THE UNIVERSITY of EDINBURGH
informatics

ipab | Institute of Perception, Action and Behaviour

AGES : 'Attentive Gaze Enabling System' for Autonomous Active Vision Avatars

by
Sharathchandra Ramakrishnan

A thesis submitted in fulfillment for the degree of
Masters of Science
Artificial Intelligence

September 2008

Dedicated to my family

Table of Contents

Declaration of Authorship

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Abstract

The following research aims to instill emergent visual attention and a perception-action cycle into autonomous active vision agents that can react to the movements of real humans, in a common virtual reality framework. The research stems from the theory of embodied cognition which supports the view that the visual perception of a character is closely tied to its action model. It is the claim of this thesis that emergent behaviour of an autonomous animated agent is not only enhanced by a localized visual perception of its world but is infact dependent on it. Many of the previous behavioural animation frameworks have used a complete scene database where the animated avatar has full knowledge of its environment, and therefore there is no room for intentional gaze behaviour and visual attention.

Proceeding from previous research in visual cognition for bottom up attention models in humans, various real time high level computer vision modules have been built to direct the attention and gaze of a humanoid towards objects of interest in the virtual environment. These include the basic low level task oriented flat shaded synthetic vision module, an optical flow based motion sensor, an innovative real time tracking module for external real world objects and people, and a simplified saliency map to enable spontaneous viewing towards conspicuous objects.

As a standalone application, the *Attentive Gaze Enabling System (AGES)*, aims to provide the real world human with real time interactive control of visual stimuli that are experienced by the avatar in the virtual world. With AGES being one of the first to propose a mixed virtual reality implementation that allows the user to objectively interact with the perception-action cycle of a virtual humanoid using a 3rd person representation of himself, further extends its future use as an experimental tool in neuropsychology and perception-action studies in humans.

Acknowledgements

I would like to thank Dr Taku Komura, my supervisor, for his directions, advice, patience and encouragement that enabled me to delve deep into research archives of computer animation and virtual environments to come up with this thesis.

Chapter 1

Introduction

There has been a growing need to represent life like animation and behaviours in real-time animated humanoids of virtual worlds right from substitutes in ergonomic evaluations, conversational agents, AI agents in simulated environments to computer games. An important aspect of realizing this goal is to grant an animated humanoid enough autonomy to animate itself based on a simulated model of its environment. The main challenge faced in this process does not just revolve around simulating believable and plausible behaviors, but is to generate emergent and unexpected behaviours, which is the primary interest of this thesis.

1.1 **Motivation and Aims:**

Research in real time character animation has evolved to a stage where realistic and smooth human like motion is possible. Researchers have also instilled AI learning algorithms, but this restricted to the generation of reactive behaviors in multi-agent scenarios. But there is a glaring assumption that most researchers have made -- an assumption that the synthetic character in a virtual environment has complete knowledge of its environment and needs no sense of visual perception based decision making.

This assumption has divided the research of character animation into different schools of interest. Most developers are solely concerned with the visual aesthetics of an animation and in synthesizing apparent 'realism' in the behaviour of characters as this finds a wide commercial use in movies and computer games.

However the subject of concern of this thesis is not to synthesize apparent animated realism, but is to explore ways in which emergent humanoid behaviour can be generated by granting the synthetic character a level of autonomy and artificial intelligence to make spontaneous decisions just like real human beings do.

The need to instill a means of visual perception in the humanoid is a first step in realizing this autonomy. Some might argue that this is absolutely unnecessary, as the rendered 3D world already has all the information that the avatar needs. But as further investigation will reveal in subsequent sections, it is this very information about the environment that a previous humanoids have prior knowledge about , that prevents autonomous and emergent behaviour.

On this note, it is apt to raise a question at this point –"How does the so-called life-like 'autonomous' humanoid know where it is in the environment if it does not look around?"

1.2 Introduction to the AGES framework:

The primary aim of the work described in this thesis is to develop an '**A**ttentive **G**aze **E**nabling **S**ystem' (**AGES**) for character animation with interactive user control to change environmental parameters of the virtual world and affect the behaviour of the animated character in real time.

The importance of visual attention and attentive gaze as an important behaviour that triggers visual perception of an environment will discussed in detail in subsequent section. The **AGES** implementation aims to instill high level vision modules in the synthetic character so it may generate spontaneous and emergent attentive behaviours within a changing environment. The various high level vision modules are designed taking into account similar processes that occur in the visual cognition of biological organisms and human beings.

Further supporting claims have been drawn from recently reaffirmed theories related to Embodied Cognition, that closely link Action to visual Perception, and lay emphasis on visual perception being the driving force for an agent to move and act.

Thus the synthetic character in the **AGES's** virtual world is an 'Active Vision Agent' which draws visual information from an unknown environment by directing its visual attention towards objects of interest.

"*Vision has a purpose and that purpose is action. Action can be practical (motor control), theoretical (creation of a purposive representation, a decision or a change in internal state) or aesthetic*" **[A Sloman].** Throughout this work, this succinct quote by Sloman that describes the relationship between agent vision and action will be considered to be the maxim for designing the visual processes of the **avatar** in the **AGES** system.

Chapter 2
Active Vision

Researchers are beginning to realize that Vision and Intelligence of an organism are not disembodied. One of the tenets of Embodied Cognition is a perception driven action model, and that an agent's perception of the environment is closely tied to the motor system of the agent and the tasks that it performs.

All previous animations of avatars navigating in an environment assumed that the avatar has complete knowledge of the environment. This is in line with the classical/cognitivist view of the mind using a rule based logic driven system that relied on a stored description approach.

This kind of framework cannot handle unexpected events like new obstacles, or user invoked changes in the environment.

2.1 Active Vision Agents Vs Passive Vision Agents:

An active vision system, present in all living organisms, controls the process of image accquisition and introduces constraints to selectively recover information from the 3D scene. *This is achieved through a mechanism of selective visual attention which is discussed in the next chapter.*

The superiority of an active vision agent when compared to a passive agent is in its ability to perform general recovery tasks (like retrieve shape from x) in a much more efficient manner.

The difference between a passive vision agent and an active vision agents is that a passive vision agent is not capable of deciding how it wants to a view a scene and is bounded by preset environmental conditions and visual parameters. A passive agent has to acquire all the visual information presented to it using computationally expensive

procedures, but cannot retrieve the exact information required to accomplish a designated task.

An active agent on the other hand, must be able to modify its visual parameters so as to extract meaningful data from a scene that can be used to solve the specific task at hand. In other words it extracts only those features from the scene that are useful with respect to solving a particular task.

A perceptual system and the world in which it operates on are inseparable. Having a 'general relationship' with the environment by simply following physical laws is an erroneous assumption to start of with.

2.2 Need for Visual Perception Based Control Of Synthetic Actors:

In most agent – object interactions in animated virtual worlds, avatars are provided with complete information about all objects in an environment and their exact positions and properties are already known in the scene graph. This information is available during the preprocessing stage for the planner to decide upon a path for the avatar to follow from a starting point to a destination.

This kind of framework is quite unrealistic as the behaviour of the agent is not life-like. In real scenarios, autonomous avatars should not have knowledge of the global scene graph, and must have a strictly local knowledge depending on their perception of the scene. For instance, achieving a realistic simulation of an avatar looking for an object located somewhere in a room must involve change in gait, position and gaze direction as the avatar looks for the object in the virtual world. This situation is very similar to the way in which humans look for objects in real life. For instance , looking for a pair of shoes lying somewhere under a bed involves bending down, and looking around for the location of the shoes, and then planning the best possible path to reach out and grab it. This is also coupled with the visual memory of the person, based on previous habits of where he usually places his shoes.

In a behaviour based animation model, as **Reynolds** put it rather succinctly-- behaviour is not only reacting to an environment but must also include the flow of information by which the environment acts on the living creature as well as ways in which the creature codes and uses this information.

As quoted by **Isla and Blumberg** who talked about 'sensory honesty' with respect to a synthetic character and raised questions about what a character must perceive and how -- "A fundamental aspect of sensory honesty is that it forces a separation between the actual state of the world and a particular character's *view* of the state of the world." **[Isla and Blumberg]**.

What is being underlined here is the importance of a system relying primarily on visual perception as a means to move and make decisions. Evidently, giving the virtual avatar it's own sense of visual perception of its environment not only makes it autonomous, but also gives the viewer a feeling of realistic animation. The following sections elaborate on supporting theories that deal with perception as the primary input to a cognitive agent and also presents the implementation details and methods undertaken to instill visual perception and an attention driven memory model in a virtual avatar.

The final aim of this work is to use the visual representation of a dynamic and unknown virtual environment to generate realistic animations of an avatar's gaze.

Chapter 3
Visual Attention

3.1 Visual Attentive Tasks Involved in an Embodied Agent:

This section outlines the various visual tasks that an embodied agent performs, and how the surrounding environment affects these processes. It is important to know how these task are affected by visual stimulus, eventually directing the gaze (in our case, head movement) of the humanoid towards the object of interest.

As discussed in **Yantis et al,** an agent's attention maybe directed by voluntary and deliberate endogenous factors, or maybe involuntarily directed by externalities and distractions in the environment, known as exogenous factors. Contrary to popular belief that we can see everything around us, in reality it is only a small amount of visual information that is processed and registered by the visual system to eventually influence the behaviour of an organism.

From previous work done in visual cognition and psychology, experiments conducted on real human beings have revealed several modes of attention and the influence of the external environment on attention. Experiments in eye-tracking have been of particular use to measuring the onset, duration and response time in an attention model. Saccades or rapid eye movements that happen when the context of attention shifts between objects are an important mode of measurement.

