Interpretation of Group Behaviour in Visually Mediated Interaction

Jamie Sherrah, Shaogang Gong, A. Jonathan Howell
and Hilary Buxton

CSRP 520

May 2000

ISSN 1350–3162

UNIVERSITY OF
SUSSEX
AT BRIGHTON

Cognitive Science
Research Papers

Interpretation of Group Behaviour in Visually Mediated Interaction *

Jamie Sherrah and Shaogang Gong
Department of Computer Science, Queen Mary and Westfield College, London E1 4NS, UK
[jamie|sgg]@dcs.qmw.ac.uk

A. Jonathan Howell and Hilary Buxton
School of Cognitive and Computing Sciences, University of Sussex, Brighton BN1 9QH UK
[jonh|hilaryb]@cogs.susx.ac.uk

Abstract

While full computer understanding of dynamic visual scenes containing several people may be currently unattainable, we propose a computationally efficient approach to determine areas of interest in such scenes. To this end, we present methods for modelling and interpretation of single- and multi-person human behaviour in real time to control video cameras for visually mediated interaction. We demonstrate that while environments containing a single person are relatively simple, interpretation of multi-person scenarios is much more difficult.

1 Introduction

Machine understanding of human motion and behaviour is currently a key research area in computer vision, and has many real-world applications. *Visually Mediated Interaction* (VMI) is particularly important to applications in video telecommunications. VMI requires intelligent interpretation of a dynamic visual scene in order to determine areas of interest for economical communication to a remote observer.

Ongoing research at the MIT Media Lab has shown some initial progress in the modelling and interpretation of human body activity [8, 13, 14]. Computationally simple view-based approaches to action recognition have also been proposed [1] and similar attempts have been made at Microsoft Research [12, 3]. However, these systems do not attempt intentional tracking and modelling to control active cameras for VMI. Previous work on vision-based camera control has been based on off-line execution of pre-written scripts of a set of defined camera actions [9]. Here we propose to model and exploit head pose and a set of "interaction-relevant" gestures for reactive on-line visual control. These will be interpreted as user intentions for live control of an active camera with adaptive view direction and attentional focus. In particular, pointing with head pose as supporting evidence for *direction* and waving for *attention* are important for deliberative camera control. The reactive camera movements could provide the necessary visual context for applications such as group video-conferencing as well as automated studio direction.

Our general approach to modelling behaviour is appearance-based in order to provide real-time behaviour interpretation and prediction [6, 5, 11]. In addition, we only use views from a single pan-tilt-zoom camera with no special markers to be worn by the subjects. We are not attempting to model the full working of the human body. Rather we will aim to exploit approximate but computationally efficient techniques. Such models should be able to support partial view-invariance, and be sufficient to recognise people's gestures in dynamic scenes. Task-specific representations need to be used to avoid unnecessary computational cost in dynamic scene interpretation [2, 10].

In this work, we define high-level models for human behaviour and use these models to interpret behaviour for VMI tasks involving first a single subject and then multiple subjects. In Section 2, we introduce the set of VMI-relevant behaviours that can already be robustly computed. We show in Section 3 how tracking and behaviour detection for a single person can be used to interpret their behaviour for control of a movable camera and demonstrate an existing real-time system. It is pointed out, however, that with only a single person in the field of view, such interpretation is quite simple. In Section 4, the possibilities for interpretation of a scene containing multiple people are examined. A framework for interpreting simultaneous behaviours of multiple people is presented and demonstrated. In Section 5 the important issues of such an approach are discussed, and the conclusion is given in Section 6.

