Active Vision Techniques for Visually Mediated Interaction

A. Jonathan Howell and Hilary Buxton

CSRP 541

November 2001

ISSN 1350–3162

UNIVERSITY OF
SUSSEX
AT BRIGHTON

Cognitive Science
Research Papers

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 videoconferencing applications. First, we present the recognition of identity, expression and head pose using Radial Basis Function (RBF) networks. Second, we address gesture-based communication and attentional focus, using 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. Third, we present methods for the gesture recognition and behaviour (user-camera) coordination in an integrated system.

Key words: Visually Mediated Interaction; Face Recognition; Gesture Recognition; Camera Control; Time-Delay Neural Networks

1 Introduction

Visually Mediated Interaction (VMI) is a process of facilitating interaction between people, either remotely or locally, using visual cues which are similar to those used in everyday interaction with other people. The aim is to enhance interaction, overcoming 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 communication and interaction. Gestures are often spontaneous but can also be intentional, where we can distinguish between verbal (sign languages) and nonverbal (pointing, emphasis, illustration) usage. In our work here we are mainly concerned with intentional, nonverbal gestures which are relevant for communication in VMI. Also, we use gaze which can provide an important cue for discourse/interaction management. In particular, gaze direction is often associated with diectic, attention-directing pointing to indicate objects or people of interest in the immediate context as part of the behavioural interaction.

We know that 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 was partially addressed by the move to active vision and dynamic models for robust tracking using sophisticated Kalman filters, as exemplified by Blake and others [1]. Recently, these have been specialised to allow the learning of complex hand dynamics [23]. More generally, research funded by British Telecom (BT) on *Smart Rooms* [38] and the ALIVE project [30] at MIT Media Lab has shown progress in the modelling and interpretation of human body activity. This used the *Pfinder* (Person Finder) system [49], which can provide real-time human body analysis. Further analysis to model the progression of ongoing activity involves techniques such as *Hidden Markov Models* (HMMs), which can be parameterised to provide information such as direction of pointing [48]. Further analysis for VMI can even involve coupled human interaction analysis using learning techniques based on deformable models [27].

Other related ongoing research using computationally simple view-based approaches to action recognition have been introduced by Bobick [3]. More recently, Pinhanez and Bobick [39] 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 attempts at Microsoft Research by Turk and Cutler [8,46] have also yielded useful results. In Pentland's group, much progress has been made in the detailed modelling and interpretation of human body activity [50]. We also have the coupled HMMsof Brand and colleagues [4] for understanding behaviour interactions, although this approach requires a great deal of training data. This is also true of parameterised HMMs [2], which can also suffer from lack of stability in the interpretation compared to deformable model tracking and analysis. More recent work by Oliver, Rosario and Pentland [36,37] has developed reliable Bayesian vision systems. Two exciting recent development are: 1) work by Galata, Johnson and Hogg using hybrid deformable and HMM behaviour models for virtual actors [15]; and 2) the action reaction learning of Jebara and Pentland [24,25], which models interactions and exploits new ideas from Support Vector Machines in conjunction with generative Bayes theory.

However, we have concentrated on developing computationally simple view-based approaches to action recognition under the ISCANIT project, which start to address the task of intentional tracking and behavioural modelling to directly drive visual interaction. In robotics, Brooks [5] emphasises the need to have this kind of perceptual grounding for behaviour, going directly from perception to action. In cognitive science (review [6], pp. 311-374), we also find that recognition of behaviour is possible with minimal perceptual information. For example, Johansson's point-light technique, in which we have access to pure movement cues, allows us to recognise human movement [26]. We can identify gender [29], emotional state [12], and types of action [11]. Human observers can do even better with friends, whom they can recognise from their gait [9]. This all suggests that human visual cognition has direct methods that are learnt for the familiar people and their behaviour. We can mimic these characteristics in subsymbolic approaches using deformable models or neural networks, although it is clear that the latter is closer to implementations in biological systems. Using animated sequences of simple geometrical shapes also demonstrates that human subjects even offer intentional descriptions of the observed movement patterns [45]. Our proposal, then, is to simply associate attention seeking pragmatic interpretation with waving gestures and zoom in on the user. This idea generalises to directional semantics for pointing gestures etc. for intentional tracking in the design of our system.

The background research here is our example-based learning techniques for face recognition [21]. The particular task considered was the recognition, in real-time, of a known group of people within indoor environments. It could not be assumed that there would be clear frontal views of faces at all times and so a key capability was to identify faces over a range of head poses. An important factor in this approach was 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 extract identity, head pose and expression information separately from the same face training data to train a computationally cheap RBF classifier for each separate recognition task [13,19]. Essentially, these adaptive methods allow us to make key inferences within our system by modelling the variability of the evidence.

In this work we again take an appearance-based approach, with each phase of a communicative gesture represented as a vector of Gabor filter coefficients. First, in Section 2, we discuss the RBF network scheme. Second, in Section 3, this is used as input to time-delay RBF (TDRBF) networks which can extract gesture and head pose information. Third, in Section 4, this data is fed to a further TDRBF network which can analyse group behaviour in order to control camera systems in an integrated system. Fourth, in Section 5, we discuss a potential design for such an integrated system. Finally, in Section 6, we draw some conclusions about our approach and further work.

