A Memory Model for Autonomous Virtual Humans

Christopher Peters Carol O' Sullivan

Image Synthesis Group, Trinity College, Dublin 2, Republic of Ireland
email: {christopher.peters, carol.osullivan}@cs.tcd.ie

Abstract

A memory model based on "stage theory", the dominant view of memory from the field of cognitive psychology, is presented for application to autonomous virtual humans. The virtual human senses external stimuli through a synthetic vision system. The vision system incorporates multiple modes of vision in order to accommodate a perceptual attention approach. The memory model is used to store perceived and attended object data at different stages in a filtering process.
The methods outlined in this paper have applications in any area where simulation-based agents are used: training, entertainment, ergonomics and military simulations to name but a few.

Keywords: virtual humans, vision, memory, virtual reality

1. Introduction

When modelling agent-object interactions in virtual environments, virtual humans are generally provided with complete access to all objects in the environment, including their precise current states, through the scene database. This is unrealistic conceptually – as real humans we know that life is not that simple. When we are getting dressed in the morning, and need to find the companion to that sock underneath our bed, we do not have the luxury of requesting its whereabouts from a scene database. Instead, we must use our intelligence, knowledge and senses to find it. Obviously, our memory of everything from what a sock looks like, to where we usually keep socks, plays a key role in the process.

If we agree that endowing an agent with the ability to follow this process would improve their level of autonomy, it may also be agreed upon that the purpose of the search is not only important in terms of functionality for the agent that initiates it, but perhaps also just as important in terms of plausibility from the point of view of those avatars who witness it.

Given that an agent is made autonomous in this way, we must then equip that agent with the ability to store useful data and disregard extraneous information. Luckily, it turns out that real humans have a very elaborate system for doing this already. Using perceptual attention, we limit our processing to restricted regions of interest in our environment in order to balance the scales between perception and cognition.

This paper combines a synthetic vision module with a memory model based on "stage theory" [1] to provide a virtual

human with a means of attending to their environment. Attention is very important with respect to memory, since it can act as a filter for determining what information is stored in memory and for how long. We focus on goal driven attention as opposed to stimulus driven attention, since the methods described here are intended for use in an autonomous prehension system.

2. Related Work

Numerous researchers have suggested the use of a virtual model of perception in order to permit agents to perceive their environment [8, 10]. An early example applies group behaviours to simulated creatures [9]. Tu and Terzopoulos [11] implemented a realistic simulation of artificial fishes. Noser et al. [7] proposed a navigation system for animated characters using synthetic vision and memory. Kuffner and Latombe [5] provide real-time synthetic vision, memory and learning, and apply it to the navigation of animated characters, using the synthetic vision system from [7].

Badler et al. [2] propose a framework for generating visual attention behaviour in a simulated human agent based on observations from psychology, human factors and computer vision. A number of behaviours are described, including eye behaviours for locomotion, monitoring, reaching, visual search and free viewing.

Hill [4] provides a model of perceptual attention in order create plausible virtual human pilots for military simulations. Objects are grouped according to various criteria, such as object type. The granularity of object perception is then based on the attention level and goals of the pilot.

3. Synthetic Vision

Our synthetic vision module is based on the model described by Noser et al. [7]. This model uses *false-colouring* and *dynamic octrees* to represent the visual memory of the character. We adopt a similar system to [5], by removing the octree structure. Rather, scene description information is encoded with a vector that contains object observation information.

The process is as follows: Each object in the scene is assigned a single, false colour. The rendering hardware is then used to render the scene from the perspective of each agent. The frequency of this rendering may be varied. In this mode, objects are rendered with flat shading in the chosen false-colour. No textures or other effects are applied. The agent's viewpoint does not need to be rendered into a particularly large area: our current implementation uses 128x128 renderings [Fig. 1]. The false-coloured rendering is then scanned, and the object false-colours are extracted.

We extend the synthetic vision module by providing multiple vision modes. Each mode uses a different palette for false-colouring the objects. The differing vision modes are useful for capturing varying levels of information detail of information about the environment. The two main vision modes are referred to as distinct mode, and grouped mode.

In the distinct vision mode, each object is false-coloured with a unique colour. The unique colours of objects in the viewpoint rendering may then be used to do a look-up of the object's globally unique identifier in the scene database. This identifier is then passed to the memory model. This mode is useful when a specific object is being attended to (Fig 1c).

