object within its camera's field of view. The object tracking behaviour produces much faster head movements, so when an attractive object is detected the robot's gaze is quickly focused on the object, as seen in Figure 1 (b). Given the robot's narrow field of view, when it is focused on something attractive it can see very little of the surrounding environment.



(a) With no attractive objects.
The robot's head pans at a constant speed while walking.



(b) With several attractive objects.
The robot fixates on two attractive objects while walking.

Figure 1: The robot's path (dashed black line) and gaze vector (white lines) while walking along a section of the model street that is marked by two "X" beacons (at the West end) and two "O" beacons (at the East end).

LABORATORY EXPERIMENTS SUPPORTED BY OVERHEAD TRACKING

A scale urban environment was modelled using cardboard boxes to simulate street walls and vista openings. Bright orange spheres were located in the street, typically placed in line with the "building façade", to simulate common urban visual attractors. Each robot was programmed to model the behaviour of a human pedestrian who walks along the street and encounters various attractors which stimulate the pedestrian's gaze, leading to opportunistic direction changes (Figure 5 (a)).

A. Laboratory Environment and Setup of the Experiments

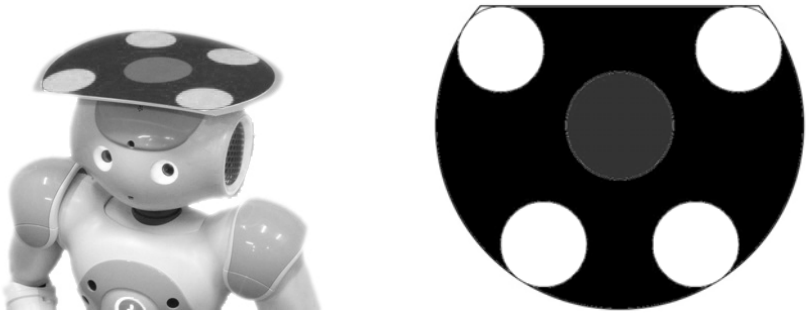
There are several external factors of the environment that decide the overall performance of the robotic system, such as the friction and terrain of the surface where the robot walks, lighting conditions, and the number of visible landmarks. To maintain the external conditions of the environment, the laboratory is equipped with different support features. The 6 by 1.5 metre area representing the model street is covered by flat, slippage-free carpet. The robot was programmed and calibrated under the laboratory's lighting conditions to search and recognise the "O" beacons and the "X" beacons that were positioned at opposite sides of the street.

1) Laboratory Setup for Tracking and Localisation: A robot in the laboratory environment can obtain its own position using a combination of visual perception, an internal world model and a set of designated landmarks. The underlying methods

include a variety of image processing tools for colour classification, edge detection, blob formation and landmark recognition, before using a localisation system to autonomously calculate the current position. The laboratory offers the advantage of running experiments with a second much more precise tracking system (within 2*cm*), using two overhead cameras that are installed on the laboratory ceiling.

a) *Tracking System*: To provide the experiments with precise ground truth data, the laboratory was equipped with an overhead tracking system capable of tracking multiple moving robots with high accuracy. The system consists of three entities: overhead cameras, an image processing server, and the robotic clients. Two Basler A601fc IEEE 1394 cameras (Basler AG, 2010), each with wide angle lens, were mounted 3 metres above the model street to observe the movement of robots during the experiment. These cameras produced images at a rate of 60 frames per second combined with a resolution of 640 by 480 pixels per camera. The recorded images were sent to the image processing server. The server used the SSL Vision software (Zickler, Laue, Birbach, Wongphati, & Veloso, 2010) to process the overhead images at 30 frames per second, in order to obtain the position of robots on the model street. Blob patterns (as prescribed by SSL Vision) were attached on the robot's head for overhead tracking (see Figure 2). Resultant positions were broadcast over the laboratory's wireless communication network to be recorded by the robots for further processing.

b) *Localisation System:* Each robot employed its own internal localisation system to autonomously determine and track its position on its internal world model. A Kalman filter was used to estimate and correct its position; this algorithm used the very noisy odometry (walking motions) and visual landmark information such as relative distances and angles to a beacon as input. From these inputs it filtered the information provided to obtain its current position on the internal world map. This position is described using three dimensions: an (x, y) position, as well as an orientation (θ). This was essential for the autonomous robot, as localisation enabled dynamic independent decision selection based on its current position.



(a) Robot with pattern (b) Blob pattern

Figure 2: Images regarding SSL software system (Zickler, Laue, Birbach, Wongphati, & Veloso, 2010)

2) Robotic Pedestrian Behaviour: The behaviour module is a program on the robot that selects the most appropriate actions or tasks it should complete based on its location and sensory information at a given point in time. In this series of experiments the behaviour selected was a model of a human pedestrian walking on a model street. The street was modelled between four beacons; two “O” and two “X” as illustrated in Figure 1. The robot was instructed to walk autonomously along the model street, between the beacons. Once the robot had reached a beacon, it was programed to turn around towards the opposing beacon and continue its walk. This process of repeatedly walking up and down the model street was sustained until the operator instructed the robot to stop.

