Tracking and Recognising Hand Gestures Using Statistical Shape Models

T Ahmad, C J Taylor, A Lanitis and T F Cootes
Dept. of Medical Biophysics, University of Manchester
email: (tna, ctaylor, lan, bim)@wiau.mb.man.ac.uk

Abstract

Hand gesture recognition from video images is of considerable interest as a means of providing simple and intuitive man–machine interfaces. Possible applications range from replacing the mouse as a pointing device to virtual reality and communication with the deaf. We describe an approach to tracking a hand in an image sequence and recognising, in each video frame, which of five gestures it has adopted. A statistically based Point Distribution Model (PDM) is used to provide a compact parameterised description of the shape of the hand for any of the gestures or the transitions between them. The values of the resulting shape parameters are used in a statistical classifier to identify gestures. The model can be used as a deformable template to track a hand through a video sequence but this proves unreliable. We describe how a set of models, one for each of the five gestures, can be used for tracking with the appropriate model selected automatically. We shown that this results in reliable tracking and gesture recognition for two 'unseen' video sequences in which all the gestures are used.

1 Introduction

Hand gestures are used extensively in communication between humans. There are also many man–machine interfaces which involve a physical action but which are effectively gestural switches, selectors and computer mice all fall into this category. The need to simplify the control of increasingly complex appliances and the wish to provide more natural means of interacting with computers have led to considerable interest in recognising hand gestures from video images. Much of the existing work is specifically motivated by the potential for application in virtual reality systems.

In this paper we describe a system for hand tracking and gesture recognition based on the use of statistical shape models. We have described previously our basic approach to modelling variable shapes and locating them in images – indeed we have used the example of a hand to provide a simple illustration of the ideas. Here, we have looked at hand gestures more seriously, investigating the ability to track a hand reliably, even when its shape varies considerably with gesture, and testing the ability to classify gestures, even though some of the distinctions are subtle. We have limited our attention to gestures seen from above over a work–surface; the models we have used are two dimensional. We have described recently how our modelling scheme can be extended to deal properly with 3D objects as observed in 2D images [1]; as a result we should be able to build on the work described here to develop a more general system.

Examples of the five gestures used in our experiments are shown in Figure 1 . The gestures are intended to be useful in mouse–less interaction with a windows–based

BMVC 1995 doi:10.5244/C.9.40

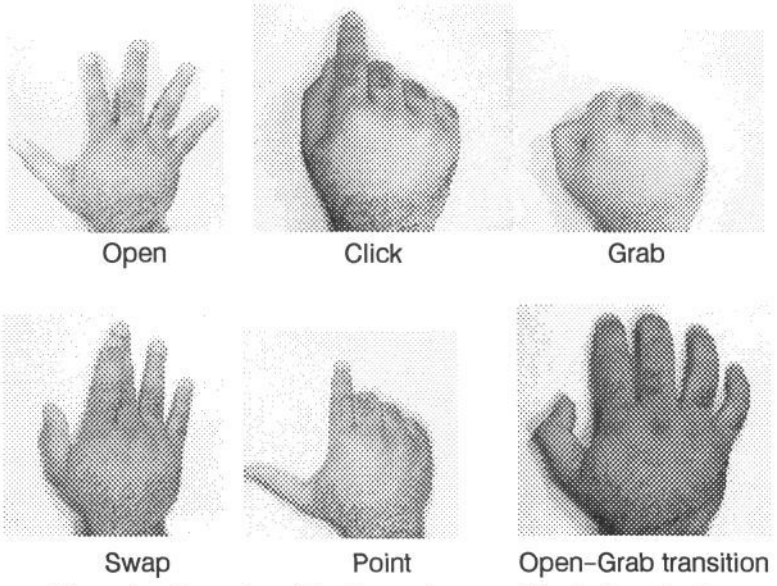

Figure 1 . Examples of the five gestures used in the investigation.

operating system and represented actions such as point, point-and-click, grab, and swap. They were selected in such a way that there were very large differences in shape between some gestures presenting a severe test for tracking methods but subtle differences between others providing a difficult classification task. Individual images of 'pure' gestures were available for model training. Continuous image sequences showing a hand moving and adopting different gestures were used for testing.

