Bottom-Up Visual Attention for Virtual Human Animation

Christopher Peters Carol O' Sullivan

Image Synthesis Group, Department of Computer Science, Trinity College Dublin, Ireland.
Email: {christopher.peters, carol.osullivan}@cs.tcd.ie

Abstract

We present a system for the automatic generation of bottom-up visual attention behaviours in virtual humans. Bottom-up attention refers to the way in which the environment solicits one's attention without regard to task-level goals. Our framework is based on the interactions of multiple components: a synthetic vision system for perceiving the virtual world, a model of bottom-up attention for early visual processing of perceived stimuli, a memory system for the storage of previously sensed data and a gaze controller for the generation of resultant behaviours. Our aim is to provide a feeling of presence in inhabited virtual environments by endowing agents with the ability to pay attention to their surroundings.

1. Introduction

Graphics programming techniques are providing ever more sophisticated representations and animations of virtual humans. Despite this, a viewer is often left with the feeling of dealing with an empty shell when viewing such agents. Something is lacking; agents do not appear to pay attention to their environment, yet they still appear to be able to sense the scene as if by telepathy. Ideally, we would like to be given the impression that they are interacting with their environment in the same way as living, thinking creatures. Although as humans we cannot directly see the thoughts of others, we do obtain an impression of thought-processes at work through the way in which other people behave. We can often tell what people are thinking by observing what they attend to as they perform their everyday activities. In this way, if the agents in our virtual environments do not appear to pay attention to anything in the scene, then our sense of presence may be severely degraded. As noted by Blumberg et al. [5], "If we provide characters with the means to express their mental state, an observer can infer their beliefs and desires". Attention and gaze are important factors in achieving this goal.

We present a system for the automatic generation of bottom-up visual attention behaviours for virtual humans. Bottom-up attention is behaviourally significant, since it constitutes a powerful alerting mechanism that allows primates to instantly become aware of unexpected predators or dangers [10]. Our attention framework is composed of four primary interacting components; a visual sensing component, an attention component, a gaze generation component and a memory component. The sensing component allows the agent to perceive the scene. Three renderings of the scene are taken. These renderings loosely approximate the scene as seen through the eyes of the agent. They allow us to conduct visibility queries for objects and also act as inputs to the visual attention and memory components. The visual attention component, taken from the field of cognitive engineering, processes the visible portion of the scene and generates a map of the most salient locations based on the features in the scene. These locations are passed to the gaze controller, which generates the actual motion. The orienting motions consist of a mixture of eye and body movements. The memory component is used by all three of the other components. It acts as a filter for storing details of important objects and also for remembering what objects have already been looked at and to what extent. In this way, the object of an agent's attention depends on both internal and external factors.

Section 2 reviews background to our research, focusing on internal models for perception and processing in order to augment movement. We describe our method for enabling the visual perception of the environment in Section 3. Section 4 covers a key component of the system: a computational model for bottom-up attention. Section 5 considers the memory component and its interactions with the other components. The gaze generator is discussed in Section 6, while Section 7 concludes with a discussion of future work and results obtained from the model.

2. Background

A number of authors have explored the use of internal perception and processing models for the animation of virtual characters. Reynolds [20] presented a behavioural model for flocks of birds and herds of animals. Based on insights from the real world, his agents, *boids*, do not have access to complete or perfect information about the world. Each boid has a spherical zone of sensitivity centred at its local origin, which means that the boid only senses nearby flock-mates. This approach approximates a sensory system while attempting to make the same final information available to the boid as its real-life counterpart would have. Tu and Terzopoulos [21] animate fishes based on a combination of their perception of the environment and internal variables such as mental state and habit. Each fish has vision and temperature sensors allowing behaviour patterns to be interrupted by reactions to environmental stimuli, such as cold water or nearby predators. The vision sensor provides a 300-degree view; any object that is not occluded and has entered this view volume is considered visible.

