VIDEO SEGMENTATION AND ANNOTATION USING GAUSSIAN MIXTURE MODELS

Roland Wilson

University of Warwick,
Coventry CV4 7AL, UK.

ABSTRACT

This paper describes a new approach to the segmentation and annotation problem using Gaussian mixture model descriptors. These have several advantages over conventional, histogram-based techniques, including: a rigorous statistical basis; the possibility of encoding spatial, colour, texture and motion features in a unified system; and the ability to trade off accuracy of representation against data volume. After a brief introduction to the class of models, results are presented to show their efficacy.

1. INTRODUCTION

The problem of segmenting and annotating digitised video data is one which is growing as fast as the very data volume represented in digital archives. Current techniques for addressing the problem tend to rely on manual intervention aided by simple descriptors of colour and texture content, often based on histograms [1, 2]. It is clear that more effective and general techniques are essential, if full use is to be made of such archive material. It is clear that to be useful, any such method must have a strong statistical element, but also take account of spatial distribution.

One interesting development in recent years has been the use of Gaussian mixture models to cope with statistical problems for which no simple parametric model exists [3, 4]. While it is well known that algorithms such as Expectation-Maximisation can lead to effective approximations in terms of a finite number of components, the general problem of mixture modelling is difficult when the number of components is unknown [3]. Alternatively, non-parametric methods, such as those based on kernel density estimation, are beset by the difficulty of choosing the right scale for the kernel.

It is shown in this paper that a suitably defined class of Gaussian mixture models can be efficiently estimated from image sequence data. The models incorporate both traditional statistical and spatial information and can be efficiently estimated in a Bayesian framework, using a multiresolution algorithm. Hence the name multiresolution Gaussian mixture model (MGMM). Moreover, the models are defined in a space whose dimension reflects the inference problem, not the image data: in colour images, a 5-D space is required (two spatial and three colour dimensions); for inferring 3-D structure from motion, typically nine dimensions are required (three spatial, three colour and three motion axes). Yet the representation has no difficulty in moving seamlessly between these spaces. A brief exposition of the models is followed by some simple experiments, which illustrate how they may be used in representing image sequences.

2. BACKGROUND

Three main elements dictate the form of representation called MGMM: ability to approximate *any* probability density in a space of arbitrary dimension; closure under affine motions and a multiresolution structure, which can be used to make computation efficient. Now, any smooth probability density function can be approximated to an arbitrary precision by a set of Gaussian functions: it is well known that the Gaussian functions are a complete set on $L^2(R^n)$. However, we can easily make a stronger claim, namely that the approximation need only involve *positive* coefficients in the expansion. To this end, we state the theorem:

Theorem 1: Let $f(.) : R^n \rightarrow R$ be any nonnegative integrable function on R^n with

$$\int_{R^n} d\vec{x}\ f(\vec{x}) = 1 \tag{1}$$

Then for any $\delta > 0$ there exists an approximation of $f(.)$ by a strictly positive sum of Gaussian functions of the form

$$\hat{f}(\vec{x}) = \sum_i\ f_i g_{\Sigma_i}(\vec{x} - \vec{\mu}_i) \tag{2}$$

of means $\vec{\mu}_i$ and covariances $\mathbf{\Sigma}_i$, such that $f_i > 0, \forall i$ and

$$\delta > \int_{R^n} d\vec{x}\ |f(\vec{x}) - \hat{f}(\vec{x})| \tag{3}$$

The other property, a crucial one for motion analysis, is the closure of the set, $\mathcal{G}^n$, of n-D Gaussian functions under affine maps $\boldsymbol{A} : R^n \mapsto R^n$

$$\boldsymbol{A}\vec{x} = \boldsymbol{L}\vec{x} + \vec{a} \tag{4}$$

where $\boldsymbol{L}$ is an invertible matrix and $\vec{a}$ a translation. Again, it is obvious that the action of $\boldsymbol{A}$ on $\mathcal{G}^n$ is closed, since

$$g_\Sigma(\boldsymbol{A}^{-1}(\vec{x} - \vec{\mu})) = g_{\Sigma_A}(\vec{x} - \vec{\mu}_A) \tag{5}$$