2 The RBF Network Scheme

The RBF network is a two-layer, hybrid learning network [32,33], which combines a supervised layer from the hidden to the output units with an unsupervised layer from the input to the hidden units. The network model is characterised by individual radial Gaussian functions for each hidden unit, which simulate the effect of overlapping and locally tuned receptive fields.

The RBF network is characterised by computational simplicity, supported by well-developed mathematical theory, and robust generalisation, powerful enough for real-time real-life tasks [42,43]. The nonlinear decision boundaries of the RBF network make it better in general for function approximation than the hyperplanes created by the multi-layer perceptron (MLP) with sigmoid units [41], and they provide a guaranteed, globally optimal solution via simple, linear optimisation. One advantage of the RBF network, compared to the MLP, is that it gives low false-positive rates in classification problems as it will not extrapolate beyond its learnt example set. This is because its basis functions cover only small localised regions, unlike sigmoidal basis functions which are nonzero over an arbitrarily large region of the input space.

Once training examples have been collected as input-output pairs, with the target class attached to each image, tasks can be learnt directly by the system. This type of supervised learning can be seen in mathematical terms as approximating a multivariate function, so that estimations of function values can be made for previously unseen test data where actual values are not known. This process can be undertaken by the RBF network using a linear combination of basis functions, one for every training example, because of the smoothness of the manifold formed by the example views of objects in a space of all possible views of that object [40]. This underlies our approach, successful in previous work with RBF networks for face recognition tasks with image sequences [21], which uses an RBF unit for each training example, and single stage pseudo-inverse calculation of weights.

2.1 The Time-Delay RBF Model

To construct a dynamic neural network, recurrent connections can be added to standard multi-layer perceptrons which then form a contextual memory for prediction over time [14,28,35]. These partially recurrent neural networks can be trained using back-propagation but there may be problems with stability and very long training sequences when using dynamic representations. Instead, we use a simple Time-Delay mechanism [47] in conjunction with an RBF network, which we term a TDRBF network, which we have previously shown can allow fast, robust solutions to difficult real-life problems [20]. Such a network can be created by combining data from a fixed time 'window' into a single vector as input. In addition, an integration layer on the TDRBF network can be used to combine results from successive time windows to provide smooth gradations between serial actions.

2.2 Gabor Filter Input Representation

Filter-based preprocessing of the images is an important intermediate step in image-based techniques, as the input representation contributes a great deal to the learnability of the task. It is important to highlight relevant parts of the information (leading to reduction in the dimensionality of input) and provide moderate invariance to normal environmental illumination [10]. We constrain the lighting to exclude strong, incidental lighting, which is very much more difficult [34]. We use a sparse arrangement of Gabor filters [18] to both suppress variation that is not important for the task, such as illumination variability, and highlight those variations that are useful, using explicit orientations and scales. This approach is used to preprocess each frame of the sequences (colour/motion information for gestures, grey-level pixels for face detection): data is sampled at three non-overlapping scales and three orientations with sine and cosine components for a total of 126 coefficients per frame.

3 Gesture Recognition

To extend our research to support *Visually Mediated Interaction* (VMI), we needed to develop person-specific and generic gesture models for the control of active cameras. A time-delay variant of the Radial Basis Function (TDRBF) network was used to recognise simple pointing and waving hand gestures in image sequences [20]. The gesture database was developed as a source of suitable image sequences for these experiments. Characteristic visual evidence can be automatically selected during the adaptive learning phase, depending on the task demands.

A set of *interaction-relevant* gestures were modelled and exploited for reactive on-line visual control. These can then be interpreted as user intentions for live control of an active camera with adaptive view direction and attentional focus. As noted in the introduction, pointing (for direction) and waving (for attention) are important for intentional control and the reactive camera movements may be able to provide the necessary visual context for applications such as group video-conferencing as well as automated studio direction.

Table 1
Body movement and behaviour definitions for the gesture database.

Gesture	Body Movement	Behaviour
pntrl	point right hand to left	pointing left
pntrr	point right hand to right	pointing right
wavea	wave right hand above head	urgent wave
waveb	wave right hand below head	non-urgent wave

Previous approaches to recognising human gestures from real-time video as a nonverbal modality for human-computer interaction have involved computing low-level features from motion to form *temporal trajectories* that can be tracked by Hidden Markov Models or Dynamic Time Warping. However, for this work we explored the potential of using simple image-based differences from video sequences in conjunction with the RBF network learning paradigm to account for variability in the appearance of a set of predefined gestures. The computational simplicity and robust generalisation of our alternative RBF approach provided fast training and on-line performance, highlighting its suitability as a source of interactive responses required by applications with active camera control.

3.1 The Gesture Database

The first database was created to provide a source of single-person gesture data [20]. It concentrated on two specific behaviours which could be used to move the camera or adapt its field of view: *pointing*, which is interpreted as a request to pass camera attention, and is implemented by zooming out and panning in the pointing direction, and *waving*, which is interpreted as a request for camera attention, and implemented by panning towards the waver and zooming in. We have two types of each behaviour, giving four gestures in all, shown in Table 1.

Four examples of each gesture from three people were collected, 48 image sequences in all. Each sequence contains 59 378×288 8-bit monochrome images (collected at 12 frames/sec for roughly 5 seconds), a total of 2832 images.