Researchers have also applied internal sensory and processing models to the animation of human characters. Renault et al. [19] and Noser et al. [15] provide a synthetic vision system allowing actors to navigate through a corridor using vision, learning and memory. The navigation problem is decomposed into solving global and local navigation problems. The local navigation algorithm uses direct input from the environment to avoid unexpected obstacles and to satisfy higher-level goals from the global navigation system. The local algorithm does not have access to the scene database: instead an internal model of the scene is constructed in an octree data structure based on the data acquired from the visual system as the actor perceives the scene. Kuffner and Latombe [13] extend this work by storing sensory information in the form of observations and allowing characters to learn about their environment using memories of such observations. This system is used for the navigation of characters through maze-like environments.

There has also been research in terms of using internal models to animate attentive behaviours for human characters. Chopra and Badler [6] propose a visual attention framework as part of wider research into controlling virtual human animation based on movement observation and cognitive modelling [3]. A number of different types of looking behaviours are implemented and the final behaviour is dependent on both top-down, volitional attention application and bottom-up, involuntary attention capture. Gillies [7] produces behavioural animation through the simulation of vision and attention. Object features are regarded as the key components in deciding what object will be attended to. Object features represent abstract and complex reasons as to why an object might be looked at, artistic appreciation, for example. In this way only top-down attention processes are considered as contributing to looking behaviours. Actors are endowed with interests and pay more attention to some properties than to others. This results in actors responding in different ways to an object.

3. Sensing

In order for virtual humans to pay attention to their environment, they must first have a means of sensing it. There are a number of approaches for providing agents with information about the environment. The most straightforward is to allow them full access to the scene database. The main limitations of this approach are realism and scalability. In terms of realism, full access to the database is the equivalent of an all-seeing being. If an agent's behaviour is partly defined by their senses and they are allowed full access to the scene database, then the resulting behaviour will appear to be implausible. In terms of scalability, the character's processing system could have a huge number of possible inputs to deal with. This could be especially true for large scenes, where the number of objects would inevitably be overwhelming.

In order to solve these problems, usually some form of filtering technique is employed to reduce the number of objects that the agent must consider [4]. The type of filtering technique used helps define the plausibility of the sensing system: for example, in the human being, the primary information flow direction is to the front. A filtering technique that prioritises information flow in this way would better approximate the system. It is also important to consider the balance between the realism with which the sensing takes place and the costs associated with the process. Our approach is to provide a balance between sensory honesty [5] and speed. In this way we ensure that the information obtained by the agent is plausible, but also compensate for certain cognitive abilities by allowing direct queries on the scene database. An example of this is our implementation of motion in the periphery: we do not use image-based motion estimation techniques, but rather query velocities directly from database objects. Our method is based on previous work by Noser et al. [15] and is monocular in nature. This synthetic vision approach uses renderings from the perspective of the agent to provide visibility information to that agent. The approach is effective in that it provides occlusion of the objects in the scene, while using dedicated graphics hardware to do the processing. Also, it is not necessary to process the scene at full detail; lighting and textures are usually disabled.

In our system, at each update of the visual system, akin to a perceptual snapshot, three renderings of the scene are taken from the viewpoint of the virtual human; a 256x256 full-scene rendering, a 128x128 distinct-mode rendering and a 128x128 grouped-mode rendering (see Figure 3). The 128x128 resolution distinct-mode rendering and 128x128 grouped-mode renderings approximate the acuity of the eye (see [17] for more details on these vision modes).

The grouped-mode rendering approximates the acuity of the eye in the periphery, where only gross object characteristics are perceived. In this mode, object colours are coded according to their group membership. The distinct-mode rendering is generated at a higher resolution over a smaller field-of-view than the peripheral rendering. In this way it coarsely approximates the area of the scene in the fovea. It ensures that elements of the scene missed by the peripheral rendering, perhaps due to size, may still be sensed. Objects and groups in the scene are assigned unique false-colours and are rendered with these. The renderings are then scanned to provide visibility information about objects and groups in the agent's field-of-view. The 256x256 full scene rendering consists of the fully rendered scene from the perspective of the virtual human, with all effects such as lighting and texturing enabled. This full rendering is used to provide information on what parts of the scene are likely to solicit attention. In order to do this, the rendering is passed on to the next phase in the process: the attention model.