Simultaneously the robot was instructed to continuously search for objects in the manner described in the implementation of gaze dynamics. Such objects included beacons to assist in self-localisation and attractive objects represented by orange spheres, which served as distractions on the model street. Upon detection of attractive objects, the robot was programed to not only trigger the subtle gaze differences, but also reduce its walking speed. The robots walk path is governed by a set of positions on the street known as waypoints, supplied from the robotic pedestrian behaviour. In order to move to a certain waypoint, the behaviour module was required to know its current position and the position of the given waypoint. The behaviour module commands the robot to reach those specific waypoints one after another, repeatedly.

3) Setup for Gaze Analysis Using an External Camera (without overhead tracking as described in Gaze Vector Estimation from Video Images): To analyse the 3D orientation of the robot’s head (without overhead cameras), a Sony Handycam HDR-HC3 was placed at one end of the model street looking directly down the street facing the robots when walking towards the “O” beacons. The camera recorded images in 1080i at 30Hz. This enables both the head’s pitch (the up–down direction of the head) and the head’s yaw (the left–right direction of the head) to be observed. This final stage of experiments in the laboratory aims at addressing the challenging real-world situation where the gaze of pedestrians will be estimated from near-frontal video recordings.

B. Results of Visual Overhead Behaviour Analysis of the Walking Robot

Before vision data from the robot and overhead tracking system could be analysed, the raw data collected required significant pre-processing. Various moving average filters were applied to reduce noise associated with the motions of the walk. These motions include swaying, slipping, and falling. The filters also reduced noise associated with calibration and alignment of overhead cameras, which resulted from switching cameras as the robot crossed to the other half of the street. All window sizes used for these filters varied proportionally to the maximum speed of the robot.

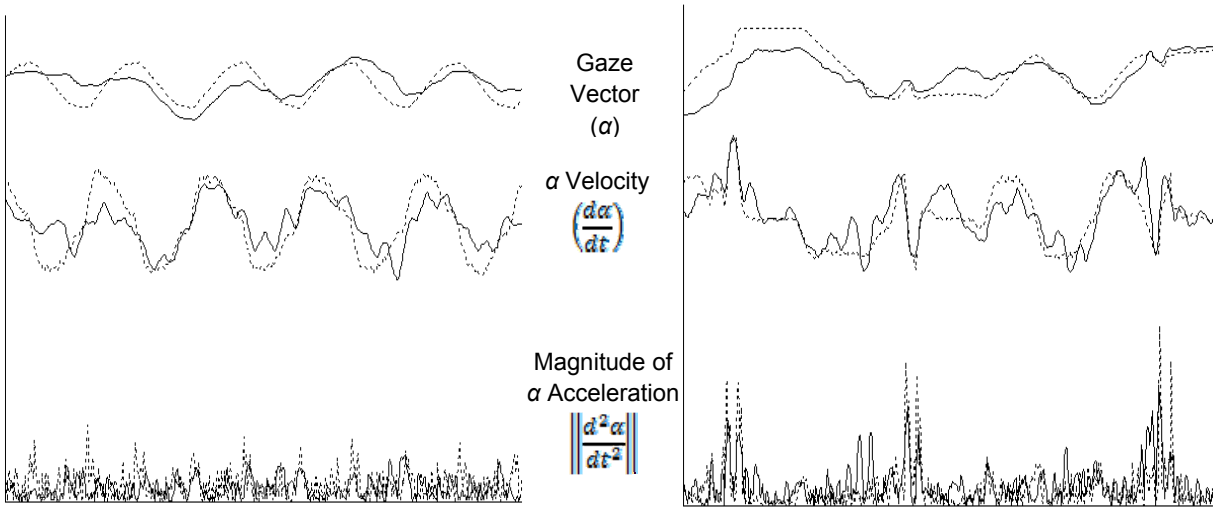
Forty traversals of the street were conducted with the inclusion of attractive objects (i.e. with distractions) and another thirty traversals were conducted where the robot

was walking under normal conditions (i.e. without distractions). The results of one series of experiments conducted are shown in Figure 3. While the robot walks under normal conditions (without distractions), the head of the robot periodically pans from left to right. This was reflected by the periodic curves of the internal gaze vector ($\alpha_{internal}$) represented by dashed lines at the top of Figure 3. It was recorded directly from the robot's neck yaw motor position sensor. This was followed closely by the external direction of gaze ($\alpha_{external}$) that was obtained from the overhead camera and is represented by the solid lines at the top of Figure 3.

In the scenario where robotic pedestrians detect attractive objects (Figure 3(b)), simultaneous changes in both $\alpha_{internal}$ and $\alpha_{external}$ signals can be observed. Whenever an attractive object is detected, the robot's head rapidly corrects its position so that the attractive object is at the centre of its field of view. As a result of this correction, a sudden change in α velocity ($d\alpha/dt$) can also be observed; this in turn produces a corresponding *'spike'* in the acceleration magnitude of α ($||d^2\alpha/dt^2||$). The robot focuses on the attractive object for a short period of time, before continuing its normal behaviour. Although under laboratory experiment conditions the robots subtle horizontal head acceleration was not directly recognisable by the eyes of a human observer, detection was possible using the described video analysis method.