3.2 Results

Experimental results showed that high levels of performance for this type of *intentional gesture recognition* can be obtained using these techniques both for particular individuals and across a set of individuals. Characteristic visual ev-

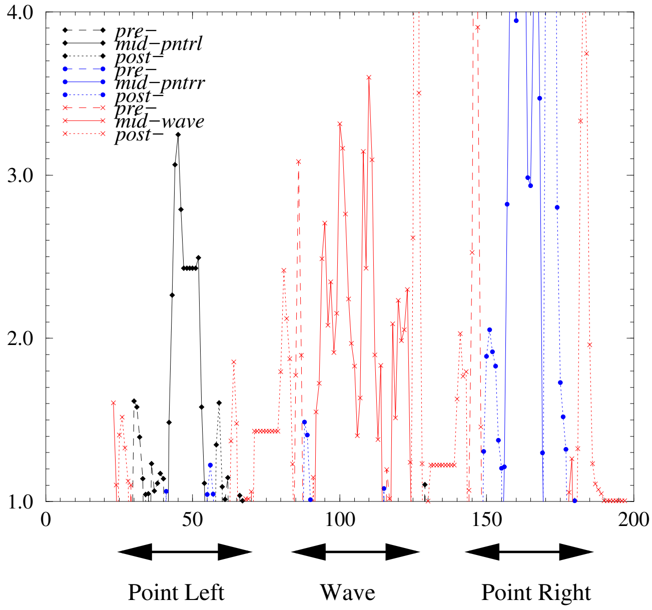


Fig. 1. Output for the multi-phase TDRBF gesture network with a test sequence with different background, lighting and person to that encountered during training. Each line represents a pre-, mid- or post-gesture class, and shows the confidence level of output when its class is the maximum, and is zero at all other times.

idence was effectively selected and can be used, if required, even to recognise individuals from their gestures [19]. Previous TDRBF network experiments had learnt certain simple behaviours based on y-axis head rotation [18], distinguishing between left-to-right and right-to-left movements and static head pose. Such tasks are simplified by constant motion, so that arbitrary short segments (2/3 frames) of the whole sequence can be used to identify the overall direction of head turning. Due to the complex motion involved in these particular gestures, characteristic parts of the complete action needed to be contained in the time window presented to the network in order that it can be recognised. The initial requirement to present the entire gesture sequence for recognition meant that event signalling could only be done retrospectively.

Subsequent work further refined the training for the gesture information to reduce the amount of data required [22]. Taking advantage of the *tri-phasic nature* of the waving and pointing gestures, each gesture can be split into a pre-, mid- and post-gesture sequence. Each of these can be trained as its own sub-gesture class. This gives some predictive power from the pre-gesture and focusses on the characteristic movement of the mid-gesture. It also gives some temporal invariance by allowing the two phases of the gesture to be different in length to each other (the original studies imposed a fixed relationship between the overall gesture length and its three phases). Rapid signalling of the gesture event can be obtained in this way, as the system does not need to wait for the post-gesture phase to occur.

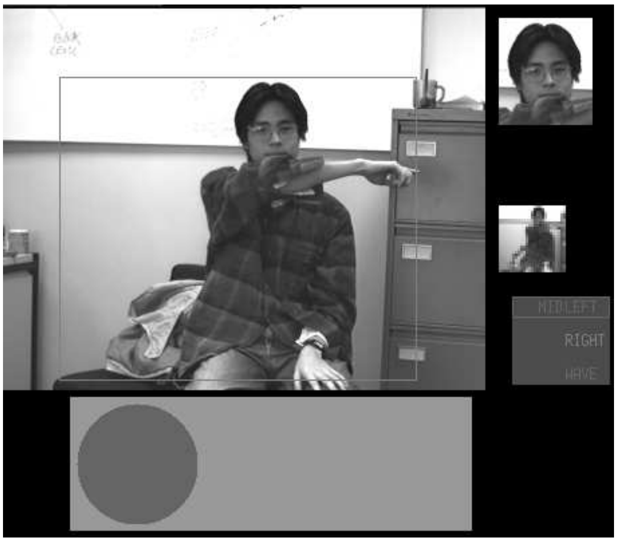


Fig. 2. Example output from the *GestureBall* application, where an orange ball is controlled on-screen via pointing and waving gestures. The system is shown at frame 48 of the test sequence, ie 10 frames after the start of the mid phase of the point left gesture, 6 frames before the end, and the ball has just moved from the centre to the left edge of the window.

Fig. 1 shows the output for a TDRBF gesture network trained with these phase classes. All mid-phase gesture events are clearly signalled, and each has a pre-phase gesture event before them. There is very little anomalous signalling, which can be dealt with either by ignoring low-confidence output, or by only accepting gestures which exhibit the phases in correct order, ie a mid-phase event must have had a matching pre-phase event beforehand. Some false pre-gesture ‘hypotheses’ are raised by the network, but these are replaced by the correct pre-gesture classification before the mid-gesture occurs. It might be expected that the pre-gesture classes will be confused, especially when they first start, as raising your hand to wave or point will initially both look very similar, but the results here indicate that they are correctly classified in time for predictive control.

An alternative way of assessing the effectiveness of ‘parsing’ gesture phase sequences for interaction based on identifying valid gesture events is demonstrated by the *GestureBall* system. This integrates the tracking and gesture recognition techniques used here so that users of the system can control a ball on-screen via pointing and waving gestures, see Fig. 2. There are three horizontal positions possible: a central, starting position, which can be returned to via the wave gesture, and left and right positions, which can be reached with the appropriate left or right point gesture. The system uses only pre- and mid-phase gestures, and registers a full gesture event only if a high confidence mid-phase of a particular gesture is encountered after a high confidence pre-phase of that same gesture. Experiments using this system with children between the ages 6–10 can show that they quickly learn gesture-result mappings and that the system is an easy, 'natural' method of controlling virtual events that has distinct advantages over the restrictive mouse/keyboard paradigm.