* Part of this work is funded by EPSRC ISCANIT project GR/L89624

2 Modelling Human Behaviour for VMI

For our purposes, *human behaviour* can be considered to be any temporal sequence of body movements or configurations, such as a change in head pose, walking or waving. However, the human body is a complex, non-rigid articulated system capable of almost infinite spatial and dynamic variations. When attempting to model human behaviour, one must select the set of behaviours to be modelled for the application at hand. Further, the level of complexity of the modelling should be concomitant with its purpose. For instance, the behaviour model required for an automatic light sensor is extremely simple: it need only recognise that motion has occurred. For the implementation of real-time systems, it is of paramount importance that only the minimum amount of information is computed to adequately model human subjects for the task at hand. In this section, salient behaviours are defined for visually mediated interaction tasks.

Implicit Behaviour

For VMI tasks, our system needs to identify regions of interest in a visual scene for communication to a remote subject. Examining the case in which the scene contains human subjects involved in a video conference, the subject(s) currently involved in communication will usually constitute the appropriate focus of attention. Therefore visual cues that indicate a switch in the chief communicator, or *turn-taking*, are most important. Gaze is quite a significant cue for determining the focus of communication, and is approximated by head pose. Gaze and other uses of body language that indicate turn-taking are generally performed unconsciously by the subject. We define *implicit behaviour* as a body movement sequence that is performed subconsciously by the subject.

We adopt head pose as our primary source of implicit behaviour in VMI tasks. Head pose at each time instant is represented by a pair of angles, yaw (azimuth) θ and tilt (elevation) ϕ. Our previous work shows that yaw and tilt can be computed robustly in real-time from 2D images of limited resolution [7].

Explicit Behaviour

Head pose information is insufficient to determine a subject's focus of attention from a single 2D view, due to loss of 3D information. Therefore it is necessary to have the user communicate explicitly with our VMI system through a set of pre-defined behaviours with vague semantics attached to them. Let us define *explicit behaviour* as a sequence of body movements that are performed consciously by a subject in order to highlight regions of interest in the scene. We use a set of pointing and waving *gestures* as explicit behaviours for control of the current focus of attention. We have previously shown that these gestures can be reliably detected and classified in real-time [6]. Specifically, a model $\mathbf{m}_i$ is maintained for each of N gestures under consideration, $i = 1, \ldots, N$, and at time t a likelihood $p(\mathbf{x}(t)|\mathbf{m}_i)$ is generated for each model that the given gesture has just been completed. These N likelihood values can be thresholded to detect a gesture, or can in themselves be considered as model outputs for explicit behaviour.

Human Behaviour

Given that both implicit and explicit behaviours can be measured from human subjects in a scene, these sources of information can be combined to form a temporal model for human behaviour. Let us define $\mathbf{b}(t)$, the *behaviour vector* of a subject at time t to be the concatenation of measured implicit and explicit behaviours. For our purposes, the behaviour vector is the concatenation of gesture model likelihoods and head pose angles:

$$\mathbf{b}(t) = [p(\mathbf{x}(t)|\mathbf{m}_1), \ldots, p(\mathbf{x}(t)|\mathbf{m}_N), \theta(t), \phi(t)]^{\mathrm{T}} \qquad (1)$$

3 Interpreting Individual Behaviour

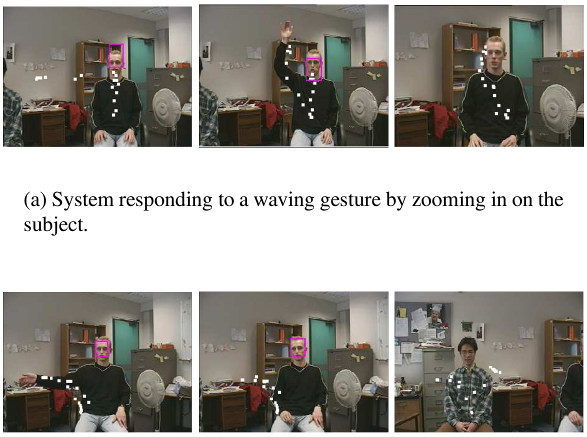
We now consider the task of interpreting the defined behaviours of a single person in the field of view. For the time being we ignore head pose and concentrate on gestures. Consider the case in which a user is remotely communicating with another person via a video camera. This scenario can be extended to cope with multiple people at either end of the communication channel by assuming only one person in the field of view at a time, but panning the camera to focus on the appropriate person. In the single-person case, there are few options for the region(s) of interest and corresponding interpretation.

