ELSEVIER

Image and Vision Computing 17 (1999) 225–231

image AND vision COMPUTING

Tracking colour objects using adaptive mixture models

Stephen J. McKenna[a,*], Yogesh Raja[b], Shaogang Gong[b]

[a]*Department of Applied Computing, University of Dundee, Dundee DD1 4HN, UK*
[b]*Department of Computer Science, Queen Mary and Westfield College, Mile End Road, London E1 4NS, UK*

Received 6 October 1997; received in revised form 16 March 1998; accepted 26 March 1998

Abstract

The use of adaptive Gaussian mixtures to model the colour distributions of objects is described. These models are used to perform robust, real-time tracking under varying illumination, viewing geometry and camera parameters. Observed log-likelihood measurements were used to perform selective adaptation. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Real-time tracking; Colour model; Gaussian mixture model; Adaptive learning

1. Introduction

Colour can provide an efficient visual cue for focus of attention, object tracking and recognition allowing real-time performance to be obtained using only modest hardware. However, the apparent colour of an object depends upon the illumination conditions, the viewing geometry and the camera parameters, all of which can vary over time. Approaches to colour constancy attempt to reconstruct the incident light and adjust the observed reflectances accordingly (e.g. Ref. [1]). In practice, these methods are only applicable in highly constrained environments. In this paper, a statistical approach is adopted in which colour distributions are modelled over time. These stochastic models estimate an object's colour distribution on-line and adapt to accommodate changes in the viewing conditions. They are used to perform robust, real-time object tracking under variations in illumination, viewing geometry and camera parameters.

Swain and Ballard [2] renewed interest in colour-based recognition through their use of colour histograms for real-time matching. Kjeldson used Gaussian kernels to smooth the histograms [3]. These colour histogram methods can be viewed as simple, non-parametric forms of density estimation in colour space. They gave reasonable results only because the number of data points (pixels) was always high and because the colour space was coarsely quantised. In the absence of a sufficiently accurate model for apparent colour, good parametric models for density estimation cannot be obtained. Instead, a semi-parametric approach has been adopted using Gaussian mixture models. Estimation is, thus, possible in a finely quantised colour space using relatively few data points without imposing an unrealistic parametric form on the colour distribution. Gaussian mixture models can also be viewed as a form of generalised radial basis function network in which each Gaussian component is a basis function or 'hidden' unit. The component priors can be viewed as weights in an output layer.

The mixture models are adapted on-line using stochastic update equations. It is this adaptation process which is the main focus of this paper. 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 can be used. These are determined off-line using an iterative algorithm. Once an object is being tracked, the model adapts and improves tracking performance by becoming specific to the observed conditions.

Finite mixture models have also been discussed at length elsewhere [4–10]. In particular, Priebe and Marchette [8] describe an algorithm for recursive mixture density estimation. It was extended to model non-stationary data series through the use of temporal windowing. Their algorithm adds new components dynamically when the mixture model fails to account well for a new data point. The approach adopted here differs in that the number of mixture components is determined using a fixed data set. These components' parameters are then adapted on-line while keeping the number of components fixed.

The remainder of this paper is organised as follows.

* Corresponding author. Tel: 0044 171 975 5230; Fax: 0044 181 980 6533.

PII: S0262-8856(98)00104-8

Gaussian mixtures for modelling objects' colour distributions are described in Section 2. In Section 3, a method for adapting the mixture models over time is given. Section 4 describes selective adaptation. Experimental results and conclusions are given in Sections 5 and 6.

2. Colour mixture models

The conditional density for a pixel, $\mathbf{x}$, belonging to an object, O, is modelled as a Gaussian mixture with m component densities:

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

where a mixing parameter $\pi(j)$ corresponds to the prior probability that $\mathbf{x}$ was generated by the jth component, $\sum_{j=1}^{m}\pi(j) = 1$. Each mixture component, $p(\mathbf{x}|j)$, is a Gaussian with mean μ and covariance matrix Σ, i.e. in the case of a two-dimensional colour space:

$$p(\mathbf{x}|j) = \frac{1}{2\pi|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j)}$$