where

$$\vec{\mu}_A = \vec{\mu} - \vec{a} \tag{6}$$

and

$$\mathbf{\Sigma}_A = \boldsymbol{L}^T \mathbf{\Sigma} \boldsymbol{L} \tag{7}$$

But now we are in a position to prove a rather interesting result, summarised as

Theorem 2: Let $f(.) \geq 0$ be an integrable function $f : R^n \rightarrow R$, as above and let $\vec{t}(.) : R^n \rightarrow R^n$ be a smooth, invertible map from R^n to itself. Then if f has an approximation as a Gaussian mixture of the form of (2), there exists an approximation of the transformed function $\boldsymbol{T}f$

$$\boldsymbol{T}f(\vec{x}) = f(\vec{t}^{-1}(\vec{x})) \tag{8}$$

of the same form, where each Gaussian component g_i is transformed according to a local affine approximation of the flow field $\vec{t}$, both approximations having integral absolute error less than $\epsilon > 0$.

The significance of the latter result should not be understated: it implies that it is not necessary to recompute the model for each frame of a sequence, but only to *move* the current model to accommodate any local motions.

3. MGMM

The key to applying these ideas is to use a sequential approach, which leads to a multiresolution tree structure: Multiresolution Gaussian mixture Modelling (MGMM). Suppose we wish to estimate the model from some data, such as an image or set of images. For example, a gray level image is modelled as a set of 3-D samples from an unknown density: 2 spatial co-ordinates and the gray level (we are not simply working with histograms or co-occurrence matrices here). If these data are denoted $\vec{X}_i, 1 \leq i \leq N$, we compute the sample mean and covariance and hence infer a single multivariate Gaussian model $g_{\Sigma_0}(\vec{x} - \vec{\mu}_0)$, where $\vec{\mu}_0, \Sigma_0$ are the sample (Maximum Likelihood) estimates for the data. Now, if they are sufficiently close to normal in distribution, this may model the data adequately. If not, then we split the data into two parts and model each part separately with one Gaussian density. This can be done using the Markov Chain Monte Carlo (MCMC) sampling technique described in [5], which treats the inference as one containing *hidden variables*, namely the class Z_i to which each data point belongs; it samples from the *posteriors* for the population size, means and covariances, assuming *conjugate* priors, whose parameters are simply those of the population as a whole. Thus the prior for the means of the two classes are Gaussian, while the covariances are drawn from a Wishart density and the population sizes from a Dirichlet distribution. Sampling for (i) the hidden variables and (ii) the corresponding population densities gives estimates of the parameters of the two Gaussians, based on the posterior estimates from the sampler $g_{\Sigma_{1j}}(\vec{x} - \vec{\mu}_{1j}), j \in [0, 1]$.

This gives rise to the following recursive estimation procedure:

1. Select class j and test its normal density approximation for 'goodness-of-fit'. If the fit is adequate, terminate, *else*

2. (a) Split class j into two components: class $j0$ and class $j1$, by sampling the hidden variables $Z_{ji}, 1 \leq i \leq N_j$.

   (b) Obtain a Bayesian estimate of the class means and covariances by sampling from the posteriors, given the prior $g_{\Sigma_j}(\vec{x} - \vec{\mu}_j)$.

Thus, given an appropriate measure of 'goodness-of-fit', a tree of Gaussians of decreasing variance is obtained, which has the property of approximating the density of the data to a prescribed accuracy, or to the extent which the data support. The question of goodness of fit is not trivial, but a simple rule is to add a fixed penalty for each split, since it increases the complexity of the description by a fixed number of parameters, giving the rule

$$\ln \frac{P(\vec{Y}|\theta_{j0}, \theta_{j1})}{P(\vec{Y}|\theta_j)} > \gamma(N_0, N_1) \tag{9}$$

where $\gamma > 0$ depends on the sizes of the two sub-classes, θ_i are the Gaussian parameters for the MGMM and the log-likelihood ratio refers to the pre- and post-split mixture models. More generally, the threshold should be a function of the population sizes, cf. [3].