For instance, the user's face is the most important area for communication. The user may wave to the camera to get its attention and have it zoom in for a close up. This is demonstrated in Figure 1(a). The user in focus may wish to prompt the camera to focus on another user in the room, but currently out of view of the camera. In this case the user would point to the other user, indicating that the camera should pan around. This is shown in Figure 1(b).

This example indicates that detection of defined singleton behaviour seems feasible. However, given the assumptions about the environment, there are only a few possible areas of interest in the scene, making interpretation trivial and not very interesting. On the other hand, the situation becomes much more complex and ambiguous when simultaneously tracking multiple people and their behaviours.

4 Interpretation of Group Activities

Although the individual interpretation of behaviours is possible by attaching pre-defined semantics in the form of



(a) System responding to a waving gesture by zooming in on the subject.

(b) System responding to a pointing gesture by panning around to another user.

Figure 1. Example of a real-time VMI system for a single person in the field of view. Each white square indicates the centroid of the motion field for a single frame. These centroids were among the features used to recognise the gestures.

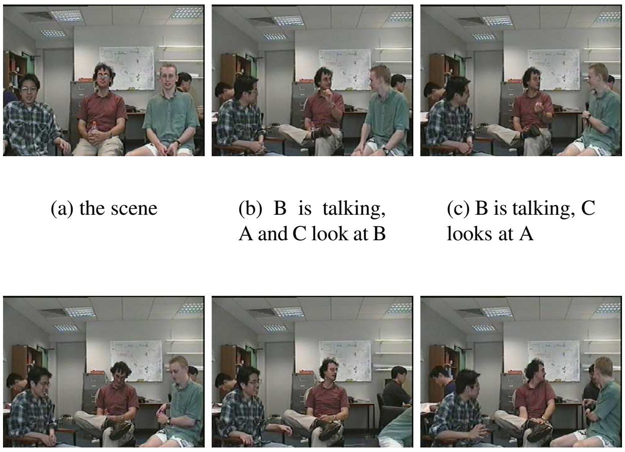
camera control commands, the case of multiple subjects is not so simple due to the combinatorial explosion of possibilities. These possibilities not only include variations in which behaviours occur simultaneously, but also in their timing and duration.

Let us now examine VMI situations which include three people simultaneously in the visual scene, communicating with each other and with the viewer of the camera footage.

4.1 Characteristics of Group Behaviour in VMI

To illustrate possible scenarios involving three people in a VMI situation, we seated three subjects in front of a video camera to have a normal conversation. In the background, two other subjects were working at their desks. The scene is shown in Figure 2(a). We label the three foreground subjects from left to right as A, B and C.

The most common behavioural pattern is for one person to speak, while the other two subjects look on. An example is shown in Figure 2(b). However, some subjects may look at the floor, or at the response of the other subjects. An example is in Figure 2(c). The person talking can exhibit a whole range of behaviours, such as emphatic hand gestures, large changes in head pose, looking at the listeners for a response, or looking at the floor while concentrating on what he is saying. When nobody is speaking, the subjects sometimes all look at the floor, as in Figure 2(d).



(a) the scene

(b) B is talking, A and C look at B

(c) B is talking, C looks at A

(d) nobody is talking

(e) C shows A and B something on the computer

(f) another subject in the background speaks

Figure 2. Examples of human behaviour during a three-person conversation.

At one point in the conversation, C wants to show A and B something of interest on the Internet; a computer sits off to the right of the view. As shown in Figure 2(e), C leans out of view and A and B both look over at the computer with keen interest. Another example of external influences of interest occurs when one of the background subjects interrupts the conversation, as shown in Figure 2(f). In this case, all three foreground subjects turn to look at the interrupter. In particular, C's head pose is a major indicator of the region of interest since he must change his pose the most.