3.3 Summary

- Simple preprocessing techniques such as frame differencing and thresholding can be effective in extracting useful motion information and segmenting gestures in time.
- Different types of TDRBF network can be trained to distinguish gestures over specific time windows, for instance person-specific gesture models (trained and tested on one person) and generic gesture models (trained on one person, tested on other people).
- The TDRBF network can distinguish between arbitrary gestures, with a high level of performance.
- Some characteristics of an individual's expression of gestures may be sufficiently distinctive to identify that person.
- The TDRBF network can learn such data both as complete gesture sequences [20] and as specific gesture phases within a tri-phasic structure [22].
- 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 by identifying pre-gesture classes.
- The validated gesture phases can be integrated into simple visual interaction interfaces such as the GestureBall application.

In summary, the Time-Delay RBF networks showed themselves to perform well in our gesture recognition task, creating both person-specific and generic gesture models. This is a promising result for the RBF techniques, considering the high degree of potential variability (present even in our constrained database) arising from different interpretations of our intentional gestures by each individual. Note that this is in addition to variability in position, lighting etc. that had to be overcome in earlier face and simple behaviour recognition work.

4 Interpretation of Group Behaviour

The methods discussed so far have allowed the implementation of a complete connectionist system for a single user. However, the implementation of a multi-user integrated system involves higher-level control by the group of participants.

Table 2
Example interpretations of camera position vectors for group interaction scenarios with three people.

Camera Position Vector	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

While full computer understanding of dynamic visual scenes containing several people may be currently unattainable, we have investigated a computationally efficient approach to determine areas of interest in such scenes. Specifically, we have devised a method for modelling and interpretation of single- and multi-person human behaviour in real time to control video cameras [44]. Such 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 to determine areas of interest for effective communication to remote users.

As we have seen, our general approach to modelling behaviour is *appearance-based* in order to provide real-time behaviour interpretation and prediction [20,44]. In addition, we only use views from a single pan-tilt-zoom camera with no special markers to be worn by the users. It should be noted that we are not attempting to model the full working of the human body. Rather our aim is to exploit approximate but computationally efficient techniques. Thus, our models are able to support partial view-invariance, and are sufficient to recognise people's gestures in dynamic scenes. Such task-specific representations need to be used to avoid unnecessary computational cost in dynamic scene interpretation [7].

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, we must select the set of behaviours to be modelled for the application at hand. For VMI tasks, our system needs to identify regions of interest in a visual scene for communication to a remote user. Examining the case in which the scene contains people involved in a video conference, the participant(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 a significant cue for determining this focus of communication, and can be approximated by head pose. *Implicit behaviour* can be defined as any body movement sequence that is performed subconsciously by the participant, and here, it is head pose that is the primary source of implicit behaviour.

However, head pose information may be insufficient to determine a participant's focus of attention from a single 2D view, due to loss of much of the 3D information. Then, it is necessary to have the user communicate explicitly with our VMI system through a set of pre-defined behaviours. *Explicit behaviour* can be defined as a sequence of body movements that are performed consciously by a participant in order to highlight regions of interest in the scene. We used a set of pointing and waving gestures as explicit behaviours for control of the current focus of attention. As we have seen, such gestures can be reliably detected and classified in real-time [20].

Our approach to modelling group interaction involves defining the *behaviour vector* of a participant to be the concatenation of measured implicit and explicit behaviours (head pose angles and gesture model likelihoods). From this a *group vector* can be defined as a concatenation of the behaviour vectors for all people present in the scene at a given time instant, and *group behaviour* is just a temporal sequence of these group vectors. Given the group behaviour, a high-level interpretation model can determine the current area of focus. In our scenarios, the region of interest is always a person so we track the head of each individual. The output need only give an indication of which people are currently attended in the high-level system and is called the *camera position vector*. This has a boolean value (0 or 1) for each person in the scene indicating whether that person is currently attended, see Table 2. This information can then be used to control the movable camera, based on the position of the people in the scene.

Given a particular group behaviour, we constructed a *scene vector*, which contains the previous camera position vector information as feedback. This allowed the current focus of attention to be maintained, even when no gestures or head turning occurred.

4.1 The Group Interaction Database

The second database [44] contains examples of group interaction in a static scene. This database contains 15 sequences, each between 240 and 536 frames in length, a total of 5485 320×240 24-bit colour images. We constrain the complexity of the data by restricting behaviour to certain fixed scenarios, shown in Table 3, and by always having three participants, who remain sitting for the complete sequence. Each scenario is a group behaviour in which the participants perform gestures and change their head pose in a fixed pre-defined

Table 3
Scenario descriptions for the Group Interaction Database, involving three participants A, B and C.

Scenario	Description
wave	Person C waves and speaks, A waves and speaks, B waves and speaks
wave-look	Person C waves and speaks, A waves and speaks, B waves and speaks. Each time someone is speaking the other two participants look at him
question	Person C waves and speaks, A and B look at C, A interjects with a question, C looks at A to answer, then looks back at camera
point	Person 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
interrupt	Person C waves and speaks, A and B look at C, a person enters from the left, A, B and C watch as the person leaves, C looks at the camera and continues speaking, A and B look at C All participants look at the camera unless stated otherwise.