An alternative view of the MGMM description is as a patchwork of affine models, each leaf node being the result of a linear regression on the data [6]. For example, in the case of a gray level image, the MGMM description gives for each class a Gaussian model, which is directly related to a least-squares approximation of the form

$$z_i(\vec{x}) = \boldsymbol{A}_i(\vec{x} - \vec{x}_i) + z_{i0} + \nu_i(\vec{x}) \tag{10}$$

where $z_i(.)$ is the gray level as a function of the spatial coordinate $\vec{x}$ for the ith class and $\nu_i(.)$ is the *residual*. The matrix $\boldsymbol{A}_i$ is easily found from the covariance matrix Σ_i for that class and $\vec{x}_i, z_{i0}$ from the mean.

Once a GMM has been built for a given scene, then it can be 'moved', as per Theorem 2, giving a motion-compensated (M-C) GMM. Indeed, if the motion field is included in the description, then we automatically get the affine approximation to the field associated with each mixture component as part of the model, via the regression in (10) above. Furthermore, motion is itself a good feature for segmentation, which helps to guide the building of the tree and identify significant changes in a dynamic scene. When the moved model no longer represents the data adequately, the likelihood will decrease significantly, indicating a change of scene.

4. EXPERIMENTS

These ideas are illustrated using three frames from a 256×256 pixel, grey level version of the Miss America sequence. In these examples, we use frame 15 of the sequence, along with an interframe motion estimate, to build a 5-D GMM. In other words, two spatial co-ordinates, two motion components and the grey level are used in the model. Although the multiresolution motion estimator yields estimates at full image resolution, its representation by Gaussian components constrains the local estimates to be affine. The GMM's in each case contained 24 components at leaf level, giving a representation of the data by a total of 480 parameters. The corresponding reconstructions are shown alongside the original images in Figure 1(a)-(d). Reconstruction PSNR's are summarised in Table 1, which shows that excellent reconstructions can be obtained from the GMM representation,

given knowledge of the classifications. In order to reconstruct frame 16, the estimated motions, obtained directly from each Gaussian component, were applied to that component in the model; the target image was then classified using the model and reconstructed using the least-squares linear fit, as in equation (10). The consistent and relatively high PSNR's, of around $30dB$, show that the model does capture significant structure in the images. Of course, this does not imply a miraculous data compression: the reconstruction requires classification of pixels,

Also shown in Table 1 are the PSNR's obtained by frame differencing (FD), with and without motion compensation, demonstrating that the motion estimator is quite effective, even when movements are large. An important feature of the GMM approach is that it is also possible to measure model fit in terms of likelihood: how likely is the image, given the model? The likelihoods for the two images, based on the motion compensated GMM, are also given in Table 1. These show that the model does a good job of approximating the data, when motion is taken into account.

In the second experiment, frame 99 was selected, as having a large displacement relative to frame 15. The motion compensated GMM again does a good job of reconstruction, as can be seen from Figure 1(e)-(f) and row 3 of Table 1. More interestingly, if the scene is altered by adding a large white square to the frame, the motion compensated GMM can still capture the image well, but the likelihood decreases quite dramatically, indicating that the square is a long way from the centroid of the component which is representing it. This shows that the GMM approach can express scene changes in a far more revealing way than simple measures, such as average squared error, with or without motion compensation.

5. CONCLUSIONS

A new approach to video sequence segmentation and annotation, based on Gaussian mixture modelling, has been presented. The new method represents a significant improvement over conventional techniques, in that it captures both statistical and spatial aspects of an image and is able to deal effectively with motion.

While there are many problems yet to be solved, this is a promising approach to a difficult, but important problem in image sequence representation. Future plans include using colour and texture information into the model, as well as exploring other measures of goodness of fit, such as the Bayes Factor.

Acknowledgement

This work was supported by the UK EPSRC. The author also wishes to thank Sony, Broadcast and Professional (Europe) for their assistance and, in particular, Dr Martin Todd.

6. REFERENCES

[1] W. Niblack, X Zhu, J.L. Hafner, D. Ponceleon, D. Petkovic, M.D. Flickner, E. Upfal, S. I. Nin, S. Sull, B. Dom, B-L. Yeo, A Srinivasan, D. Zivkovic, and M. Penner, "Updates to the QBIC system," in *Proc. SPIE Conf. on Storage and Retrieval for Image and Video Databases,*, 1997, vol. 3312.

[2] Y. Rui, T. S. Huang, and S-F. Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," *J. Visual Commun. and Image Rep.*, vol. 10, pp. 39–62, 1999.

