Appeared in 2nd International Symposium on Smart Graphics, ACM Int. Conf. Proc. Series, pp 124-128, Hawthorn NY USA, June 2002

Jeremiah: The Face of Computer Vision

Richard Bowden
CVSSP, School of ECM
Guildford
Surrey
+44 (0)1483 689838
r.bowden@surrey.ac.uk

Pakorn Kaewtrakulpong
Vision and VR Group
Brunel University
Uxbridge
+44 (0)1895 274000
pakorn.kaewtrakulpong@brunel.ac.uk

Martin Lewin
Vision and VR Group
Brunel University
Uxbridge
+44 (0)1895 274000
martin.lewin@brunel.ac.uk

ABSTRACT

This paper presents a humanoid computer interface (Jeremiah) that is capable of extracting moving objects from a video stream and responding by directing the gaze of an animated head toward it. It further responds through change of expression reflecting the emotional state of the system as a response to stimuli. As such, the system exhibits similar behavior to a child. The system was originally designed as a robust visual tracking system capable of performing accurately and consistently within a real world visual surveillance arena. As such, it provides a system capable of operating reliably in any environment both indoor and outdoor. Originally designed as a public interface to promote computer vision and the public understanding of science (exhibited in British Science Museum), Jeremiah provides the first step to a new form of intuitive computer interface.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces - Graphical user interfaces (GUI), Input devices and strategies, Interaction styles. I.3.6 [**Computer Graphics**]: Methodology and Techniques - Interaction techniques. G.3 [**Probability and Statistics**]: Multivariate statistics. I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding - Intensity, color, photometry, and thresholding, Motion. I.4.6 [**Image Processing and Computer Vision**]: Pixel classification. I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – Color, Motion, Tracking.

General Terms

Algorithms, Security, Human Factors.

Keywords

Human Computer Interaction, Public Understanding of Science, Interactive Virtual Humans, Artificial Life, Computer Vision.

1. INTRODUCTION

The requirements for vision systems capable of locating and tracking humans within an environment vary from intelligent visual surveillance to smart environments and workspaces. Computer vision provides a powerful mechanism for sensing humans, as it is a non-intrusive method of providing substantial information about the location, activities and even intentions of individuals. The benefits have driven extensive research into visual surveillance and behavioral analysis. These systems often provide monitoring with no feedback to the user. However, the development of such technology provides the tools for a new form of interaction with computers; a more natural interface with which to interact.

People natural interact with each other regularly and seamlessly where non-verbal communication forms a considerable amount of this interaction. Such natural interaction is extended to other objects that mimic or resemble humans as people have a natural tendency to humanize objects, toys or virtual agents. It therefore seems an obvious solution to humanize the computer interface as predicted by such creations as HAL from 2001. To this end, researchers are attempting to develop such systems [3, 4,9,13,14,16]. Previous work by authors such as Waters et al [14] and Maes et al [9] has produced similar interactive/responsive systems, the major difference between these approaches and our system is in the vision system used that provides a robust and flexible system that can cope with any environmental conditions.

This paper presents a humanoid interface 'Jeremiah', to what is effectively a visual surveillance system capable of locating and monitoring the movement of humans within any environment. As such it provides an artificial life which has been used within an artistic performance, investigating the interaction between the live and virtual performer. As a public demonstration of computer vision and virtual humans within the remit of increasing the public understanding and awareness of science and to assess the requirements of such interactive systems. Jeremiah has recently been demonstrated at the National Science Museum, London, UK as a exhibit for the general public.

This paper attempts to describe the complete system that is 'Jeremiah', its components and structure. Section 2 gives a general overview of the individual components that constitute Jeremiah. Section 3 describes the computer vision system that is at the heart of his operation. Section 4 presents the graphics system that generates feedback to the user in a natural and intuitive way. Section 5 discusses the emotion engine that generates expressions for feedback generated from visual stimulus. Section 6 discusses the operation of the complete system and future work currently underway to provide a fully interactive virtual human.

2. SYSTEM OVERVIEW

Jeremiah is based around two basic subsystems, a graphics system that constitutes the head and a vision system that allows him to see. There is also a simple emotion engine that responds to visual stimuli via expressions or emotions.

