Active Vision Techniques for Visually Mediated Interaction

A. Jonathan Howell and Hilary Buxton
School of Cognitive and Computing Sciences,
University of Sussex, Falmer, Brighton BN1 9QH, UK

Abstract

In this paper we introduce adaptive vision techniques used, for example, in video-conferencing applications. Radial Basis Function (RBF) networks have been trained for gesture-based communication with colour/motion cues to direct face detection and capture 'attentional frames'. These focus the processing for Visually Mediated Interaction via an appearance-based approach with Gabor filter coefficients used as input to time-delay RBF networks. We use these methods for behaviour (user-camera) coordination in an integrated system.

1. Introduction

Visually Mediated Interaction (VMI) facilitates interaction between people using visual cues similar to those used in everyday communication. The aim is to overcome limitations due to, for example, distance or disability. This involves many visual competences such as recognising facial expression, gaze, gesture and body posture which are all used in human interaction. Gestures are often spontaneous but can also be intentional, where we can distinguish between verbal (sign languages) and nonverbal (pointing, emphasis, illustration). In our work here we are mainly concerned with intentional, nonverbal gestures which are relevant for VMI. Also, we use gaze which provides an important cue for discourse/interaction management. In particular, gaze direction is associated with attention-directing pointing to indicate objects or people of interest in the scene.

Robust tracking of non-rigid objects such as human faces and bodies involved in machine analysis of this kind of interactive activity is difficult due to rapid motion, occlusion and ambiguities in segmentation and model selection. This is addressed by the move to active vision and dynamic models, e.g. learning to track complex hand dynamics [15]. More generally, research funded by British Telecom (BT) on *Smart Rooms* [20] at MIT Media Lab has shown progress in the modelling and interpretation of human body activity, e.g. the *Pfinder* (Person Finder) system [26]. This can provide real-time human body analysis with further work to model the progression of ongoing activity using techniques such as *Hidden Markov Models* (HMMs), which can be parameterised to provide information such as direction of pointing [25]. Further analysis for VMI can even involve coupled human interaction analysis using learning techniques based on deformable models [18].

Other recent related research using computationally simple view-based approaches to action recognition have been introduced by Bobick [2]. Pinhanez and Bobick [21] have developed a PNF network approach using the temporal terms *(past, now, fut)* for human action detection, which allows fast performance compared to equivalent evaluations of Allen's interval logic. Similar approaches at Microsoft Research by Turk and Cutler [7, 24] have also yielded useful results. In Pentland's group, much progress has been made in the detailed modelling and interpretation of human body activity [27]. We also have coupled HMMs [3] for understanding behaviour interactions and parameterised HMMs [1]. More recent work [19] has developed reliable Bayesian vision systems. Two further developments are: 1) work by Galata, Johnson and Hogg using hybrid deformable and HMM behaviour models for virtual actors [9]; and 2) the action-reaction learning of Jebara and Pentland [16].

We have concentrated on developing computationally simple view-based approaches to action recognition, which address the task of using intention in behaviour modelling to directly drive VMI. In robotics, Brooks [4] emphasises the need to have this kind of perceptual grounding for behaviour, going directly from perception to action. In cognitive science (review [5, pp. 311–374]), we also find that recognition of behaviour is possible with minimal perceptual information, e.g. Johansson's point-light technique allows us to recognise human movement [17]. Even animated sequences of simple geometrical shapes are interpreted using intentional descriptions [23].This suggests that human visual cognition has direct methods that are learnt for behaviour interpretation and control. We can mimic these characteristics in subsymbolic approaches using neural networks. Our proposal, then, is to directly associate an attention seeking pragmatic interpretation with waving gestures and zoom in on the user. This idea generalises to directional semantics for pointing gestures for intentional tracking in

the design of our system.