4. Attention

The human information processing system is a limited resource system [1]. In terms of information, our environment contains a great deal more than we could possibly hope to process. Nonetheless, we seem to manage despite the limits of our cognitive resources. A key factor in helping us to do this is our ability to deploy our senses to likely areas of important activity. The notion of an attentive system would appear to be obvious; "…everyone knows what attention is" said famous 19th century psychologist and philosopher William James. Yet the concept of attention is probably somewhat more illusive, as is highlighted by the more cautious approach of more recent research… "there may not even be an 'it' there to be known about" [16].

From our point of view, attention is a vitally important concept if we decide to restrict the sensory information of our agents. Agents that either do not pay attention to their surroundings or pay attention to the wrong things at the wrong times will look mechanistic and clumsy. Worse still, they run the risk of not being able to meet with the demands that their environment may produce. Consider what would happen to the agent who didn't pay attention to the vehicles on the road when crossing it! Under such circumstances, random attending behaviour will not suffice; motivation must be considered.

When considering attention, visual attention would appear to be a good place to start; the visual modality plays a key role in human sensing [18]. In designing such a system, it is useful to look to the fields of psychology and cognitive engineering, where attention has been the focus of a great deal of research. The two-component framework theory of attention appears to be a good starting point for such investigations. According to this theory, attention can be divided into top-down and bottom-up components. Top-down, voluntary attention occurs when we have a task in mind and direct our attention to focus our resources on accomplishing that task. In contrast, when considering bottom-up, or involuntary attention, locations in our environment appear to grab or demand our attention. In this way, our attention is drawn to salient parts of our view. These are normally viewed as external events, flickering lights or peripheral movement for example. Human attentive behaviour appears to be composed of a complex mixture of both bottom-up and top-down attention processes [10].

In this paper, we consider a model of bottom-up visual attention. Our method is suitable for modelling attention when there is no task at hand (e.g. spontaneous looking), or for interrupting task-level attention with potentially important events. The bottom-up model of attention that we use has been provided by Itti, Koch and Niebur [8, 9, 11] and has been shown to be effective with both natural [10] and rendered [23] scenes. Based on a biologically plausible architecture proposed by Koch and Ullman [12] and by Nieber and Koch [14], it attempts to mimic the low-level, automatic mechanisms responsible for attracting our attention to the salient locations in our environment and closely follows the neuronal architecture of the earliest hierarchical levels of visual processing.

The attention model itself processes an input image, calculating local contrast for intensity, orientation and colour features respectively. These feature types are computed in a centre-surround fashion, providing a biologically plausible system that is sensitive to local spatial contrast rather than amplitude in a given feature. An input image is decomposed into four constituent channels, one for intensity, one for orientation and two for colour. Each channel is used as the first level in constructing a dyadic image pyramid, which is a set of images where each successive image is a filtered and decimated version of its predecessor. For the intensity and orientation channels, a Gaussian filter is applied. The orientation channel is filtered with Gabor filters of angles 0, 45, 90 and 135 degrees. Feature maps representing centre-surround differences are then obtained from the filtered images. Centre-surround processing is a relative measurement that calculates how much a part of an image *pops out* from its neighbouring area. This could be a matter of a black dot on a white background, or a single diagonal line surrounded by vertical ones. The feature maps for each feature (intensity, colour and orientation) are then combined respectively into three conspicuity maps. Each conspicuity map provides a measurement of scene areas that pop out for that feature type. Combining the three conspicuity maps produces a unified