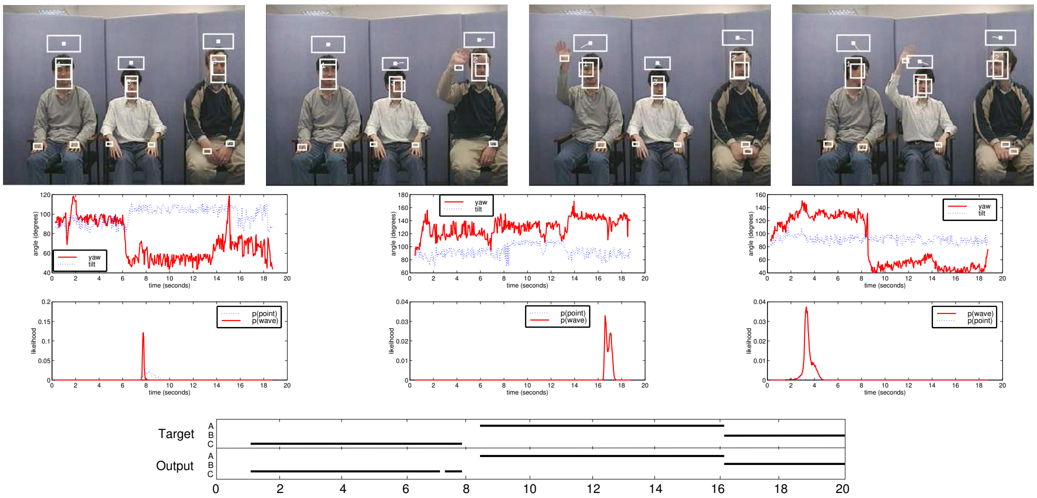


Fig. 3. Results for group interaction behaviour recognition using the **wave-look** scenario (see Table 3), individuals are labelled A, B and C from left to right: Example frames from sequence (top), plots showing pose angles and gesture likelihoods (middle), and target/output camera position vectors (bottom) (from [44]).

order. The exact timing of the events varies between different instances of the same scenario, but the focus of attention switches from one region to the next in the same order.

4.2 Results

To learn the transformation from scene vector to camera position vector, we developed an effective Time-Delay RBF Network, trained on half of our sequence database and tested for generalisation on the other half [44].

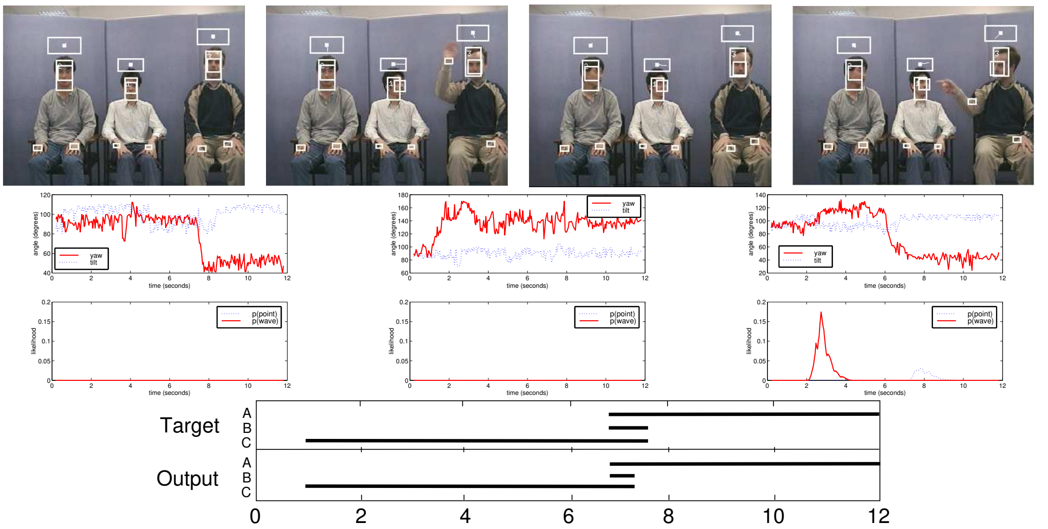


Fig. 4. Results for group interaction behaviour recognition using the **point** scenario (for details see Fig. 3).

Figs. 3 and 4 show examples of the system output for two example scenarios: **wave-look** and **point**. The top sections of each figure 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. The top sections of each figure 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.

The bottom section of Figs. 3 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, although the transition points from one focus of attention to another do not always exactly coincide. However, these transition points are highly subjective and very difficult to determine with manual coding, so this result is not surprising and the results give switches of attention that are acceptable at the perceptual level.

4.3 Summary

- A framework has been devised for tracking people and recognising their group behaviours in VMI contexts. This requires high-level information about group and individual interaction in a ‘scene vector’ to learn a ‘camera control vector’, specified by a temporal model.
- The scene vector provides ongoing probabilities of the dynamic head-pose and gesture phases for interacting participants and the camera control vector provides reactive direction and zoom.

- Pre-defined gestures and head pose of several individuals in the scene can be simultaneously recognised for interpretation of the scene.
- A scene vector-to-camera control transformation can be performed via a TDRBF network, using example-based learning.