The background research here is our view-based learning techniques for face recognition real-time, of a known group of people within indoor environments [13]. A key capability was to identify faces over a range of head poses and our approach exploited the flexibility of the example-based *Radial Basis Function* (RBF) network learning approach, which allowed us to reformulate the training in terms of the specific classes of data we wished to distinguish. For example, we could categorise head pose or expression information separately from identity by training RBF classifiers for each separate task [8]. Similarly, our approach to gesture recognition uses time-delay variants of RBF networks [12]. Essentially, these adaptive methods allow us to make key inferences within our system by modelling the variability of the evidence.

2. Capturing the Attentional Frame

Our techniques here use colour/motion cues from the image sequence to identify and track the head. Once we know the position and size of the head, we define an *attentional frame* around the person. The attentional frame is a 2-D area around the focal user that contains all the body movement information relevant to our application, which is all movement of the head and right arm. To allow people to move closer or further away from the camera, this information is normalised for size (relative to head size) around an arbitrary standard position from the camera.

Our main priority is to find *real-time solutions* for the application. Therefore, we use two computationally cheap pixel-wise processing techniques on our image: thresholded frame differencing, giving motion information, and Gaussian mixture models, giving skin colour information. These are combined to give a binary map of moving skin pixels within the image, and we use local histogram maxima to identify potential 'blob' regions. A box, which is large enough to contain the head at all distances in our target range, is fitted over the centroid of each of these regions. Fig. 1(a) shows how each box is centred on the centroid of each maximum, with the inner lines showing the standard deviation of the pixels along the *x*-axis from that centroid. It can also be seen that the hands are ignored in this example, as they are too low down to be included in a face-size 'blob'.

A robust approach to head tracking using colour/motion blobs is what we call *temporal matching*: the tracker only considers blobs from the current frame which have been matched to nearby blobs from previous frames. This excludes any anomalous blobs that appear for one frame only in an image sequence, and promotes those that exhibit the greatest temporal coherence. Having found the position and size of the head, we extract the attentional frame from around the person.

(a) (b)

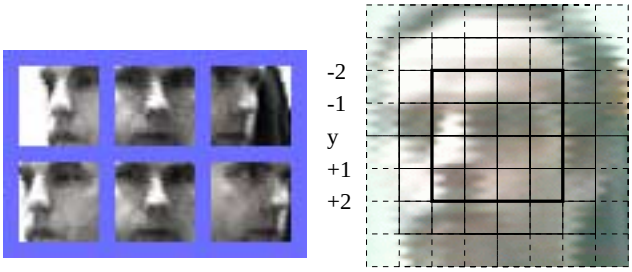
Figure 1. Colour/motion cues position an attentional frame around a person: (a) a box is centred around each colour/motion 'blob', (b) an 'attentional frame' is drawn around the person relative to the head.

3. Pose-Invariant Face Detection

The previous section described how we isolated small areas of moving skin-tones from the overall image. This reduces computation and network size, by allowing the face detector to work only within a small subset of the full spectrum of possible objects typically encountered in an office environment. Specifically, we consider the restricted form of face detection where we need to distinguish a face only from other moving skin-tone blobs (typically hands).

In order to perform effective face recognition, we need to identify the position of the central face area (eyes, nose, mouth), rather than the entire skin area on the head (which also includes forehead, neck, ears, etc). Our face detection task, therefore, is to distinguish centred faces from both non-centred faces and other moving skin-tone blobs. We train RBF networks with examples of both to provide a continuous 'face/non-face' output, with a level of confidence based on the difference between the two output values from the network [11]. This level of confidence allows discarding of low-confidence results where data is noisy or ambiguous.

Our training examples take variable head-pose into account, so the central face region of a person can be recognised at all normal physiological pose positions. Facial information is only visible on a human head from (roughly) the front $\pm 120^\circ$ of x- and y-axis movement, and z-axis movement is physiologically constrained to around $\pm 20^\circ$ (when standing or sitting). The face region is centralised on the nose, rather than the face, for all profiles, as this allows non-occluded face information to remain roughly in the same position, see Fig. 2(a). This has previously been shown to more useful for pose-varying face recognition we then easily determine a coarse estimate of head-pose, such as left, frontal or right, from the output grid. This qualitative level of head-pose is very useful for



(a) (b)

Figure 2. (a) Two methods for segmenting 25×25 pose-varying face data, (b) the grid system for detecting potential faces within a potential 'head blob' region of the image.

