Probabilistic Hierarchical Framework for Clustering of Tracked Objects in Video Streams

Riad Hammoud, Roger Mohr

Probabilistic Hierarchical Framework for Clustering Tracked Objects in Video Streams [1] [2]

R. Hammoud and R. Mohr
INRIA Rhône-Alpes and GRAVIR-CNRS
655 avenue de l'Europe, 38330 Montbonnot Saint Martin, FRANCE
Phone: (33) 4 76 61 52 35 Email: riad.hammoud@inrialpes.fr

I*rish* **M***achine* **V***ision and* **I***mage* **P***rocessing Conference, pages 133-140*
September, 2000

Abstract Currently it is impossible to automatically achieve high level video indexing as suggested for instance by MPEG-7. Part of the problem is the identification of moving objects in the framework of a video. It represents such a challenging problem due to the variable appearance of objects over time. This paper studies how to automatically classify tracked objects in a video based on estimated Gaussian mixture density functions of each tracked object. In the color feature space, the variability of each tracked object is modeled separately by a Gaussian mixture where the appropriate number of Gaussians is determined by the Integrated Classification Likelihood criterion. Then, the Ascendant Hierarchical Classification technique is used to identify the clusters of tracked objects. Thus, the problem is different to conventional classification in that the underlying measurements are not feature vectors but density estimates. The proposed framework is evaluated on the *Avengers* TV movie, segmented into 2749 individual objects of seven different clusters.

Keywords: Appearance modeling, Gaussian mixture distribution, Kullback distance, Video indexing and MPEG-7, Unsupervised classification.

1 Introduction

The forthcoming MPEG-7 standard provides the framework for efficient representation, processing, and retrieval of visual information. This includes many recent multimedia applications such as hyper/interactive video and video content-based searching and navigation [18] [10]. Yet many problems must still be addressed and solved before this technology can emerge. An important problem is the content classification into a number of categories [6]. The categorization of content will help the user speeding up the search.

1.1 The problem addressed and the proposed approach

This paper focuses on the problem of automatic classification of detected objects in a video stream. Such work requires a set of tasks to be processed robustly: *Object selection*,

[1] This work is supported by Alcatel CRC Grant Alcatel-Inria No. 198G098.
[2] Demos of this work are available at `http://www.inrialpes.fr/movi/people/Hammoud/`

Object modeling, and *Object matching and grouping*. These three steps are handled in the following way:

• **Object selection** If objects belong to a limited set, like in the face localization problem, a particular model can be implemented as is done in [22]. However, for the general case of any kind of objects appearing in a video, there is no other alternative than outlining them and from there a tracking might be performed. The scene motion is the only general hint that can be used for objects extraction [13] [8]. To do this step, the current framework implements the method of [8] which detects moving areas in successive images and from there the tracking is performed.

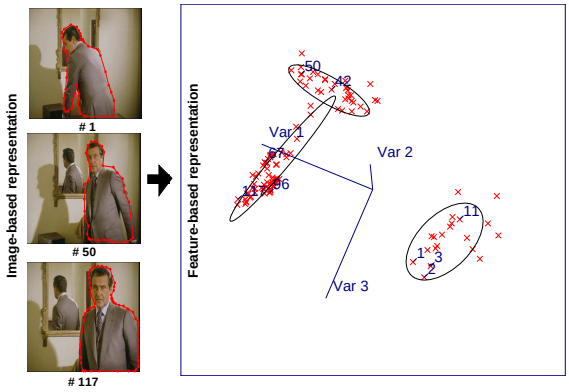

Figure 1: *Illustration of the content variation of a non-rigid tracked object and its modeling in the three principal components of the RGB histogram space by a mixture of three Gaussians. Three key-frames among* 117 *occurrences of the object are depicted and the covariance ellipses are also displayed.*