We have been able to show 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 applications, 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 that we do. Unfortunately, there is no reason to believe that current computer architectures are capable of such reasoning and our simple temporal models fall far short of full intentional semantics. Therefore, a significant issue in future work will be the feasibility of learning generalised temporal structures and default behaviours from sparse data.

5 Towards an Integrated System

In this section we present our work towards a complete connectionist system for understanding the visual aspects of human interaction which could be used, for example, in video-conferencing applications. First, we present methods for face detection and capture of attentional frames to focus the processing for Visually Mediated Interaction. This frame can be used for recognising the various gesture phases that can then be used to control the camera systems in the integrated system, as discussed in previous sections.

5.1 Capturing the Attentional Frame

Our techniques here used colour/motion cues from the image sequence to identify and track the head. Once we know the position and size of the head, we can 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 our application. Therefore, we used two computationally cheap pixel-wise processing techniques on our image: thresholded frame differencing, giving motion information, and Gaussian mixture models [31], giving skin colour information. These were combined

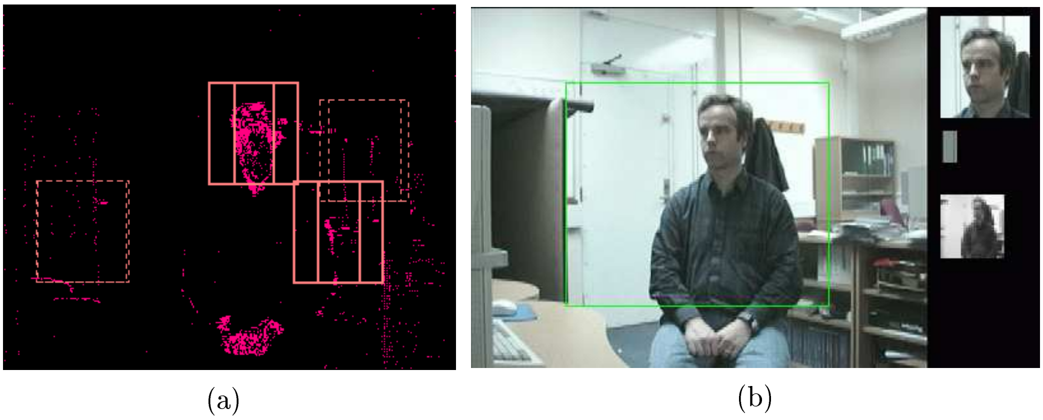


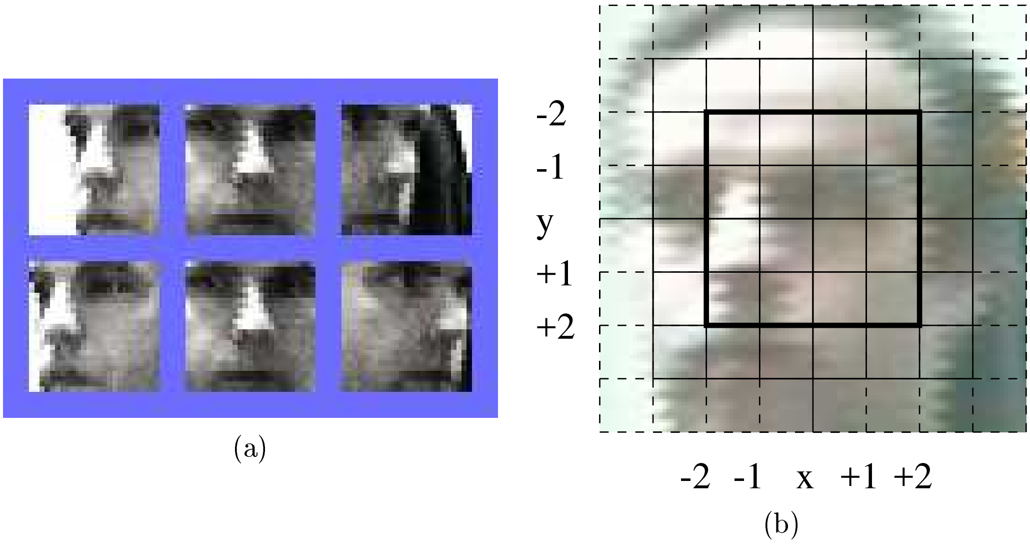
Fig. 5. Use of colour/motion information to position an attentional frame around a person: (a) a box is centred around each colour/motion 'blob', the inner vertical lines representing the standard deviation of the pixels along the x-axis, giving a width measure, (b) having identified which box contains the head (the uppermost one in (a)), an attentional frame box is drawn around the person relative to the head position, and sized according to head width. The top right image shows the image area inside the head box, bottom right the resampled area of the image inside the attentional frame.

to give a binary map of moving skin pixels within the image, and we used local histogram maxima to identify potential 'blob' regions. A box which was large enough to contain the head at all distances in our target range was then fitted over the centroid of each of these regions. Fig. 5(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 can extract the attentional frame from around the person.

5.2 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 can consider the restricted form of face detection where we



(a)

(b)

Fig. 6. (a) Two methods for segmenting 25×25 pose-varying face data: (top row) nose-centred, (bottom row) face-centred, the former being used for experiments here, (b) the grid system for detecting potential faces within a potential 'head blob' region of the image: each area tested is represented by a 4×4 box, the thick line shows the central position ($x, y = 0$), normal line and dashed lines the outer positions 1 and 2 spaces out from the centre. In this case, a maximum output would be expected at $x = -1, y = +1$, which indicates a head-pose slightly down and turned to the right.