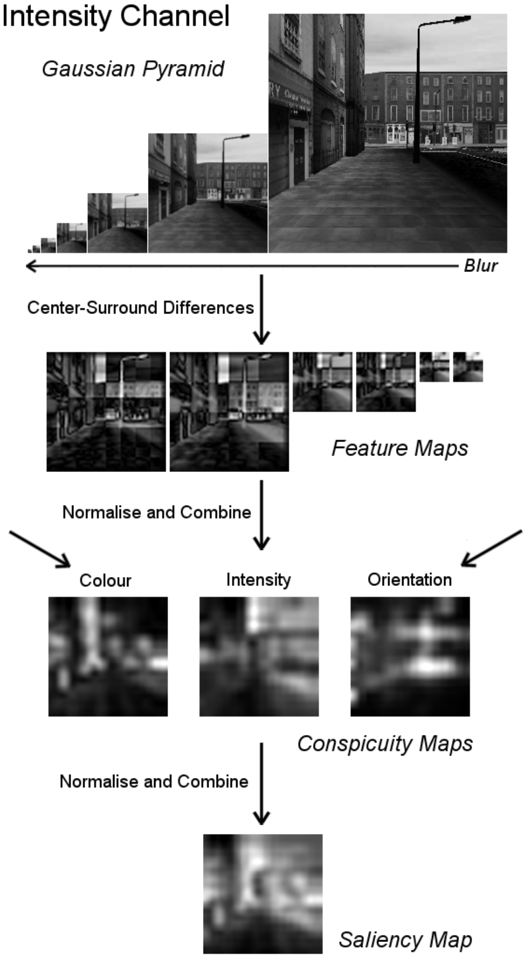

Figure 1. A general schematic of the model of visual attention.

measurement of pronounced parts of the entire scene. This result, a 16x16 saliency map, is the primary output of the attention model (see Figure 1). Combination strategies across feature modalities are of particular significance to the model. We will not discuss these here, apart from mentioning that we use a global non-linear normalisation operator; this provides higher weightings for those maps containing fewer, thus more pronounced, areas of saliency. The interested reader is referred to [9] and [10] for more details.

The final stage in the bottom-up attention model is a winner-take-all (WTA) network of simple integrate-and-fire neurons. The WTA network finds the maximum of the saliency map at any time, which corresponds to the current most salient location in the scene. Inhibition of return (IOR) may be implemented for spatial locations by reducing the chances of success of previous winners. This ensures that the focus of attention visits numerous parts of the scene and does not remain fixed on the single most pronounced location (see [10] for details). In our implementation, there are two options for IOR. The image-based option inhibits locations in the scene. While this is useful for static scenes, there are problems when the viewer or the scene moves, since old IOR locations may be invalidated. To handle this, we use an object based IOR mechanism. Every object in the scene is provided with an uncertainty level, which is a measure of the completeness of an agent's mental representation of the object. A high uncertainty level indicates that the object has not been attended to before, while a low uncertainty level signifies that the agent has a relatively complete representation of the object. The locations of objects with low uncertainty levels are inhibited in the saliency map in order to represent the reduced importance of parts of the scene that the agent has already been familiarised with. This solves the problem of a moving scene or viewer, since uncertainty levels are available in terms of a global reference frame.

The attention component carries out two tasks. First of all, it directly provides gaze locations and information for gaze generation. These are important, not only so that the agent can pay attention to a salient part of the environment, but also so that it can *be seen* to pay attention. Secondly, as a result of providing gaze locations to the gaze controller, it controls what sensory information persists in the agent's short-term memory. In this way, a feedback loop is achieved; what an agent pays attention to will have an effect on what is stored in its memory and the information that an agent has in memory will have an effect on what it decides to pay attention to.

5. Memory

Memory is an important feature for both humans and agents and is especially important when agents have only restricted access to the scene database. We base our system of memory on what is called the stage theory of memory [2]. This is an influential theory of memory from the field of cognitive psychology that proposes information storage in 3 stages: sensory memory (STSS), short-term memory (STM) and long-term memory (LTM).

The STSS is a short duration memory area where information from the senses is converted into signals that the brain can understand. This memory has a very fast rate of decay, so items in it will be lost unless they enter the next stage of memory. In practice, our STSS stage takes the fovea and peripheral renderings (see Section 3) and extracts view components from them. A view component consists of the unique object or group colour extracted from the rendering as well as screen-space bounding box information. Essentially, at this stage, we have the identifiers of objects and groups that are visible, but they have not yet been resolved.