A variety of approaches to hand tracking and gesture recognition have been reported previously. Baudel and Beaudouin-Lafom [2], Cipolla et al [3], and Davis and Shah [4] all describe systems based on the use of a passive 'data glove' with markers that can be tracked relatively easily between frames. Cipolla et al recover 3D structure from image sequences but do not attempt to classify gestures. Davis and Shah track the motion of the finger tips and perform simple classification based on vectors representing their trajectories. Rehg and Kanade [5] describe a system which does not require special markers. They use a 3D articulated hand model which they fit to stereo data but do not attempt gesture recognition. Blake et al [6] describe a tracking system based on a real-time 'snake'; their system can deal with arbitrary pose, but treats the hand as a rigid object. Heap [7] shows that a 2D statistical model, of the same form as those used in our work, can be used to track a hand in real-time against a cluttered background. The experiments he describes do not involve very large variations in shape and no classification results are presented, though an approach to classification is proposed. The work described here was carried out roughly contemporaneously with Heap's. A similar approach is adopted but different aspects of the problem are explored. A more detailed description of the work is given in Ahmad's MSc thesis [8].

In the remainder of the paper we show how a 'multi-gesture' hand model can be built and used for gesture classification. We show that the model can be used for tracking hands in image sequences but that this approach is not particularly reliable. We describe how reliable tracking can be achieved using a set of models, one for each gesture, and selecting between them automatically. We present experimental results which show that image sequences can be interpreted successfully using multi-model tracking in combination with a multi-gesture model for classification.

2 Multi-Gesture Model

Robust tracking and gesture 'understanding' can most easily be achieved using a model-based method. Since we wish to deal with a variety of gestures, and indeed different instances of the same gesture will not be identical, a method of modelling which can deal with variable shape is required. We have used Point Distribution Models (PDMs), deformable templates generated by performing a statistical analysis on a set of training examples. We describe below how a multi-gesture model, capable of fitting to examples of all five target gestures, was created and used for gesture classification.

2.1 Model Training

We created a PDM from a training set of hand outlines comprising 14 examples of each of the five gestures. A set of 89 landmark points were placed in equivalent positions on each example. 'Primary' landmarks were placed at the tips of the fingers and the creases between them, others were equally spaced along the boundary segments between the primary landmarks. For gestures such as 'grab' and 'point', where some fingers are closed, the finger tip landmarks were placed on the knuckles, giving the effect of short fingers. A PDM is generated by performing a least squares alignment of the members of the training set, as represented by their landmark points, followed by a Principal Component Analysis of the vectors formed by concatenating the ordinates (x_i,y_i) of the aligned landmark points for each example; the details have been described elsewhere [9]. The result is a vector representing the mean shape and a set of basis vectors representing the main modes of variation around the mean. Any of the training examples and similar new examples can be approximated using

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{1}$$

Where	$\mathbf{x} = \{x_0, y_0, \ldots x_n, y_n\}^T$	is an instance of the shape model,
	$\bar{\mathbf{x}}$	is the mean shape,
	$\mathbf{P} = \{\mathbf{p}_1 \ldots \mathbf{p}_t\}$	is a matrix of basis vectors,
	$\mathbf{b} = \{b_1 \ldots b_t\}^T$	is a vector of shape parameters.

Given $\mathbf{x}$ we can compute $\mathbf{b}$ using

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{2}$$

For the multi-gesture hand model 95% of the variability in the training set could be explained using six basis functions (modes), 99% using 11 modes. The mean shape and the main modes of variation are shown in Figure 2.

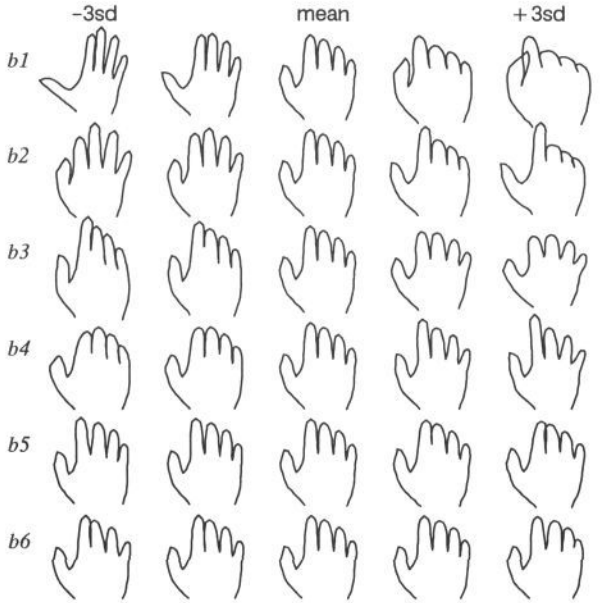

Figure 2 . Modes of variation for the multi-gesture model. Each row shows the effect of varying one of the shape parameters keeping the others at zero.