The other primary vision mode is called grouped vision mode. In this mode, objects are false-coloured with group colours, rather than individual colours. Objects may be grouped according to a number of different criteria. Some examples of possible groupings are brightness, luminance, shape, proximity, and type. The grouped vision mode is useful for lower detail scene perception (Fig 1d, Fig 1e).

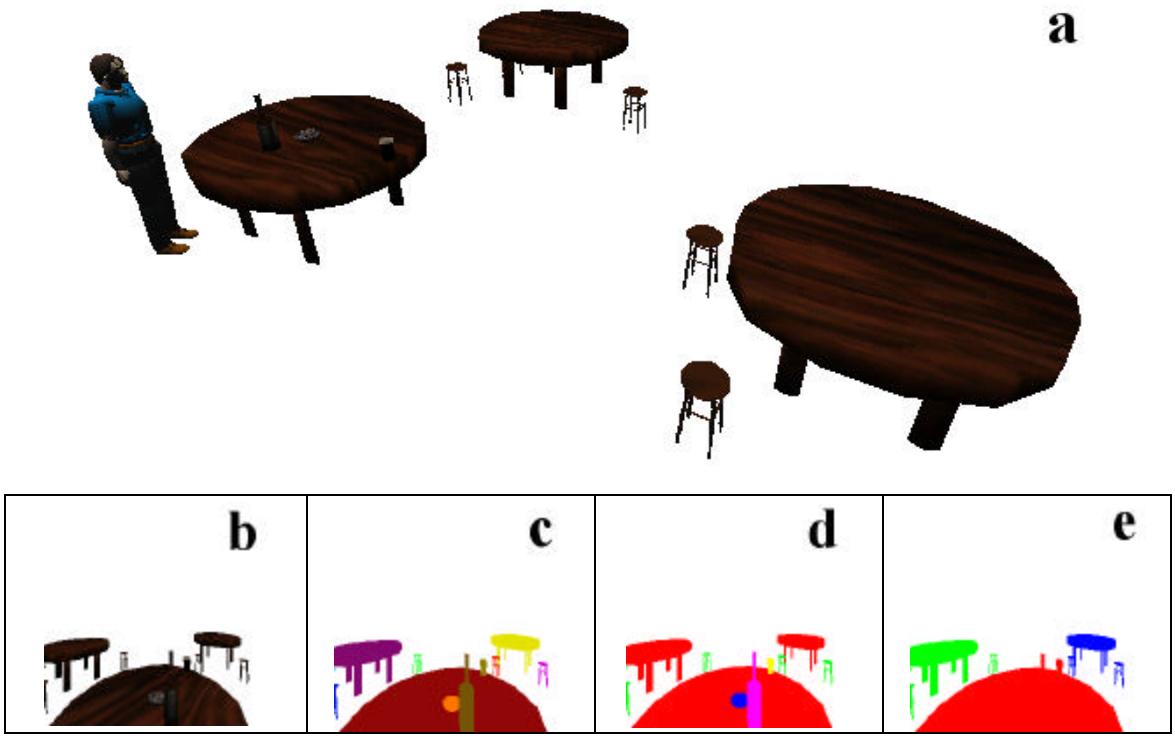

Fig 1(a) A perspective view of a scene containing the agent and a number of objects. (b) – (e) Views as seen from the perspective of the agent with no false colouring applied (b), false colouring according to object id (c), false colouring according to object type (d) and false colouring according to object proximity (e).

The information acquired by the virtual human under the above circumstances is referred to as an observation. In our implementation, the precise position of an object or group in the environment is not stored as part of an observation unless a certain amount of attention has been given to it. Rather, an approximation of the object's location in spherical coordinates with respect to the agent's viewing frame is used. During the scanning process, bounding boxes are assembled for each object based on the object's minimum and maximum x and y coordinates extracted from the view specific rendering, and the object's minimum and maximum z coordinates extracted from the z-buffer for that view. The object's position is then estimated to be the centre of this bounding box. This process has the overall effect of making accurate judgements about the positions of partially occluded objects more difficult. Also, estimates made about the distance to the centre of the object will vary depending on the obliqueness of the object with respect to the viewer.