However, this work does not try to simulate eye tracking our agents and focuses more on controlling the gaze of our humanoid. When drawing the attention of the humanoid to an object, the gaze orientation of the head is roughly directed along a vector connecting the agent's head to the coordinates of the object of attention in the agent's viewspace. (This is dealt with in the subsequent section). Anyway, keeping this is mind, let us now review the various attention modes present in an agent:

3.1.1 Endogenous Attention and 'scanpaths':

If an agent's attention is drawn by goal-specific factors, or in other words in a task driven situation, the attention of the agent is a voluntary and deliberate endogenous attention, which is essentially a top-down model of attention. What is of interest to us here is that way in which the agent visually scans the environment while on a task. Such spatial movements of the agent's gaze and head correspond to '**scanpaths**' in space that have found be indicative of a global problem solving strategy **[Stark and Choi , Yarbus]**. They found that an internal cognitive map guides the strategy of the visual task. Scanpaths were thus looked at as some sort of 'foveal sampling' that was used to validate reality with the cognitive map and therefore is a top down model of attention.

3.1.2 Exogenous Attention:

Externalities or 'Distractors' in the environment are called exogenous events **[Jonides et al**]. They come into play especially when the agent is experiencing divided attention while engaged in a task like visual search. A classic instance of this is an implementation involving tracking of multiple moving objects while the agent is looking for a particular object. **This presents a challenging problem which has been attempted in this project.** For this a motion sensing behavior function is encoded into the synthetic vision module, such that an object moving in the agent's peripheral field of view acts as a distractor. This enables a mode of divided attention which basically increases the response time of the agent in attending to deliberate task assigned objects in the scene. A more detailed account of simulating this realistic behaviour in an agent is discussed in the implementation section of the vision module.

3.1.3 Spontaneous Looking or Free Viewing:

The Endogenous Attention driven scan paths, mentioned at the beginning of this section is a top-down control model of attention with deliberate task driven intent.

However in the absence of deliberate intent in human beings, attention seems to be directed towards local regions containing the most conspicuous objects in a scene and is called Free Viewing. This type of attention is bottom-up **[Kahneman]** in the field of psychology and visual cognition. Free viewing or spontaneous looking in the absence of deliberate intent or a task is a very subjective behaviour varying from individual to individual.

However, the most conspicuous object in a scene or image that is bound to attract the attention of a typical human being can be computed by contructing a saliency map of the scene, which is proposed to be implemented in the synthetic actors vision module. Usually areas of high local feature contrasts may be used to find the most conspicuous object in the scene as used by **[Chopra et al].** This thesis attempts a more sophisticated approach using an innovative saliency map, particularly suited to the requirements of computer graphics.

Spontaneous looking in the absence of a task directive is of particular interest to this thesis, which has tried to generate realistic and spontaneous animations of agent gaze when an avatar is introduced to an unknown virtual environment. Details mentioned in the implementation section will make it more clear as to how kinematic behaviour model is implemented. The salient vision module to enable automated free viewing and gaze is activated when there is no task driving the agent, although other navigation tasks such as walking maybe parallely executed.(in this implementation, by user control due to lack of a generic path planner).

3.2 Attention Selection:

The problem of attention selection in living organisms has been an area of active and interesting research for neuroscientists and psychologists.

Researchers have observed an interesting feature, that in the visual processing of dynamic scenes most living organisms use a serial processing technique **[M. Heisenberg, R. Wolf].** Particular regions or objects in a scene are selected by either their behavioural importance or by visual cues.
Humans identify objects and analyze their spatial relationships by a process that involves either rapid saccadic movement of the eye to bring an object within the focus of the foveal region or by covert shifts of attention. An object is then said to be "attended" if it enters short term memory for a long enough time to be reported as "present" by the higher cognitive functions of awareness. Thus visual attention is closely linked to visual awareness [**Crick F, Koch C.]**.

Contrary to the obvious belief that the visual processing in humans must be parallel, it is infact a serial implementation**[M. Heisenberg, R. Wolf]**. Considering the amount of information that has been found to flow in the optic nerve of primates – 10^8 bits per second – it is almost impossible to realize this processing serially. *But the process that nature uses to get around this problem is to selectively process only certain parts of the input image by an order of preference, and shift the focus of attention from one region to another serially.* So how does this selection happen?

Previous research in the field of visual cognition in human beings, that have also been applied to real robots have attempted to solve this problem of attention selection using both top-down and bottom-up models of attention.

A top-down approach with 'variable selection criteria' is driven by cognitive or voluntary control which selects an object of attention by shifting the 'spotlight of attention'. A much faster bottom-up approach relies solely on the visual properties of a scene to select conspicuous salient objects in a scene by building a saliency map,

mostly using centre surround mechanisms. The **AGES** system in its current stage has just begun to implement a saliency map based bottom up model of attention selection to select objects of attention towards which the gaze of the humanoid must be directed towards during conditions of free viewing when the humanoid is in an idle behaviour state.

The next chapter describes the implementation of the vision modules that provides visual information to enable gaze control as well as convey object information to the animated humanoid.

Chapter 4
Synthetic Vision Modules

4.1 Previous work in Synthetic Vision:

There have been many frameworks in the past that have instilled sensory perception, in particular, synthetic vision,to animated avatars.

The seminal work by [**Reynolds]** presents a distributed behavioural model to simulate flocking behaviour among virtual birds or boids. Similar to individual particles in a particle system, the virtual 'birds' did not have complete information about their surrounding environment and it was this imperfect knowledge of their environment that produced unpredictable behaviour. Although the system did not simulate any real sensory mode used by living organisms, each boid had a spherical zone of sensitivity around it and flocking behaviours with neighbouring boids were specified using physical constraints like velocity, attracting and repulsing forces for collision avoidance, global and local heading directions. The result of this model was an interesting flocking behaviour, visually akin to those seen in real world bird formations. It was concluded that the simulated 'flocking' behaviour was not only enhanced by a limited and localized knowledge about the environment but also was dependent on it.

'Artificial Fishes' developed by [**Tu and Terzopoulous]** consists of a self animating behavioural model giving rise to emergent behaviours similar to the Reynolds' Boids model but had more elaborate physical models driven by a perception model. The artificial fish had two on board sensors, a visual sensor and a temperature 'sensor' to perceive the environment and induce actions. The vision sensor extracted useful information about neighboring fishes from the 3D world -- such as colours, sizes, distances and identities of objects, as well as instantaneous velocities of objects of interest. The fish's vision sensor had access to the geometry, material property, and illumination information that was taken directly from the graphics pipeline used for

rendering, with a scope of visibility in a 300 degree spherical angle. At every time step of the simulation, an intention generator generates behaviour routines based on the inputs from the visual sensor.

[Renault et al] used a synthetic vision system for high level animation of humanoid characters. The character used the output of the vision system to move in a corridor and at the same time avoid approaching characters and objects. The synthetic vision module rendered the scene from the character's point of view and details extracted from the environment are stored in a 2D array whose elements consisted of a vector holding the pixel at that point, the distance from the character's eye to the pixel and an object identifier of the object at that position. In the synthetic vision view port, objects are not rendered with their original colour but each object is assigned a unique colour code so as to differentiate and represent object properties using colour values.

An innovative extension to above described framework was published in the work of [**Noser at al],** in which was added, a memory and learning mechanism. Their synthetic vision module was used by an avatar to explore an unknown environment (built at runtime using L-Systems) and to build mental models and cognitive maps. After this the avatar could successfully plan a path to a specified location and navigate through the environment. The implementation produces a flat shaded rendered scene from the avatar's point of view using the z-buffer drawing capabilities of the underlying graphics hardware. Objects are false coloured with a colour code called Vision_Id which is set to be unique for different objects, such that semantic information about the properties of the object can be stored and retrieved. For local navigation and low level operations such as obstacle avoidance the z-component of the distance from a pixel to the avatar was used. However, global planning required the exact 3D position of an object and this was obtained by inverting the modelling and projection transforms.

[Noser et al] used an octree data structure to internally represent the visual memory of the avatar, the advantage being that occupancy grid models like octrees inherently represent the 3D graph structure of the environment (visual memory representation will be discussed in detail in subsequent sections.) An elegant modification to the above described internal memory representation was suggested by **Kuffner et al** who used the object geometry along with a list of Object IDs and their most recently observed states to be to be represented as a tuple.

4.2 Contributions of AGES:

The above discussed previous work represent low level synthetic vision modules used typically for task based behavior, namely, top bottom execution in Endogenous models.

Although AGES does have such a low level vision module implemented for task based behaviours, the innovative vision modules present in AGES are for exogenous attention and motion sensing using optical flow, a real world vision & attention interface, and finally spontaneous free viewing module using a saliency map approach.

This is in line with the motivation of this thesis that seeks to take advantage of recent developments in graphics hardware, as it feels the time has come for animated avatars to have full blown and advanced purposive high level vision modules in order to attain autonomy in unknown and dynamic virtual environments.

There has not yet been an attempt in previous research to scale such high level vision modules into a single animation framework that operates in real time with user interactivity.

The synthetic vision modules implemented in **AGES** makes use of recent developments made in graphics hardware that allow us to perform asynchronous read back from the GPU, rather than synchronous read back performed in the early days of [**Kuffner at al],** which resulted in a transfer overhead between the CPU and GPU. As mentioned earlier, the aim of the avatar's vision module is to supply the avatar with a high level abstract visual representation of its environment and it's events. All the problems associated with the vision computation of real robots like distance detection, pattern recognition and noisy images maybe skipped.**[Thalmann et al].**

The synthetic vision modules used by the Active Vision Avatar of AGES have been categorized into 3 vision modules based on the behaviour state and context of the avatar –

i) A low level vision module to provide basic Object Functional Information.

ii) A motion sensing and tracking module using optical flow.

ii) A Real World Vision module to direct gaze and attention towards objects in the real world.

iii) A Spontaneous Gaze or Free Viewing vision module

We shall now see how each of these vision modules have been implemented and how they depend swtich based on the avatar's behaviour state. (As mentioned earlier the various high level vision modules operating in humans is based on the context of the human and what he or she expects in a given situation)

4.3 Synthetic Vision Modules Details:

4.3.1 Object Functional Information Vision Module:

This vision module is active when the humanoid is in a task oriented behaviour. This could be in a situation when:

a) the user is manually exploring and directing the gaze of the avatar towards the environment while it learns/ makes a visual memory representation.

b) when the agent is on a task driven behaviour, like looking for a particular object to launch a motion.

c) Alternatively, this vision module is invoked after a higher level vision module has directed the avatar's gaze and attention towards a particular object.

