Tracking and Segmenting People in Varying Lighting Conditions using Colour

Yogesh Raja, Stephen J. McKenna and Shaogang Gong
Machine Vision Laboratory, Department of Computer Science
Queen Mary and Westfield College, London E1 4NS
E-mail: {jpmetal,stephen,sgg}@dcs.qmw.ac.uk

Abstract

Colour cues were used to obtain robust detection and tracking of people in relatively unconstrained dynamic scenes. Gaussian mixture models were used to estimate probability densities of colour for skin, clothing and background. These models were used to detect, track and segment people, faces and hands. A technique for dynamically updating the models to accommodate changes in apparent colour due to varying lighting conditions was used. Two applications are highlighted: (1) actor segmentation for virtual studios, and (2) focus of attention for face and gesture recognition systems. A system implemented on a 200MHz PC tracks multiple objects in real-time.

1 Introduction

Colour has been used in machine vision for tasks such as segmentation [11] and recognition [4, 12]. Colour offers many advantages over geometric information for these problems such as robustness under partial occlusion, rotation in depth, scale changes and resolution changes [12]. Furthermore, the use of colour enables real-time performance on modest hardware platforms. Here, we show how colour is used for fast, robust detection, tracking and coarse segmentation. Multiple colours may be modelled without the difficulties encountered by some other techniques [4, 14]. The framework can track multi-coloured objects [8] and multiple objects can be tracked simultaneously.

Two applications in particular have motivated this work. The first is the increasing requirement in the broadcasting industry for a method of superimposing actors onto artificially generated studios. Currently, these "virtual studio" applications are enabled through the use of "chroma-keying", a technique which segments non-blue regions of video images. This necessitates painstaking preparation of the environment. An alternative is desired that will minimise the effort required. Secondly, the work is motivated by the need for computationally efficient and robust focusing of attention for face and gesture recognition systems.

In Section 2, a technique for modelling colour using Gaussian mixture models is described and compared with more traditional histogram approaches. Section 3 describes a real-time, colour-based tracking framework. Section 4 introduces a method for updating mixture models and Section 5 shows how this ability to adapt is controlled to prevent the build-up of errors over time. Section 6 describes the application of the framework for virtual studios. In Section 7, a system is described which can track multiple objects and its use for face and gesture recognition systems is illustrated. Finally, a discussion is given in Section 8.

2 Modelling Colour Distributions

Methods have been proposed for modelling skin-coloured regions for face detection and tracking (e.g. [3, 10]). In particular, a system constructed by Wren et al. [14] known as "Pfinder" enabled tracking of a person in a controlled environment with a static camera. Each pixel in an image had an associated feature vector comprising spatial and colour components. These feature vectors were clustered to yield a collection of "blobs" defined by spatial and spectral similarity. The collection of blobs constituted a representation of the person. This limited tracking to a person with homogeneously coloured regions against a non-changing background.

Swain and Ballard [12] described a scheme which used histograms for modelling the colours of an object and two algorithms for matching such models with image regions. The colour space was quantised through the histogram's structure which comprises a number of "bins". An algorithm known as "histogram intersection" was used for matching image histograms with model histograms. In order to find an object in an image, "histogram backprojection" was used.

In this paper, probability densities are estimated from the colours of the background and peoples' clothing, head and hands. The assumption is that a person of interest in an image will form a *spatially contiguous* region in the image plane (i.e. the parts of a person will not be fragmented due to partial occlusion). With an additional assumption that the set of colours for the person and background are relatively distinct, the pixels belonging to the person may be treated as a statistical distribution in the image plane. A person's position and size may then be computed by estimating the mean and variance of this distribution.

2.1 Modelling the Colour of a Person using Mixture Models

Although colour histograms can be used to estimate densities in colour space, the level of quantisation imposed on the colour space influences the resulting density. If the number of bins n is too high, the estimated density will be "noisy" and many bins will be empty. If n is too low, density structure is smoothed away. Histograms are effective only when n can be kept relatively low and where sufficient data are available. A potentially more effective "semi-parametric" approach for colour density estimation is to use Gaussian mixture models. In this approach, a number of Gaussian functions are taken as an approximation to a multi-modal distribution in colour space and conditional probabilities are then computed for colour pixels [6, 8]. The conditional density for a pixel, $\mathbf{x}$, belonging to an object, $\mathcal{O}$, is modelled as a Gaussian mixture with m components:

$$p(\mathbf{x}|\mathcal{O}) = \sum_{j=1}^{m} p(\mathbf{x}|j)\pi(j)$$

where a mixing parameter $\pi(j)$ is the prior probability that $\mathbf{x}$ was generated by the j^{th} component, $\sum_{j=1}^{m} \pi(j) = 1$. Each mixture component, $p(\mathbf{x}|j)$, is a Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Expectation-Maximisation (EM) provides an effective maximum-likelihood algorithm for fitting such a mixture to a data set [1, 9].