Smart Graphics '02, June 11-13, 2002, Hawthorne, NY, USA.

A video camera is placed which views the workable region in front of the display device. This normally involves a wide-angle lens. A typical multi-user installation is demonstrated in Figure 1. Objects are segmented from the image stream using online background subtraction using a mixture of gaussians to model color variation at a per pixel level within the scene [7,8]. This produces a binary segmentation of foreground objects within the field of view. Shadows are removed from the segmentation by comparing the variation in intensity and chromatisity with those learnt by the model. Foreground objects are then extracted by searching for connected regions within the image. Objects are connected temporally between frames using a simple nearest neighbor scheme to provide inter-frame correspondence and produce a recent trajectory history. Objects are then sorted according to how interesting they are. This 'interest' is determined from the size and velocity of objects over the last *n* frames. A specific object of interest is randomly selected from the scene using 'interest' as probabilistic weighting. Assuming the optical axis of the camera is aligned with the direction of projection, the object position is converted to a direction gaze angle using a pinhole camera model. The head will then turn its interest to gaze at the object while simulating both physical constraints for head/eye movement and incorporating mass, momentum and ambient movement. The emotional response of Jeremiah's head is also driven from visual stimulus. There are 4 key emotions: happiness, sadness, anger and surprise. The strength of an emotion is then visualized by linearly interpolating the facial model towards the strongest emotion.

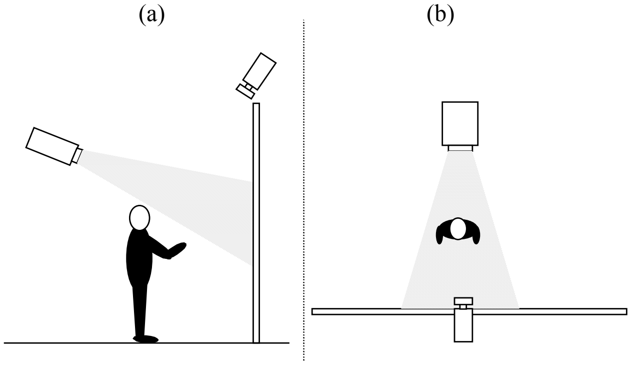

Figure 1. Shows the basic set-up of Jeremiah using a projector mounted high above the audience and a camera that can see both performance space and audience. (a) side view, (b) plan view

3. JEREMIAHS VISION SYSTEM

Jeremiah's vision system is based upon a background segmentation approach developed as part of robust intelligent visual surveillance system [7,8] and is based upon the work of Stauffer and Grimson [5,11,12]. This allows a static background scene to be learnt dynamically, providing the ability to probabilistically label both background and moving foreground objects while allowing subtle alterations to the scene content such as global changes in lighting, sensor noise, moved furniture or variable objects such as trees, vegetation or VDU clicker.

Background subtraction involves calculating a reference image, subtracting each new frame from this image and thresholding the result. What results is a binary segmentation of the image that highlights regions of non-stationary objects. The simplest form of the reference image is a time-averaged background image. This method suffers from many problems due to changing scene structure and therefore a successful approach must constantly re-estimate the background model.

Within the computer vision community there has been a wealth of research performed on this task. For a thorough review of this work the interested reader is directed to [7,8].

The system attempts to model the variation of background by building a Gaussian mixture model of the color distribution for each pixel in the image. This mixture model is iteratively refined for each new frame to allow for subtle changes in scene content. The Gaussian mixture model is learnt using a variation of incremental expectation maximization and produces a probability density function which can be used to probabilistically label the pixels of a new image as either foreground or background[1].

The system produces a binary segmentation of background and foreground regions as shown in Figure 2. However, as the shadow of an object can significantly alter the color distribution of a background pixel, some heuristic is required to determine the effect of these cast shadows. This is done by assuming that a shadow will not significantly alter the chromatisity of a pixel merely change the intensity/brightness. We use an effective computational colour model similar to the one proposed by Horprasert et al [6,7].