[3] S. J. Roberts, D. Husmeier, I. Rezek, and W. Perry, "Bayesian Approaches to Gaussian Mixture Modelling," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. 20, no. 11, 1998.

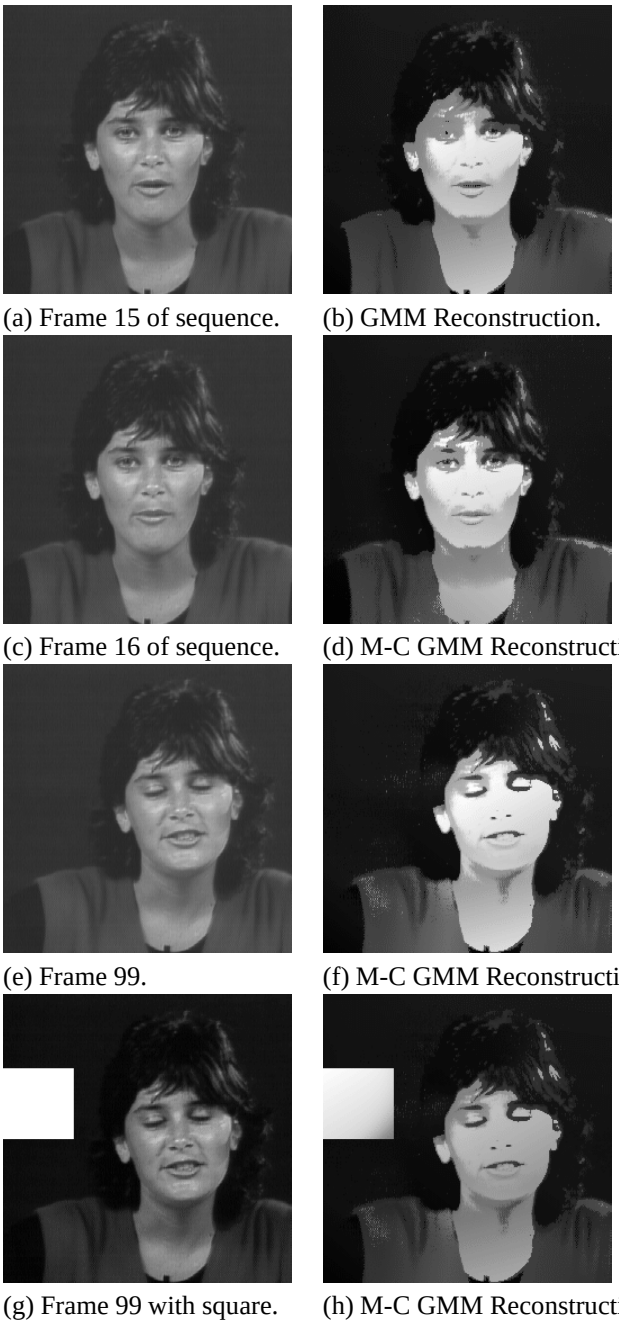
[4] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian Mixture Density Modeling, Decomposition and Applications," *IEEE Trans. Image Processing*, vol. 5, pp. 1293–1301, 1996.

[5] C. P. Robert, "Mixtures of Distributions: Inference and Estimation," in *Markov Chain Monte Carlo in Practice*, S. Richardson W. R. Gilks and D. J. Spiegelhalter, Eds. Chapman & Hall, 1996.

[6] L. Breiman, J. H. Friedman, R. A. Ohlsen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

Frame	Frame Diff. PSNR (dB)	M-C FD PSNR (dB)	GMM PSNR (dB)	GMM Log-likelihood
15	-	-	31	-10496
16	34	34	32	-9775
99	18	27	29	-12050
99 with square	13	13	19	-12087

Table 1. PSNR and likelihood figures for frames 15, 16 and 99 of the Miss America sequence, and an artificially produced scene change.



(a) Frame 15 of sequence. (b) GMM Reconstruction.

(c) Frame 16 of sequence. (d) M-C GMM Reconstruction.

(e) Frame 99. (f) M-C GMM Reconstruction.

(g) Frame 99 with square. (h) M-C GMM Reconstruction.

Fig. 1. Reconstruction of frames 16 and 99 from motion compensated GMM of frame 15.