2.2 Automatic Model Order Selection

A constructive algorithm for automatic model order selection has been implemented. Firstly, two data sets are obtained: a training set and a validation set. The validation set is chosen such that a resulting model will generalise accordingly. A mixture model is initialised with a small number of components, typically one. Model order is then adapted by performing the following steps: (1) EM is applied to fit the training data; (2) The log-likelihood for the validation set is computed; (3) The component with lowest responsibility for the training set is split into two separate components by performing Singular Value Decomposition on its covariance matrix Σ:

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}$$

The diagonal matrix $\boldsymbol{\Lambda}$ contains the singular values for component j. The two new components both have the covariance matrix obtained by halving the largest value in $\boldsymbol{\Lambda}$ and multiplying out the above equation again to obtain $\boldsymbol{\Sigma}_{new}$. The means are found by shifting along the principal axis.

These steps are then repeated and a plot of the log-likelihood for the validation set is obtained. It is assumed that the optimal model order corresponds to the peak in this log-likelihood plot.

3 Tracking without Dynamic Updates

Under stable lighting conditions, person tracking and segmentation can be effectively performed using predetermined mixture models in two-dimensional hue-saturation space which permits some level of robustness against limited brightness changes. Probabilities are computed for pixels and the size and position of the object are estimated from the resulting distribution in the image plane [8]. Object position at frame t is taken to be the mean $\boldsymbol{m}_t = (m_x, m_y)$ and object size is estimated from the standard deviation $\boldsymbol{\sigma}_t = (\sigma_x, \sigma_y)$.

More precisely, for a given frame t, the object position $\boldsymbol{m}_t$ is estimated as an offset from the position $\boldsymbol{m}_{t-1}$:

$$\boldsymbol{m}_t = \boldsymbol{m}_{t-1} + \frac{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{m}_{t-1})}{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})}$$

where $\mathbf{x}$ ranges over all image coordinates in the region of interest and $\boldsymbol{\xi}_{\mathbf{x}}$ is the colour point at image position $\mathbf{x}$. To improve accuracy, probabilities $p(\boldsymbol{\xi}_{\mathbf{x}})$ are thresholded. Values lower than the threshold are taken to be background and are consequently set to zero in order to nullify their influence on the estimation of $\boldsymbol{m}_t$ and $\boldsymbol{\sigma}_t$.

The size of the object is estimated by computing the standard deviation of the image probability density:

$$\boldsymbol{\sigma}_t = \sqrt{\frac{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})\{(\mathbf{x} - \boldsymbol{m}_{t-1}) - \boldsymbol{m}_t\}^2}{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})}}$$

This model can be used for tracking with partial occlusion, scale changes and a moving camera.

4 Dynamically Updating the Colour Model

Colour appearance is often unstable due to changes in both background and foreground lighting. The colour constancy problem has been addressed mainly through the formulation of physics-based models of light formation (e.g.[2]). Here, we adopt a statistical approach in which colour distributions are estimated over time. Under the assumption that lighting conditions change smoothly over time, the model is adapted to reflect the changing appearance of the object being tracked [7]. At each frame, t, a new set of pixels, $X^{(t)}$, is sampled from the object and used to update the mixture model[1]. These new data sample a slowly varying non-stationary signal. Let $\psi^{(t)}$ denote the sum of the responsibilities for the data in frame t, $\psi^{(t)} = \sum_{\mathbf{x}\in X^{(t)}} p(j|\mathbf{x})$. The parameters are first estimated for each mixture component, j, using only the new data, $X^{(t)}$, from frame t:

$$\boldsymbol{\mu}^{(t)} = \frac{\sum p(j|\mathbf{x})\mathbf{x}}{\psi^{(t)}}, \qquad \pi^{(t)} = \frac{\psi^{(t)}}{N^{(t)}}$$