Figure 2(a) and (b) shows the results of applying this background subtraction technique to a live image stream. The segmentation runs at full frame rate (25Hz) on a moderately specified desktop PC (1.2GHz PIII) with either a low cost frame capture card or using video for windows and the Microsoft Vision SDK[2]. Figure 2(a) shows the segmentation with pixels deemed static background denoted by white, foreground objects denoted by black and shadows denoted by grey pixels. Figure 2(b) shows this segmentation with shadows removed. Note that the image successfully delineates between the foreground objects (people) and background (floor and walls). The presence of flowers in the middle of the scene (used within the performance) are classified as background as they have remained static for a long period of time (>40 secs). This allows large changes in scene content to be made and results in Jeremiah loosing interest in static objects such as furniture after 40 seconds of inactivity.

[1] For the interested reader the parameterization of the background model uses 5 mixture components per pixel set at a size of 2.5 standard deviations with 80% of the image expected to represent background scene and a learning rate of .001 (i.e. it takes 1000 frames (40 Secs) for a static object to be considered as background). For an explanation of the parameterization see [7 8]. Although the parameterization is to an extent scene specific we have found through experimentation that this parameterization provides consistent results across all installations.

[2] research.microsoft.com

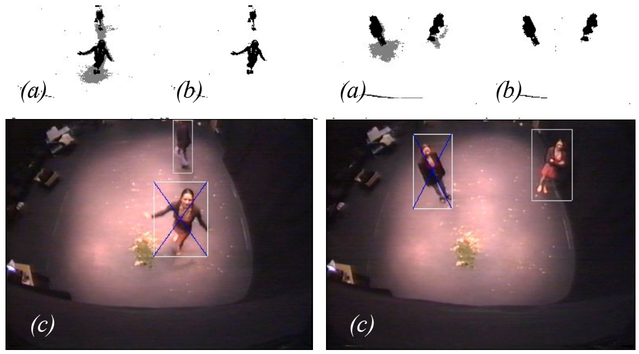

Figure 2. Demonstrates Jeremiah segmenting objects from his field of view for 2 different frames within a sequence. The segmentation (a) shows the results of the background subtraction with shadows shown in grey. The binary segmentation (b) shows these results with shadows removed. (c) Shows the result once all foreground objects have been extracted and sorted according to interest. The intensity of the bounding box represents how interesting the objects are and the x denotes the current object of fixation.

Foreground objects are extracted from the binary segmentation using connected component analysis. The image is searched and connected pixels recursively labeled as individual objects obj_n^t where x_n^t & y_n^t are the centeroid of the n^{th} connected region at time t and the area calculated as the number of pixels constituting the region denoted by a_n. Speed of movement (pixels per frame) for each object is then estimated where

$$s_n^t = \min_{i=0}\left(d(obj_n^t, obj_i^{t-1})\right)$$

and d is the Euclidean distance. The weighted combination of object area and speed provide an interest factor for the object,

$$I_n = a_n + \omega s_n^t$$

The speed of movement obviously introduces a bias for the movement of close objects due to perspective projection. However, the object area a_n already imposes this bias so the effects of perspective projection serve only to reinforce this bias.

A new object of interest is selected at each frame, randomly from those segmented through MonteCarlo selection using the normalized interest factor to weight the selection. The position of the object is converted to a gaze angle using a simple pinhole camera model as shown in Figure 3.

Assuming the optical axis of the camera is aligned with the normal to the display plane, similar triangles allow a gaze angle to be calculated from the position of the object and the focal length of the camera lens. This is possible as we are not concerned with the distance of an object from the camera merely the angle the head must be turned to align with it. Therefore

$$\theta_y = -\tan^{-1}\left(\frac{kx_n^t}{f}\right) \text{ and } \theta_x = -\tan^{-1}\left(\frac{ky_n^t}{f}\right)$$

where k is the pixel scaling of the CCD array.