Expectation-maximisation (EM) provides an effective maximum-likelihood algorithm for fitting such a mixture to a data set $X^{(0)}$ of size $N^{(0)}$ [4,11,12]. The EM algorithm is iterative with the mixture parameters being updated in each iteration. Let ψ^{old} denote the sum of the posterior probabilities of the data evaluated using the old model from the previous iteration, $\psi^{\text{old}} = \sum_{\mathbf{x}\epsilon X^{(0)}} P^{\text{old}}(j|\mathbf{x})$. The following update rules are applied in the order given:

$$\mu_j^{\text{new}} = \frac{\sum P^{\text{old}}(j|\mathbf{x})\mathbf{x}}{\psi^{\text{old}}} \qquad \pi_j^{\text{new}} = \frac{\psi^{\text{old}}}{N^{(0)}}$$

$$\Sigma_j^{\text{new}} = \frac{\sum P^{\text{old}}(j|\mathbf{x})(\mathbf{x}-\mu_j^{\text{new}})^T(\mathbf{x}-\mu_j^{\text{new}})}{\psi^{\text{old}}}$$

where all summations are over $\mathbf{x}\epsilon X^{(0)}$. Bayes' theorem gives the posterior probabilities:

$$P(j|\mathbf{x}) = \frac{p(\mathbf{x}|j)\pi(j)}{p(\mathbf{x}|O)}$$

EM monotonically increases the likelihood with each iteration, converging to a local maximum. The resulting mixture model will depend on the number of components m and the initial choice of parameters for these components. Initially, the following simple procedure was used. A suitable value for m was chosen based upon visual inspection of the object's colour distribution. The component means were initialised to a random subset of the training data points. All priors were initialised to $\pi = 1/m$ and the covariance matrices were initialised to $\sigma\mathbf{I}$, where σ was the Euclidean distance from the component's mean to its nearest neighbouring component's mean. In practice, this initialisation method invariably provided good performance for the tracking application described in this paper. Alternatively, a constructive algorithm which uses cross-validation to automatically select the number of components m and their parameters has been used [13]. In this method, disjoint training and validation sets are used. A single Gaussian is first fit to the training set. The number of components is then increased by iterating the following steps: (1) the likelihood of the validation set given the current mixture model is estimated; (2) the component with the lowest responsibility for the training set is split into two separate components with equal covariance matrices and principal axes equal to the original component's principal axis; (3) the iterative EM algorithm is run. These three steps are repeated until the validation likelihood is maximised or is considered to be sufficiently large.

Most colour cameras provide an RGB (red, green, blue) signal. In order to model objects' colour distributions, the RGB signal is first transformed to make the intensity or brightness explicit so that it can be discarded in order to obtain a high level of invariance to the intensity of ambient illumination. Here the HSI (hue, saturation, intensity) representation was used and colour distributions were modelled in the two-dimensional hue-saturation space. Hue corresponds to our intuitive notion of 'colour' whilst saturation corresponds to our idea of 'vividness' or 'purity' of colour. At low saturation, measurements of hue become unreliable and are discarded. Likewise, pixels with very high intensity are discarded.

It should be noted that the HSI system does not relate well to human vision. In particular, the usual definition of intensity as $(R + G + B)/3$ is at odds with our perception of intensity. However, this is not important for the tracking application described here. If in other applications it was deemed desirable to relate the colour models to human perception then perceptually based systems like CIE $L^*u^*v^*$ and CIE $L^*a^*b^*$ should be used instead of HSI.

Gaussian mixture models have been used to perform real-time object tracking given reasonably constrained illumination conditions. The resulting tracking system is surprisingly robust under large rotations in depth, changes of scale and partial occlusions [14,15]. However, in order to cope with large changes in illumination conditions in particular, an adaptive model is required.

3. Adaptive colour mixture models

A method is presented here for modelling colour dynamically by updating a colour model based on the changing appearance of the object. Fig. 1 illustrates a colour mixture model of a multi-coloured object adapting over time. While the components' parameters are adapted over time, the number of components is fixed. The assumption made here is that the number of components needed to

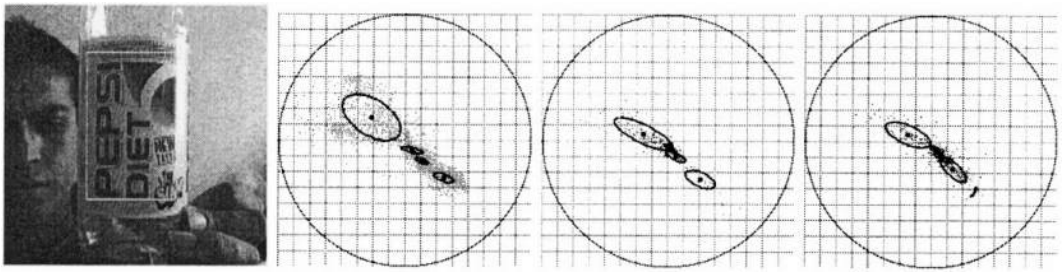

Fig. 1. A mixture model superimposed onto plots of a bottle's colour distribution. Hue corresponds to angle and saturation to the distance from the centre. Ellipses show the four Gaussian components. The leftmost plot shows the original mixture model. The remaining two plots show the model adapting to the illumination and viewing conditions.