It is evident that this low level vision module comes into play when the avatar is in an endogenous state. *It is of lesser interest to us, as this vision module comes into play after a high level vision module has already established the gaze of an avatar towards a particular object.*

However it is an important part of the avatar's vision repertoire, as it is useful to spontaneously launch an action towards an object by knowing its functional behaviour, as well as keep a visual memory representation of the unknown environment.

Description:

To represent the actor's vision in this module, we render a flat shaded, unlit model of the 3D world which is obtained by asynchronous read back from the graphics pipeline using the pixel buffer object, and maybe used to drive the control mechanism of the avatar.

The basic information that can be processed from this vision module are :

1. **Depth Perception :** The depth of the pixel related to an object is directly obtained from the z-buffer, and the 3D position of the object in the avatar's field of view is obtained by inverting the modeling and projection transforms.
2. **Object Functional Recogniton :** By using false coloured objects, one can convey object function and identity through the vision module (ie. The objects are rendered in the agent's viewport using predefined colours which define the object type or object function)

Since the calculation of the visibility of an object is a rendering operation, methods to speed up the rendering of complex scenes such as scene-graph management, hierarchal level of detail approximation and frame-to-frame coherency may be adopted. The size of the rendered viewport does not have to be very large and the resulting image pixels maybe scanned , in order to obtain a list of visible objects in that frame.

Details:

Proceeding from the synthetic vision module implemented by **Noser at al,** for an image of size (dim_length x dim_length), in the normalized vision volume, pixel (i,j) is represented by a normalized point :

$$p_{norm}[0] = -1 + 2(j + 0.5)/dim\_length$$

$$p_{norm}[1] = -1 + 2(i + 0.5)/dim\_length$$

$$p_{norm}[2] = -1 + 2\, Z_{screen} / Z_{max}$$

$$p_{norm}[3] = 1$$

where Z_{screen} is the Z buffer value of the pixel(i,j) with $0 <= Z_{screen} <= Z_{max}$

In order to obtain the 3D position of a point in the real scene, the modeling and projection transforms are inverted and multiplied by the normalized point p_{norm} :

$$preal = pnorm . (M.P)^{-1}$$

M -- Modeling matrix

P -- Projection matrix

p_{norm} – Normalized point and p_{real} is the point of the real scene.

4.3.1.2 Visual Memory Representation:

The false coloured vision module is used by the avatar to make a visual note of the spatial distribution of objects in a unknown environment along with object properties including object function.

One of the implementations proposed in this thesis is to encode object functional information within virutal objects and hardwire it to the agent's behavioral state machine to launch an action from the pose graph, towards the particular object. This is discussed in the section which couples our perception-action system into a bigger picture.

In the process of scanning the environment using the false coloured visual port, an avatar can construct a visual memory representation of the scene. This kind of memory scheme was first described by [**Noser et al]** who used object level spatial subdivison techniques like an octree data structure to store positions of the object, and this was an efficient method to plan an accurate path planner in an environment.

However an octree implementation is not scalable for larger environments, and we would just like to rely upon the geometry of the environment to maintain an array list indexed by Object Ids. Using an array indexed by functional object ID was an approach also used by **Kuffner et al** to provide a synthetic character with an internal world model. AGES adopts a similar representation of object properties by the use of tuples to represent observations :

If O denoted the set of all objects present in the environment, and an agent maintains a set M of observations from the output of the vision module, then these observations are represented as a tuples $< OB\_ID_i , P_i , T_i , t_i >$.

Where:

OB_ID_i -- Object ID of the object i

P_i – Properties of the object i -- These properties can be semantic information about the object and are used by higher level behavioural modules to affect the

behaviour state of the avatar. In the **AGES** system, the property represented by objects relate to an Action cycle. For instance, mushrooms were color coded with the Object ID that has a hard-wired correspondence to a behaviour state that generates an 'EAT' action.

T_i – Transformation of the object i – The transformation represents the orientation and position of the object at the time of observation.

t – time stamp of observation

M is set to represent the character's visual memory of O. To begin with M is Empty, and as the agent observes the environment, it uses the flat shaded synthetic vision module to scan pixels and extract the false color Ids of objects present in the scene. After each scan a set S of all scanned OB_IDs present in the scene is returned. The next step is to create a tuple <OB_ID, Pi, Ti, and t > for each of the objects. If a tuple containing an object already exists then this object was already observed previously and is not added to M. If no tuple in M exists with a particular OB_ID, then the observed object is new and is added to the character's visual memory M.

We shall return to the application of this low level flat shaded synthetic vision module again in the section that deals with the bigger picture of coupling perception and action into a vision driven behavioural model.

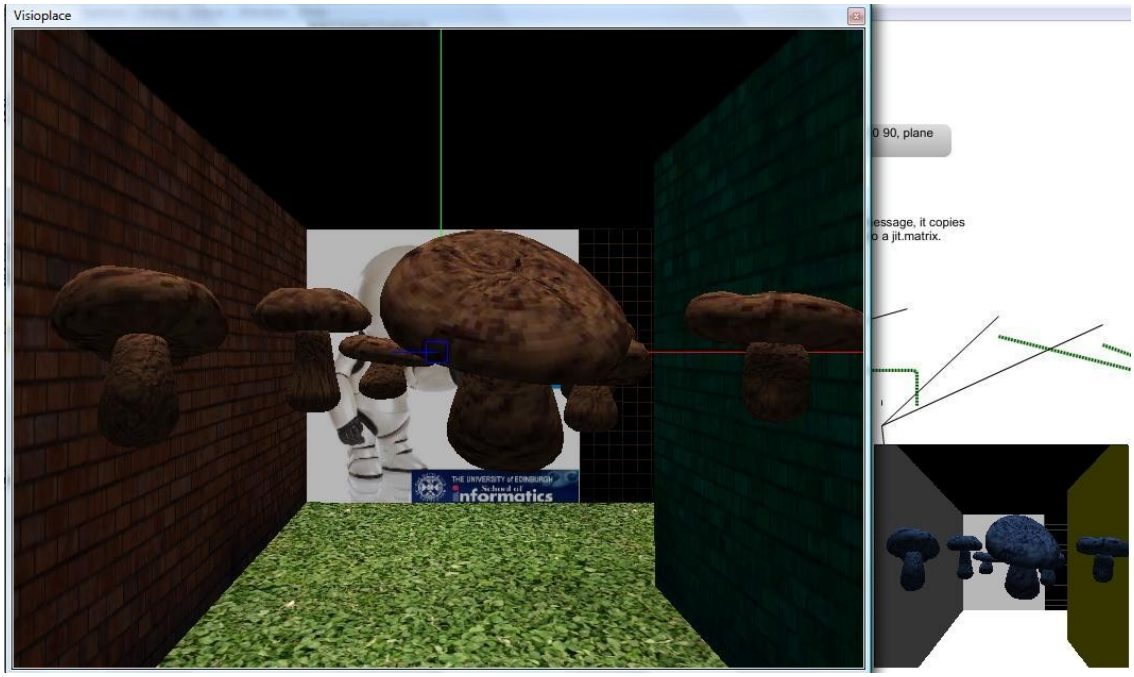


Figure 1

The above figure shows a flat shaded Object Functional Information Vision Module, implemented in the AGES environment using synch GPU readback

Mushrooms are false colored with blue to display object function --> Action 'EAT'

4.3.2 Motion Sensing Vision Module Using Optical Flow :

This section describes the high level vision module of AGES that enables the avatar to detect a pattern of apparent motion of an object by the relative motion between it's virtual camera and the plane of the motion, or in other words visually represent a vector field of an image changing with respect to time.

Motion of objects in the environment that capture the attention of an organism, in this case our agent, are exogenous events and could be looked at as external distractors. Let us consider a stationary agent looking at a static scene and suddenly a moving object comes into its field of vision. The animation would look unrealistic if a humanoid does not pay attention or respond to this sudden event.

Previous animations have prior knowledge of such dynamics in an environment, or have access to the 3d space coordinates from the motion script of the moving object, and so manually hard-wiring the camera coordinates to the moving object created an 'illusion' of an attentive avatar.

However in this system, the environment in unknown and unpredictable, and the only access to environmental events is through real time GPU read-back, and so the problem is almost similar to that of computer vision for real robots.

In order to capture the attention of the agent towards this object we must use an efficient real time vision method to sense the motion and it's direction.

Although one might think a simple algorithm to track the change in average color of the pixel matrix output of the synthetic vision module is enough to detect a motion, it will not be enough in a realistic 3D scene where a stationary wall might have the same color as a moving ball. Also, in order to get more information about the position and direction of motion of the object, such a simple method will not suffice.

I would like to mention here that it is essential to know the big difference between motion perceived in the image and theoretical projection of velocities onto the image. For instance, a shining rotating smooth sphere without any surface markings, although in a state of motion cannot be perceived to move, as constant illumination causes no change in image intensity over time.

The most natural approach to solve this problem is to compute the optical flow as it can be used to infer not only motion of objects but also their structure. Optical flow has been used in the past to provide a simple and elegant means for automated synthetic perception based low level navigation and obstacle avoidance in computer graphics, most notably by **[Blumberg et al]** who constructed a simple first order per frame motion energy model while navigating a 3d scene.

There are several methods to compute optical flow used in complex vision computations in robotics that include correlation methods, feature tracking, energy based methods and gradient based approaches. Feature based approach involves robust feature detection prior to the tracking which is not feasible in a 3d framework. Besides optical flow using edge detected features return motion flow that are only perpendicular to the edges.

This system adopts a gradient based approach using spatial temporal partial derivatives, an open source HS algorithm based on the work of **[Horn and Schunk]** , which is much more suited to this system's requirements. Their method estimates motion at every point in the image, and this is well suited to direct a stationary avatar's gaze to any moving object or motion that requires attention.