The optical axis is seldom aligned with that of the display device and often the offset is quiet considerable. This obviously breaks the assumptions on which the model is based but the addition of a scaling and offset terms are sufficient to compensate. This is possible as the system need only produce visually accurate results. The perception of gaze direction is more important than numerical accuracy. The gaze direction is therefore calculated as

$$\theta_y = offset_x - scale_x \times \tan^{-1}(kx_n^t)$$

and similarly for θ_x where *offset* and *scale* are calculated at installation using a simple alignment procedure. The constant k is fixed for a common lens configuration and any discrepancy between k and f at installation are factored into *scale.* The alignment procedure involves manually turning the head to look at a known object in the field of view at predefined positions, namely the center of the working volume and the horizontal and vertical extremities. From this, the head orientation and position of the object within the image can be used to estimate the unknown parameters.

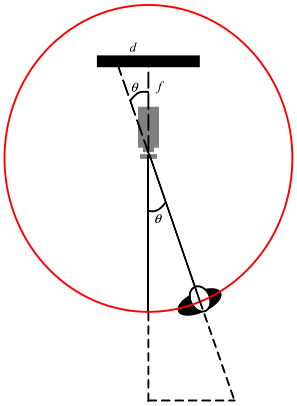

Figure 3. Extracting a gaze direction angle for any object within the field of view using a weak perspective model and simple trigonometry

4. THE VIRTUAL HEAD

Jeremiah is based upon DECface [15] and consists of a simple mesh representing the face with an underlying bone model that allows the mesh to be deformed to reflect the movement of the underlying structure. This provides a lifelike facial avatar, which can be animated to produce varying facial expressions.

Figure 4 shows DECface and the 4 basic expressions (*b-e*) used within the system. Each of these expressions is represented as a target position for the underlying bone structure and linear interpolation used to produce varying strength of any expression.

Figure 5 shows Saul[3] who is Jeremiah's replacement. Saul is an eigen model [1] built from motion keyframes. However, he operates in a similar fashion to Jeremiah and uses linear interpolation at a vertex level to morph between the predetermined target positions for his key emotions.

[3] The Animation Key frames for Saul were provided courtesy of BBC Imagineering, UK

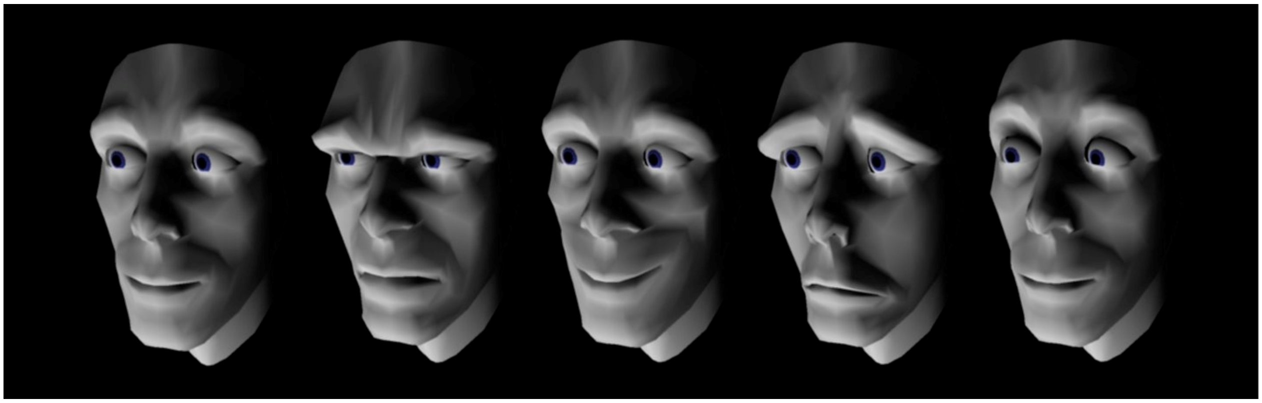

Figure 4. Jeremiah (DECface) and the basic emotions, (a) the default face, (b) anger, (c) happiness, (d) sadness, (e) surprise.