The $\alpha_{external}$, its velocities and accelerations, were compared with the truth data obtained from within the robot ($\alpha_{internal}$), and its velocities and accelerations respectively. All traversals in the experiment obtained a positive correlation, i.e. the external gaze angles are associated with the internal truth values, with $d\alpha/dt$ obtaining the strongest correlation (M = (r = 0.77, p < 0.001), SD = 0.10). This was closely followed by α signals (M = (r = 0.76, p < 0.001), SD = 0.15). The $||d^2\alpha/dt^2||$ achieved a moderate correlation value (M = (r = 0.42, p < 0.001), SD = 0.10). From the correlation values, it is apparent that the $\alpha_{external}$, $d\alpha_{external}/dt$ and subsequently $||d^2\alpha_{external}/dt^2||$ are significantly affected by noise and as a result some loss of signal is unavoidable.

In Figure 3, $||d^2\alpha/dt^2||$ signals were plotted at the bottom as the dashed and solid lines, representing $\alpha_{internal}$ and $\alpha_{external}$ respectively. Although the internal and external $||d^2\alpha/dt^2||$ was moderately correlated, it is of particular significance to observe the strong reliability of these signals for the detection of visually attractive objects. The majority of data collected in the experiments with distractions showed distinct changes in $||d^2\alpha/dt^2||$ when the robots head engaged a visually attractive object.



(a) One sample without distractions (b) One sample with distractions

Figure 3: Sample of streetwalking experiment data collected from the robot (- - - - Internal), superimposed with post-processed data from the overhead camera (——— External). It is of particular significance to observe that in the scenario with distractions (b), spikes in the acceleration signals correspond to the robot distracted by an attractive object.

ROBOTIC PEDESTRIANS IN URBAN SPACE

When autonomous robots leave the laboratory and are placed in real world environments they cannot rely on information provided by overhead cameras and are required to be equipped with alternative tools. For the robotic pedestrians of the present project two questions are of central importance. First, how can a robot determine its current location and orientation (if it is not equipped with GPS)? Second, how can the robot detect and decide what are the visually attractive or "salient" objects in the environment?

For the first question, in order for a robotic pedestrian to calculate its current position in a real world environment, it would rely on its internal localisation system (see *Laboratory Environment and Setup of the Experiments 1 (b)*), since the use of an overhead camera would not be possible. To calculate this position, a robotic pedestrian would require a map of the environment, information about the location of several visually detectable landmarks and walking motions performed since last estimate. The position estimates are obtained through its movements about the map from the walking motions performed. To increase the estimate accuracy over time, this estimate is frequently corrected when visual features are recognised.

When moving to the outside environment the visual landmarks used within the laboratory may be replaced using uniquely colour coded cylindrical beacons. The advantages of these cylindrical beacons are both their ease of recognition and an invariant shape with respect to viewing position. These beacons must be placed in known positions around the map, ideally spread out so that they can be seen from a large range of positions. An example of a map designed for an outdoor environment

can be seen in Figure 4. So long as the robot moves within the mapped environment and within visual range of the markers, the visual feedback from the beacons will allow an accurate positional estimate.

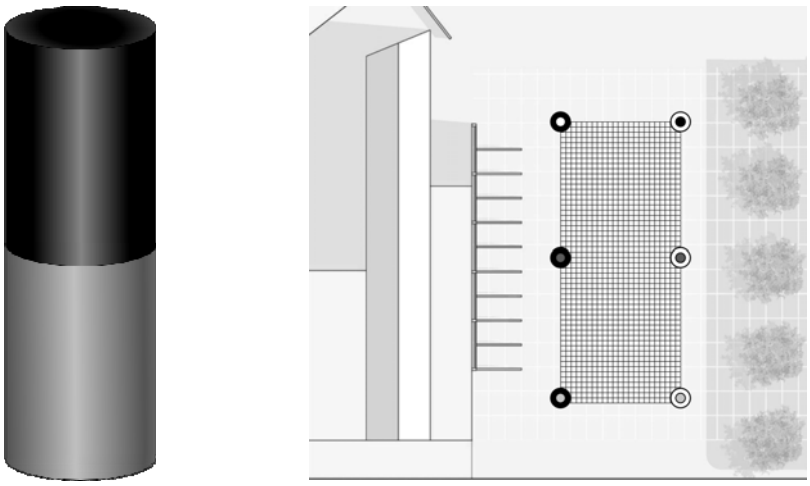


Figure 4– (Left) Example of a two-coloured beacon, this is the beacon seen on the top left position in the map. (Right) Example of uniquely identifiable visual beacons placed on a mapped area. The outer circle represents the lower colour, while the inner circle represents the upper colour.

For the second of these questions, in the laboratory experiment attractive objects or facades were labelled by orange spheres that robots could detect by their colour and shape. In a real environment robots would have to detect salient objects from the environment themselves. Computational attention provides mechanisms of extracting useful and interesting information. Commonly known as “Saliency Detection” or “Interest Point Detection”, computational attention helps to identify the regions of sensory input that stand out from their neighbourhood and attract the attention of the subject. Traditional approaches to saliency detection work on 2D images where small regions of the image which differ significantly from adjacent areas by their intrinsic properties like colour, intensity and orientation are identified. With the advent of portable 3D cameras, the availability of depth information allows a robot to distinguish objects by different shape and curvature analysis methods. Objects that exhibit an exceptionally high variation in their shape properties when compared to neighbouring structures in space can be regarded as salient. For further details of available saliency detection methods, the reader is referred to (Frintrop, Rome, & Christensen, 2010).