• **Tracked object modeling** The large set of *individual objects* detected in a real video sequence (i.e. 129600 *individual objects* in a film of 90 minutes, 24 frames per second and 1 segmented object per frame) corresponds in fact to a small set of real different *tracked objects* like persons, for instance; each *tracked object* is a collection of continuous *individual objects* (occurrences) within the same shot. In most cases these tracked objects (car, walking person, ...) are moving and they are acquired in poorly constrained dynamic scenes (a real movie film). Therefore the data set is contaminated by various source of "noise": low-resolution, large-scale changes, 3D-rotations, variable illumination and occasionally occlusions. Recognition based upon isolated individual objects of this kind, and using classical recognition methods, is highly inconsistent and inaccurate. In order to overcome these difficulties, the large intra-shot variability of each tracked object is modeled. Here the color feature space is used and it is modeled by a multi-modal distribution using a mixture of Gaussians. Such a modeling consists in capturing the content variation by identifying for each tracked object a set of clusters; each cluster is modeled by a single Gaussian and it groups a set of similar views of the tracked object within a shot. Figure 1 illustrates an example of a moving tracked object modeled by a mixture of three Gaussians in the RGB feature space. These three Gaussians correspond to the three principal situations of the object in the shot (back, side and face). Moreover, such a compact representation

also allows a decrease in the complexity of the indexing process required for matching in such a large collection of tracked objects.

- **Matching and grouping similar objects** Many clustering techniques exist in the literature [14] but they can not be applied directly on density distributions. The problem of grouping tracked objects is different here to conventional clustering in that the underlying measurements are not feature vectors but density estimates. Two distance measures can be computed on density distributions: the Kullback-Leibler divergence [15] and the Bhattacharyya distance ([7], page 99). Based on this proximity data the Ascendant Hierarchical Classification (AHC) is used to identity clusters of tracked objects. The final number of clusters is determined in an interactive way by the user; this interaction is required as no full automatic procedure would provide perfect identification on such noisy data.

1.2 Relevant works

The closest work related to this paper is probably the one conducted on the problem of anemia by Cadez et al. [3]. They propose a powerful approach to classify patients into two classes: normal and iron deficient. This discriminative classification approach is called *hierarchical* because the Gaussian mixture classifier is used twice. In the first time, the large set of measurements ($40,000$ blood cells) extracted on each patient are modeled by a mixture of Gaussians. In the second time, the parameters of the mixture of two Gaussians (two classes of patients) are estimated in the space of densities (parameters space). This work is the closest to the contribution presented in this paper; it differs mainly by the input data (tracked objects; color distributions) and some related criterion related to mixture density estimation (number of Gaussian components is not fixed). Also, the classification technique we use is the AHC.

Modeling with Gaussian mixture is now becoming very popular. Rosales [19], McKenna et al. [16] and Hammoud at al. [9] use the Gaussian mixture model to recognize human actions, face colors and non-rigid moving objects in videos, respectively. Then, they use the Gaussian mixture classifier to identify the appropriate class of new entities (action, face or object). The modeling of the variability of objects in the feature space improves the recognition accuracy rate even in the presence of complex scenes [11].

1.3 Paper organization

This paper presents a first study on clustering with this kind of model and the experimental results contain strong trends. Experiments are done on a video sequence of 1938 frames, extracted form the standard video, *Avengers* TV film, given by the National Institute of Audiovisuel in France (INA). It was segmented into 2749 individual objects which correspond to 29 tracked objects of seven different clusters.

In section 2 we provide an overview of the first step of the proposed framework: object acquisition and low-level characterization. Intra-shot variability modeling and clustering of densities are the two main parts and they are described in sections 3 and 4 respectively. Experiments and test results of the approach are presented in section 5. Subsection 5.2

discusses the performance and how this approach can be used in a pragmatic implementation of semi-automated video description. Finally, section 6 gives some conclusions and perspectives.

2 Detecting, tracking and characterizing objects

In order to index video and allow a non-linear navigation in its structure, entities like moving objects have to be detected. First videos are segmented into shots. In our framework we use the method of [2]; It relies on a robust, multi-resolution and incremental estimation of a 2D affine motion model between successive frames, accounting for the global dominant image motion.

Video segmentation. The object acquisition is made firstly by extracting moving areas with consistent motion analysis. Then these entities are tracked within the shot. For this step, a large research effort has already been made [13] [8]. The motion approach is widely used. In the presented framework we use the method of [8] to detect and track moving objects. Static objects are more easily tracked with standard tools. The result is a collection of individual images of a tracked object within a shot.