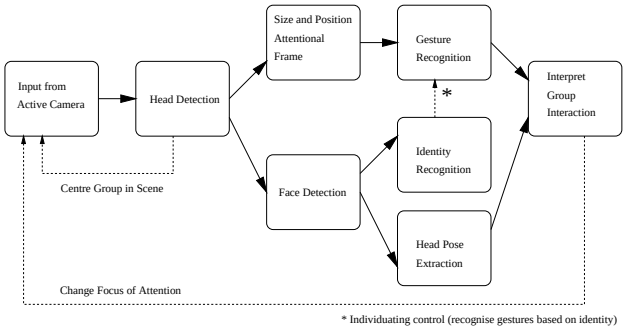


Figure 3. A block diagram outlining the integrated system (from [14]).

our group interaction analysis [22]. The 'non-face' class is taken from a larger grid to encourage face detection only where the image was accurately aligned on the face, such as within the dotted lines in Fig. 2(b), and from around the centroids of 'distractor' moving skin/colour regions, e.g. hands, within each frame.

4. The Integrated System

The design for the complete integrated system is seen in Fig. 3, where the input from the active camera is first processed to detect heads and position attentional frames, then face, gesture and pose classification, followed by the interpretation of group interaction.

A complete video-conferencing active camera control system requires high-level interpretation of group and individual interaction [22]. As we have seen, we propose a system for behavioural control, whereby gesture and head pose information, contained in a 'scene vector', is provided for this interpretation to take place. This allows the system to provide camera control information via a learnt mapping onto a 'camera control vector' representation. The scene vector provides head-pose and gesture probabilities for the people in the field of view, and the camera control vector determines the focus of attention in terms of which users are included in the processed scene. If individuated control of the system is required, then we need to identify who these people are (from a small known group), as shown in Fig. 3. Two extra stages, therefore, are needed: gesture and (pose invariant) identity recognition. Practical techniques for tackling these tasks in real-time, using the RBF and TDRBF networks are taken from [12, 13].

5. Summary

- We can use colour/motion cues to effectively segment and track human heads in image sequences.
- An attentional frame can be extracted relative to the head position and size to allow the real-time recognition of hand gestures through time.
- By extracting colour/motion regions from the overall image, the face detection task is greatly simplified.
- A face detection network can be used to give a qualitative estimate of head-pose for predictive control using implicit behaviour.
- Splitting multi-phasic gestures into separate phase classes not only gives more precise timing of gesture events, but also allows the gesture recognition network to provide prediction hypotheses for behaviour control.

We have fully integrated real-time recognition, tracking and on-line intentional control for single users, but there are still some outstanding problems for multiple interacting users. We can control attentional switching for multiple users in known scenarios, e.g. 3 people sitting and passing control in an orderly fashion [22]. A major issue with this kind of example-based learning approach to multi-participant behaviour interpretation is the feasibility of collecting sufficient data. The multiplicity of possible events increases exponentially with the addition of extra participants and the combinatorics can only be captured at the level of examples used for training. The use of high-level models such as *Bayesian Belief Networks* (BBNs) can provide a combination of hand-coded *a priori* information with machine learning to ease training set requirements. This is because the BBNs model the decomposition of the problem and it is the model parameters (conditional probabilities) that are learnt so that higher level inferences can be made from low level visual evidence (see, for example, [6]).