$$\boldsymbol{\Sigma}^{(t)} = \frac{\sum p(j|\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_{t-1})^T(\mathbf{x}-\boldsymbol{\mu}_{t-1})}{\psi^{(t)}}$$

where $N^{(t)}$ denotes the number of pixels in the new data set and all summations are over $\mathbf{x} \in X^{(t)}$. The mixture model components then have their parameters updated using weighted sums of the previous recursive estimates, $(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1}, \pi_{t-1})$, estimates based on the new data, $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \pi^{(t)})$, and estimates based on the old data, $(\boldsymbol{\mu}^{(t-L-1)}, \boldsymbol{\Sigma}^{(t-L-1)}, \pi^{(t-L-1)})$:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \frac{\psi^{(t)}}{D_t}(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}_{t-1}) - \frac{\psi^{(t-L-1)}}{D_t}(\boldsymbol{\mu}^{(t-L-1)} - \boldsymbol{\mu}_{t-1})$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t-1} + \frac{\psi^{(t)}}{D_t}(\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}_{t-1}) - \frac{\psi^{(t-L-1)}}{D_t}(\boldsymbol{\Sigma}^{(t-L-1)} - \boldsymbol{\Sigma}_{t-1})$$

$$\begin{aligned}\pi_t &= \pi_{t-1} + \frac{N^{(t)}}{\sum_{\tau=t-L}^{t} N^{(\tau)}}\left(\pi^{(t)} - \pi_{t-1}\right) \\ &\quad - \frac{N^{(t-L-1)}}{\sum_{\tau=t-L}^{t} N^{(\tau)}}\left(\pi^{(t-L-1)} - \pi_{t-1}\right)\end{aligned}$$

where $D_t = \sum_{\tau=t-L}^{t} \psi^{(\tau)}$. The following approximations are used for efficiency:

$$\psi^{(t-L-1)} \approx \frac{D_{t-1}}{L+1}$$

$$D_t \approx (1 - 1/(L+1))D_{t-1} + \psi^{(t)}$$

The parameter L controls the adaptivity of the model[2].

During the processing of a sequence, new samples of data for adaptation are gathered from a region of appropriate aspect ratio centred on the estimated object centroid. It is assumed that these data form a representative sample of the objects' colours. This will hold for a large class of objects.

In order to boot-strap the tracker for object detection and re-initialisation after a tracking failure, a set of predetermined generic object colour models which perform reasonably in a wide range of illumination conditions are used. Once an object is being tracked, the model adapts and improves tracking performance by becoming specific to the observed conditions.

5 Selective Adaptation

An obvious problem with adapting a colour model over time is the lack of ground-truth. Any colour-based tracker can lose the object it is tracking due, for example, to occlusion. If such errors go undetected the colour model will adapt to image regions which do not correspond to the object. This is clearly undesirable. In order to help alleviate this problem, observed log-likelihood measurements were used to detect erroneous frames. Colour data from these frames were not used to adapt the object's colour model.

The adaptive mixture model seeks to maximise the log-likelihood of the colour data over time. The normalised log-likelihood, $\mathcal{L}$, of the data, $X^{(t)}$, observed from the object at time t is given by:

$$\mathcal{L} = \frac{1}{N^{(t)}} \sum_{\mathbf{x}\in X^{(t)}} \log p(\mathbf{x}|\mathcal{O})$$

At each time frame, $\mathcal{L}$ is evaluated. If the tracker loses the object there is often a sudden, large drop in the value of $\mathcal{L}$. This provides a way to detect tracker failure. Adaptation is then suspended until the object is again tracked with sufficiently high likelihood.

A temporal median filter was used to compute a threshold, T. Adaptation was only performed when $\mathcal{L} > T$. The median, ν, and standard deviation, σ, of $\mathcal{L}$ were computed for the n most recent above-threshold frames, where $n \leq L$. The threshold was $T = \nu - k\sigma$, where k was a constant.

Figures 1 and 2 illustrate the use of the mixture model for face tracking and the advantage of an adaptive model over a non-adaptive one. In this sequence, the apparent colour of the person's face changes significantly upon entering the laboratory due to the difference in lighting conditions to that of the corridor. In Figure 1, a non-adaptive model was trained on the first image of the sequence and used to track throughout. It was unable to cope with the varying conditions and failure eventually occured. In Figure 2, the model was allowed to adapt and successfully maintained lock on the face. The small box outlines the region from which the data for adaptation were sampled.