2.2 Gesture Classification

We have shown above that examples of hand shapes representing the gestures of interest can be reconstructed reasonably accurately using a model controlled by eleven **b** parameters. Given a shape **x**, equation 2 can be used to find the vector **b** of shape parameters which results in the model instance which best approximates **x**. These parameters form a compact description of the given shape and can be used as the basis for gesture classification. The examples in the training set can be used to estimate the distribution of **b** values for each gesture class, i, in terms of its mean and covariance $\mathbf{K}_i$. Test shapes can be classified by computing their shape parameters **b** and assigning them to the nearest class using the Mahalanobis distance metric

$$D_i = (\mathbf{b} - \overline{\mathbf{b}_i})^T \mathbf{K}_i^{-1} (\mathbf{b} - \overline{\mathbf{b}_i}) \qquad (3)$$

Table 1 shows the result of applying this classification scheme to the training data. It can be seen that most of the examples are classified correctly though there is some confusion between the most similar classes 'open' and 'swap'. The overall error rate is 5.7% though, of course, the results for unseen data which are presented later are a more meaningful indication of classification accuracy.

3 Tracking Gestures

The classification method described above assumes that the landmark points defining the shape to be classified are known. This is true for the training data, which was manually annotated, but for new image sequences it is necessary to locate the hand and fit the model in each frame. We can achieve this by using the hand PDM to create an Active Shape Model (ASM). The details have been described elsewhere [9] but we

Table 1. Results of gesture classification on the training data.

Gesture	Classification				
	Open	Grab	Swap	Point	Click
Open	12	0	2	0	0
Grab	0	14	0	0	0
Swap	2	0	12	0	0
Point	0	0	0	14	0
Click	0	0	0	0	14

outline the method below and present results obtained for a test sequence using the multi-gesture model. These results are not particularly impressive; we show why and describe a modified approach for which much better results are obtained.

3.1 Image Search with the Multi-Gesture Model

ASM search involves projecting an instance of a PDM into an image and modifying it iteratively to better fit the data. The model instance is initialised by choosing a pose (position, scale and orientation) and **b** vector (typically **0**). At each model point a grey-level profile is collected, from the image, in a direction perpendicular to the model boundary. A search along each profile finds a candidate for the true position of that model point using the best fit to a grey-level model derived from the training set. The model points attempt to move to these new locations by changing the pose and **b** vector. The modification to **b** required to approximate the model points to their new proposed positions can be computed directly using equation 2. The modified model is reprojected into the image and iteration continues until a stable result is obtained. By modifying the pose and shape parameters, rather than the model points directly, the model is constrained to find solutions which are similar in shape to those in the training set. In the experiments described below we used a multi-resolution version of the algorithm which is generally more robust and leads to more accurate solutions in fewer iterations [10].

Two image sequences were obtained, each showing a hand moving and changing continuously between gestures; all the target gestures were represented. Sequence 1 contained 454 frames which, to form the basis of assessment of the automatic system, were manually classified as belonging to one of the five gesture classes (267 frames) or a transition between gestures (187 frames). Similarly, Sequence 2 contained 357 frames showing a more rapidly changing hand with 183 gesture frames and 174 transition frames. Tracking was performed by initialising the model in the centre of the image with the mean shape and scale for the first image in a sequence, then using the pose and shape solution found for each frame to initialise the search in the next. The results for Sequence 1 are summarised in Table 2. These show that many of the frames are classified correctly but that a significant number (22%) fail. On observing the operation of the tracker it was clear that the primary reason for these errors was a

Table 2. Classification results for video Sequence 1 using the multi-gesture model.

Gesture	Classification				
	Open	Grab	Swap	Point	Click
Open	60	27	0	0	0
Grab	0	92	0	0	0
Swap	0	0	21	0	0
Point	0	30	0	13	1
Click	0	0	0	0	23

failure to track the hand correctly. This is illustrated clearly in Figure 3 which shows the model fit value, a measure of the evidential support for the solution, plotted against frame number. A low value of the fit function implies a good fit. The fit value suddenly increases at about frame 250 and stays high until tracking recovers at about frame 370. Since the model was not fitting correctly to the data for the intervening frames classification was generally incorrect.