accurately model an object's colour does not alter significantly with changing viewing conditions. However, the parameters of the model will definitely need to adapt. An initial mixture model is obtained by running the EM algorithm discussed in the previous section. Each mixture component then has an 'initial' parameter set (μ_0, Σ_0, π_0). In each subsequent frame, t, a new set of pixels, $X^{(t)}$, is sampled from the object and can be used to update the mixture model[1]. These colour pixel data are assumed to sample a slowly varying non-stationary signal. Let $\psi^{(t)}$ denote the sum of the posterior probabilities of the data in frame t, $\psi^{(t)} = \sum_{\mathbf{x}\epsilon X^{(t)}} P(j|\mathbf{x})$. The parameters are first estimated for each mixture component, j, using only the new data, $X^{(t)}$, from frame t:

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

$$\Sigma^{(t)} = \frac{\sum P(j|\mathbf{x})(\mathbf{x}-\mu_{t-1})^T(\mathbf{x}-\mu_{t-1})}{\psi^{(t)}}$$

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

$$\mu_t = \mu_{t-1} + \frac{\psi^{(t)}}{D_t}(\mu^{(t)} - \mu_{t-1}) - \frac{\psi^{(t-L-1)}}{D_t}(\mu^{(t-L-1)} - \mu_{t-1})$$

$$\Sigma_t = \Sigma_{t-1} + \frac{\psi^{(t)}}{D_t}(\Sigma^{(t)} - \Sigma_{t-1}) - \frac{\psi^{(t-L-1)}}{D_t}(\Sigma^{(t-L-1)} - \Sigma_{t-1})$$

$$\pi_t = \pi_{t-1} + \frac{N^{(t)}}{\Sigma_{T=t-L}^{t} N^{(\tau)}}(\pi^{(t)} - \pi_{t-1}) - \frac{N^{(t-L-1)}}{\Sigma_{T=t-L}^{t} N(\tau)}(\pi^{(t-L-1)} - \pi_{t-1})$$

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

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

This yields a recursive expression for D_t:

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

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

4. Selective adaptation

An obvious problem with adapting a colour model during tracking 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. In order to 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, $\mathcal{L}^{(t)}$, of the data, $X^{(t)}$, observed from the object at time t is given by:

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

At each time frame, $\mathcal{L}^{(t)}$ is evaluated. If the tracker loses the object there is often a sudden, large drop in its value. This provides a way to detect tracker failure. Adaptation is then suspended until the object is again tracked with sufficiently high likelihood. A temporal filter was used to compute a threshold, T_t. Adaptation was only performed when $\mathcal{L}^{(t)} > T_t$. 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 set to $T = \nu - k\sigma$, where

[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 [4,16].

Fig. 2. Eight frames from a sequence in which a face was tracked using a non-adaptive model. The apparent colour of the face changes due to: (i) varying illumination; and (ii) the camera's auto-iris mechanism which adjusts to the bright exterior light.

k was a constant. In all the experiments described here, $k = 1.5$, $n = 2f$ and $L = 6f$, where f denotes the frame rate in Hz.

5. Experiments

The adaptive mixture modelling described in the previous two sections was integrated with an existing colour-based tracking system [14,15] implemented on a standard 200 MHz Pentium PC platform with a Matrox Meteor framegrabber. This system performs tracking at approximately $f = 15$ Hz. The tracker estimates the centroid, height and width of the object. 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 object's colours. This will hold for a large class of objects.

Figs. 2 and 3 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 illumination conditions coupled with the camera's auto-iris mechanism resulted in large changes in the apparent colour of the face as the person approached the window. Towards the end of the sequence the face became very dark, making hue and saturation measurements unreliable. In Fig. 2, 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 occurred. In Fig. 3, the model was allowed to adapt and successfully maintained lock on the face.