In camera-centered coordinates, each point on a 3D surface moves along a 3D path X (t). When projected onto the image plane each point produces a 2D path $x(t) \equiv (x(t), y(t))T$, the instantaneous direction of which is the velocity $dx(t)/dt$. **[B K P Horn et al].** The 2D velocities for all visible surface points is often referred to as the 2D motion field. The objective of optical flow is to compute an approximation to the motion field in a time varying image intensity.**[B K P Horn et al].**

In order to perceive motion we must see a changing pattern projected onto a 'screen'.

Also, a moving object must give rise to a constantly changing brightness pattern. The optical flow algorithm of Horn and Schunk senses and segments the motion of the object using a spatio-temporal derivative of an evolving image brightness function which partially determines optical flow. The assumption made is that the brightness of any part of the imaged world changes very slowly, so that the total derivative of the brightness is zero.

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0,$$

where $I(x, y, t)$ is the image brightness function.

The in depth details of this algorithm is out of the scope of our discussion and can be obtained from the list of references.

Some simplifying assumptions have been made while designing the virtual environment that have influenced the operation of this vision module. The design of the virtual environment is in line with the the assumptions and restrictions made by **[Horn and Schunk]**. Firstly, they assumed a simplified world where the apparent brightness of objects can be directly identified with the movement of surfaces in the scene. They also assumed that the image surface is flat to avoid variations in brightness due to shading effects and that incident illumination is constant over the surface. The vision module is designed to identify moving objects in a static background with high contrast. The functionality is also extended to motion sensing in well defined contexts that have a pre-decided physical location within the environment.

Design:

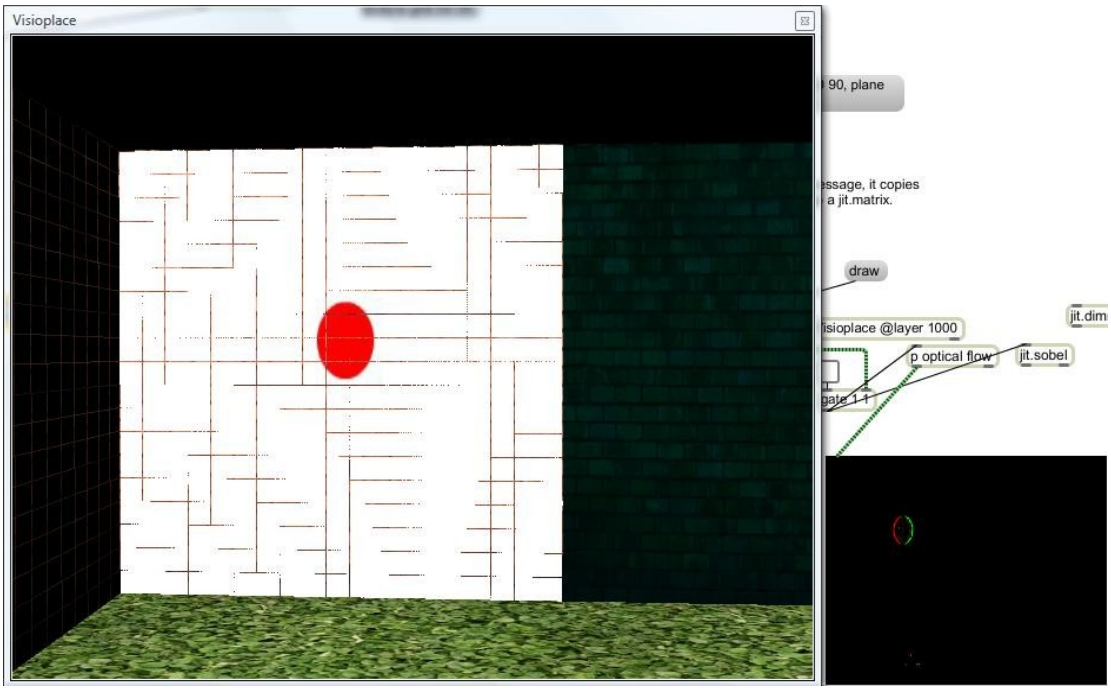
In line with the above constraints, I have designed the virtual world such that, the object to be tracked is a flat unshaded red ball moving along the white background of a plane. To reiterate the fact that the humanoid is in an unknown virtual environment and has no access to the coordinates of the ball, the system actually projects a video of the

moving ball onto one of the planes in the virtual world.

As there are different vision modules required for different contexts, the optical flow vision module is activated in only certain regions in the 3d room. This is a simplifying assumption needed to achieve real time performance as it it not possible to run all 3 high level vision modules from the GPU at the same time.

The output of this vision module, could be used to signal a higher level action module. For instance a sudden moving car in the environment, would be sensed by the optical flow module, so the avatar may move away or wave. Shown below is a simulation in this system which tracks the position of a moving ball in a static environment using optic flow. It is to be noted that tracking is a full blown function of this module, although it may just be used to detect motion or direct attention.

Figure 2-- Optical flow based motion sensor implemented in the AGES environment.



Note how only the moving object is tracked

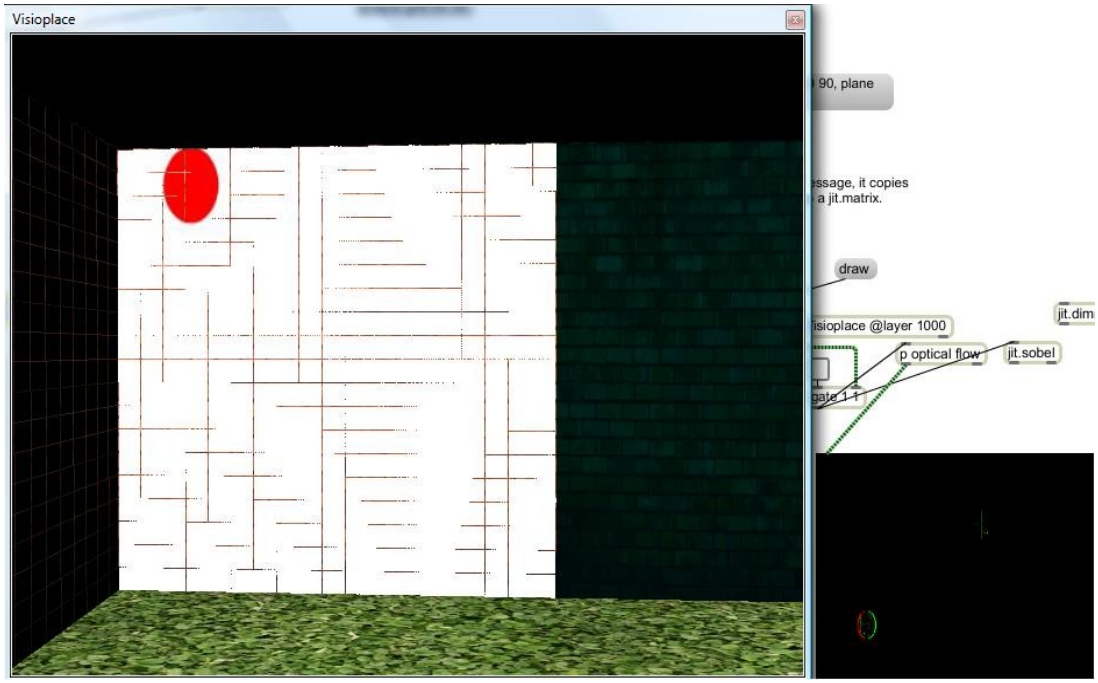
Figure Comments:

Left **-->** Moving ball in real world frame

Right --> Synthetic vision module -->Optical flow detects motion field and trajectory

FIGURE 3

2 Frames later than Figure 2



4.3.3 The Real World Vision Module:

This section provides an innovative framework that has not yet been attempted in most virtual world animations.

This is a real time high level vision module that provides the avatar with ready to use attentive information related to dynamic moving objects in the real world so that it may direct gaze towards it.

So far we have seen animated moving objects, and simple geometric objects that can be easily scanned using buffer read-back and segmented, but when it comes to making the avatar react with real world scenes rendered into the virtual world, it is a unique meeting place for computer vision and computer graphics based humanoids.

Motivation :

This idea has no immediate reference or background information, but one of the previous works that has been a motivation is the 'Videoplace' invented by Myron Kreuger, the researcher who coined the term – 'Virtual Reality'. It was way back in 1989 that he combined real time computer vision and graphics to interactively control virtual characters using gestures. Unfortunately the touch screen industry lobby was too powerful, and the WIMP (Windows Icons Menus and Pointer) interface still continues to exist.

Given the processing power today's underlying hardware, I feel that the future of interactive animation and humanoid behaviour lies in combining external reality into the virtual world.

Implementation:

The implementation consists of a real time external camera feed, captured as a video matrix and rendered directly into the virtual world and thus it becomes part of the virtual world, as a virtual object with plane information. It is to be noted that this content does not have to be a camera feed but can be any arbitrary video information.

The objective is to enable complete user interactivity and control, to generate realistic animation with totally unpredictable objects of attention introduced by the user.

As again, we use the underlying GPU hardware to capture the rendered scene from the avatar's point of view. But since we are interacting with real world components, applying the ordinary algorithms that we have applied so far on the GPU synthetic vision module will fail to generate any meaningful and usable information.

Thus the contents of the real world matrix has been preprocessed in real time to segment blobs belonging to the same region and a tracking point has been provided to the avatar so it may direct its gaze towards it.

As seen in the screen shots, the preprocessing algorithm performs edge detection and grows areas towards boundaries to segment all continuous regions as the same blob. The tracking point is placed at the centre of mass/centroid of the blob and a circle with area approximately equal to the area of the blob is drawn around it. This information of the vision module provides the avatar with tracking points corresponding to different objects of attention.
In this present system the avatar randomly chooses between the red tracking points and directs it's attentive gaze towards it.

.