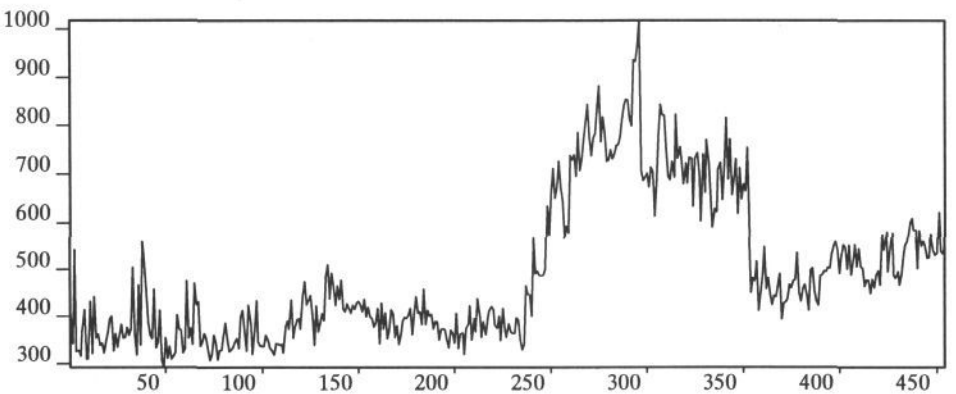

Figure 3 . Model fit value plotted against frame number for tracking using the multi-gesture model in Sequence 1.

The reason for this failure is fairly easy to understand. The key to all model-based schemes is specificity - the ability to allow only solutions which are consistent with the form of solution expected. An ASM uses a flexible template but one which can only vary in ways found in the training set. In this case, however, we have introduced several classes of shape into the same model. These form separate clusters in shape (**b**) space, indeed we rely on this for classification. Shapes which lie between the clusters are also allowed; some will represent transitions between target gestures but others will not represent any feasible gesture. Thus our model lacks specificity and can find 'illegal' solutions. Once the current solution gets too far away from the correct solution it becomes difficult for the system to recover. This hypothesis was verified by partitioning the training set and building separate models of each gesture. These models were used to fit to the corresponding gestures in the sequence. The

results were very good but this did not, of course, represent a practical way forward since the appropriate model to use could only be known once the gesture had been classified.

3.2 Tracking with Weighted Models

We have already seen that a multi-gesture model can be used to produce a reliable classification as long as the model has been fitted accurately to an image. This suggests the idea of using single gesture models to track reliably, using the multi-gesture classifier to select the appropriate model. This is possible because, given a shape generated by one model, it is straightforward to find the model parameters which give the best approximation to that shape for a second model, by using equation 2. Thus we can use one model for tracking and another for classification. Similarly we can easily switch between using different models for tracking on the basis of classification results.

The scheme we have just outlined would be perfectly feasible if only pure gestures were present in video sequences. In practice, tracking would fail during transitions between gestures because the single gesture models would not be able to fit successfully to these frames. The solution we adopted was to use weighted models which were trained using some examples of every gesture but many more of their target gesture. The idea was to exploit the specificity of single-gesture models whilst introducing enough additional variability to cope with starting the transition to each of the other gestures. The models we used in our experiments used the training examples of the target gesture 10 times and the remaining training examples once. Five such weighted models were generated, one for each gesture. The mean shape and main modes of variation of the 'open-weighted' model are shown in Figure 4.

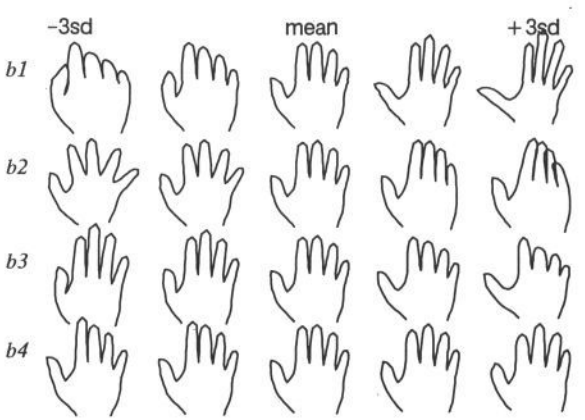

Figure 4 . Modes of variation for the 'open-weighted' model

Tracking using weighted models in conjunction with the multi-gesture model to classify gestures and automatically select the appropriate model for the subsequent frame proved much more successful than tracking using the multi-gesture model. An example of a transition between two gestures is shown in Figure 5 where the