Fig. 4 illustrates the advantage of selecting when to adapt. 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.

Fig. 5 depicts the tracking of a multi-coloured item of clothing with adaptation performed in every frame. Although tracking was robust over many frames, erroneous adaptation eventually resulted in failure. Fig. 6 shows the last four frames from the same sequence tracked correctly using selective adaptation.

6. Conclusions

Objects' colour distributions were modelled using Gaussian mixture models in hue-saturation space. An adaptive learning algorithm was used to update these colour models over time and was found to be stable and efficient. These adaptive models were used to perform colour-based object

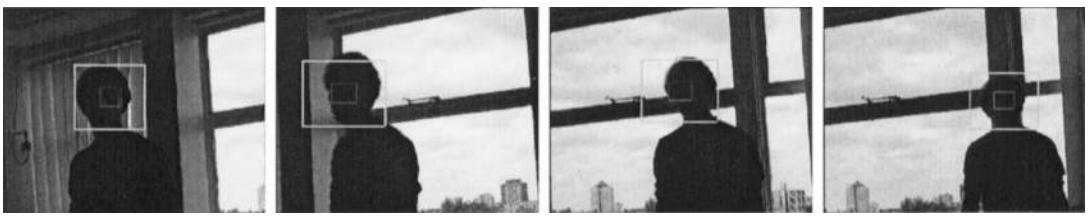

Fig. 3. The sequence depicted in Fig. 2 tracked with an adaptive colour model. Here, the model adapts to cope with the change in apparent colour. Only the last four images are shown for conciseness. Performance in previous frames was similar.