For gaze vector estimation and trajectory analysis in a real world situation, alternative techniques to those that use an overhead camera are required. The following two subsections describe an approach for estimation of the gaze direction and pedestrian behaviour trajectory analysis using video recordings from a frontal perspective.

Gaze Vector Estimation from Video Frontal Images

Gaze estimation is a widely studied area (Balasubramanian, Ye, & Panchanathan, 2007; Ono, Okabe, & Sato, 2006; Ren, Rahman, Kehtarnavaz, & Estevez, 2010),

where numerous different techniques have been trialled. These include template matching, detector arrays, regression, and support vector machines. In a survey of gaze estimation algorithms and techniques (Murphy-Chutorian & Trivedi, 2009), it was reported that manifold learning methods obtained the most accurate results for gaze estimation, where biased manifold embedding (Balasubramanian, Ye, & Panchanathan, 2007) achieved the best result with an mean absolute error of 1.44^o in the yaw axis.

In (Wong, Chalup, Bhatia, Jalian, Kulk, & Ostwald, 2011) we presented results that used biased isomap (Balasubramanian, Ye, & Panchanathan, 2007) along with out-of-sample extensions (de Silva & Tenenbaum, 2003) to obtain the gaze vector of a single image. This involved calculating a 2-dimensional non-linear manifold (Figure 7) of robot head pose images (Figure 5) that were recorded during walking experiments (Figure 6). This dataset consisted of 518 images of a real robot head looking in different directions; ±90^o in the yaw direction and -35^o to 30^o in the pitch direction, neighbouring images for both directions were 5^o apart. An out-of-sample extension algorithm was used in Figure 8 to embed an unseen robot head sample (marked by 'x') as a new point at the correct location relative to the other images into the manifold. The gaze direction associated with the newly embedded sample was then estimated from the neighbouring points on the manifold.

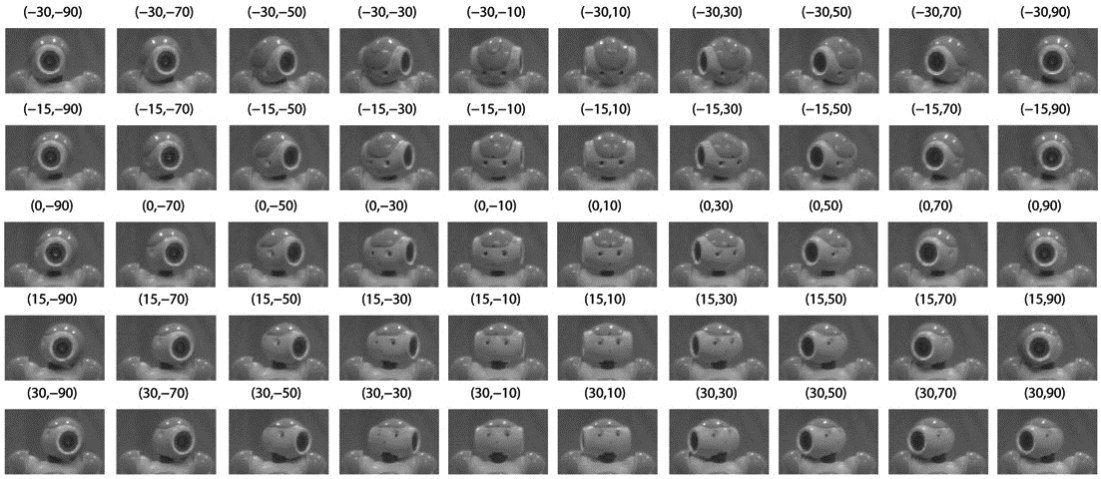


Figure 5: Selected images from the robot head pose dataset.

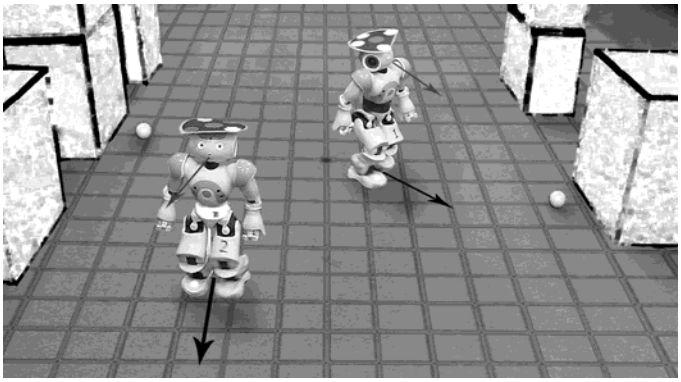


Figure 6: Robots walking down the model street. Gaze direction and movement direction are indicated by arrows.