In addition to these possibilities, there are many other individual behaviours that complicate the situation. For example, coughing, scratching, gesticulating, nodding, briefly interjecting a word or two, crossing the legs or putting the hands behind the head all count as motion events, but may have no real communicative significance for our target application.

Clearly the range of possibilities and ambiguities present a problem for any single visual cue, no matter how accurately it can be computed. For example, head pose would fail in many cases in the example presented here because subject A is already facing B and C, and tends to shift gaze rather than by turning his head. It is only by fusing different visual cues that we can hope to successfully interpret the scene.

4.2 High-Level Group Behaviour Interpretation

Now we describe a methodology for machine understanding of group behaviours. Given the complexities of the unconstrained multi-person environment described above, we examine a more constrained situation. We assume a fixed number N of people who remain in the scene at all times. Let us define the *group vector* to be the concatenation of the N behaviour vectors of these people at time t:

$$\mathbf{g}(t) = [\mathbf{b}_1(t)^{\mathrm{T}}, \mathbf{b}_2(t)^{\mathrm{T}}, \ldots, \mathbf{b}_N(t)^{\mathrm{T}}]^{\mathrm{T}} \tag{2}$$

The group vector is an overall description of the scene at a given time instant. Now we define a *group behaviour* as a temporal sequence of group vectors, $[\mathbf{g}(t_1), \mathbf{g}(t_2), \ldots, \mathbf{g}(T)]$. Given a group behaviour, we introduce a *high-level interpretation model* to determine the current area of focus. Since the region of interest is almost always a person and we track the head of each individual, the output need only give an indication of which of the N people are currently attended to. Therefore we define the output of the high-level system to be the *camera position vector*:

$$\mathbf{c}(t) \quad = \quad [f_1, f_2, \ldots, f_N] \tag{3}$$

where f_i is a boolean value (0 or 1) indicating whether person i is currently attended to. An interpretation can then be placed upon $\mathbf{c}(t)$ to control the movable camera, examples are given in Table 1. Clearly such a scheme would require two cameras, one to frame the whole scene for tracking of all individuals, and the other for taking close-up shots. Here we use a “virtual camera” by cropping focal regions from the global image. The high-level interpretation model must transform a recent history of group vectors into a camera vector for the current scene. However, without the feedback to retain the previous focus of attention, the system will lack the context to correctly interpret behaviour. For instance, if a subject waves to gain focus of attention, the camera vector must remain on the subject until another subject attracts attention. Without feedback, the subject would lose the focus of attention as soon as the gesture has ended. Therefore the general form of the high-level interpretation system $F()$ is:

$$\mathbf{c}(t) = F(\mathbf{g}(t), \mathbf{c}(t-1)) = F(\mathbf{s}(t)) \tag{4}$$

where $\mathbf{s}(t)$ is the *scene vector* at time t, defined as the concatenation of the current group vector and previous camera vector, $\mathbf{s}(t) = [\mathbf{g}(t), \mathbf{c}(t-1)]$.

Given this model, the high-level interpretation system must perform the translation from behaviours to focus of attention based on a fusion of external semantic definition and statistics of behaviours and their timings. The semantics may come from a set of rules, but an exhaustive specification of the system would be infeasible due to the multiplicity of possible co-occurring behaviours and their timings. We take a supervised learning approach: the system is trained on a set of example group behaviours, with the aim of generalising to new group behaviours. To learn the transformation from scene vector to camera position vector, we used a Time-Delay RBF Network [4], trained on half of our sequence database and tested on the other half.