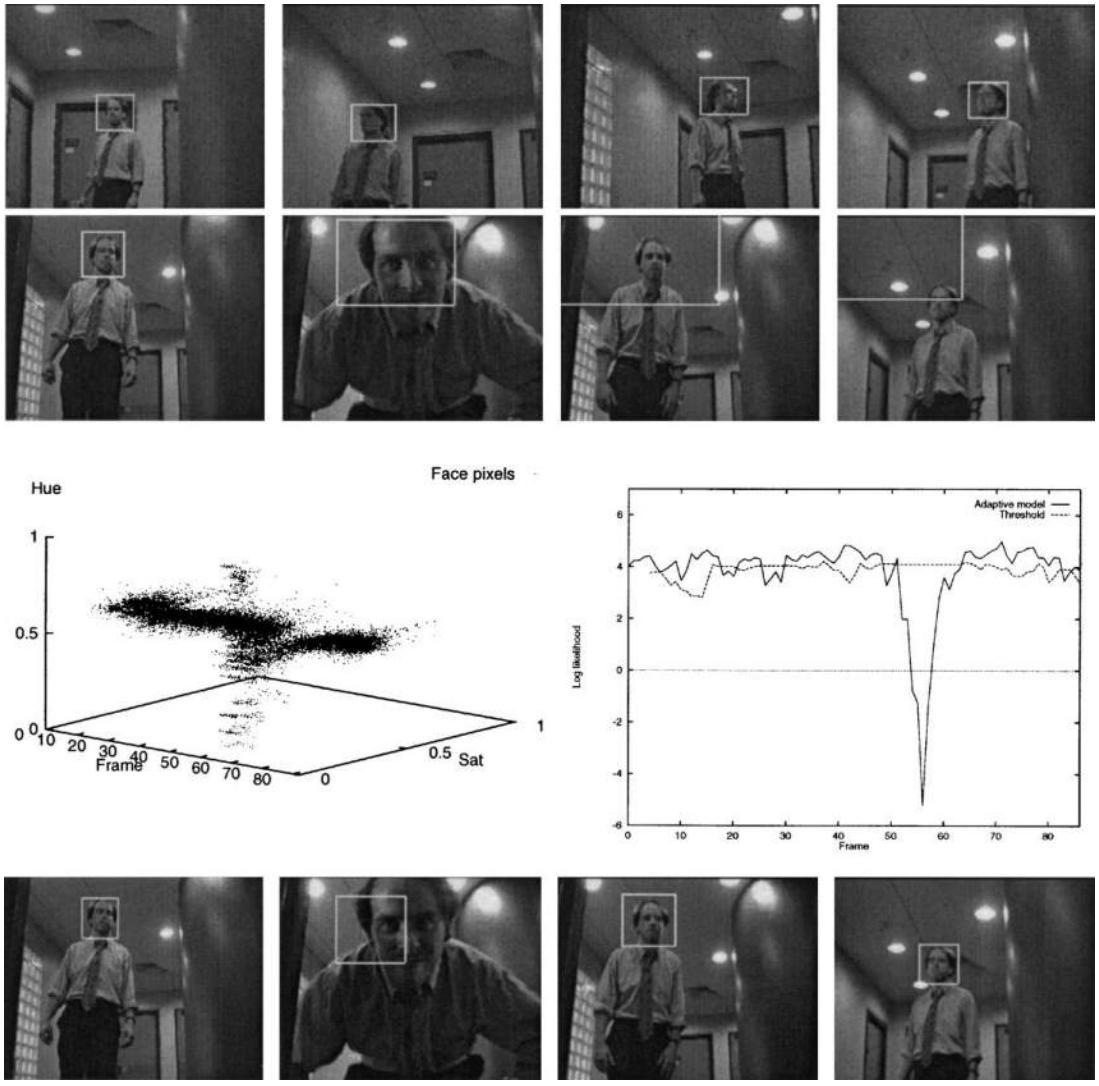

Fig. 4. At the top are frames 5, 15, 25, 35, 45, 55, 65 and 75 from a sequence. There is strong directional and exterior illumination. The walls have a fleshy tone. At around frame 55, the subject rapidly approaches the camera which is situated in a doorway, resulting in rapid changes in illumination, scale and auto-iris parameters. This can be seen in the three-dimensional 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.

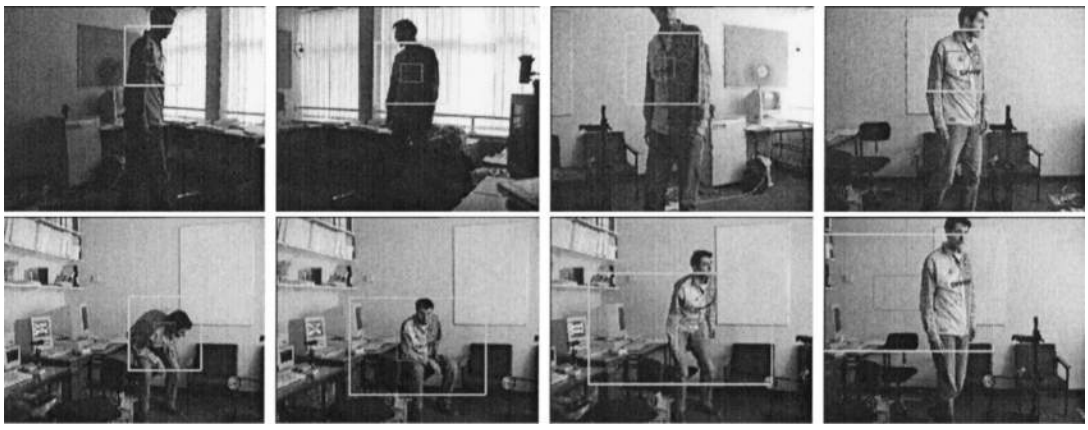

Fig. 5. A green, yellow and black shirt tracked using the adaptive mechanism. Eventually, tracking inaccuracies cause the model to adapt erroneously and the system fails.

Fig. 6. The sequence shown in Fig. 5 tracked using selective adaptation. The shirt was correctly tracked throughout. Only the last four frames are shown for brevity.

tracking in real-time under varying illumination, viewing geometry and camera parameters. Outlier detection based on a normalised log-likelihood statistic was used to detect tracking failures. This adaptive scheme outperformed the non-adaptive colour models.

Topics for further work include: (i) emphasised co-operation with other visual cues during periods when colour becomes unreliable; (ii) adaptive modelling of background scene colours; and (iii) adaptive model order selection, i.e. adaptation of the mixture size during tracking.

Acknowledgements

S.J.M. was supported by EPSRC grant IMV GR/K44657 whilst at Queen Mary and Westfield College. Y.R. was supported by an EPSRC/BBC CASE studentship.