In both cases the response and general animation of the head is the same. The eyes and face are maintained separately. The eyes have a 2.5-degree convergence and are constantly rotated to face the object of interest. The head has a probability of 0.02 that it will randomly turn to follow this eye movement. If the difference between eye and face angle exceeds 25 degrees horizontally or 10 degrees vertically this probability that the head will follow becomes 1.0. This produces very realistic results where the eyes may flick between adjacent objects but turn the head if the objects become too separated. Although the movement of the eyes is instantaneous, the head movement has momentum terms which means the face will lag behind rapid eye movement and may, depending upon velocity/momentum, overshoot and correct the gaze direction. Some ambient movement is randomly introduced and blinking is also randomly triggered at an average rate of once every four seconds. This is reduced to once every six seconds when ‘angry’ and increased to once every three seconds for ‘happiness’ to reflect emotion in a similar manor to humans[10].

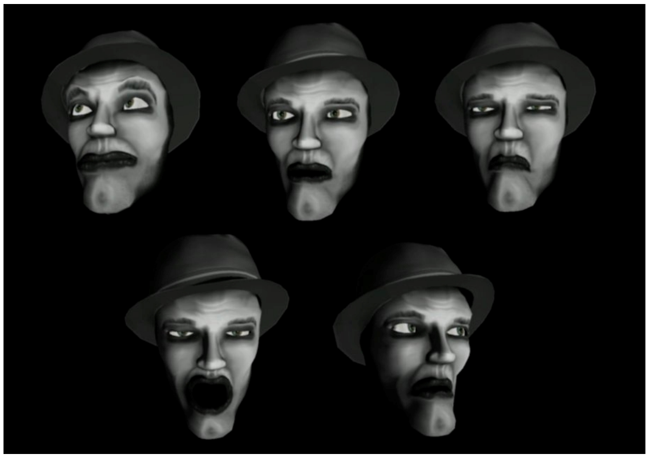

Figure 5. Random faces generated from the Saul eigen model.

5. THE EMOTION ENGINE

The emotion engine probabilistically determines the current state of emotions from simple parameters extracted from objects of interest within the visual field. Each of the four emotions is represented by a probability of 0 to 1. All emotions reduce linearly over time and specific events or activities increase selected emotions. Any emotion in excess of 0.5 is animated by the various mechanisms of the two facial avatars.

Jeremiah likes visual stimulus - high rates of movement make him happy. Jeremiah likes company - no stimulus makes him sad. Jeremiah doesn’t like surprise - high rates of change in the size of objects make him surprised. Jeremiah doesn’t like to be ignored - if objects exist but don’t move then he assumes they are ignoring him and hence gets angry. If Jeremiah experiences too much of any particular stimulus he will get bored with it and reduce its influence on himself.

This simple set of rules allows chaotic behavior in a similar fashion to flocking simulations where a set of rules in a complex environment can produce what appears to be emergent behavior.

6. FINAL SYSTEM AND FUTURE WORK

Although Jeremiah is quiet simple in both construction and operation he captures peoples attention through his life-like responses and apparent awareness of events. People want to believe that he is more than he is and as such read far more into his direct response to their activities. Although he only possesses 4 key emotions his behavior suggests more subtle understanding than he actually possesses. His behavior is similar to a child watching, learning and responding to the world. The random elements and nature of the vision system mean that he never responds to the same stimulus in the same way twice. However, his power lies in the robust underlying vision system which is capable of constantly adjusting to variation in scene structure and content. If an object is placed infront of Jeremiah he will remain interested in it until such time as his vision system learns that it has become static. When this happens it ceases to be of any interest and becomes included in the background model. If the object is then removed Jeremiah automatically remembers what the scene looked like before the object was present (due to the ability of the system to model multiple colors for each pixel) and therefore suffers no confusion with anti-objects or ghosts which would result from simpler background subtraction methods. Fore example, the system camera can often see the projector that produces the head but the model is capable of learning the variation of light from the projector and discounting it as static background.

Given a number of objects of interest Jeremiah will share his attention between them unless through relative size or motion one specific object becomes the dominant object of interest.

It is impossible to provide results which demonstrate the realism of the interactions which can develop in a static paper format however the interested reader is encouraged to visit http://www.ee.surrey.ac.uk/Personal/R.Bowden/jeremiah/jeremiah.html where movies are available that demonstrate him in operation.