[1] Throughout this paper, superscript $^{(t)}$ denotes a quantity based only on data from frame t. Subscripts denote recursive estimates.

[2] Setting $L = t$ and ignoring terms based on frame $t - L - 1$ gives a stochastic algorithm for estimating a Gaussian mixture for a stationary signal [1, 13].

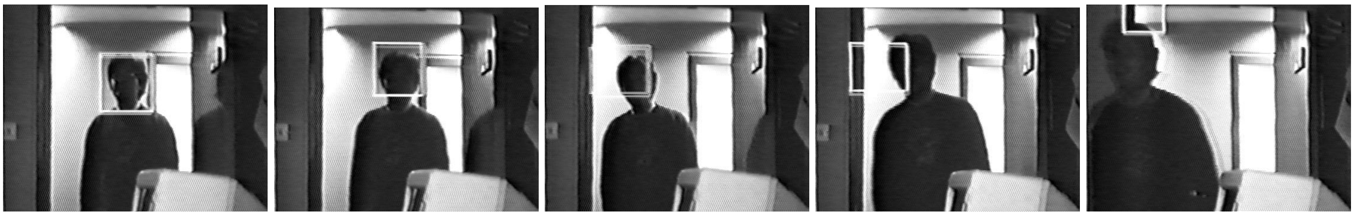

Figure 1. Five frames from a sequence in which a face was tracked using a non-adaptive model. Apparent colour changed due to different spectral composition of lighting in the corridor and the laboratory.

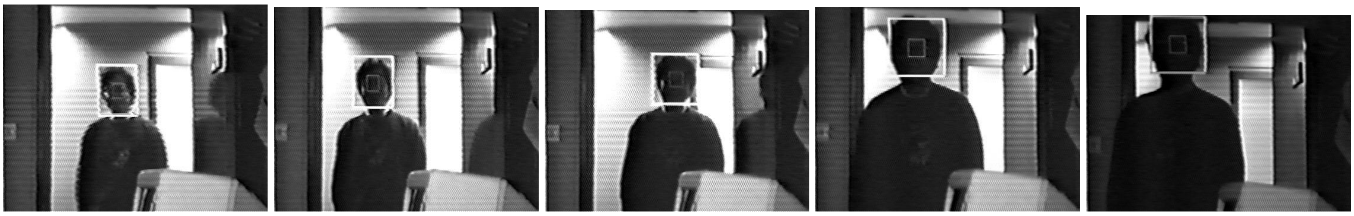

Figure 2. The sequence depicted in *Figure 1* tracked with an adaptive colour model. Here, the model adapts to cope with the change in apparent colour.