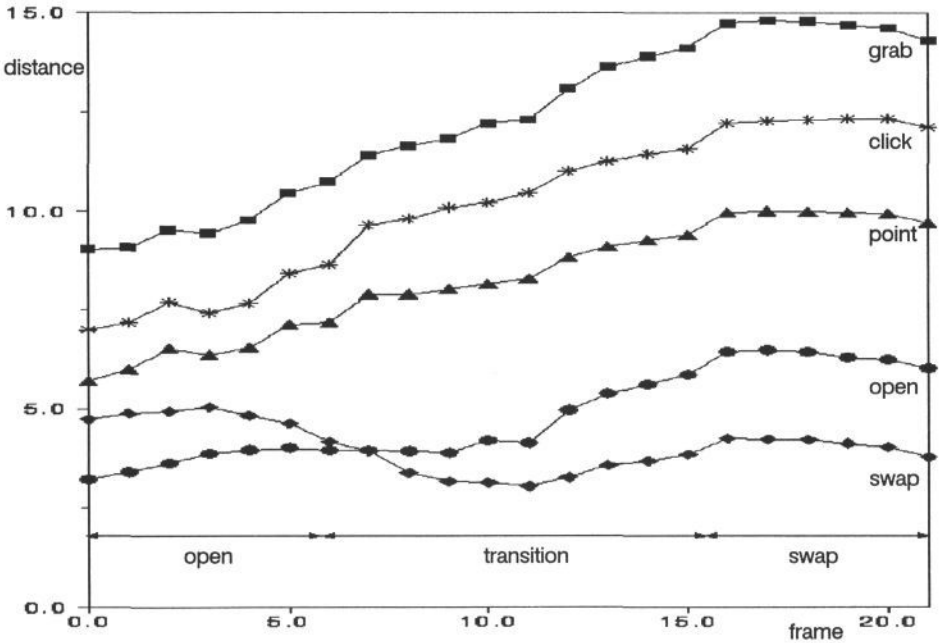

Figure 5. Mahalanobis distance to each gesture class plotted against frame number. A transition from the 'open' model to the 'swap' model takes place at frame 9.

Mahalanobis distance to each class is plotted against frame number; the 'swap' model was selected instead of the 'open' model at frame 9.

Although this new approach was generally more reliable, tracking occasionally failed. When this happened the effect was catastrophic because an arbitrary model would be selected and, if it was incorrect, would stand no chance of fitting in subsequent frames. To deal with this problem a recovery mechanism was implemented. The model fit value was calculated for each frame. If several successive frames (4 in our experiments) gave a fit value above a chosen threshold recovery was initiated. During recovery an attempt was made to fit each of the weighted models to the current image – the model giving the best (lowest) fit value was selected. The choice of threshold was not critical.

4 Results for Image Sequences

Classification results obtained for Sequence 1 using the scheme described above are shown in Table 3. All the gesture frames were classified correctly; very similar results were obtained without error recovery. Table 4 summarises the results for Sequence 2 both with and without error recovery. Again, all the gesture frames were classified correctly with the full scheme, but in this case performance was much worse without recovery. The effect of recovery is illustrated in Figure 6 which shows fit value plotted against frame number for Sequence 2, both with and without recovery; the positions where recovery takes place, when enabled, are marked. Without recovery, tracking fails at around frame 225 and does not recover until around frame 320; with recovery, tracking is maintained throughout.

Table 3. Gesture classification results for video Sequence 1 using weighted models with automatic model selection and error recovery.

Gesture	Classification				
	Open	Grab	Swap	Point	Click
Open	87	0	0	0	0
Grab	0	92	0	0	0
Swap	0	0	21	0	0
Point	0	0	0	44	0
Click	0	0	0	0	23

Table 4. Gesture classification results for video Sequence 2 using weighted models with automatic model selection with and (without) error recovery.

Gesture	Classification				
	Open	Grab	Swap	Point	Click
Open	122 (108)	0 (14)	0	0	0
Grab	0	33	0	0	0
Swap	0 (2)	0 (1)	13 (10)	0	0
Point	0	0	0	8 (3)	0 (5)
Click	0	0	0	0	7

5 Conclusions and Discussion

We have shown that we can successfully track hands and classify gestures in video sequences. The task was made difficult by including examples of both radically and subtly different gestures. The former make it difficult to create a specific model which will track reliably, the latter make classification difficult. Our solution uses a separate statistical model for each gesture for tracking and a multi-gesture model for classification and selection of the appropriate tracking model.