6. Conclusions and Further Research

It is clear that there are many potential advantages of Visually Mediated Interaction with computers over traditional keyboard/mouse interfaces. For example, removing

system-dependant IT training and allowing the user a more intuitive form of system direction. However, there are still many challenges for integrating multi-user interaction analysis and control due to the ambiguities and combinatorial explosion of possible behavioural interactions. We have demonstrated how our connectionist techniques can support real-time interaction by detecting faces and capturing 'attentional frames' to focus processing. To go further we will have to build our VMI systems around the task demands which include both the limitations of our techniques and potentially conflicting intentions from users. Connectionist techniques are generally well suited to this kind of situation as they can learn adaptive mappings and have inherent constraint satisfaction.

Further research is taking two main directions: 1) the development of gesture-based control of animated software agents in the EU Puppet project; and 2) the development of context-based control in more complex scenarios in the new EU ActIPret project. The first extends the use of action selection and dynamic control functions in gesture-based interfaces where pointing can indicate the current avatar and movement patterns can control animation parameters. The second involves recognition of complex behaviours and activities that consist of a sequence of events that evolve over time [10]. As yet there has been little work that combines automated learning of behaviours in different contexts. In other words, it is usually only simple, generic models of behaviour that have been learnt rather than learning when and how to apply more complex models in a context sensitive manner.

Acknowledgements

The authors gratefully acknowledge the invaluable discussion, help and facilities provided by Shaogang Gong, Jamie Sherrah and Stephen McKenna and funding under the EPSRC ISCANIT and EU ActIPret projects.

References

[1] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proc. ICCV*, pp. 382–388, Cambridge, MA, 1996.

[2] 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.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *Proc. ICPR*, San Juan, Puerto Rico, 1997.

[4] R. A. Brooks. From earwigs to humans. *Robotics and Autonomous Systems*, 20:291–304, 1997.

[5] V. Bruce and P. Green. *Visual Perception*. Lawrence Erlbaum, London, 1990.

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

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

[8] S. Duvdevani-Bar, S. Edelman, A. J. Howell, and H. Buxton. A similarity-based method for the generalization of face recognition over pose and expression. In *Proc. FG*, pp. 118–123, Nara, Japan, 1998.

[9] A. Galata, N. Johnson, and D. C. Hogg. Learning variable length Markov models of behaviour. *Computer Vision & Image Understanding*, 81:398–413, 2001.

[10] R. J. Howarth and H. Buxton. Conceptual descriptions from monitoring and watching image sequences. *Image & Vision Computing*, 18:105–135, 2000.

[11] A. J. Howell. Face recognition using RBF networks. In R. J. Howlett and L. C. Jain, editors, *Radial Basis Function Networks 2*, pp. 103–142. Physica-Verlag, 2001.

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

[13] A. J. Howell and H. Buxton. Learning identity with radial basis function networks. *Neurocomputing*, 20:15–34, 1998.

[14] A. J. Howell and H. Buxton. RBF network methods for face detection and attentional frames. *Neural Processing Letters*, 15:1–15, 2002.

[15] M. Isaard and A. Blake. A mixed-state condensation tracker with automatic model-switching. In *Proc. ICCV*, pp. 107–112, Bombay, India, 1998.

[16] A. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *Proc. ICVS'99*, Las Palmas de Gran Canaria, Spain, 1999.

[17] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[18] N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *Proc. ICPR*, pp. 866–871, 1998.

[19] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modelling human interactions. In *Proc. ICVS'99*, Las Palmas de Gran Canaria, Spain, 1999.

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

[21] C. Pinhanez and A. F. Bobick. Human action detection using PNF propagation of temporal constraints. In *Proc. ICPR*, Santa-Barbara, CA, 1998.

[22] J. Sherrah, S. Gong, A. J. Howell, and H. Buxton. Interpretation of group behaviour in visually mediated interaction. In *Proc. ICPR*, pp. 266–269, Barcelona, Spain, 2000.

[23] R. H. Thibadeau. Artificial perception of actions. *Cognitive Science*, 10:117–149, 1986.

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

[25] A. D. Wilson and A. F. Bobick. Recognition and interpretation of parametric gesture. In *Proc. ICCV*, pp. 329–336, Bombay, India, 1998.

[26] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. PAMI*, 19:780–785, 1997.

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