Appendix A

Here we derive the update equations for the adaptive mixture model components. For each mixture component, let μ_t and Σ_t be the mean and the covariance matrix estimated from the $L+1$ most recent time-slots:

$$\mu_t = \frac{\sum_{\tau=t-L}^{t} \sum_{\mathbf{x}\in X^{(\tau)}} p(j|\mathbf{x})\mathbf{x}}{\sum_{\tau=t-L}^{t} \psi^{(\tau)}}$$

$$\Sigma_t = \frac{\sum_{\tau=t-L}^{t} \sum_{\mathbf{x}\in X^{(\tau)}} p(j|\mathbf{x})(\mathbf{x}-\mu_{\tau-1})^T(\mathbf{x}-\mu_{\tau-1})}{\sum_{\tau=t-L}^{t} \psi^{(\tau)}}$$

The above expressions are both of the form:

$$\theta_t = \frac{\sum_{\tau=t-L}^{t} \theta^{(T)}\psi^{(T)}}{D_t}$$

where θ_t denotes either μ_t or Σ_t. A recursive expression for θ_t is derived as follows:

$$\theta_t = \frac{1}{D_t}\left(\sum_{\tau=t-L-1}^{t-1} \theta^{(\tau)}\psi^{(\tau)} + \theta^{(t)}\psi^{(t)} - \theta^{(t-L-1)}\psi^{(t-L-1)}\right)$$

$$= \frac{1}{D_t}\left(\theta_{t-1}\sum_{\tau=t-L-1}^{t-1} \psi^{(\tau)} + \theta^{(t)}\psi^{(t)} - \theta^{(t-L-1)}\psi^{(t-L-1)}\right)$$

$$= \frac{1}{D_t}\left(\theta_{t-1}\sum_{\tau=t-L}^{t} \psi^{(\tau)} - \theta_{(t-1)}\psi^{(t)} + \theta_{(t-1)}\psi^{(t-L-1)} + \theta^{(t)}\psi^{(t)} - \theta^{(t-L-1)}\psi^{(t-L-1)}\right)$$

$$= \theta_{t-1} + \frac{\psi^{(t)}}{D_t}(\theta^{(t)} - \theta_{t-1}) - \frac{\psi^{(t-L-1)}}{D_t}(\theta^{(t-L-1)} - \theta_{t-1}) \quad \text{(A1)}$$

The update expression for the component priors π_t is obtained similarly to Eq. (A1). If the number of data points is the same in every time frame (i.e. $N^{(\tau)} = N$, for all τ) then we have:

$$\pi_t = \pi_{t-1} + \frac{\pi^{(t)} - \pi^{(t-L-1)}}{L-1}$$

References

[1] D.A. Forsyth, Colour constancy and its applications in machine vision, Ph.D. thesis, University of Oxford, 1988.
[2] M.J. Swain, D.H. Ballard, Colour indexing, International Journal of Computer Vision (1991) 11–32.
[3] R. Kieldsen, J. Kender, Finding skin in color images, in: 2nd International Conference on Automatic Face and Gesture Recognition, Killington, Vermont, USA, 1996.
[4] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York, 1995.
[5] B.S. Everitt, D.J. Hand, Finite Mixture Distributions, Chapman and Hall, New York, 1981.
[6] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Applications to Clustering, Marcel Dekker Inc., New York, 1988.
[7] C.E. Priebe, Adaptive mixtures, Journal of the American Statistics Association 89 (427) (1994) 796–806.
[8] C.E. Priebe, D.J. Maxchette, Adaptive mixtures: recursive nonparametric pattern recognition, Pattern Recognition 24 (12) (1991) 1197–1209.
[9] C.E. Priebe, D.J. Marchette, Adaptive mixture density estimation, Pattern Recognition 26 (5) (1993) 771–785.
[10] D.M. Titterington, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.
[11] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B-39 (1977) 1–38.
[12] R.A. Redner, H.F. Walker, Mixture densities, maximum likelihood and the EM algorithm, SIAM Review 26 (2) (1984) 195–239.
[13] Y. Raja, S.J. McKenna, S. Gong, Colour model selection and adaptation in dynamic scenes, in: European Conference on Computer Vision, Freiburg, Germany, June 1998.

[14] S. McKenna, S. Gong, Y. Raja, Face recognition in dynamic scenes, in: British Machine Vision Conference, University of Essex, UK, 1997, pp. 140–151.
[15] Y. Raja, S. McKenna, S. Gong, Segmentation and tracking using colour mixture models, in: Asian Conference on Computer Vision, Hong Kong, 1998, pp 607–614.
[16] H.G.C. Traven, A neural network approach to statistical pattern classification by 'semiparametric' estimation of probability density functions, IEEE Transactions on Neural Networks 2 (3) (1991) 366–378.