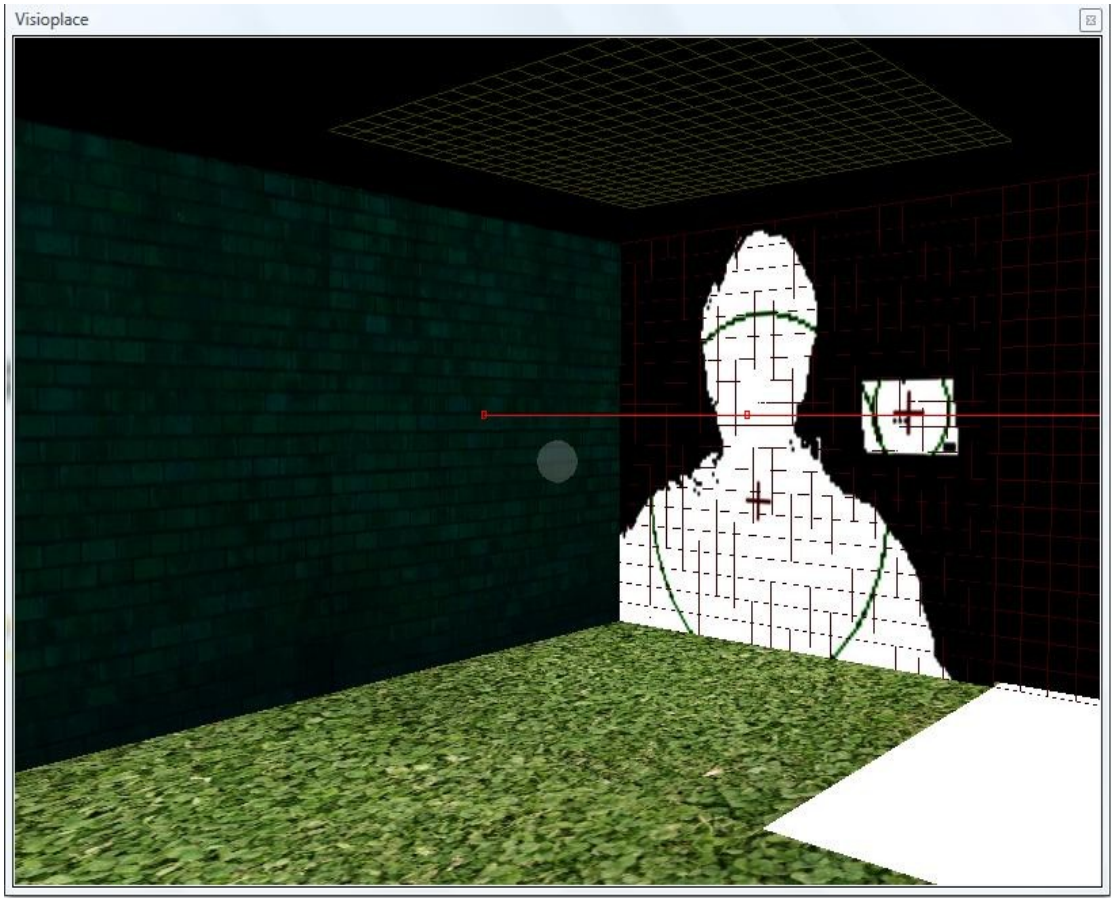


Figure 4 : Real time video feed inside the virtual world of AGES. There are two objects of attention given to the avatar, one is a person (me) and the other is a picture in the background.

4.3.4 Spontaneous Free Gazing

The Attentive Vision Module :

As discussed before,when there is no task directed behaviour routine, the agent adopt a spontaneous viewing mechanism.

There have been many approaches to represent visual attention. **Culhane et al** developed a data driven approach for visual attention with a hierarchal representation of the input stimuli and an attention mechanism that selected the regions of the hierarchy from the most abstract level down to the lowest level. Instead of using a spotlight to illuminate the top most region, an attention beam that illuminates the entire hierarchy passed through the regions of greatest interest and inhibited regions with lesser interest, by computing 'pass zones' and 'block zones'. This is based on a classical top- down approach to the attention theory and as discussed in the last chapter, is slower and not suited to our needs of real time rendering.

To automate this feature of attention selection, a synthetic vision based approach that uses a bottom-up attention model using a saliency map is proposed, where visual stimulus is used to control selective attention.

Saliency Map:

The construction of saliency maps to implement attention selection in robots has been a widely used method in computer vision applications. The output of the saliency map represents the conspicuous nature of various objects in a scene as scalar values in the visual field.

From the spatial distribution of these scalar locations it is possible to guide the attention of the humanoid towards the object with greatest conspicuity. Therefore the very nature of automated visual attention maybe acheived to make the gaze behaviour of the autonomous humanoid in an unknown virtual environment seem as realistic as possible.

Previous work:

The use of a saliency map to tackle the attention selection problem using computer vision techniques in robotics are many.

Itti and Koch combined information from different early vision components like colour, intensity and orientation to construct a saliency map using a gaussian pyramid and a gabor filter. Although this is considered the seminal work for the construction of saliency maps, it is not suited for interactive frame rates demanded by real time computer animation.

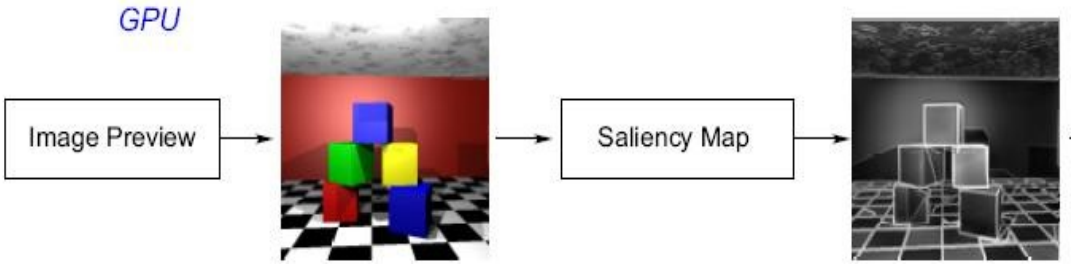
A GPU based saliency map was developed by **[Peter, Kurt et al] a**nd was used to for an innovative application in hi-fidelity selective rendering of 3D scenes. They proposed a saliency map to find out regions of an image that users are most likely to look at, (ie. the most attention seeking parts of the image) and rendered only those parts of the image with high quality and the rest of the scene with much lower quality.

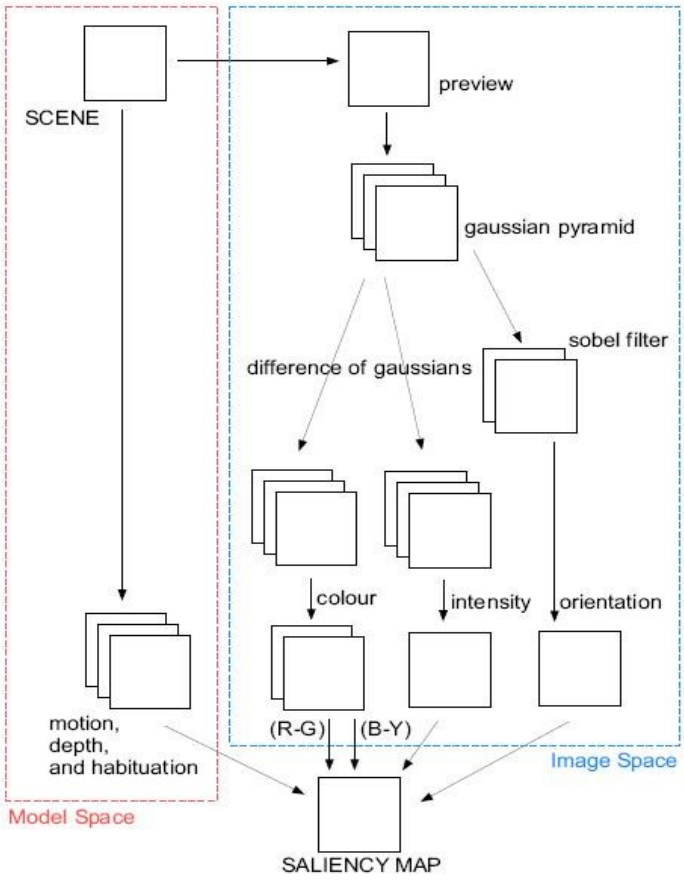
Their method takes advantage of the high performance capability of modern GPUs, and contructs the 2D saliency map in a matter of milliseconds and is very suited to real time animated frameworks. In addition they also added motion information to enhance the factors of selectivity in the saliency map.

The realization of lower resolution representation of out 3d scene using a Gaussian Pyramid, is out of the scope of this thesis. However, in the current implementation, a few steps have been taken to generate a saliency map, based on the technique presented briefly below.

This has been discussed in the future implementation scetion. In the current implementation an attempt to create a saliency map has been made although due to the complexity of creating a gaussian pyramid ,within the time frame assigned submitting this report, a working implementation of a saliency map is still work in progress. The current implementation has integrate a real time sobel filter to provide edge orientation information for the construction of the saliency map. The use of a a sobel edge filter is seen the previous work of **[Peter, Kurt et al]** who used it to construct a saliency map

for hi-fidelity selective rendering of photorealistic images.





The above figures show the saliency map constructed by **[Peter, Kurt et al]** to determine conspicuous objects for hi-fidelity selective rendering, a model that the AGES system wishes to reuse for its saliency map to enable spontaneous viewing for the avatar.