We constrain the complexity of the task by restricting the group behaviours to certain fixed scenarios. Let us define a *scenario* to be a group behaviour in which the subjects perform gestures and change their head pose in a fixed pre-defined order. The exact timing of the events will vary between different instances of the same scenario, but the focus of attention should switch to the same regions at approximately the same times. Descriptions of example scenarios involving three subjects are given in Table 2. Several examples of each scenario were collected, and training examples were labelled by hand with a camera position vector for each scene vector. A high-level system consisting of a recurrent RBF network was trained on these examples and then tested on a different set of test instances of the same scenarios.

Figures 3–6 show examples of the system output for two example scenarios: **wave-look** and **point**. Figures 3 and 5 show temporally-ordered frames with boxes framing the head, face and hands being tracked. In each frame, head pose is shown above the head with an intuitive dial box. Figures 4 and 6 show the head pose angles (top) and gesture likelihoods (middle) for persons A, B and C (from left to right). One can see the correspondence of peaks in the gesture likelihoods with gesture events in the scenario.

$\mathbf{c}(t)$	interpretation
`[0,0,0]`	frame whole scene
`[1,0,0]`	focus on subject A
`[0,1,1]`	focus on subjects B and C using a split-screen effect

Table 1. Example of possible interpretations of camera position vectors for three people.

The bottom sections of Figures 4 and 6 show the training signal, or target camera vectors, traced above the actual output camera vectors obtained during tests with the trained RBF network. It can be seen that the network follows the general interpretation of group behaviour, though the exact points of transition from one focus of attention to another do not always coincide. These transition points are highly subjective and very difficult to determine, even with manual coding.

5 Discussion

We have shown an example of how multi-person activity scenarios can be learned from training examples and inter-

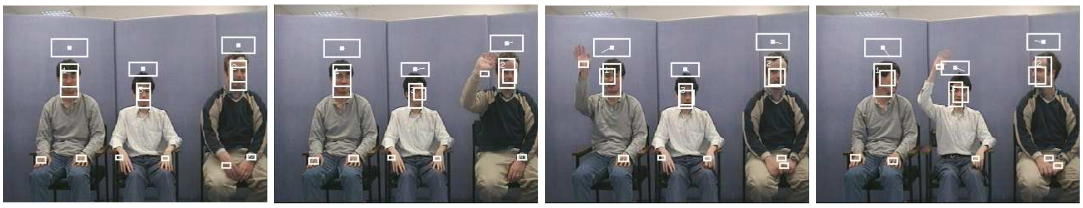


Figure 3. Frames from wave-look sequence. Individuals are labelled A, B and C from left to right.