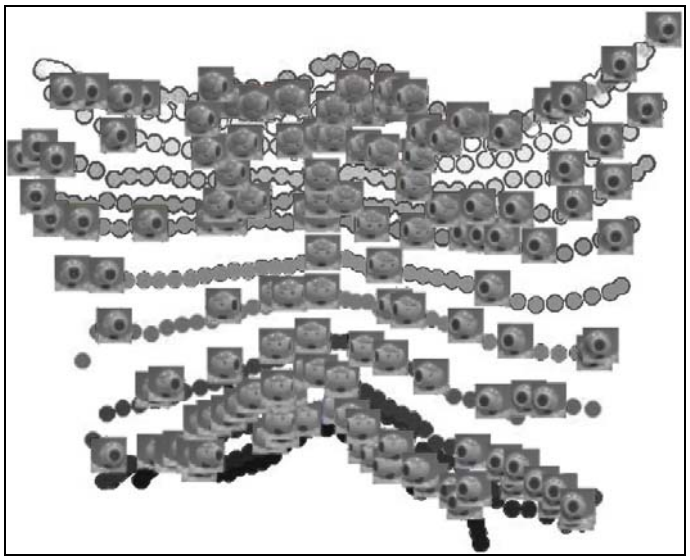


Figure 7: Gaze manifold obtained through Biased Isomap

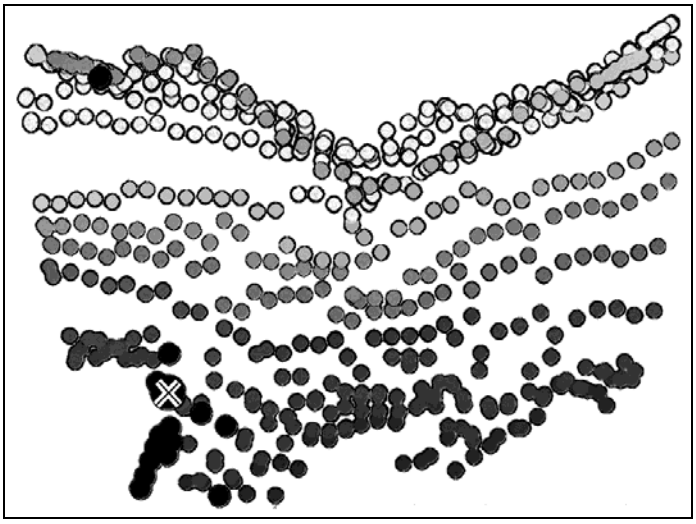


Figure 8: Out-of-sample extension: The point marked with an "X" represents an unseen sample embedded into the gaze manifold from Figure 7. The smaller black points represent the close points obtained through out-of-sample extension, with the "X" marking the closest point.

The resulting manifold shown in Figure 7 is a 2-dimensional surface patch. It appears to be slightly distorted because it is calculated only from a small subset of all possible head rotations. In order to obtain a manifold with a topology that more accurately represents the head pose dynamics we used the simulated head shown in Figure 9 for generating a dataset that comprises 5402 images of the simulated head as it rotates in 3D through full revolutions about the yaw and pitch axis in steps of 5 degrees between neighbouring images. Application of Isomap (de Silva & Tenenbaum, 2003), revealed a torus-like manifold that represents the cross product S^1 x S^1 of the two full rotations (Figure 10). The head roll axis parameter was left constant at zero in our experiments because it has no influence on the direction of the gaze vector.

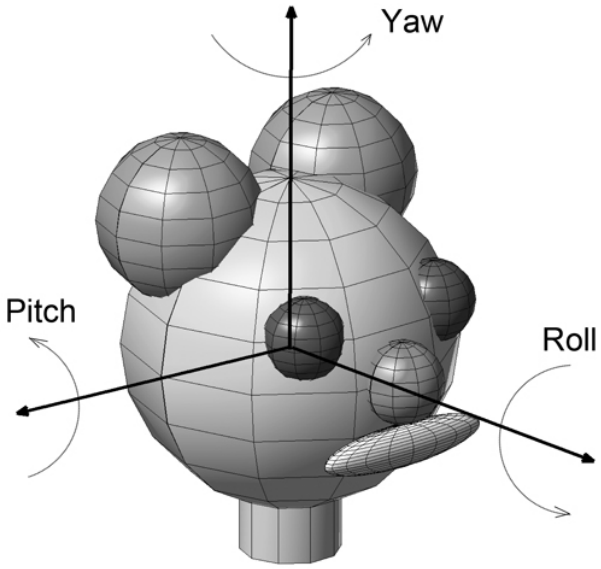


Figure 9: A simulated head was rotated about the pitch and yaw axis to form a dataset of 5402 head images.

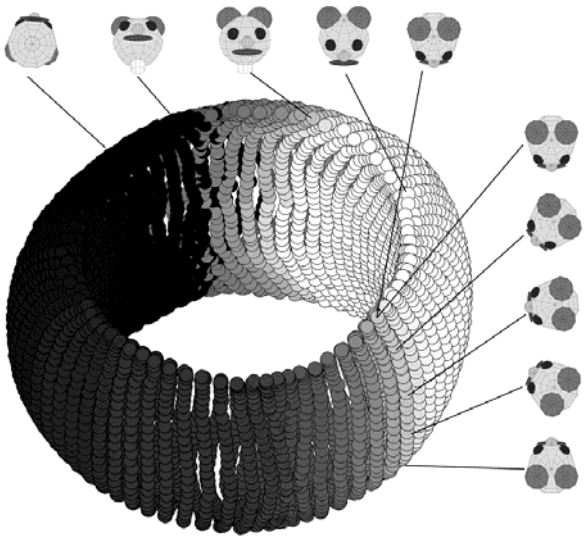


Figure 10: The torus manifold was obtained by Isomap using the complete simulated head dataset. Each point represents an image. The displayed example images illustrate rotations about the yaw and pitch axis.