The next stage of memory is the STM. The STM relates to our thoughts at any time: when an item enters the STM it could be said that we are *aware* of it. In contrast to the STSS however, there is a longer persistency for items in the STM. In our implementation, all objects that are in the fovea are passed from the STSS into the STM. Since the fixation of the fovea is determined by the attention component, this means that it ultimately gets to decide what objects will enter the STM (see Figure 2). Note that both the STM and STSS are only capable of storing objects; salient *locations* are not explicitly stored. Although the attention component is image-based, the memory system is object-based; any objects that occupy salient locations are entered into the STM. When items enter the STM, they are resolved into observations, in a manner similar to [13]. In our system, an observation consists of a resolved handle to the object in the scene database, as well as its transformation matrix at the time of observation, an uncertainty level and the world-time that the observation occurred at. Essentially, this allows the agent to keep a record of objects that it noticed in the environment, as well as when it saw them, where they were and how much is known about them. In the spirit of sensory honesty, we do not allow the agent access to all of the object type data and only under certain circumstances allow access to its state data. In this way, we seek to balance the costs of replicating database information and the realism of the sensing abilities of the agent.

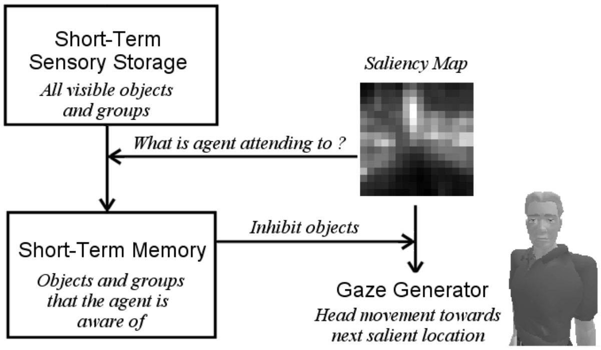
Short-Term Sensory Storage
All visible objects and groups
Saliency Map
What is agent attending to ?
Short-Term Memory
Objects and groups that the agent is aware of
Inhibit objects
Gaze Generator
Head movement towards next salient location

Figure 2. Interactions between the memory, attention and gaze components. The final gaze locations may be dependent on a function of scene uncertainty and saliency.

The final stage of the stage memory model is long-term memory. Although the LTM is not used in our current implementation, we envisage many possible uses for it. However, these uses are outside the scope of this paper and would be more applicable to a full model of attention.

6. Gaze Generation

The gaze generation component brings together all of the other components in order to automatically generate gaze animations. It controls and mediates between the other components as well as providing high level commands such as 'look at car5'. Three basic types of looking behaviour are implemented that vary in terms of the duration of the gaze and the amount of motion generated for the animation. These types are termed *glance*, *look* and *stare*. A glance animation is meant for short animations and places most of the emphasis of the motion on the eyes. This gives the effect of the agent looking out of the corner of his eye and is useful for less salient objects that nonetheless solicit some attention. A look animation lasts an intermediate amount of time and equally distributes the motion between the eyes, head and shoulders. Finally, a stare animation lasts the longest and biases the motion towards the head and shoulder areas. This motion is used when the agent is paying close attention to something in the environment.

These gaze animations are important to the system, since they serve functional purposes as well as aesthetic ones. This is because the agent's perception of the environment is updated in a snapshot manner. Essentially, the gaze generator requests the next most salient object from the attention model. When this is returned, the gaze generator looks towards that location. Once the location has been fixated, a further perceptual snapshot takes place. These perceptual snapshots do not necessarily entail updates of the bottom-up attention model. For example, quick glances to the periphery may resolve the salient location without an attention update; the gaze manager can continue generating gazes to salient locations based on the original attention information, provided the scene has not changed in any major ways.

7. Discussion

We tested the visual attention model on two scenes; a bar scene and a street scene (see Figure 4). Sample animations are available at the following url: http://isg.cs.tcd.ie/petersc/casa. A number of timing profiles were taken to evaluate our implementation of the attention model. These timings were averaged over 50 tests on a Dell Dimension 8200 2Ghz Intel Pentium 4 Processor, with 512 MB RAM fitted with an Nvidia Geforce 4 Ti 4600 graphics accelerator (see Table 1).