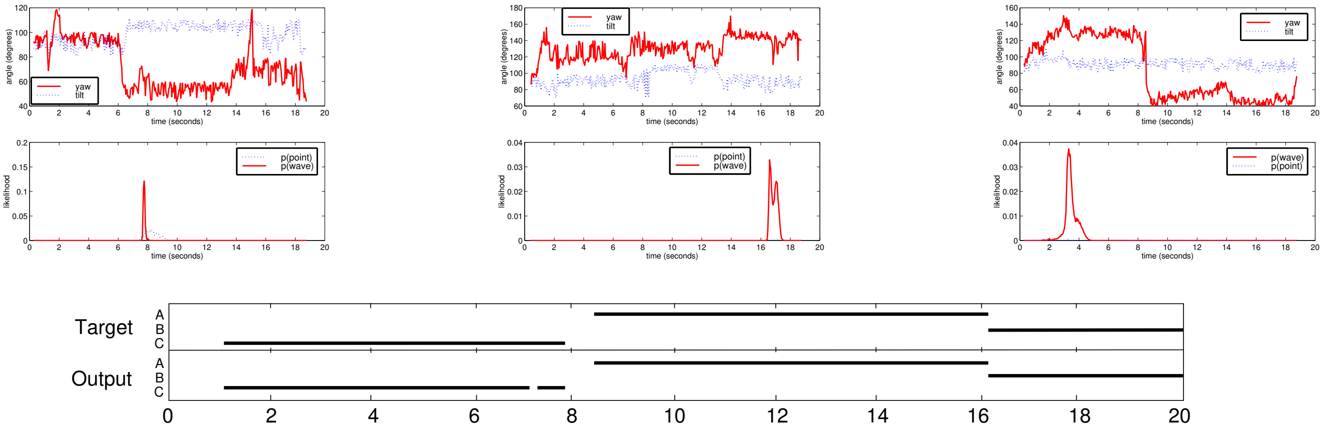


Figure 4. Results for wave-look scenario. Plots show pose angles (top) for persons A, B and C from left to right, gesture likelihoods (middle) and target/output camera position vectors (bottom).

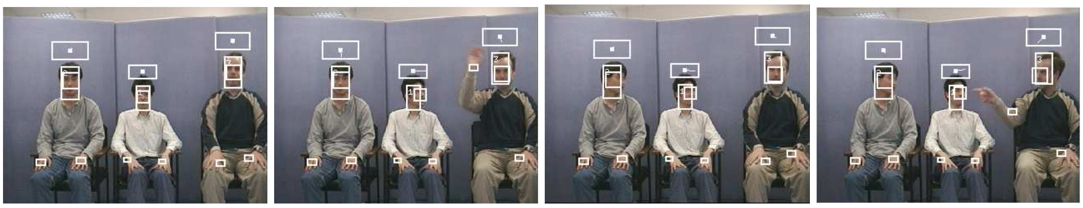


Figure 5. Frames from point sequence. Individuals are labelled A, B and C from left to right.

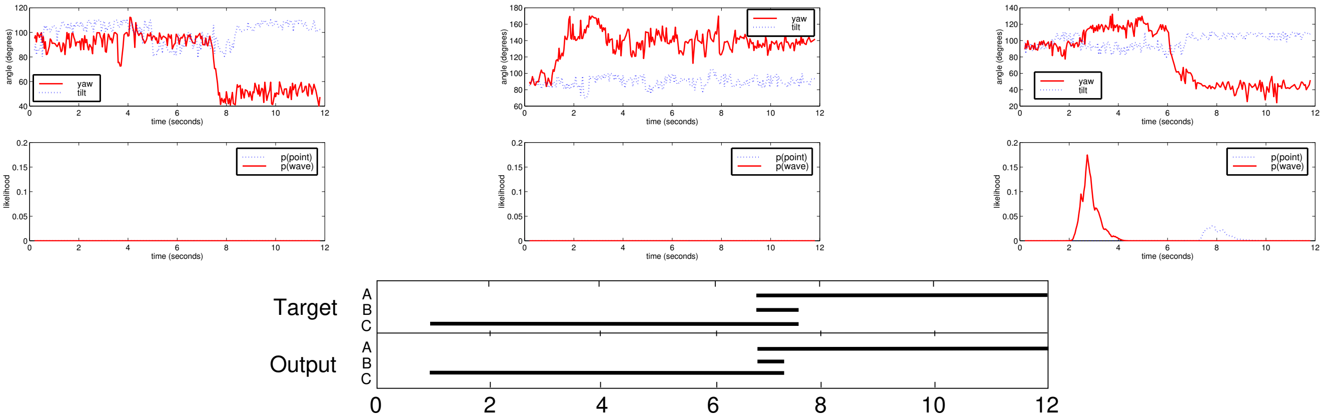


Figure 6. Results for point scenario. Plots show pose angles (top) for persons A, B and C from left to right, gesture likelihoods (middle) and target/output camera position vectors (bottom).

scenario	description
wave-look	C waves and speaks, A waves and speaks, B waves and speaks. Each time someone is speaking the other two subjects look at him
point	C waves and speaks, A and B look at C, C points to A, C and B look at A, A looks at camera and speaks

Table 2. The example scenarios described in temporal order of their behaviours. All subjects are looking at the camera (forward) unless stated otherwise.

polated to obtain the same interpretation for different instances of the same scenario. However for the approach to scale up to more general application, it must be able to cope with a whole range of scenarios. The approach implicitly requires such a system to extrapolate to novel situations in the same way as a person. However, there is no reason to believe that current computer architectures are capable of such reasoning. Therefore a significant issue addressed in this paper and in future work is the feasibility of learning correlated temporal structures and default behaviours from sparse data.

