Segmentation and Tracking Using Colour Mixture Models

Yogesh Raja, Stephen J. McKenna and Shaogang Gong

Dept. Computer Science, Queen Mary and Westfield College, London.
E-mail: jpmetal@dcs.qmw.ac.uk

Abstract. A system is described that provides robust and real-time focus-of-attention for tracking and segmentation of multi-coloured objects. Gaussian mixture models were used to estimate the probability densities of object foreground and scene background colours. Tracking was performed by fitting dynamic bounding boxes to image regions of maximum probability. Two scenarios are presented: (1) real-time face tracking based upon a skin colour model and (2) dynamic body segmentation for virtual studios based upon combined foreground and background models.

1 Introduction

This work was initially motivated by a requirement in the broadcasting industry for an effective method for segmenting moving people from image sequences in order to perform superimposition onto virtual studios. Currently, the state of the art in virtual studio superimposition involves the use of "chroma-keying" techniques which perform colour-based segmentation to replace blue regions in an image with an alternative image. These techniques require painstaking preparation of a studio so that the background is entirely covered in blue material. Care is also taken to ensure that actors have no blue colours about their appearance. A new system for performing the task of segmentation without the need for this preparatory effort is desired and the work presented here offers a contribution to this end. This work also has implications for areas such as teleconferencing, vision-based man-machine interfaces and face recognition.

Colour has been used in machine-based vision systems for tasks such as segmentation [9] and recognition [2, 4, 10]. Colour cues have been shown to offer several significant advantages over geometric information for certain tasks in visual perception, such as robustness under partial occlusion, rotation in depth, scale changes and resolution changes [10]. Furthermore, colour processing can often utilise efficient algorithms yielding real-time performance on standard hardware.

The techniques presented here use colour as a cue for object localisation, segmentation and tracking. Multi-coloured objects are modelled using colour mixtures which estimate probability density functions in colour space. Whilst a single colour, e.g. skin tone, can be adequately modelled as a Gaussian distribution, multiple colours can be modelled using a mixture of Gaussians. Additionally, modelling scene background enables classification of pixels as object or background by computing posterior probabilities.

Two scenarios are described in this work. Firstly, detection and tracking of human faces was performed using a face colour model. A relatively simple colour model was used and real-time performance was obtained on a standard PC platform. Secondly, the more difficult task of segmenting humans from video sequences was considered for a virtual studio application. More complex, scene-specific colour models incorporating both human body foreground and scene background were used. Pixels were classified as foreground (object) or background by computing posterior probabilties from the two mixture densities. Background pixels were subsequently replaced with an alternative background.

The remaining parts of this paper are organised as follows. Colour modelling is discussed in Section 2 with emphasis on the use of Gaussian mixtures. Section 3 desribes a colour tracking framework. Section 4 presents the human face tracking application. Body tracking and segmentation for virtual studios are presented in Section 5. Finally, Section 6 gives conclusions and future work.

2 Statistical Colour Mixture Models

A major difficulty with using colour cues in machine vision is the *colour constancy* problem which arises due to variation in colour values brought about by lighting changes. This is particularly apparent in RGB (red, green, blue) space. Intensity is distributed throughout all three parameters, rendering colour values highly sensitive to scene brightness. A simple approach to colour constancy is to use the HSV colour space which consists of hue angle (H), colour saturation (S) and brightness (V). In order to obtain a limited level of intensity invariance, colours can be modelled in HS-space.

