Interpretation of Group Behaviour in Visually Mediated Interaction

Jamie Sherrah, Shaogang Gong, A Jonathan Howe and Hilary Buxton

C P

May

I N

UNIVERSITY OF

Cognitive Science Research Papers

[illegible] *

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

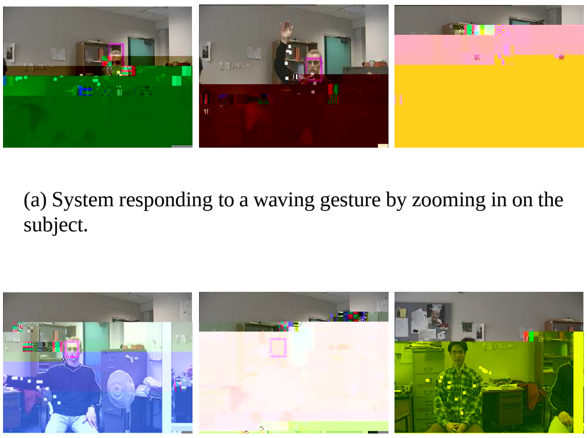
[illegible]

[illegible]

[illegible]

[illegible]

For our purposes, [illegible] can be considered to



(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

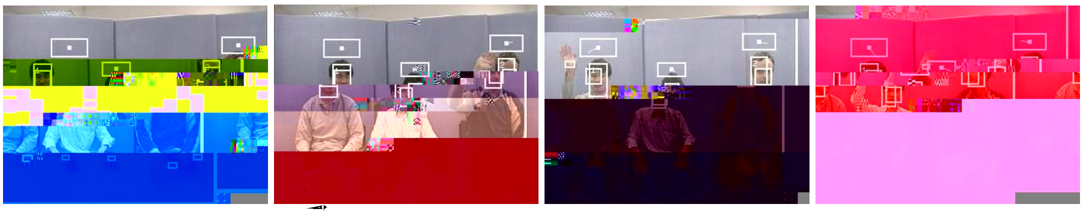
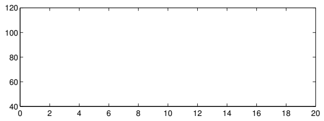


Figure 3. Frames from [illegible] sequence. Individuals are labelled A, B and C from left to right.



scenario	description
[illegible]	C waves and speaks, A waves and speaks, B waves and speaks. Each time someone is speaking the other two subjects look at him
[illegible]	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 *[illegible]* 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.

[illegible]

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.

[illegible]

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

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

[3] R. Cutler and M. Turk. View-based interpretation of real-