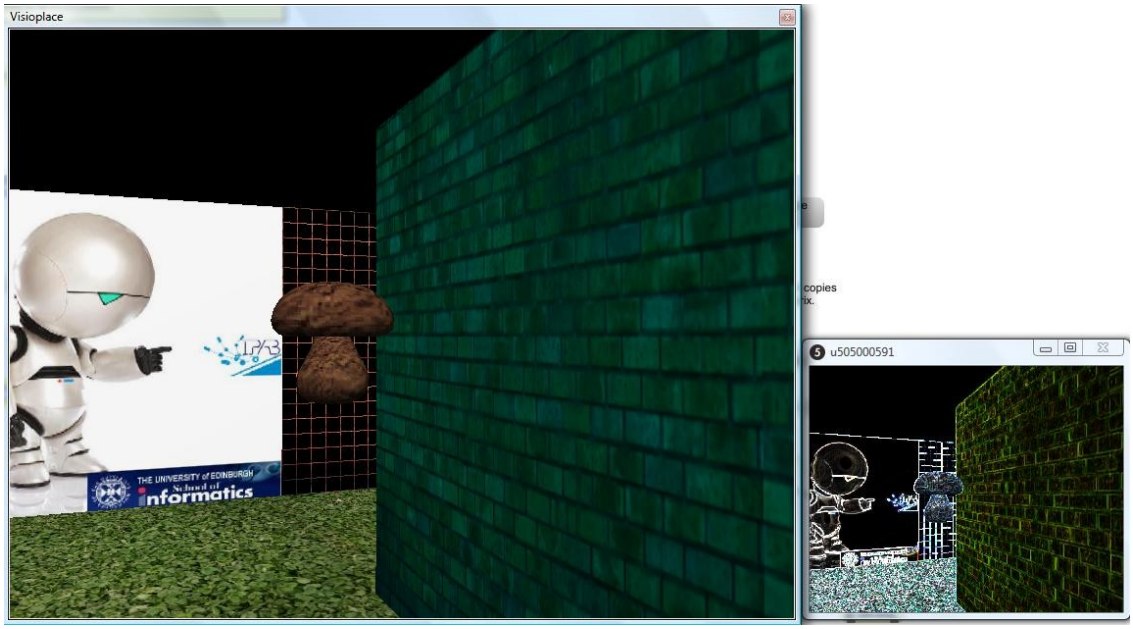


Figure 5: Saliency Map Module (Work In progress)

GPU read back on the right show edge detection using the Sobel filter

Chapter 5

'Through the Eyes Control' for Character Animation

Establishing points of attention, and tracking moving objects of interest both within and outside the virtual environment is one thing. Using this information to accurately direct the gaze of the humanoid towards these objects is the other side of the coin. The kinematic control of the humanoid's gaze is an important module in the behavioural system of the avatar, especially because our agent is visually stimulated.

The drawbacks of most animation systems is that there is very little interaction between the visual perception of the humanoid and the body movement of the avatar. Simple animations that have involved avoiding obstacles, where the agent has to turn his body in order to modify his path does not necessarily have to rely on real time visual input as most often the path planner is precomputed and the agent 'seems' to detect the obstacle and walk around the obstacle in real time. However the animation seems unrealistic when a moving object is introduced in the environment and the avatar simply gazes into the distance as though nothing happened.

However for the **AGES** system, the aim is to realistically control the gaze of an attentive avatar towards arbitrary dynamic objects or locations in the 3D scene which are of attentive interest to the avatar. Further the movement and presence of objects in an unknown virtual environment must catch the visual attention and gaze of the avatar in real time. Thus it intends to provide the real time control module that directly couples the visual perception of the environment with the avatar's gaze.

5.1 Structural Framework:

In a vision induced character animation system such as the **AGES,** it is essential that along with the correct orientation of an avatar's gaze, the corresponding changes in the gait of the avatar are also taken into account to make the animation seem as realistic as possible.

There has been considerable research done in the field of animation that deals with the production of realistic body movements that correspond to actions of the avatar. The notable work of **[Courty and Marchand]** provides a detailed account of the various requirements for character animation in an image induced action framework.

There are many segments of the body which must be taken into account when generating vision induced animation. Obviously the pivotal part of the skeletal framework lies in the orientation of the eyes, as it specifies the direction in which the avatar is looking at the world. The neck, torso and alignment of the spine are next in line. In fact the lateral and vertical movement of the neck joint controls the direction of head movement and in most cases is more important than eye movement as description of the eye is very rarely provided in most character animations.

In certain task driven situations when the character has to look closely towards a distant object, the animation must include intentional bending, either forwards or backwards, as the character appears to peer towards an object. Simply changing the orientation of the head for looking from side to side, makes the animation seem rigid and unrealistic. Therefore we can consider the kinematic framework of a character to be composed of linked chains. In our case the root of this chain is in the eye, as visual perception through the eyes is directly connected to the kinematic framework of the avatar.

The choice of structural hierarchy of a skeleton can be more complicated allowing degrees of freedoms for many components, but this comes at a high computational cost.

There are many ways to represent the body of the avatar and control its movements. One may consider control models based on physically realistic parameters such as

dynamic or optimal control. Though such models strive to generate life like animations, they are extremely difficult to realize as they have complicated and abstract input parameters like muscles and twists.

The use of inverse kinematics has been a popular method to generate animations. **[Courty and Marchand]** used an inverse kinematic approach, where the node of the articular chain was the eyes of the avatar which had 2 degrees of freedom. Movements of the eye where transferred down the to the rest of the chain using the inverse Jacobian of the chain. The advantage of such a process is that it enables a closed loop control model wherein each frame has a new visual update which issues a new movement to the structural framework of the avatar. Thus the system is realistic and reactive to environmental changes on a frame basis. However the disadvantages of inverse kinematic control is that the solution is local and does not guarantee successful reach of the goal.

Another approach that has been used by many researchers is motion capture data obtained from real humans and to apply space constraint techniques to modify motions and create the most appropriate ones. With a number of motion capture data formats being supported by rendering engines, the kind of animation that is possible is most realistic.

*Before proceeding further to justify a choice of skeletal framework, it is worth recalling what the objective of the **AGES** project is in the first place. What the system is focussed on is to develop high level real time vision and gaze control modules that have not been developed or thought of before by researchers of real time animation frameworks , and to use the meaningful output of such modules to provide control information to generate appropriate behaviours. Besides the design of the **AGES** environment allows users to ineractively change scenes, wallpapers, virtual objects and grab the avatar's attention and call the avatar to a particular location by providing hand gestures. Thus most of the avatar's behaviour is predefined, and there is no need to generate new sophisticated motions by warping existing motions, which many cite as being one of the drawbacks related to motion capture data. In view of the above factors, the choice of skeletal representation for the character animation in AGES is hierarchies supported by motion capture formats.*

5.2 The Gaze Controller :

In the perception-action cycle of the AGES system, the function of the vision modules is to provide the avatar with a processed image of the environment that clearly shows regions and objects of interests as identifiable pixels and track points.

This processed visual information is the only input given the gaze controller which has to then modify the character's gait and orientation of head towards the object of interest in order to simulate realistic and attentive visual gaze, the eventual aim of the AGES framework.

Previous work in gaze generation by **[Lee, Marsella, Traum et al]** has enabled the generation of avatar gaze to reflect the internal state of the avatar. Their approach however was used to simulate realisitic conversational agents, and they considered gaze to be a window into the 'mind' of the avatar. This was further extended by **[Lance et al]** to closely relate saccadic eye movements to the corresponding shift in gaze to reflect cognitive intent and emotional state of the avatar.
Clearly, the above mentioned gaze controllers reflect the internal cognitive process of an agent, as opposed to the vision inspired reflexive gaze control required by the humanoid of the AGES framework.

However the method of **[Lance at al]** that reuses two motion capture data corresponding to the same gaze, one emotionally expressive and an another plain, and compute the Gaze Warping Transformation to produce a gaze that combines these two based on the Motion Warping algorithm forwarded by [Popovic et al] is an interesting way of combining motions.

*The AGES system has used a similar concept to attempt and formulate a **new** method to generate arbitrary gaze directions from a set of vertical, horizontal and rotating motions of the head and corresponding body segments obtained through motion capture.*

[[[[[[Simplifying Assumptions: We will not be considering saccading eye movements of the avatar, and instead only the motion of the head and the connected segments of the body will be animated to display the shift in attention experienced by the avatar. As a simplifying assumption for the initial implementation in the AGES system, we will consider all points of attention and movement of objects projected on a 2D plane, that is parallel to the viewing plane of the humanoid.**]]]]]**

5.2.1 Implementation Details:

All the synthetic vision modules that operate on the 3D scene return reusable information as a 2D matrix from which the rough location of a point of interest were obtained. The look-at position of the virtual camera in the AGES system was directed towards this point in order to represent the shift in gaze of the avatar in a first person view. As long as the gaze direction vector of the avatar is aligned along the look at position, a realistic gaze is generated, as depth information is redundant due to the 2D constraint.

But the actual physical shifting of the head of an avatar such as to orientate its gaze towards this look-at position is not a trivial task. *At the first stage of implementation let us consider a stationary avatar looking at a scene with a new point of attention that has emerged from the output of the vision module.*

This is shown in a screen shot from the 'real world vision module' that has just returned a tracking point of a object introduced by the user that, has been projected onto a plane in the virtual world of AGES. The red tracking point of the blob is searched for by scanning the GPU matrix. The lookat position of the camera is then set to the coordinates of the track point.