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 trained 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 [19]. This level of confidence allows discarding of low-confidence results where data is noisy or ambiguous.

Our training examples need to 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 ±120° of x- and y-axis movement, and z-axis movement is physiologically constrained to around ±20° (when standing or sitting) [18]. 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. 6(a). This has previously been shown to more useful for pose-varying face recognition [19]. We can then easily determine a coarse estimate of head-pose, such as left, frontal or right, from the output grid. This

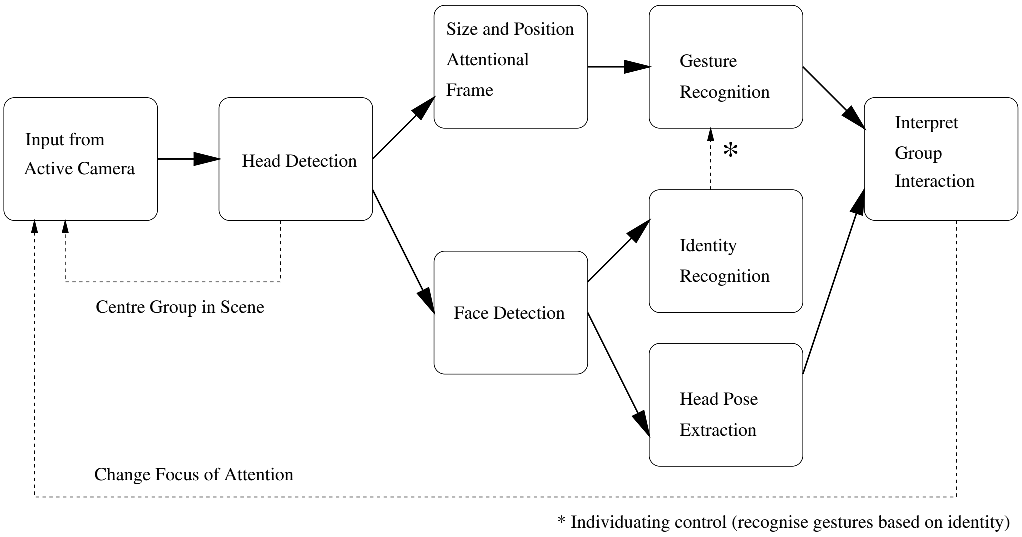


Fig. 7. A block diagram outlining the integrated system (from [22]).

qualitative level of head-pose was found to be very useful for group interaction analysis [44].

Therefore, the position of the nose was determined manually in order to extract specific centred/non-centred face patches from each image for training. Around this, a 3×3 grid of 25×25 face images was extracted from each frame for the 'pro-face' class. This corresponds to the regions within the the solid lines in Fig. 6(b). The 'non-face' class data was of two types:

- from a larger grid outside the 3×3 face grid (to encourage face detection only where the image was accurately aligned on the face), such as within the dotted lines in Fig. 6(b), and
- from around the centroids of 'distractor' moving skin/colour regions, eg. hands, within each frame.

5.3 The Integrated System

The design for the complete integrated system is seen in Fig. 7, where the input from the active camera is first processed to detect heads and position the 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 [44]. 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. 7. Two extra stages, therefore, are needed: gesture and (pose invariant) identity recognition. Section 3 above discussed practical techniques for tackling these tasks in real-time, using the RBF and TDRBF networks [20,21].

To complete our integrated system, we need to pass this gesture and head-pose information, with identity if appropriate, to a higher-level interpretation network [44], as discussed in Section 4. In addition, we have to adapt our system to cope with multiple people in the scene, which increases the complexity of the low-level processing stage. There will be more head blobs to find, but by assigning attentional frames to each person, and analysing each of these separately, it is hoped that problems due to occlusion from other members of the group will be kept to a minimum. This will allow a full implementation of the multi-user system with generalised attentional switching.

5.4 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 explicit behaviour control.

Although it has been possible to fully integrate real-time recognition, tracking and on-line intentional control for single users, there are still some outstanding problems for multiple interacting users. We can control attentional switching for multiple users in known scenarios eg 3 people sitting and passing control in an orderly fashion as in Table 3. As mentioned earlier, 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. Therefore, it is difficult to know which scenarios to collect in order to evenly populate the space of possible scenarios with the training set. The use of high-level models such as *Bayesian Belief Networks* (BBNs) might 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, [7]).

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, we have also seen that 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 (e.g. the GestureBall application) extends the use of symbolic (action selection) and mimetic (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 [16,17]. 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 under the EPSRC-funded ISCANIT project during the development and construction of the gesture database and in collaborative work with the group interaction experiments, and also by Mike Scaife and Yvonne Rogers at the Interact Lab at the University of Sussex, for the *GestureBall* application.

References

[1] A. Blake and A. Yuille. *Active Vision*. MIT Press, 1992.

[2] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proceedings of International Conference on Computer Vision*, pages 382–388, Cambridge, MA, 1996.

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

[4] M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, San Juan, Puerto Rico, 1997.

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

[6] V. Bruce and P. Green. *Visual Perception: Physiology, Psychology and Ecology.* Lawrence Erlbaum Associates, London, 1990.

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

[8] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 416–421, Nara, Japan, 1998. IEEE Computer Society Press.