Each gaze image has a corresponding point on the torus that represents the gaze direction of the head shown in the image. Short video sequences of head movements correspond to trajectories or paths on the torus surface. Similarly the robot head pose dataset can be seen as a 2-dimensional sub-manifold that is embedded in the torus manifold.

Manifold alignment (Zhai, Li, Chang, Shan, Chen, & Gao, 2010) was used to align the robot head dataset (Figure 7) with the simulated head dataset (Figure 10). Manifold alignment techniques aim to learn a mapping between different manifolds, to unite local systems to form a global coordinate system. In our case, the images of the robot head dataset were aligned to obtain correspondences with corresponding samples of

the simulated head dataset forming a new common co-ordinate system (Figure 11). Through the process of alignment, the robot heads obtained labels that were derived from the neighbouring points of the simulated head dataset. The error of alignment could be calculated by comparing the newly obtained labels with the truth data of the robot head dataset. In Figure 11, the correspondence labels achieved a root mean squared error (RMSE) of 3.37^{o} pitch, while the yaw direction obtained an RMSE of 70.65^{o} after outlier removal. These results show that the pitch direction had a much lower error rate compared to the yaw direction; this is a result of using unbalanced data. That is, the number of samples that represent each pitch angle is much greater than the number of samples that represents each yaw angle. The RMSE of pitch and yaw over the total manifold is 0.93% and 19.6%, respectively.

The torus manifold consists of a larger number of samples covering the complete discrete range of full rotations (±180^{o} for both pitch and yaw axis) compared to the subset of samples in the discrete range of -35^{o} to 30^{o} pitch, ±90^{o} yaw for the robot manifold; the geodesic distances on the torus (Figure 10) more accurately represent the head pose dynamics than the distances on the surface patch (Figure 8). In the combined manifold that results of the alignment of the two datasets (Figure 11) the good properties of the torus manifold are induced to the robot dataset. Each sample of the robot dataset has now a higher density of neighbours from both datasets and the distances between them are more accurate. Therefore the labels of out-of-sample extensions can be estimated more precisely.

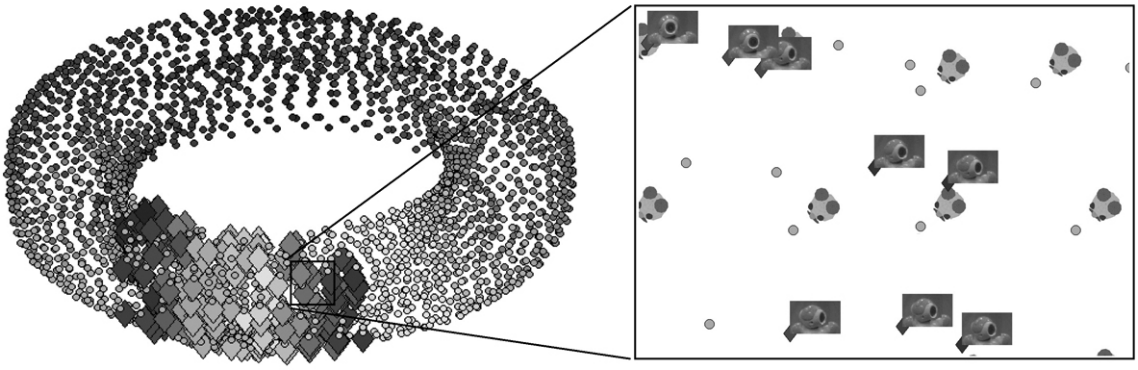


Figure 11: Robot head dataset (diamond points) was aligned with the simulated head dataset (circular points). The magnified section of the torus shows diamonds and points that are labelled with their corresponding real robot and simulated head images, respectively. These heads images show that a robot head has an approximate neighbour close by from the simulated head dataset which has the simular gaze direction.

Video sequences of pedestrian head images can be aligned with the torus manifold to estimate the gaze vector for each image in the sequence. The alignment method can align sets of images of heads recorded using different robotic or human pedestrians. Future experiments will apply the described approach to real world pedestrian gaze vector estimation using videos of pedestrians walking through an urban scene.

Temporal Behaviour Analysis using Outlier Detection

For analysing temporal gaze and movement dynamics, a system is used that was previously employed for analysing simulated agents that move through computer models of urban environments (Jalalian, Chalup, & Ostwald, 2011). This analysis system utilises one-class support vector machines (SVMs) (Schölkopf, Platt, Shawe-Taylor, & Williamson, 2001; Vert & Vert, 2006) to identify when and where in the considered area an agent shows abnormal changes in its gaze direction and movement trajectory. The approach is based on statistical learning, which can cope well with noisy signals and can generalise to unknown inputs (Vapnik, 1998).

First the system learns a statistical model characterising normal behaviour, based on sample observations of regular agent movement without the impact of significant visual attractions in the environment. Irregular behavioural characteristics of the robot caused by spotting of visually attractive objects can then be detected by the system as outliers.