An observation is represented as a tuple that is composed of the following components:

objID	globally unique identifier of the object
objAzi	azimuth of object
objEle	elevation of object
objDis	distance to object
t	time stamp

A specific object will have at most a single observation per agent. The observation will match the last perceived state of the object, although it must be noted that this may not correspond with the actual current state of

the object. Observations are also stored for groups of objects, using a similar process, where groups are bounded and their positions calculated as above. Finally, it should be noted that when observations are stored as memories (see section 4), their coordinates are expressed in Cartesian rather than spherical coordinates.

4. Memory Model

We base our system of memory on what is referred to as "stage theory" by Atkinson and Shiffrin [1]. They propose a model where information is processed and stored in 3 stages: sensory memory (STSS), short-term memory (STM) and long-term memory (LTM).

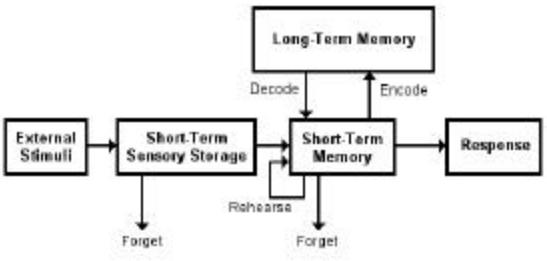

Fig 2 The adopted memory model from [1].

Short-term sensory storage (STSS) is a short duration memory area where a variety of sources of information (e.g. light, smell, sound, etc) are converted into signals that the brain can understand. Since this memory has a very fast rate of decay, it is essential that information be attended to in order to transfer it to the next stage of processing (short-term memory). Our model of STSS only takes account of the visual modality and is derived from the viewpoint rendering discussed previously. Observations extracted from this rendering comprise the STSS. We allow a large number of observations to be stored in the STSS, although it should be noted that only visually sensed items will make it into this memory, and many of these items will be groups of objects rather than individual objects. The STSS is updated with each refresh of the viewpoint rendering.

Short-term memory (STM) relates to our thoughts at any given moment in time. It is created by attention to an external stimulus or internal thoughts. Short-term memory is limited both in duration and by the number of units of information that can be processed at any one time. Research suggests that the STM can process between 7±2 and 5±2 units or chunks of information [6]. These units correspond to letters, numbers, and also larger units such as words and phrases.

Our model allows a maximum of 8 units of storage, where we define a unit as either an object observation or a group observation. Memory entries are removed from the STM under two conditions: they are displaced by newer memories when the STM is full, and they also decay over time (forgetting). The default time allotted to each memory in the STM module is 20 seconds, after which it decays. In the case where the memory entry is rehearsed however, we extend the time allotted to the memory to 20 minutes. Rehearsal occurs when attention is paid to a specific object over a period of time. In general, we assume that the more an item is attended, the longer it will be allowed to stay in the STM. Because we use a goal-directed attention approach, the items that are attended to (and thus, would be expected to occupy the STM) will be those relating to the goal. Take, for example, the goal of searching for the brown bottle object in a scene. At the end of this search, we would expect the STM to contain other bottles that the agent attended, the group containing the brown bottle, and finally the brown bottle object itself.

Long-term memory (LTM) allows long-term storage of information, and generally allows this information to be recalled provided suitable cues are available, although it may take several minutes or even hours to do this. We assume that memories that are stored in the LTM do not expire. Although there are numerous ways to transfer

memories from the STM to the LTM, we assume that only repeated exposure is necessary to encode them into the latter.

Attention is modelled in the system by using the different vision modes to control the detail of the information acquired. When the agent becomes attentive towards an object, that object is rendered in the distinct vision mode mentioned earlier. In this mode, the full object data may be obtained, including its globally unique identifier. The pre-attentive agent state is modelled using the 'group by proximity' vision mode. In this mode, individual objects are not discerned, but rather the states of whole groups of objects are perceived. This type of filtering allows the virtual human, as well as the real human, to reduce large amounts of perceptual data into a manageable size. The 'group by type' vision mode could be viewed as being part of the attention acquiring process. It operates with finer granularity than the 'group by proximity' mode, and is suitable for goal-directed requests by object type (e.g. "take a bottle").