2.1 Modelling the Foreground

Colour histograms [10] are a simple non-parametric method for modelling. In a histogram, the density at a point in a colour space quantised into n bins is approximated by the fraction of pixels which fall into the corresponding bin. If n is too large, the estimated density will be "noisy" and many bins will be empty. If n is too small then the distribution's structure will be "smoothed" away. The use of histograms for estimating colour densities is only possible because n can be kept relatively small and because there are many data points (pixels) available. A potentially more effective "semi-parametric" technique for colour density estimation is the use of Gaussian mixture models. The conditional density for a pixel, $\boldsymbol{\xi}$, belonging to an object $\mathcal{O}$ is modelled as a mixture with m component densities:

$$p(\boldsymbol{\xi}|\mathcal{O}) = \sum_{j=1}^{m} p(\boldsymbol{\xi}|j)P(j) \tag{1}$$

where a mixing parameter $P(j)$ corresponds to the prior probability that pixel $\boldsymbol{\xi}$ was generated by component j and where $\sum_{j=1}^{m} P(j) = 1$. Each mixture

component is a Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ, i.e. in the case of a 2D colour space:

$$p(\boldsymbol{\xi}|j) = \frac{1}{2\pi|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu}_j)^T \Sigma_j^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}_j)} \tag{2}$$

Expectation-Maximisation (EM) provides an effective maximum-likelihood algorithm for fitting such a mixture to a data set [1, 8]. Fig. 1 shows an example of a Gaussian mixture model of a multi-coloured object in HS-space.

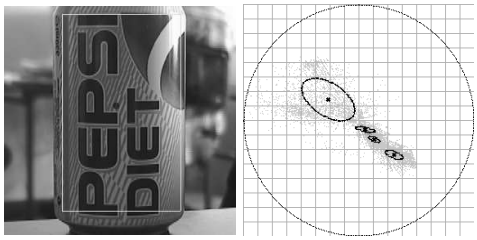

Fig. 1. A multi-coloured object and its Gaussian mixture model in HS-space. The mixture components are shown as elliptical contours of equal probability.

Outlier points, which can be caused by image noise and specular highlights, have little influence upon the mixture model. Once a model has been learned it can be converted into a look-up table for efficient on-line indexing of colour probabilities.

2.2 Modelling the Background

In virtual studios, it is 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 $P(\mathcal{O}|\boldsymbol{\xi})$:

$$P(\mathcal{O}|\boldsymbol{\xi}) = \frac{p(\boldsymbol{\xi}|\mathcal{O})P(\mathcal{O})}{p(\boldsymbol{\xi}|\mathcal{O})P(\mathcal{O}) + p(\boldsymbol{\xi}|\mathcal{S})P(\mathcal{S})} \tag{3}$$

The probability of misclassifying a pixel is minimised by classifying it as the class with the greatest posterior probability. A pixel is therefore classified as object foreground if and only if $P(\mathcal{O}|\boldsymbol{\xi}) > 0.5$. 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.

Alternatively, a single combined colour distribution can be estimated using both object and background data. If a Gaussian mixture model is used, the Gaussian components can be treated as basis functions which form the hidden layer in a neural network. The output layer of this network is trained to classify data as either object or background. This is a form of Hyper Basis Function (HyperBF) network [6]. The k^{th} output unit computes a function $f_k(\boldsymbol{\xi})$:

$$f_k(\boldsymbol{\xi}) = \sum_{j=1}^{m} w_j p(\boldsymbol{\xi}|j) \tag{4}$$

There are two output units: one representing the object class and the other the background. The output layer weights w_j can be determined by applying Singular Value Decomposition (SVD) [7]. Each input pixel is assigned to the class represented by the output unit with the highest activation.

3 Tracking Using Colour Models

The tracking dynamics involve estimating the position, width and height of the object. This box provides a focus of attention for further processing. The position and size of the box are found by computing the mean $\boldsymbol{\mu}^t = (\mu_x, \mu_y)$ and standard deviation $\boldsymbol{\sigma}^t = (\sigma_x, \sigma_y)$ of the local colour probability distribution within a rectangular search area centred on $\boldsymbol{\mu}^{t-1}$ in the image domain at time t. The dimensions of this search area are determined by scaling the dimensions of the bounding box at time $t-1$. The experiments presented in this paper were performed with search areas $\frac{3}{2}$ times the height and width of the bounding box.