Another issue with the machine learning approach to multi-subject behaviour interpretation is the feasibility of collecting sufficient data. The multiplicity of possible events increases exponentially with the addition of extra subjects. Therefore it is difficult to know which scenarios to collect beforehand in order to evenly populate the space of possible scenarios with the training set. Also, the training set needs to be manually labelled which is extremely time consuming. There are several avenues of investigation which may yield solutions to these problems. The use of high-level models such as Bayesian belief networks allows a combination of hand-coded *a priori* information with machine learning to ease training set requirements. Clustering techniques could be used to select the most important scenarios before hand-labelling. Adaptive training could be adopted so that an inadequate training set is used initially, and the system is manually "corrected" afterwards during operation.

Since this system relies on several independent components, the overall probability of failure of at least one component is always quite high. This has consequences for the high-level interpretation system. First, the system must be able to cope with missing or noisy inputs, such as a head tracker that has lost lock. It is likely that not all low-level information is required to determine the focus of attention. Second, the system outputs may be fed back to the low-level sub-systems to guide them in their processing, ie. indicating what to look for. Such properties would imbue the system with some semblance of real intelligence.

6 Conclusion

The key issues have been explored and a framework presented for tracking people and recognising their correlated group behaviours in VMI contexts. Pre-defined gestures and head pose of several individuals in the scene can be simultaneously recognised for interpretation of the scene. When there is only a single person present in the view, interpretation of behaviour can be quite trivial to achieve computationally. In the presence of multiple people, however, ambiguities arise and a high-level interpretation of the combined behaviours of the individuals becomes essential.

References

[1] A. F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Proc. Royal Society London, Series B*, 352:1257–1265, 1997.

[2] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78:431–459, 1995.

[3] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Proc. FG'98*, pp. 416–421, Nara, Japan, 1998.

[4] A. J. Howell and H. Buxton. Recognising simple behaviours using time-delay RBF networks. *Neural Processing Letters*, 5:97–104, 1997.

[5] A. J. Howell and H. Buxton. Learning gestures for visually mediated interaction. In *Proc. BMVC*, pp. 508–517, Southampton, UK, 1998. BMVA Press.

[6] S. J. McKenna and S. Gong. Gesture recognition for visually mediated interaction using probabilistic event trajectories. In *Proc. BMVC*, pp. 498–507, Southampton, UK, 1998.

[7] E. Ong, S. McKenna, and S. Gong. Tracking head pose for inferring intention. In *European Workshop on Perception of Human Action*, Freiburg, June 1998.

[8] A. Pentland. Smart rooms. *Scientific American*, 274(4):68–76, 1996.

[9] C. Pinhanez and A. F. Bobick. Approximate world models: Incorporating qualitative and linguistic information into vision systems. In *AAAI'96*, Portland, Oregon, 1996.

[10] J. Sherrah and S. Gong. Exploiting context in gesture recognition. In *Proc. CONTEXT'99*, 1999.

[11] J. Sherrah and S. Gong. Fusion of 2-D face alignment and 3-D head pose estimation for robust and real-time performance. In *Proc. Int. W'shop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems*, pp. 24–30, 1999.

[12] M. Turk. Visual interaction with lifelike characters. In *Proc. FG'96*, pp. 368–373, Killington, VT, 1996.

[13] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. In *Proc. FG'96*, pp. 51–56, Killington, VT, 1996.

[14] C. R. Wren and A. P. Pentland. Dynamic models of human motion. In *Proc. FG'98*, pp. 22–27, Nara, Japan, 1998.