5. Implementation

The implementation of memory is split into a number of separate memory modules: one each for the STSS, STM and LTM. Each memory module is based on memory duration, capacity and a rehearsal value. Unlike the other memory modules, the STSS module also contains the view-port rendering. Each module contains a list of memory entries. A 'memory entry' contains an observation, and other information such as how many times the memory has been rehearsed, and when the last rehearsal took place. The LTM module contains encode (add memory), decode (retrieve memory) and recall (query memory) functions. When an item is retrieved from LTM, it is moved into the STM, overwriting anything currently in the STM. This is useful for modelling a context switch, where the agent's focus of attention is changed. The class hierarchy for the memory system is shown in Fig 3.

Our implementation of the goal driven memory and attention process is summarised as follows:

A goal command is given to the virtual human. This goal command contains the globally unique identifier of the object that attention is to be directed towards. If the object is already memorised in the STM or the LTM, then the observation information is extracted, and the virtual human will become attentive towards (look at) the object and update its perception of the object using the distinct vision mode. If the object was memorised in the STM, this procedure is regarded as a rehearsal.

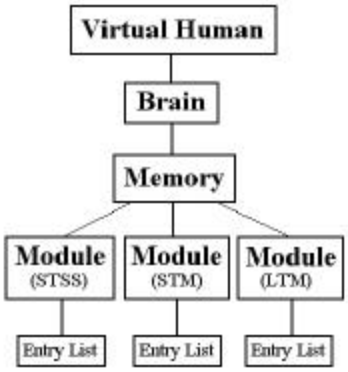

Fig 3 Class hierarchy for implementation of memory

If the object is not in the STM or the LTM, then the agent's perception of the environment will be rendered using the 'group by proximity' vision mode (currently, agents do not initiate an active search of their surroundings; they only search the groups in their view at the time the task is issued). They will then go through the groups in the STM one by one, and render them using the 'group by type' vision mode. If an object of the same type as the requested object is there, then they will become attentive towards the object and will check to see if it is the goal object. If it is not, the search will continue through other objects of similar type in the group, and in the case where there are no more, the search will proceed to other groups. If it is the goal

object, the perceived state of the object is entered in the STM.

The memory model was implemented on the ALOHA animation system [3], an animation system for the real-time rendering of characters. This system uses the OpenGL API on a Windows platform.

6. Conclusions and Future Work

We have presented a memory model that uses a synthetic vision module in order to acquire information about a virtual environment. The granularity at which this information processed by an agent is determined by the use of multiple vision modes. As mentioned, the intended purpose for the memory model is for the implementation of an attention-based prehension system for virtual humans. Aside from modelling virtual human prehension, work will also focus on a realistic visual search algorithm.

References

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

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

[3] Giang T, Mooney R, Peters C, O'Sullivan C, "ALOHA: adaptive level of detail for human animation", Eurographics 2000, Short Presentations, 2000.

[4] Hill RW, "Perceptual Attention in Virtual Humans: Towards Realistic and Believable Gaze Behaviours", Simulating Human Agents, Fall Symposium, 2000.

[5] Kuffner J, Latombe JC, "Fast synthetic vision, memory, and learning models for virtual humans", Proc. of Computer Animation, IEEE, pages 118-127, 1999.

[6] Miller GA, "The magical number seven, plus or minus two: Some limits on our capacity for processing information", Psychological Review, 63, pages 81-97, 1956.

[7] Noser N, Renault O, Thalmann D, Thalmann NM, "Navigation for digital actors based on synthetic vision, memory and learning", Computer Graphics, 19, pages 7-19, 1995.

[8] Renault O, Thalmann NM, Thalmann D, "A vision-based approach to behavioural animation", Visualization and Computer Animation, Vol. 1, pages 18-21, 1990.

[9] Reynolds CW, "Flocks, herds and schools: A distributed behavioural model", Computer Graphics, 21(4), pages 25-34, 1987.

[10] Tu X, "Artificial animals for computer animation: biomechanics, locomotion, perception, and behaviour", PhD thesis, University of Toronto, Toronto, Canada, 1996.

[11] Tu X, Terzopoulos D, "Artificial fishes: Physics, locomotion, perception, behaviour", Proc. SIGGRAPH '94, pages 43-50, 1994.