DISCUSSION AND CONCLUSION

The problem of pedestrian simulation for complex urban and architectural environments cannot be solved using the current available software and conceptual models, almost all of which are driven by the expectation of strategic or expedient reactions. Such models may be effective for simulating people evacuating a building in emergency conditions, or leaving a sports arena after an event, but they are unable to predict the way people will react under less-directed conditions. (Whyte, 1980) famously noted that without a clear and immediate strategic goal, (for example to be at a bus stop in time for the morning commute to work), pedestrians will be drawn to interact with any number of urban attractors including seating, coffee shops, newspaper stands and fountains (van Schaick & van der Spek, 2008). These interactions are both spontaneous and opportunistic and without developing a sense of how they are first seen while walking, it is impossible to predict how people will behave. Video analysis of pedestrians using urban environments has some potential to solve this problem, and a purely statistical model of pedestrian dynamics in a single space may be constructed, but without a more detailed understanding of the factors shaping such a model, it cannot be extrapolated to other urban spaces. The key factor that must be solved in order to progress research in this field is the relationship between human gaze and human movement. The present paper uses robots to offer a novel approach to beginning to solve this problem.

Robots are ideal for this research because they allow for “actual” pedestrian data (parameter readings taken directly from their control systems) to be compared with the data collected through video recordings of their actions. This allows for a calibration process of the video analysis system to be undertaken in a relatively controlled environment.

Experiments in the laboratory environment allow robot movement and gaze tracking through overhead cameras. This set-up still features the “real world” challenges of

lighting (including shadows) and surface textures but these are at least consistent in the laboratory. However, despite all of these beneficial conditions, the problem of gaze detection remains a complex one and the experiments in the laboratory demonstrated the impact that high levels of visual “noise” had on the results. Furthermore, there were some specific problems with unexpected robot behaviour; for example, sometimes a robot lost orientation and deviated from its expected walking direction. Another methodological challenge was that the robot’s walk is omnidirectional; the body normal does not always coincide with the walk direction. Therefore $\alpha_{internal}$ and $\alpha_{external}$ typically do not represent exactly the same angle, although α velocities show similar behaviour. However, despite these challenges, it was evident that the visual impact of urban attractors could be detected, as represented in the spikes in the chart for the magnitude of acceleration.

For gaze estimation outside the laboratory pilot experiments were conducted where frontal video recordings were evaluated using the technique of manifold alignment. The global structure of the set of real head/gaze images was strengthened through alignment with an artificially generated torus-shaped head rotation dataset. The non-linear torus structure that underlies the head rotation data would not be obtainable from real recordings alone. It is our currently best alternative to overhead tracking when outside the laboratory. As robot heads and human heads can be processed alike by this approach the system after being calibrated using robots can be employed for human gaze estimation from frontal video recordings.

With an estimate of the critical gaze parameter available pedestrian dynamic behaviour analysis can be conducted in existing or computer simulations of planned urban environments. The study of (Jalalian, Chalup, & Ostwald, 2011) describes a system that was designed to detect visually attractive objects or sightlines in urban environments through analysis of pedestrian behavioural data (that includes the gaze parameter).

ACKNOWLEDGMENT

This project was supported by ARC DP1092679 “Modelling and predicting patterns of pedestrian movement: using robotics and machine learning to improve the design of urban space”. The authors are grateful all members of the Newcastle Robotics Laboratory who contributed to the NUbot software system that is part of RoboCup research. Overview of author contributions to the paper: Corresponding author, robot gaze vector detection and data analysis (ASWW); paper concept and project supervision (SKC); overhead tracking and saliency detection (SB); pedestrian simulations and data analysis (AJ); robotic motor control of walk and head dynamics (JK); localisation (SN); architectural and urban context, discussion and co-supervision of research (MJO).

REFERENCES

Aschwanden, G., Haegler, S., Halatsch, J., Jeker, R., Schmitt, G., & Gool, L. V. (2009). Evaluation of 3D City Models Using Automatic Placed Urban Agents. *9th International Conference on Construction Applications of Virtual Reality* (pp. 165-176). Sydney: University Of Sydney.

Auchincloss, A. H., & Diez Roux, A. V. (2008). A New Tool for Epidemiology: The Usefulness of Dynamic-Agent Models in Understanding Place Effects on Health. *American Journal of Epidemiology, 168*(1), 9-12.

Balasubramanian, V. N., Ye, J., & Panchanathan, S. (2007). Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on* (pp. 1-7). Minneapolis: IEEE.

Bandini, S., Manzoni, S., & Vizzari, G. (2009). Crowd Behavior Modeling: From Cellular Automata to Multi-Agent Systems. In D. Weyns, & A. M. Uhrmacher (Eds.), *Multi-Agent Systems: Simulation and Applications.* (pp. 301-324). Boca Raton: CRC Press.

Bandini, S., Manzoni, S., & Vizzari, G. (2009). Modeling, Simulating, and Visualizing Crowd Dynamics with Computational Tools Based on Situated Cellular Agents. In H. Timmermans (Ed.), *Pedestrian behavior: Models, Data collection, and Applications* (pp. 45-62). Bingley: Emerald Group Publishing Ltd.

Basler AG. (2010, November 30). *Basler A600 Specifications*. Retrieved November 30, 2010, from Basler Vision Technologies: http://www.baslerweb.com/beitraege/unterbeitrag_en_23042.html

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Clifford, S. (2009). *Introduction to algorithms, Third Edition.* Cambridge: MIT Press.

de Silva, V., & Tenenbaum, J. B. (2003). Global Versus Local Methods in Nonlinear Dimensionality Reduction. (S. T. S. Becker, & K. Obermayer, Eds.) *Advances in Neural Information Processing Systems, 15*, 721-728.

