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. We present methods for modelling and interpretation of multi-person human behaviour in real time to control video cameras for visually mediated interaction.

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) requires intelligent interpretation of a dynamic visual scene in order to determine areas of interest for fast and effective communication to a remote observer.

Recent work at the MIT Media Lab has shown some initial progress in the modelling and interpretation of human body activity [7, 9]. Computationally simple view-based approaches to action recognition have also been proposed [1, 2]. However, these systems do not attempt intentional tracking and modelling to control active camera views 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 [8]. 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 evidence for *direction* and waving for *attention* are important for deliberative camera control. The reactive camera movements should provide the necessary visual context for applications such as group video-conferencing and automated studio direction.

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. 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 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, some salient behaviours are defined for visually mediated interaction tasks.

Implicit Behaviour

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 [6].

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 [5, 4]. 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 are thresholded to detect a gesture, or are in themselves considered as model outputs for explicit behaviour.

Human Behaviour
Given that both implicit and explicit behaviours are 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:

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

3 Interpretation of Group Activities

Although the individual interpretation of behaviours is possible by attaching pre-defined semantics in the form of 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. Clearly the range of possibilities and ambiguities present a problem for any single visual cue, no matter how accurately it can be computed. It is only by fusing different visual cues that we may successfully interpret the scene.

3.1 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:

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

The group vector is an overall description of the scene at a given time instant. Let us 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*:

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

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. Such a scheme would require at least 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:

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

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 [3], 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. 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 1–4 show examples of the system output for two example scenarios: **wave-look** and **point**. Figures 1 and 3 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 2 and 4 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.

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.

The bottom sections of Figures 2 and 4 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 transitions are highly subjective and difficult to determine, even to the human eye.

4 Conclusion

Some key issues in visual interpretation of group behaviours in single views have been explored, and a framework has been presented for tracking people and recognising their correlated group behaviours in VMI contexts. Predefined gestures and the head pose of several individuals in the scene can be simultaneously recognised for scene interpretation. In the presence of multiple people, ambiguities arise and a high-level interpretation of the combined behaviours of the individuals becomes essential.

We have shown examples of how multi-person activity scenarios can be learned from training examples and interpolated 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, and extrapolate to novel situations in the same way as a person. A significant issue raised in this paper and for future work is the feasibility of learning correlated temporal structures and default behaviours from sparse data.

Since this system relies on several independent components, the overall probability of failure of at least one component is always quite high. Therefore the high-level interpretation system must be able to cope with missing or noisy inputs. The system outputs may be fed back to the low-level sub-systems to guide their processing.

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] 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.

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

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

[5] 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.

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

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

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

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

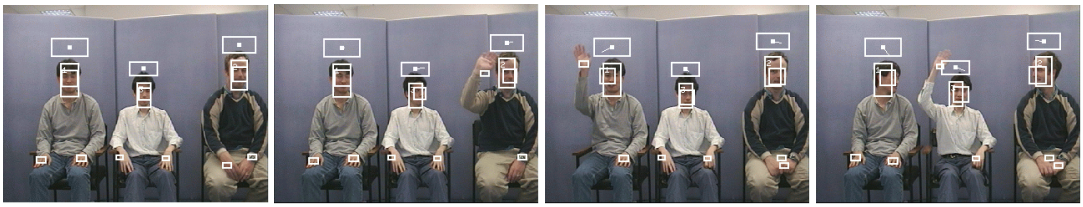

Figure 1. Frames from wave-look sequence. Individuals are labelled A, B and C from left to right.