The saliency generation section of the attention algorithm runs in constant time with respect to the number of objects in the scene and had an average run time of 0.28217 seconds in the street scene. This is the time taken between passing the input retinal image to the attention model and receiving the saliency map (excluding processing conducted by the winner-take-all network). Although the creation of the saliency map generation routine averaged 92% of the total time, it should also be pointed out that the attention algorithm does not have to be run every frame. When a new gaze location is needed but the attention model does not need to be updated, only the sensing and winner-take-all network components need to be updated to obtain a new salient location. Since the retinal (full scene rendering) image does not need to be taken in this case, further savings are made. In the street scene, which contained 1405 polygons, the retinal image averaged 6.7% of the total sensing time. The savings would be greater for more complex scenes, bearing in mind that the fovea and periphery views (which must always be updated) are simpler renderings. Due to the nature of the synthetic vision model, these renderings could be speeded up using a multitude of visibility techniques. In our implementation, we distribute the attention processing over a number of frames in order to provide updates of the scene at interactive rates enabling a real-time visualisation of the simulation.

Table 1. Percentage of time taken for the sensing components (including object resolution), saliency map generation, and winner-take-all activation.

Sense (%)	3.0454
Saliency (%)	92.0264
Winner (%)	4.928

We currently use the attention model with the goal of creating gross gaze movements, which include the eyes, head and shoulders. In doing this, however, bottom-up attention provides only part of a full solution for generating plausible gaze motions. For example, the current inhibition-of-return technique that we use, based on object uncertainties, will tend to visit most of the scene as oppose to cycling between the most salient locations. While this appears to be realistic for covert attention activities, people tend to look repeatedly at interesting or informative parts of an image [22].

Further additions to the attention model and gaze controller are envisaged. Time varying stimuli, such as flicker, have been accounted for in the updated model outlined in [10] and provide important contributions to bottom-up attention. These features can be added to our system with minor modifications. Perhaps the most significant future improvement will be an integrated attention model featuring a top-down attention component. This could allow us to consider subtle factors such as object novelty and task relevance when planning gaze motions. A simple first approach may be to use the long-term memory to store a field representing whether the agent had encountered an object-type before. Objects that had not been encountered before may solicit more attention. Task-relevant attention could be achieved by increasing the importance of certain object types.

More work would be beneficial on the management of the saliency map generator in order to provide a more scalable system. In our implementation, the computation of the final orientation conspicuity map took on average 82.2% of the total calculation time for the saliency map. There are perhaps cases where only an intensity channel and colour channel computation would suffice.

References

[1] J.R. Anderson. *Cognitive psychology and its implications*, W.H. Freeman and Company, New York, 1995.

[2] R. Atkinson and R. Shiffrin. "Human memory: a proposed system and its control processes". In K. Spence and J. Spence (eds.), *The psychology of learning and motivation: advances in research and theory*. Vol. 2, New York: Academic Press, 1968.

[3] N. Badler, D. Chi and S. Chopra. "Virtual human animation based on movement observation and cognitive behavior models", *Computer Animation Conference 1999*, Geneva, Switzerland, 1999, pp. 128-137.

[4] C. Bordeux, R. Boulic and D. Thalmann. "An efficient and flexible perception pipeline for autonomous agents", *Eurographics '99*, Vol. 18, No. 3, 1999, pp. 23-30.

[5] B. Blumberg, R. Burke, D. Isla, M. Downie and Y. Ivanov. "Creature smarts: the art and architecture of a virtual brain", In *Proc. Game Developers Conference*, 2000, pp. 147-166.

[6] S. Chopra and N. Badler. "Where to look? Automating attending behaviors of virtual human characters", In *Autonomous Agents and Multi-Agent Systems*, Vol. 4 (1/2), 2001, pp. 9-23.

[7] M. Gillies. "Practical behavioural animation based on vision and attention", PhD Thesis, Technical Report TR522, University of Cambridge Computer Laboratory, 2001.

[8] L. Itti, C. Koch and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, Vol. 20, No. 11, 1998, pp. 1254-1259.

[9] L. Itti and C.Koch. "A comparison of feature combination strategies for saliency-based visual attention systems", *Conference on Human Vision and Electronic Imaging IV*. SPIE, Vol. 3644, 1999, pp. 373-382.

[10] L. Itti. "Models of bottom-up and top-down visual attention", California Institute of Technology, PhD Thesis, 2000.

[11] L. Itti and C. Koch. "A saliency-based search mechanism for overt and covert shifts of visual attention", *Vision Research*, Vol. 40, 10-12, 2000, pp. 1489-1506.