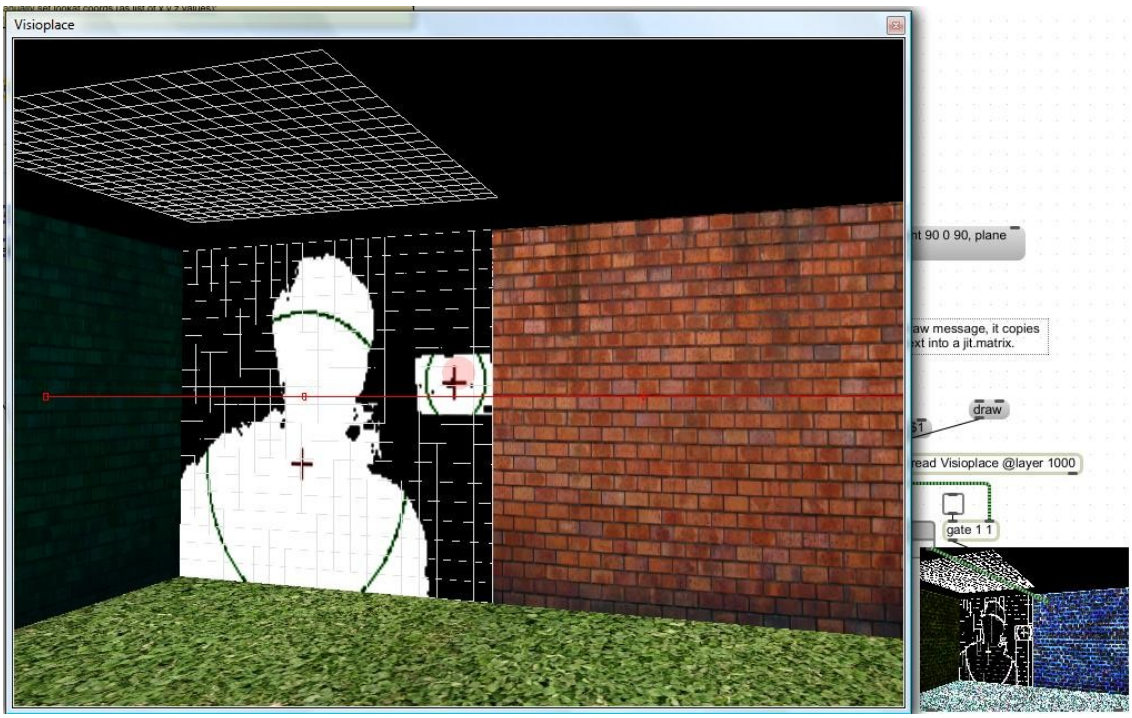


Figure 6: Look at position (red circular bullseye) oriented towards the red tracking point, whose coordinates are obtained by scanning the GPU matrix.

Extending this framework when a humanoid with a skeletal structure is present:
Let the original position of the avatar's look-at position be **(x1, y1)** as projected onto the plane where the tracking point of the object is present.

Let **(x2, y2)** be the new position towards which the attentive gaze vector of the avatar must be orientated.

Now, the motion capture data that have to be obtained are:

1) lateral movement of the avatar's head from left to right
2) the vertical movement of the avatar's head from top to down,
3) circular motion of the avatar human's head around the neck

Considering a cone of a vision of 45 degree, the max and minimum amplitudes of the lateral and vertical movements must be scaled to some values.

This motion capture data will consist of the changes in other segments of the body as well, to provide realistic animation.
The idea is to combine two motions of the head (lateral and vertical corresponding to x and y displacements required to reach a new look-at point) to generate an intentional gaze motion towards that direction.

From **[Popovic et al],** A Motion Warping curve provides a transformation from the keyframes of an animation curve **x(t) defined as a set of value, frame pairs, (x_i , t_i)** to those of a new motion **x' (t') defined as the set of pairs (x_i' , t_i')** through the use of 2 functions :

Let **X (t)** be motion frames corresponding to **vertical up and down motions** of the head.
Let **X' (t)** be motion frames corresponding to the **lateral right to left movement** of the head.

Function 1 : Given t_i', calculate the frame t_i in the vertical motion curve to obtain x_i.

$$t_i = g(t_i')$$

$$g(t_i') = c(t_i')*t_i'$$

where given an actual frame time t_i', in the lateral movement, **g(t)** determines the corresponding frame t_i in the vertical up and down movement through a time scaling parameter $c(t_i')$.

Function 2:

$$X'(t_i) = X(t_i) + b(t_i)$$

here $b(t_i)$ is a spatial offset parameter that transforms $X(t_i)$ into $X'(t_i)$.
The final Gaze Warping Transformation is an **m * n** set of **(c, b)** pairs, where **m** is the number of degrees of freedom in the animated body, and **n** is the number of key-frames in the animation.

5.2.2 Need for an alignment algorithm :

It was observed by researchers that certain artifacts like the head popping out of the body frame were observed when warping two similar head motions because the motion

curves of the two motions at a given time instant where not aligned, ie. Corresponded to slightly different spatial locations of the head.

The values used for alignment are found by the function of v(t) , the three dimensional velocity vector of the head.

There are other factors that need to be considered, like the prevention of body parts moving outside the frame of the body. This may have to be addressed by introducing a rigid skeleton in an inverse kinematics system, to place constraints on the independent data received from the motion sensors. *Anyhow this section of realistic kinematic gaze control is a topic of ongoing research with respect to the AGES framework and shall not be discussed in detail here*.

Chapter 6
Runtime Animation Framework:

6.1 Outset :

This chapter deals with stitching together the various modules that are required by the avatar to exhibit attentive gaze, into a behavioural model that instantiates the different modules to generate an animation during runtime with user interactivity.

There are basically 4 modules that operate in tandem in order to generate realistic attentive gaze activated behaviour in the humanoid of the AGES framework:

1. **Context Switcher** – This is a high level state switching module that activates and controls all the required modules needed to generate behaviour at different contexts. These contexts are automatically activated based on the physical location of the avatar, and alternatively may also be activated by the user at runtime.
2. **Vision Module –** The vision module is responsible for providing the avatar with processed ready to use visual information that is used by the avatar to direct its attention towards objects and points of interesting. A detailed account of the vision modules has already been presented in Chapter 4.
3. **Attentive Gaze Controller** – The gaze controller receives information from the vision module active in a current context and directs the gaze of the avatar towards an object of interest. Although the orientation of the physical head of the avatar is handled by a common kinematic gaze controller, separate image processing is required in each context in order to determine the coordinates of the track points in the image plane. This track point, the final desired look at point, is given to the kinematic gaze controller in order to orientate the gaze towards a new point of interest.
4. **Behavioral FSM :** The behavioural FSM contains the various behavioural states of the avatar, and each behaviour corresponds to a series of motion graphs. This technique has been followed by many researchers who have built a

framework for real time behavioural animation in the past. This is the only pre-computed aspect in the AGES system and has to be physically designed to generate pre-specified behaviours in different contexts.

All the modules operate on a closed loop per frame basis, and at the end of a the execution of the behaviour, the control is transferred back to the vision module.
A closer look at the different contexts that couple the the different modules to generate interesting animations of the humanoid, is dealt with in the next section.

6.2 Contextual Behaviours:

The physical design of the virtual world of AGES is an inseparable part of the behavioural animation model. By an 'unknown' virtual environment we mean that the avatar has no prior knowledge about the nature of visual stimuli it is going to experience and has no information about the objects present in the virtual world. However, we define different contexts based on the physical location of the avatar, so that the appropriate vision modules and associated behaviours may be activated. Therefore it is essential to know the physical design the AGES setup.

The avatar enters a virtual room, with four contexts defined in two rooms of the environment, each with different behaviours. The following sections gives a brief recap of these contexts and mentions ideas for various animations that now may be actuated with ease.

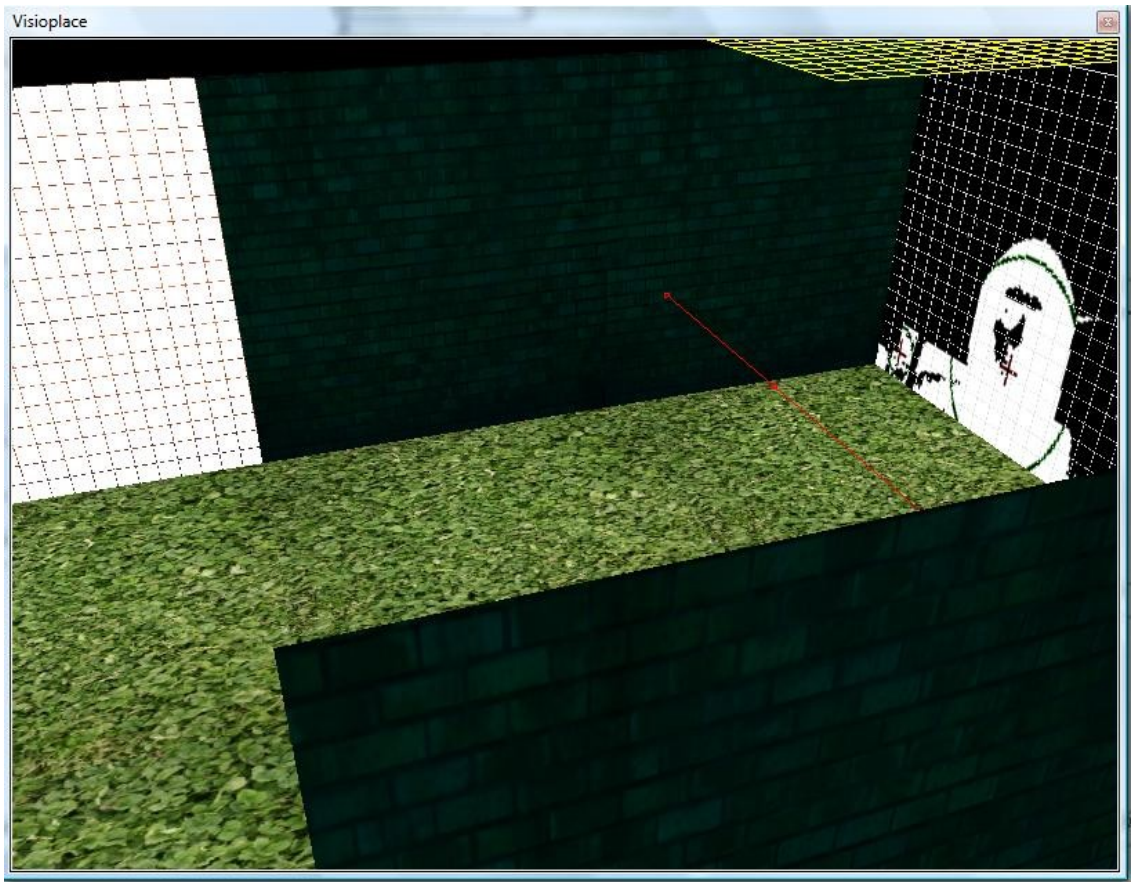


Figure 8: Top Right : Real world interaction context.
Left : Motion Sensing Context

6.2.1 OBJECT FUNCTION CONTEXT: The entrance of the virtual world has object models, chosen and rendered by the user at runtime. In this context the vision modules that will be active are the OBJECT FUNCTION VISION module (the basic low level flat shaded synthetic vision module) and will be activated by the user when he directs the avatar's gaze towards objects. The avatar then makes a note about their functions to launch an action. This is a task driven endogenous attention model where the gaze is not automated, but guided by the user.