For a given time frame t, the box position $\boldsymbol{\mu}^t$ is estimated as an offset from the position $\boldsymbol{\mu}^{t-1}$:

$$\boldsymbol{\mu}^t = \boldsymbol{\mu}^{t-1} + \frac{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}^{t-1})}{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})} \tag{5}$$

where $\mathbf{x}$ ranges over all image coordinates in the region of interest and $\boldsymbol{\xi}_{\mathbf{x}}$ is the HS colour vector at image position $\mathbf{x}$. To improve accuracy, probabilities $p(\boldsymbol{\xi}_{\mathbf{x}})$ are thresholded. Probabilities lower than the threshold are taken to be background and are consequently set to zero in order to nullify their influence on $\boldsymbol{\mu}^t$ and $\boldsymbol{\sigma}^t$. The size of the bounding box is estimated by computing the standard deviation of the image probability density:

$$\boldsymbol{\sigma}^t = \sqrt{\frac{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})\{(\mathbf{x} - \boldsymbol{\mu}^{t-1}) - \boldsymbol{\mu}^t\}^2}{\sum_{\mathbf{x}} p(\boldsymbol{\xi}_{\mathbf{x}})}} \tag{6}$$

In the next two sections, the above colour model was applied to both face and human body tracking/segmentation. Face tracking requires only a simple foreground model whereas body tracking/segmentation benefits from the use of combined foreground and background models.

4 Face Tracking Using a Skin Colour Model

Human face tracking has a wealth of possible applications such as security, teleconferencing and human-computer interfaces. A fast and robust method for isolating faces for subsequent recognition is a prerequisite for a fully automated face recognition system. The tracking technique presented here provides a useful component for such systems [5].

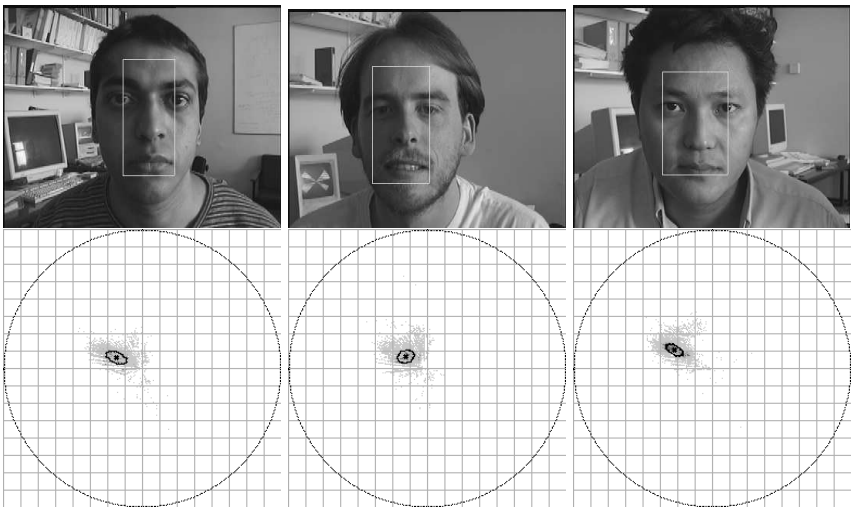

Fig. 2. The tight clustering of skin colour for three types of skin colour is illustrated here. The top row shows the face regions used to build the mixture models. The bottom row illustrates the colour distributions in HS-space.

Skin colour was modelled by collecting colour samples from images of faces. The face colours occupied a relatively compact area of HS-space (e.g. as found by [3]) and, in fact, a single Gaussian with full covariance was often sufficient to model the distribution of skin colour (Fig. 2). Fig. 3 shows a sequence of a face being tracked with a moving camera against a cluttered background. The tracker's ability to deal with changes in scale, large rotations in depth and partial occlusion are all clearly demonstrated.

This tracking system was implemented on a standard PC with a 200MHz Pentium processor, a Matrox Meteor colour frame grabber and a Sony EVI-D31 active camera with pan/tilt actuators and a zoom lens. The active camera can be driven by maintaining the mean position of the image probability distribution at the centre of the image. The tracking process is performed at approximately 15 frames per second. Tracking is robust without the use of temporal prediction. However, a recursive filter such as a Kalman filter might yield some improvement in performance and in particular help prevent the tracker "jumping" from one face to another. Problems are inevitably caused by large changes in the spectral

Fig. 3. A face is tracked against a cluttered background while the camera pans, tilts and zooms.

composition of scene illumination. In particular, it has been found necessary to use two models, one for interior lighting and one for exterior natural daylight.