The results of this investigation highlight the need to model shape space distributions in more sophisticated ways. In ASM search we currently treat the distribution of allowed shapes in shape space as unimodal, allowing any solution which falls within a given Mahalanobis distance of the mean shape. In our example this gave rise to a lack of specificity which caused tracking to fail. Ideally we should generalise this approach, perhaps using a mixture model in shape space. Unfortunately it is not obvious how to modify the ASM search algorithm to take into account a more complicated form of distribution. We have recently described a more general form of shape model based on a neural net performing non-linear Principal Component Analysis [11]. This may offer a partial solution but we have yet to test this hypothesis.

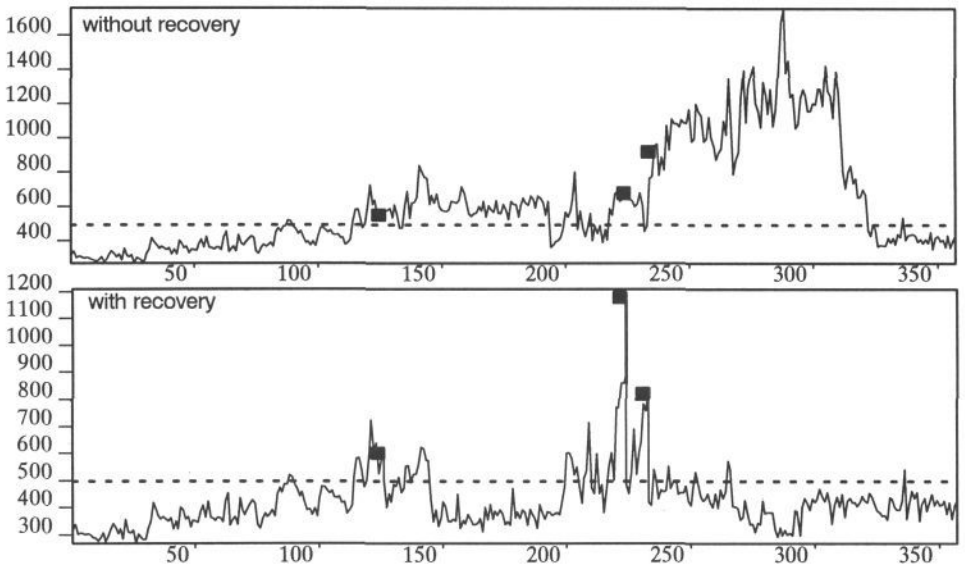

Figure 6 . Model fit value plotted against frame number for Sequence 2, both with and without recovery. The points where recovery takes place (when enabled) are marked.

6 References

1. Cootes TF, Di Mauro EC, Taylor CJ, Lanitis A. "Flexible 3D models from uncalibrated cameras", To appear in the Proc. of BMVC, 1995.
2. Baudel T, Beaudouin-Lafom M. "Charade: remote control of objects using gestures", Com ACM, pp 28–35, 1993.
3. Cipolla R, Okamoto Y, Kuno Y. "Robust structure from motion using motion parallax", in Proc. ICCV, IEEE Computer Society Press, 1993, pp 374–382.
4. Davis J, Shah M. "Recognising hand gestures", in Proc. ECCV 1, Springer Verlag, Berlin, 1994, pp 331–340.
5. Rehg JM, Kanade T. "Visual tracking of high dof articulated structures: an application to human hand tracking", in Proc. ECCV 2, Springer Verlag, Berlin, 1994, pp 35–45.
6. Blake A, Curwen R, Zisserman A. "A framework for spatiotemporal control in the tracking of visual contours", IJCV 11, pp 127–145, 1993.
7. Heap A. "Robust real-time hand tracking and gesture recognition using smart snakes", Internal Report, Olivetti Research Ltd, 1995.
8. Ahmad T. "A model-based hand gesture recognition system", MSc Thesis, University of Manchester, 1994.
9. Cootes TF, Taylor CJ, Cooper DH, Graham J. "Active Shape Models - Their Training and Application", Computer Vision and Image Understanding, 61, pp 38–59, 1995.
10. Cootes TF, Taylor CJ, Lanitis A. "Active Shape Models: Evaluation of a Multi-Resolution Method for Improving Image Search", in Proc. BMVC, BMVA Press, 1994, pp 327–336.
11. Sozou PD, Cootes TF, Taylor CJ, Di Mauro EC. "Non-linear point distribution modelling using a multi-layer perceptron", To appear in the Proc. of BMVC, 1995.