A high level task script specifies what objects the avatar should look for. For instance in our sample simulation the virtual world is populated with mushrooms , false colored

with an Object ID, that corresponds to the 'EAT' behaviour that should be encoded in a state of the Behaviour FSM that corresponds to the motion sequence for 'JUMP – GRAB – EAT'

We will not go into details as to how the behaviour FSM executes motions from the motion graph. Many such behavioural frameworks exist, and for details one may refer to the elegant work of [Shum et al] or [Lau and Kuffner et al].

6.2.2 THE MOTION SENSING CONTEXT:

Defined in the top right part of the second room as seen in the top view snapshot presented in **Figure #,** this context is used by the avatar to detect motion and direct an attentive gaze towards a moving object. The Optical Flow based vision module that provides the avatar with the necessary vector field of a segmented object has been discussed in detail in Chapter 4. A flag set by the optical flow motion sensor directs the attention of the avatar towards a moving object. This module demonstrates the emergent attention displayed by the active vision avatar in the AGES system. Previous behavioural animations lack the ability to display such attention when something moves in a scene (for instance rapidly flashing traffic lights), and few animations models which do have basic motion sensing capability have not yet used a gradient based optical flow module in real time. In the AGES system the user can introduce suddenly moving object animations at will by simply selecting a video clip, which will direct the attention of the avatar so that gaze appears synchronized with the movement of the object.

6.2.3 THE REAL WORLD INTERACTION CONTEXT:

An artificial reality portal exists in the bottom corner of the second room where the avatar interacts with a virtual plane rendered in the graphics world that contains a real time feed of external objects and people using a real camera input.

As one of the most exciting contexts in the virtual world of AGES, the real world vision module supplies the avatar with blobs corresponding to multiple objects moving in the real world, and the centroid cross-hair of each blob serves as a point of attention to direct the avatar's gaze.

The work of the gaze controller is simplified in this case as the preprocessed blobs provide accurate tracking points for multiple stationary and moving objects in the real world.

User's can interactively play with the avatar, and various animated behaviours maybe customized into the avatar, like jumping and reaching for a person's hand that may contain an object.
This particular context would prove to be one of the most visually engaging interactive animations of the AGES system.

Chapter 7

Conclusions

This thesis has presented the implementation of the Attentive Gaze Enabling System (AGES), an innovative standalone virtual environmental that enables realistic attentive reactions of an Active Vision agent in an unknown virtual world where the user has the ability to interact and introduce visual stimuli to the avatar in real time.

- It is essential for an Active Vision agent to be able to modify its visual parameters so as to extract meaningful features from a scene in order to execute a task at hand. *The high level vision modules that were implemented as part of AGES do exactly that-- they extract those selected features from the 3D scene and present them in an easily usable pixel matrix form to the gaze control module of the avatar.*

- Changing different visual stimuli, by introducing different moving objects for the optical flow module, and introducing objects from the real world through the blob detection module have successfully directed the visual attention and gaze direction of the avatar towards these objects of interest. *Therefore it is evident that the AGES system can generate emergent behaviour based purely on visual input, thus satisfying the criterion for an autonomous Active Vision Agent system.*

- *The artificial reality created by projecting real time camera feed directly into the 3D rendered world after pre-processing blob centroids of real world objects, is a form of mixed virtual reality that has not been attempted in traditional computer animations before.* The ability of the avatar to interact with such an image from the external world is an example of innovative integration of computer vision and computer graphics using the same underlying hardware.

  Apart from engaging the user to interact with the avatar, this mixed reality module enables users to easily direct the attention of the avatar to arbitrary objects present in the real world.

7.1 Future Directions

- In the current stage of implementation, the AGES framework simulates the gaze direction of the avatar from a first person's perspective, ie. from the avatar's viewpoint. An immediate future implementation is already underway to render a skeletal hierarchy format to simulate realistic humanoid gaze.

- Currently it is not certain if the AGES virtual world will have a single path planning walking humanoid. An interesting alternative would be to have a humanoid located at each of the different contextual locations in the virtual world, while the user navigates in first person mode to these locations. This could result in very engaging user interaction , especially in the mixed reality module, where the user is present in first person as well as in third person in front of the avatar and can induce interesting attentive avatar behaviour.

- The implementation of the saliency map to enable spontaneous viewing, by allowing the avatar to direct its attention towards the most conspicuous object in the scene is still not robust enough. The current module does however include the real time implementation of a Sobel filter previously used to construct saliency maps used for hi-fidelity selective rendering, and this extension to provide selective attention to the avatar of the AGES framework has not been used before in other works.

  Computer animated characters have come a long way ever since the seeds of 3D computer graphics were sown, and there has been an ever increasing need to have not just realistic behaviour but intelligent life-like agents that can react spontaneously to their dynamic environment, and at the same time engage the user into a virtual augmented reality experience by combining real time computer vision techniques.

It is hoped that AGES will soon evolve into a hi-level tool box for simulating varied behaviours of autonomous active vision agents that can react interactively with users.

REFERENCES :

C. Reynolds, Flocks herds, and schools: A distributed behavioral model. In Proc. SIGGRAPH '87, Computer Graphics 21(4), 25-34 (1987).

Tu, X., and Terzopoulos, D. 1994. Artificial fishes:Physics, locomotion, perception, behavior. In *SIGGRAPH'94 Conference Proceedings*. Orlando, FL: ACMSiggraph

D. Isla, B. Blumberg. New Challenges for Character-Based AI for Games. AAAI SpringSymposium on AI and Interactive Entertainment, Palo Alto, CA, March 2002.

Noser, N., O. Renault, D. Thalmann, and N.M. Thalmann.Navigation for digital actors based on syntheticvision, memory and learning. Computer and Graphics,Vol. 19, pages 7-19, 1995.

Yantis, .Stimulus-driven attentional capture and attentional control setting,. *J. Exp. Psychol. Hum. Percept. Perform.* vol. 19, no. 3, pp. 676.681, 1993.

L. Stark and Y. Choi, .Experimental metaphysics: the scanpath as an epistemological mechanism,.in W. Zangemeister, H. Stiehl, and C. Freksa (eds.), *Advances in Psychology: Visual Attention and Cognition*, Elsevier: North-Holland, Chap. 1, 1996.

A. L. Yarbus, *Eye Movements and Vision*, Plenum Press: New York, 1967.

J. **Jonides**, .Voluntary versus automatic control over the mind's eye movements,. *Attention Perform.*
vol. 9, pp. 187.203, 1981.

D. **Kahneman,** *Attention and Efforts*, Prentice-Hall: Englewood Cliffs, NJ, 1973.

M. *Heisenberg*, R. *Wolf*, Vision in *Drosophila*. *Studies of Brain Function* (Springer-Verlag, Berlin, 1984).

S. Culhane and J.K. Tsotsos. An attentional prototype for early vision. In *Proceedings of Computer Vision (ECCV '92)*,volume 588 of *LNCS*, pages 551–562, Berlin, Germany, May 1992.

Crick F, Koch C. Abstract. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. Nature . 1998

[6] L. Itti, J. Braun, D.. Lee, and C. Koch. A model of early visual processing. In *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

B. Blumberg. Go With the Flow: Synthetic Vision for Autonomous Animated Creatures.Proceedings of the First International Conference on Autonomous Agents (Agents'97),ACM Press 1997, pp.538,-539.

B. K. P. Horn. *Robot Vision*. MIT Press, Cambridge, Massachusetts,1986.

Berthold K.P. **Horn** and Brian G. **Schunck.** Determining Optical Flow. Artificial

Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Christopher Peters and Carol O' Sullivan. Bottom-Up Visual Attention for Virtual Human Animation,Proceedings of the 16th International Conference on Computer Animation and Social Agents (CASA.03 2003 IEEE.

Jina Lee, Stacy **Marsella,** David **Traum, Jonathan Gratch, and Brent Lance** The Rickel Gaze Model: A Window on the Mind of a Virtual Human

Peters, C.**, Pelachaud,** C.**, Bevacqua,** E.**, Mancini,** M.**, Poggi, I.**: A model of attention and interest using gaze behavior. vol. 3661, pp. 229–240. Springer, Heidelberg (2005)

Chopra-Kullar, S., **Badler**, N.: Where To Look? Automating Attending Behaviors of Virtual Human Charactors. In: Proc. of 3rd Annual Conf. on Autonomous Agents, pp. 9–23 (1999)

Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge
(1976)

2008 Lance, B., Marsella, S. A Model of Gaze for the Purpose of Emotional Expression in Virtual Embodied Agents. In Padgham, Parkes, Mueller and Parsons (eds.): Proceedings of Autonomous Agents and Multi-Agent Systems (AAMAS) 2008.

Witkin, A., Popovic, Z. 1995. Motion Warping. Proceedings of SIGGRAPH, ACM Press, New York (1995).

A Sloman. 1989. On designing a visual system. ***Journal of Experimental and Theoretical AI*** **1989, pp. 289-337.**

Y. Aloimonos, ``Active Vision Revisited," in Y. Aloimonos (Ed.), *Active Perception*, Lawrence Erlbaum Associates, 1993.

LEE, J., AND LEE, K. H. 2004. Precomputing avatar behavior from human motion data. Proceedings of 2004 ACM SIG- GRAPH/Eurographics Symposium on Computer Animation, 79–87.

Courty,N. Marchand,E. Arnaldi,B. Through the eyes control of a virtual humanoid France Telecom R&D, IRISA /INRIA Rennes;

Peter Longhurst, Kurt Debattista,Alan Chalmers A GPU based saliency map for high-fidelity selective rendering, University of Bristol

Kuffner, J. and J.C. Latombe. Perception-Based Navigation for Animated Characters in Real-Time Virtual Environments. The Visual Computer: Real-Time Virtual Worlds, 1999.