Jeremiah is a first step towards a more intuitive and cognitive interface with computers and has demonstrated that people find him very intuitive and natural to interact with. Saul is currently being developed to completely replace Jeremiah. In addition to the emotional key frames Saul has visemes which correspond to the phonemes of speech and provide the mechanism to lip sync animation to synthesized speech. In addition, voice recognition is now a well understood and a successful tool that can be incorporated into the system. Furthermore, we will incorporate biometric recognition and authentication to provide a personalized experience. Current research is also concentrating on a full interactive body model that will extend our previous work on markerless motion capture [2] and human animation [1] to produce a fully interactive human. We are currently extending our previous work into sign language and gesture recognition to develop a sign language translator to convert one national sign language to another. Here Saul will provide the facial animation and emotions along with a body avatar to convey sign back to a deaf human.

As a public demonstrator Jeremiah has been very successful in communicating computer vision applications/research and has recently been installed in the London Science Museum. For this simple initial demonstrator, the public perception has demonstrated that computer interfaces which people can relate to, not only capture imaginations but are more readily accepted as a method for conveying information.

7. REFERENCES

[1] Bowden R, Learning Statistical Models of Human Motion, IEEE Workshop on Human Modelling, Analysis and Synthesis, CVPR2000, Hilton Head Island, 16 July 2000.

[2] Bowden, R., Mitchell, T., A., Sarhadi, M., Non-linear Statistical Models for the 3D Reconstruction of Human Pose and Motion from Monocular Image Sequences. Image and Vision Computing, 18(9), pp729-737, June 2000.

[3] Breazeal, C. and Scassellati, B., "A context-dependent attention system for a social robot". In Proc. Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99). Stockholm, Sweden. 1146—1151, 1999.

[4] Devin, V., E., Hogg, D., C., Reactive Memories: An Interactive Talking-Head. In: proc British Machine Vision Conference, Machester 2001.

[5] Grimson, W., Stauffer, C., Romano, R., Lee, L., Using adaptive tracking to classify and monitor activities in a site. in Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231). IEEE Comput. Soc. 1998. 1998.

[6] Horprasert, T., Harwood, D., Davis, L.S. a statistical approach for real-time robust background subtraction and shadow detection. in IEEE ICCV'99 FRAME-RATE WORKSHOP. 1999.

[7] KaewTraKulPong, P. Bowden, R., Adaptive Visual System for Tracking Low Resolution Colour Targets. in The British Machine Vision Conference. University of Manchester. 2001.

[8] KaewTraKulPong, P., Bowden, R., An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. in 2nd European Workshop on Advanced Video-based Surveillance Systems.Kingston upon Thames. 2001.

[9] Maes, P., Darrell, T., Blumberg, B., Pentland, A. The ALIVE system: Wireless, full-body interaction with autonomous agents, In: ACM Multimedia Systems (1996)

[10] Moore, G., Talking Heads: Facial Animation in The Getaway. Gamasutra Feature, April 2001, http://www.gamasutra.com/features/20010418/

[11] Stauffer C, Grimson W. E. L. Adaptive background mixture models for real-time tracking. in Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149). IEEE Comput. Soc. Part Vol. 2, 1999.

[12] Stauffer C, Grimson W. E. L., Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000. 22(8): p. 747-57.

[13] Thórisson, K. R. Layered Modular Action Control for Communicative Humanoids. Computer Animation '97, Geneva, Switzerland, June 5-6, 134-143. 1997.

[14] Waters, K., Rehg, J., Loughlin, M., Kang, S. B., Terzopoulos, D. Visual Sensing of Humans for Active Public Interfaces, In: Computer Vision for Human-Machine Interaction, Cipolla, R. and Pentland, A. (eds), Cambridge University Press (1998) 83-96

[15] Waters, K.: A muscle model for animating three-dimensional facial expressions. In: Computer Garphics (SIGGRAPH '87), 21(4) (July 1987), 17-24.

[16] Wren, C., Azarbayejani, A., Darrell, T., Pentland, A., Pfinder: Real-Time Tracking of the Human Body, published in IEEE Transactions on Pattern Analysis and Machine Intelligence July 1997, vol 19, no 7, pp. 780-785