Egan, C., Willis, A., & Wincenciak, J. (2009). The effects of a distractor on the visual gaze behavior of children at signalized road crossings. *Journal of Vision, 9*(8), 371.

Frank, L. D., Engelke, P. O., & Schmidt, T. L. (2003). *Health and Community Design: The Impact of the Built Environment on Physical Activity.* Washington DC: Island Press.

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational Visual Attention Systems and Their Cognitive Foundations. *ACM Transactions on Applied Perception, 7*(1), 1–39.

Gao, Y., & Gu, N. (2009). Complexity, human agents, and architectural design: A computational framework. *Design Principles and Practices: An International Journal, 3*(6), 115-126.

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., et al. (2009). Mechatronic design of NAO humanoid. *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on* (pp. 769-774). Kobe: IEEE.

Jalalian, A., Chalup, S. K., & Ostwald, M. J. (2011). Architectural evaluation of simulated pedestrian spatial behaviour. *Architectural Science Review, 54*(2), 132-140.

Magnenat-Thalmann, N., & Thalmann, D. (2005). Virtual humans: Thirty years of research, what next? *The Visual Computer, 21*, 1-19.

Magnenat-Thalmann, N., Jain, L. C., & Ichalkaranje, N. (Eds.). (2008). *New Advances in Virtual Humans: Artificial Intelligence Environment. Studies in Computational Intelligence* (Vol. 140). Berlin: Springer.

Maim, J., Haegler, S., Yersin, B., Mueller, P., Thalmann, D., & Gool, L. V. (2007). Populating ancient pompeii with crowds of virtual romans. *The 8th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST'07)* (pp. 26-30). Brighton: Eurographics.

Molnár, P., & Starke, J. (2001). Control of distributed autonomous robotic systems using principles of pattern formation in nature and pedestrian behavior. *IEEE Transactions on Systems, Man, and Cybernetics, 31*(3), 433–435.

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head Pose Estimation in Computer Vision: A Survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31*(4), 607-626.

Musse, S. R., Kallmann, M., & Thalmann, D. (1999). Level of autonomy for virtual human agents. *Proceedings of 5th European Conference on Advances in Artificial Life, ECAL '99* (pp. 345–349). Lausanne: Springer.

Ono, Y., Okabe, T., & Sato, Y. (2006). Gaze Estimation from Low Resolution Images. *Pacific-Rim Symposium on Image and Video Technology, PSIVT'2006* (pp. 178-188). Hsinchu: Springer.

Ren, J., Rahman, M., Kehtarnavaz, N., & Estevez, L. (2010). Real-time head pose estimation on mobile platforms. *Journal of Systemics, Cybernetics and Informatics, 8*(3), 56-62.

Saarloos, D., Kim, J.-E., & Timmermans, H. (2009). The Built Environment and Health: Introducing Individual Space-Time Behavior. *International Journal of Environmental Research and Public Health, 6*(6), 1724-1743.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J. S., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation, 13*(7), 1443-1471.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science, 290*(5500), 2319-2323.

Valenti, R., Staiano, J., Sebe, N., & Gevers, T. (2009). Webcam-based visual gaze estimation. *Image Analysis and Processing, 15th International Conference* (pp. 662-671). Vietri sul Mare: Springer.

van Schaick, J., & van der Spek, S. C. (Eds.). (2008). *Urbanism on Track: Application of Tracking Technologies in Urbanism.* Amsterdam: IOS Press.

Vapnik, V. N. (1998). *Statistical Learning Theory.* New York: J. Wiley.

Vert, R., & Vert, J.-P. (2006). Algorithms, Consistency and Convergence Rates of One-Class SVMs and Related. *Journal of Machine Learning Research, 7*(6), 817–854.

Whyte, W. H. (1980). *The Social Life of Small Urban Spaces.* New York: Project for Public Spaces.

Wong, A. S., Chalup, S. K., Bhatia, S., Jalian, A., Kulk, J., & Ostwald, M. J. (2011). Humanoid Robotics for Modelling and Analysing Visual Gaze Dynamics of Pedestrians Moving in Urban Space. In R. Hyde, D. Hayman, & D. Canbrera (Ed.), *From principles to practice in architectural science. Anzasca 2011, 45th Annual Conference of the Australian and New Zealand Architectural Science Association.* Sydney: University Of Sydney.

Zhai, D., Li, B., Chang, H., Shan, S., Chen, X., & Gao, W. (2010). Manifold Alignment via Corresponding Projections. In F. Labrosse, R. Zwiggelaar, & Y. &. Liu (Ed.), *Proceedings of the British Machine Vision Conference* (pp. 3.1-3.11). Aberystwyth: BMVA Press.

Zickler, S., Laue, T., Birbach, O., Wongphati, M., & Veloso, M. (2010). SSL-vision: The shared vision system for the robocup small size league. (J. Baltes, M. G. Lagoudakis, T. Naruse, & S. S. Ghidary, Eds.) *RoboCup 2009: Robot Soccer World Cup XIII. Lecture Notes in Artificial Intelligence (LNAI 5949)*, 425–436.