5 Multi-coloured Object Tracking and Segmentation

In this section, colour mixture models are used to track multi-coloured objects. If a multi-coloured object consists of several distinct and differently coloured patches, then it may be beneficial to decompose the object and 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. [1]

An example of tracking performance is shown in Fig. 4, where a soft drinks can was located and tracked robustly under changing background, scale, rotation in depth and occlusion. Only the colour distribution of the object was modelled.

The distributions in colour space formed by multicoloured 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. The third image in Fig. 4 illustrates this problem. The bounding box is overly large as a result of colours similar to the can lying within the search space. In many situations, however, the colour distribution of the background scene can also be modelled. Fig. 5 shows an example in which both the person and the background have been modelled. Pixels were classified as object or background using equation (3) with

[1] The number of Gaussian components to use is currently determined empirically.

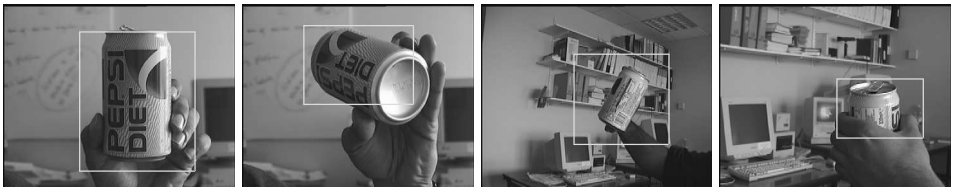

Fig. 4. Using only a foreground Gaussian mixture model, tracking was robust against a cluttered background with changing camera position and zoom, differing orientations of the object and large occlusions. Colours similar to the can on the bookshelf in the third image lie within the search space and consequently contribute to the estimation of the bounding box size.

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.

6 Conclusions and Further Work

A general framework was presented for modelling the colour distributions of multi-coloured objects using Gaussian mixtures and for using these models to perform tracking and segmentation. Expectation-Maximisation provided an effective algorithm for training the mixtures. The method has been shown to work consistently in two quite different scenarios. Firstly, real-time face tracking was performed using a simple foreground colour model which was robust under changing camera position and zoom. Secondly, body tracking and segmentation were performed by combining foreground and background colour models for use in a virtual studio application.

Pixel-wise classification based only upon colour information is obviously insufficient in order to guarantee perfect segmentation. However, it is often surprisingly effective. In addition, the image of posterior probabilities provides a rich source of information and could be combined with other visual processes. To this end, current work is being done to incorporate shape constraints into the system to exploit the results obtained from the segmentation technique.

References

1. C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.
2. G.D. Finlayson. Colour object recognition. Master's thesis, Simon Fraser Univ., 1992.

Fig. 5. Segmentation results. The top row shows three images from a sequence. The second row illustrates the segmentation accuracy (performed using a multi-resolution approach). The third row shows the reconstructed (superimposed) sequence.

3. M. Hunke and A. Waibel. Face locating and tracking for human-computer interaction. In *28th Asilomar Conf. on Signals, Systems and Computers*, 1994.
4. J. Matas, R. Marik, and J. Kittler. On representation and matching of multi-coloured objects. In *IEEE ICCV*, pages 726–732, 1995.
5. S. J. McKenna, S. Gong, and Y. Raja. Face recognition in dynamic scenes. In *BMVC*, 1997.
6. T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of The IEEE*, 78(9), September 1990.
7. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.
8. R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.
9. W. Skarbek and A. Koschan. Colour image segmentation - a survey. Technical report, Technical University of Berlin, 1994.
10. M. J. Swain and D. H. Ballard. Colour indexing. *IJCV*, pages 11–32, 1991.

This article was processed using the LaTeX macro package with LLNCS style