[9] J. E. Cutting and L. T. Kowlowski. Recognition of friends by their walk. *Bulletin of the Psychonomic Society*, 9:353–356, 1977.

[10] J. G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, 36:1169–1179, 1988.

[11] W. H. Dittrich. Action categories and the perception of biological motion. *Perception*, 22:15–22, 1993.

[12] W. H. Dittrich, T. Troscianko, S. E. A. Lea, and D. Morgan. Perception of emotion from dynamic point-light displays represented in danse. *Perception*, 25:727–738, 1996.

[13] 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 *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 118–123, Nara, Japan, 1998. IEEE Computer Society Press.

[14] J. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

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

[16] R. J. Howarth and H. Buxton. Attentional control for visual surveillance. In S. Maybank and T. Tan, editors, *ICCV Workshop on Visual Surveillance.* IEEE Computer Society Press, 1997.

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

[18] A. J. Howell. *Automatic face recognition using radial basis function networks.* PhD thesis, University of Sussex, 1997.

[19] A. J. Howell. Face recognition using RBF networks. In R. J. Howlett and L. C. Jain, editors, *Radial Basis Function Networks 2: New Advances in Design*, pages 103–142. Physica-Verlag, 2001.

[20] A. J. Howell and H. Buxton. Learning gestures for visually mediated interaction. In P. H. Lewis and M. S. Nixon, editors, *Proceedings of British Machine Vision Conference*, pages 508–517, Southampton, UK, 1998. BMVA Press.

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

[22] A. J. Howell and H. Buxton. RBF network methods for face detection and attentional frames. *Neural Processing Letters*, 15, 2002 (In Press).

[23] M. Isaard and A. Blake. A mixed-state condensation tracker with automatic model-switching. In *Proceedings of International Conference on Computer Vision*, pages 107–112, Bombay, India, 1998. IEEE Computer Society Press.

[24] A. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *Proceedings of International Conference on Vision Systems (ICVS'99)*, Las Palmas de Gran Canaria, Spain, 1999.

[25] T. Jebara and A. Pentland. On reversing Jensen's Inequality. In *Advances in Neural Information Processing Systems*, volume 13, Denver,Colorado, 2000.

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

[27] N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pages 866–871. IEEE Computer Society Press, 1998.

[28] M. I. Jordan. Serial order: A parallel, distributed processing approach. In J. L. Elman and D. E. Rumelhart, editors, *Advances in Connectionist Theory: Speech*. Lawrence Erlbaum, Hillsdale, NJ, 1989.

[29] L. T. Kozlowski and J. E. Cutting. Recognising the sex of a walker from a dynamic poit-light display. *Perception and Psychophysics*, 12:575–580, 1977.

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

[31] S. J. McKenna, S. Gong, and Y. Raja. Face recognition in dynamic scenes. In A. F. Clark, editor, *Proceedings of British Machine Vision Conference*, pages 140–151, Colchester, UK, 1997. BMVA Press.

[32] J. Moody and C. Darken. Learning with localized receptive fields. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of 1988 Connectionist Models Summer School*, pages 133–143, Pittsburgh, PA, 1988. Morgan Kaufmann.

[33] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294, 1989.

[34] Y. Moses, Y. Adini, and S. Ullman. Face recognition: the problem of compensating for illumination changes. In J. O. Eklundh, editor, *Proceedings of European Conference on Computer Vision, Lecture Notes in Computer Science*, volume 800, pages 286–296, Stockholm, Sweden, 1994. Springer-Verlag.

[35] M. C. Mozer. Neural net architectures for temporal sequence processing. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Predicting the Future and Understanding the Past*, pages 243–264. Addison-Wesley, Redwood City, CA, 1994.

[36] N. Oliver, B. Rosario, and A. Pentland. Graphical models for recognising human interactions. In *Advances in Neural Information Processing Systems*, Denver, Colorado, 1998.

[37] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. In *International Conference on Vision Systems*, Gran Canaria, Spain, 19989.

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

[39] C. Pinhanez and A. F. Bobick. Human action detection using PNF propagation of temporal constraints. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, Santa-Barbara, CA, 1998.

[40] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[41] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[42] D. A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1, pages 305–313, San Mateo, CA, 1989. Morgan Kaufmann.

[43] M. Rosenblum and L. S. Davis. An improved radial basis function network for autonomous road-following. *IEEE Transactions on Neural Networks*, 7:1111–1120, 1996.

[44] J. Sherrah, S. Gong, A. J. Howell, and H. Buxton. Interpretation of group behaviour in visually mediated interaction. In *Proceedings of 15th International Conference on Pattern Recognition*, pages 266–269, Barcelona, Spain, 2000.

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

[46] M. Turk. Visual interaction with lifelike characters. In *Proceedings of International Conference on Automatic Face & Gesture Recognition*, pages 368–373, Killington, VT, 1996. IEEE Computer Society Press.

[47] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, 37:328–339, 1989.

[48] A. D. Wilson and A. F. Bobick. Recognition and interpretation of parametric gesture. In *Proceedings of International Conference on Computer Vision*, pages 329–336, Bombay, India, 1998. IEEE Computer Society Press.

[49] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 19:780–785, 1997.

[50] C. R. Wren and A. P. Pentland. Dynamic models of human motion. In *Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition*, pages 22–27, Nara, Japan, 1998. IEEE Computer Society Press.