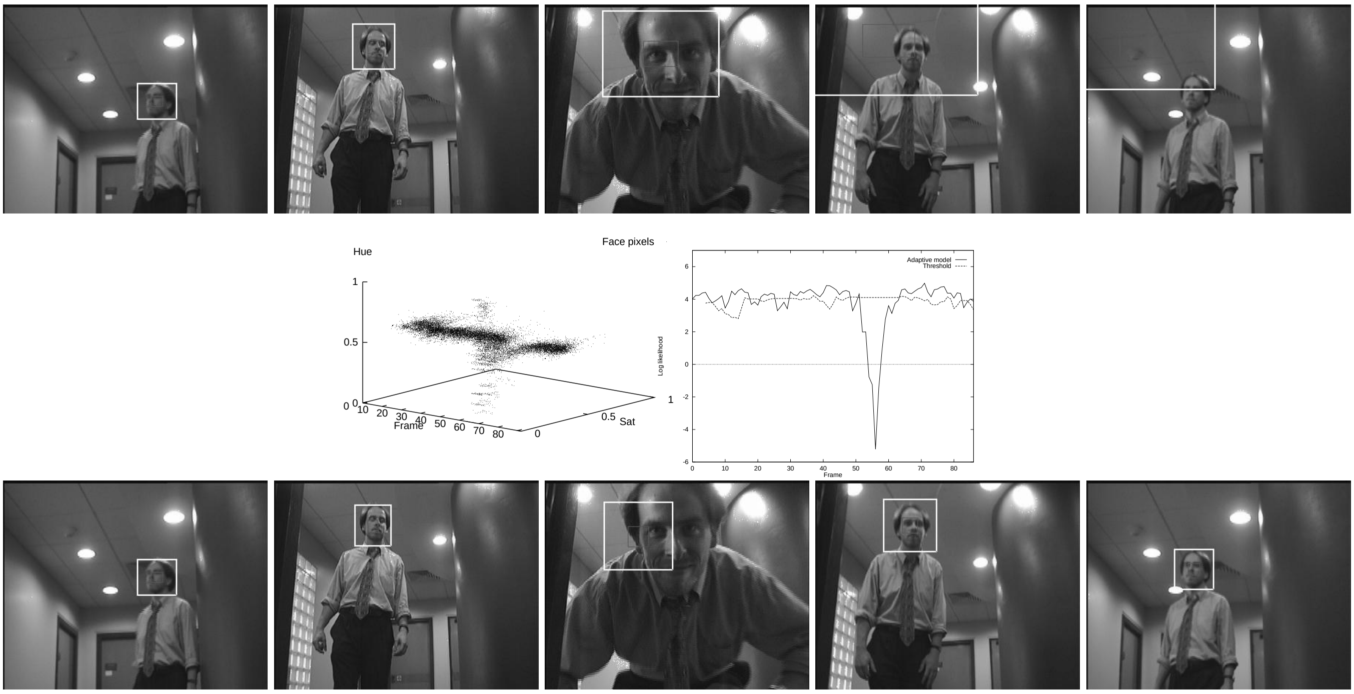

Figure 3. At the top are frames 35, 45, 55, 65 and 75 from a sequence with strong directional and exterior illumination. The walls have a fleshy tone. At around frame 55, the subject rapidly approaches the camera situated in a doorway, resulting in rapid changes in illumination, scale and auto-iris parameters. This can be seen in the 3D plot of the hue-saturation distribution over time. In the top sequence, the model was allowed to adapt in every frame, resulting in failure at around frame 60. The lower sequence illustrates the use of selective adaptation. The right-hand plot shows the normalised log-likelihood measurements and the adaptation threshold .

Figure 3 illustrates the advantage of selective adaptation. The person moved through challenging tracking conditions, before approaching the camera at close range (frames 50-60). Since the camera was placed in the doorway of another room with its own lighting conditions, the person's face underwent a large, sudden and temporary change in apparent colour. When adaptation was performed in every frame, this sudden change had a drastic effect on the model and ultimately led the tracker to fail when the person receded into the corridor. With selective adaptation, these sudden changes were treated as outliers and adaptation was suspended, permitting the tracker to recover.

6 Modelling Foreground and Background in Virtual Studios

Colour mixture models were used to track and segment people wearing multi-coloured clothing. If the appearance of a person consists of several distinct and differently coloured patches, then it may be beneficial to model each patch using a separate colour model (see e.g. [4]). Furthermore, if such colour patches are relatively homogenous, each can be directly modelled using a single Gaussian thus avoiding the need for the iterative EM algorithm. However, many objects cannot be thus decomposed and their colours are better modelled using a mixture.

The distributions in colour space formed by multi-coloured objects are multi-modal and can span wide areas of the colour space. Thresholding probabilities generated by a foreground model alone is often ineffective due to severe overlap between background and foreground colour distributions. In virtual studios, it is computationally desirable to model the colour distribution of the background scene in addition to the objects to be tracked. Given density estimates for both the object, $\mathcal{O}$, and the background scene, $\mathcal{S}$, the probability that a pixel, $\boldsymbol{\xi}$, belongs to the object is given by the posterior probability $P(\mathcal{O}|\boldsymbol{\xi})$:

$$P(\mathcal{O}|\boldsymbol{\xi}) = \frac{p(\boldsymbol{\xi}|\mathcal{O})P(\mathcal{O})}{p(\boldsymbol{\xi}|\mathcal{O})P(\mathcal{O}) + p(\boldsymbol{\xi}|\mathcal{S})P(\mathcal{S})}$$

The probability of misclassifying a pixel is minimised by classifying it as the class with the maximum posterior probability. The prior probability, $P(\mathcal{O})$, was set to reflect the expected size of the object within the search area of the scene $[P(\mathcal{S}) = 1-P(\mathcal{O})]$. This approach has the advantage that object and scene models can be acquired independently. In the virtual studio scenario, this means that a single background scene model can be acquired and subsequently used with many different people. Figure 4 shows an example in which both the person and the background have been modelled. Pixels were classified as person or background by computing their posterior probabilities with the prior probabilities set to $P(\mathcal{S}) = P(\mathcal{O}) = 0.5$. A multi-resolution approach was taken in which segmentation was performed in a coarse-to-fine manner. Once the position and size of the bounding box had been estimated, superimposition of the object onto an alternative background sequence was performed. Only pixels inside the search area of the tracker were classified. All pixels outside this area were rendered as background.