[12] C. Koch and S. Ullman. “Shifts in selective visual attention: toward the underlying neural circuitry”, *Human Neurobiology*, Vol. 4, No.2, 1985, pp. 19-227.

[13] J. Kuffner and J.C. Latombe. “Perception-based navigation for animated characters in real-time virtual environments”, *The Visual Computer: Real-time Virtual Worlds*, 1999.

[14] E. Niebur and C. Koch. “Computational architectures for attention”, In *The Attentive Brain*, MIT Press, Cambridge, Mass., 1998, pp. 164-186.

[15] N. Noser, O. Renault, D. Thalmann and N.M. Thalmann. “Navigation for digital actors based on synthetic vision, memory and learning”, *Computer and Graphics*, Vol 19, 1995, pp. 7-19.

[16] H. Pashler, *The psychology of attention*, MITPress, Cambridge, 1998.

[17] C. Peters and C. O’ Sullivan. “A memory model for autonomous virtual humans”, In *Proc. Eurographics Ireland*, 2002, pp. 21-26.

[18] M.I. Possner, M.J. Nissen and R.M. Klein, “Visual dominance: an information-processing account of its origins and significance”, *Psychological Review*, vol. 83, 1976, pp. 151-171.

[19] O. Renault, D. Thalmann and N.M. Thalmann. “A vision-based approach to behavioural animation”, *Visualization and Computer Animation*, Vol. 1, 1990, pp. 18-21.

[20] C.W. Reynolds. “Flocks, herds and schools: A distributed behavioural model”, *Computer Graphics*, Vol. 21, No. 4, 1987, pp. 25-34.

[21] X. Tu and D. Terzopoulos. “Artificial fishes: physics, locomotion, perception, behaviour”, In *Proc. Siggraph 1994*, 1994, pp. 43-50.

[22] A.L. Yarbus. *Eye movements and vision*, Plenum Press, New York, 1967.

[23] H. Yee, S. Pattanaik and D. Greenberg. “Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments”, *ACM Transactions on Graphics*, Vol. 20, No. 1, 2001, pp. 39-65.

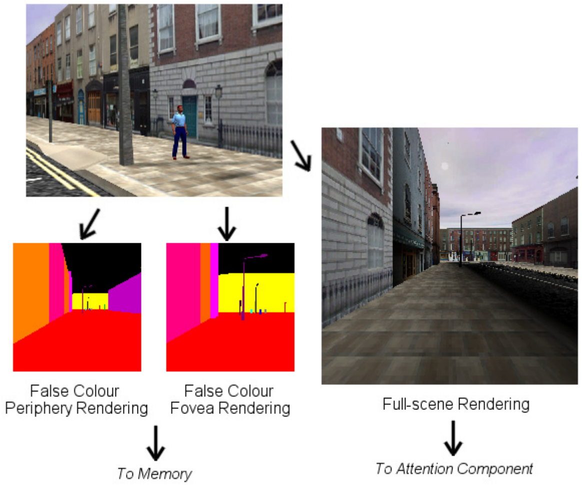

Figure 3. The synthetic vision module renders the scene from the perspective of the agent.

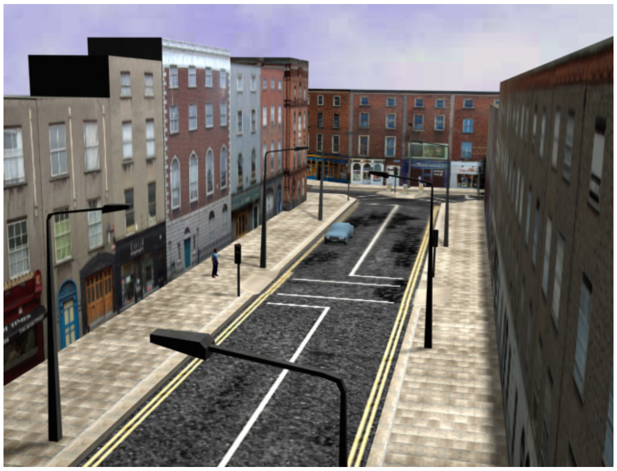

Figure 4. Depiction of the street scene that is used to test the attention system.