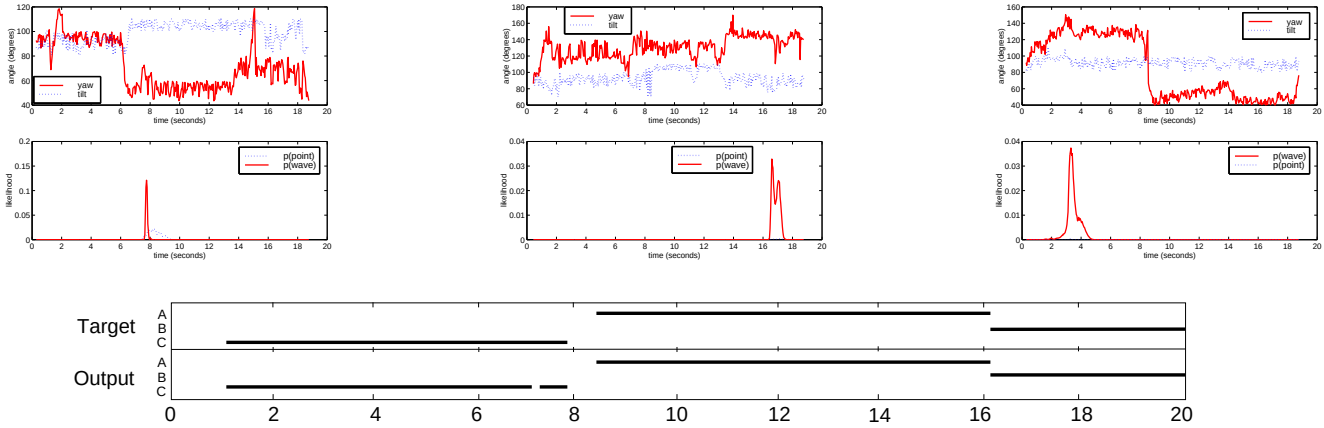

Figure 2. Results for wave-look scenario. Plots show pose angles (top) for persons A, B and C from left to right, gesture likelihoods (middle) and target/output camera position vectors (bottom).

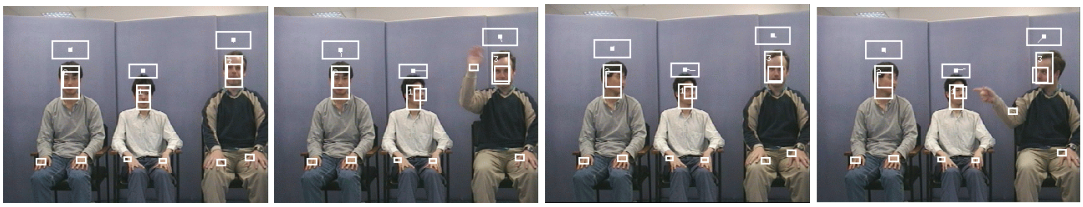

Figure 3. Frames from point sequence. Individuals are labelled A, B and C from left to right.

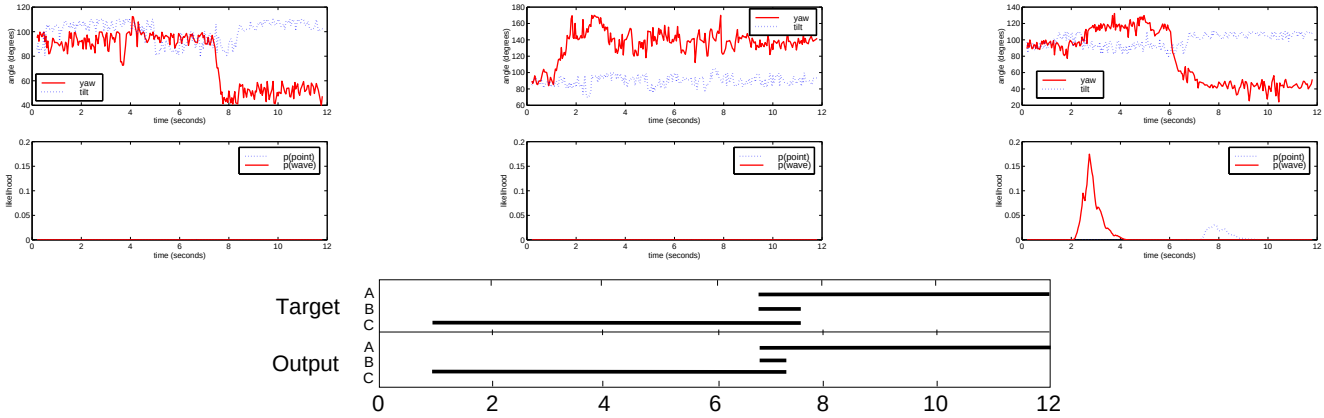

Figure 4. Results for point scenario. Plots show pose angles (top) for persons A, B and C from left to right, gesture likelihoods (middle) and target/output camera position vectors (bottom).