7 Focus of Attention for Face and Gesture Recognition

A system for tracking multiple objects based on motion was reported at FG'96 [5]. A similar approach is adopted here for detecting and tracking multiple objects based on their colour. In particular, a skin colour model can be used to efficiently focus attention on faces and hands for face and gesture recognition processes. Figure 5 shows such a system running on a 200MHz PC at a frame rate of around 15Hz. Regions of high colour probability are grouped and matched from frame to frame using time-symmetric matching to maintain a consistent list of objects. Objects are dynamically initialised and deleted from the list depending upon confidence measures derived from recent colour evidence. Kalman filters can be used to track each object.

8 Discussion

We have described how mixtures of Gaussians can be used to model the statistical distribution of colours of a person. A framework has been implemented that tracks and segments regions based on the probability distribution they induce in the image plane. A method has been described that dynamically updates the mixture model to accommodate lighting changes, controlled by a mechanism for detecting tracking errors. A combined foreground and background model has been used for coarse segmentation and virtual studio superimposition. A multi-object skin colour tracker was used for focus of attention for face and gesture recognition. Current work is concentrating on fusion of motion, colour and shape for detection, tracking and dynamic segmentation.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Cambridge University Press, 1995.

[2] D. A. Forsyth. *Colour Constancy and its Applications in Machine Vision*. PhD thesis, University of Oxford, 1988.

[3] R. Kjeldsen and J. Kender. Finding skin in color images. In *2nd Int. Conf. on Auto. Face and Gest. Recog.*, 1996.

[4] J. Matas, R. Marik, and J. Kittler. On representation and matching of multi-coloured objects. In *ICCV*, pages 726–732, 1995.

[5] S. McKenna and S. Gong. Tracking faces. In *2nd Int. Conf. Face and Gesture Recognition*, pages 271–276, 1996.

[6] S. McKenna, S. Gong, and Y. Raja. Face recognition in dynamic scenes. In *BMVC*, 1997.

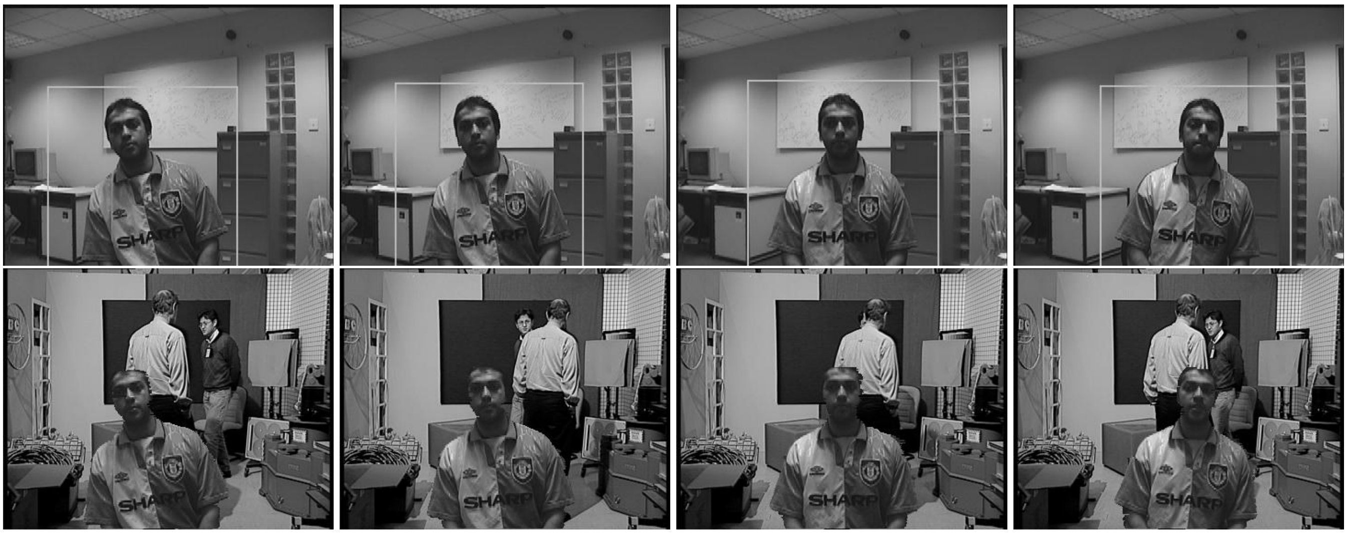

Figure 4. Segmentation results. The top row outlines the tracked region for segmentation and the second row illustrates superimposition onto an alternative sequence.

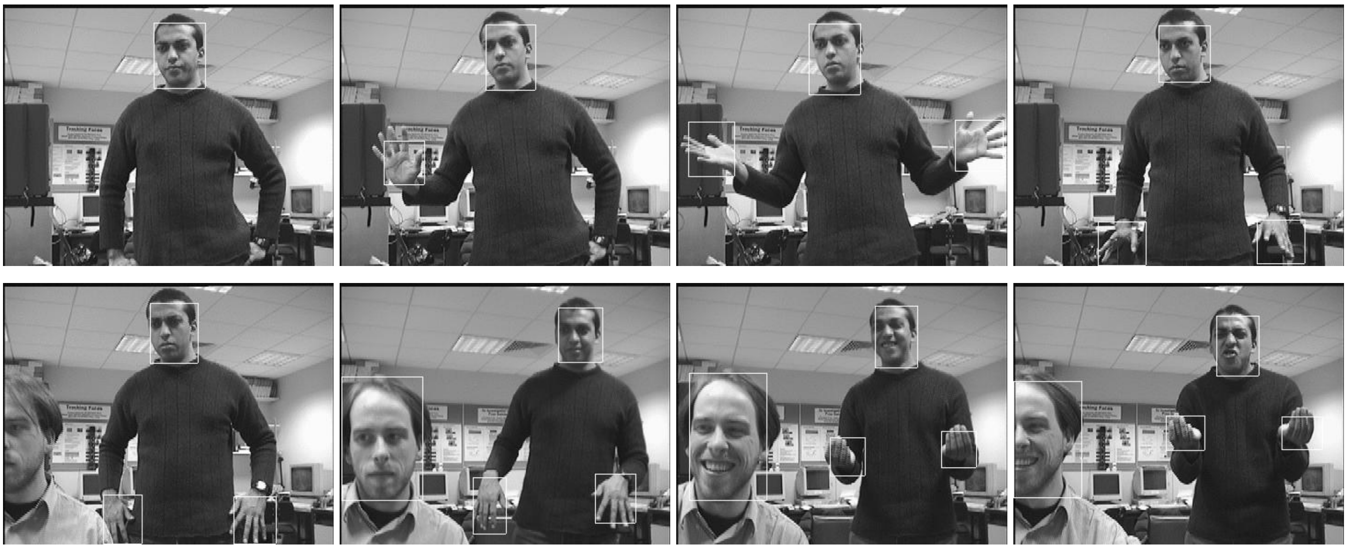

Figure 5. Focus of attention for face and gesture recognition using a multi-object colour tracker. The system runs at around 15Hz on a 200MHz PC.

[7] S. McKenna, Y. Raja, and S. Gong. Object tracking using adaptive colour mixture models. In *ACCV*, 1998.

[8] Y. Raja, S. McKenna, and S. Gong. Segmentation and tracking using colour mixture models. In *Asian Conference on Computer Vision*, Hong Kong, January 1998.

[9] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[10] D. Saxe and R. Foulds. Toward robust skin identification in video images. In *2nd Int. Conf. on Auto. Face and Gest. Recog.*, 1996.

[11] W. Skarbek and A. Koschan. Colour image segmentation - a survey. Technical report, Tech. Univ. of Berlin, 1994.

[12] M. J. Swain and D. H. Ballard. Colour indexing. *IJCV*, pages 11–32, 1991.

[13] H. G. C. Traven. A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density functions. *IEEE Trans. Neural Networks*, 2(3):366–378, 1991.

[14] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder:real-time tracking of the human body. *IEEE PAMI*, 19(7):780–785, 1997.