Object characterization. Many different types of features could be computed on the segmented objects. As the objects are deformable and very variable in time (for instance see figure 2), the geometric information (i.e contours, ...) may be irrelevant. In this paper, the color distribution is considered as it was shown to be a powerful source of information for object recognition [23] since colors of object correlate strongly with object identity. A simple and effective recognition scheme is to represent and match objects on the basis of color-metric histograms as proposed by Swain and Ballard [23]. Thus, the color histogram is independent of many imaging conditions which are represented strongly in the experimental database of this paper e.g. the orientation of a scene, 3D rotations and the absence (or occlusion) of some of the colors. However, color histograms do not distinguish between two different images with similar global color distributions. So this approach could be extended in many different direction (correlograms, joint histograms) [5]. As this is not the focus of the paper, we rework this kind of descriptors.

3 Intra-shot variability modeling

During the tracking within a shot, the appearance of an object is variable due to all the imaging conditions listed in the introduction. This visual variability produces a multi-modal distribution of features in the low-level feature space; each object is represented by a single point in the feature space whose coordinates are the values of the feature vector (see figure 1 for instance). Let Y to be the set of feature vectors collected during the tracking of an object and y_i be the feature vector of dimension d that characterizes the occurrence i.

The distribution of Y is modeled as a joint probability density function, $f(y \mid Y, \theta)$ where θ is the set of parameters for the model f. Let $\mathbf{x} = (x_1, ..., x_n)$ be the complete data set with $x_i = (y_i, c_i)$ and c_i is the class label of x_i (i.e. we have $c_i = j$ if and

only if j is the mixture component from which y_i arises). For the rest of this paper, we assume that f can be approximated as a mixture of $J-$Gaussians density function: $f(x|\theta) = \sum_{j=1}^{J} p_j \varphi(x|\mu_j, \Sigma_j)$ where the p_j's ($0 < p_j < 1$ and $\sum_{j=1}^{J} p_j = 1$) are the mixing proportions and where $\varphi(x|\mu_j, \Sigma_j)$ is the Gaussian density parameterized by the mean μ_j and the covariance matrix Σ_j. The vector of parameters to be estimated, for each tracked objects separately, is $\theta_j = (p_j, \mu_j, \Sigma_j)$, for $j = 1, \ldots, J$.

The standard EM algorithm [17] is used to estimate the parameters of mixture of Gaussians while the Integrated Classification Likelihood [1] criterion is used to choose the appropriate number of Gaussians. Each cluster is approximated by a Gaussian density and represents a set of similar views of the tracked object.

The EM algorithm. The Expectation-Maximization (EM) algorithm is a well-known iterative technique for maximum likelihood approximation of the whole data. It is based on the idea that there is a set of missing variables that relates the data to the unknown clusters. If those missing variables were known, the maximum likelihood distribution would be easy to calculate.

Each iteration of EM consists of an Estimation (E) and a Maximization (M) step. The E-step evaluates a probability distribution, $t_{ij}(\theta^m)$ (e.g. at iteration 'm') for the data given the model parameters from the previous iteration.

$$t_{ij}(\theta^m) = \frac{p_j^m \varphi(x_i, \mu_j^m, \Sigma_j^m)}{\sum_{\ell=1}^{J} p_\ell^m \varphi(x_i|\mu_\ell^m, \Sigma_\ell^m)}. \tag{1}$$

The M step then finds the new parameter set that maximizes the probability distribution.

$$p_j^{m+1} = \frac{1}{n}\sum_{i=1}^{n} t_{ij}(\theta^m) \ , \ \ \mu_j^{m+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^m)\, y_i}{\sum_{i=1}^{n} t_{ij}(\theta^m)} \text{ and } \ \Sigma_j^{m+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^m)(y_i - \mu_j^{m+1})(y_i - \mu_j^{m+1})^{\mathsf{T}}}{\sum_{i=1}^{n} t_{ij}(\theta^m)}. \tag{2}$$

The E and M steps are iterated until convergence or the maximum number of iterations is reached. The algorithm is run many times in order to be independent from the initial solution where the initialization step is done using the K-means algorithm.

Choosing mixture components' number. Many criteria are proposed in the literature [21] [1] for selecting the best number of Gaussians with a known Gaussian model (in this paper it is the general model with no constraints on θ). They are based on the following idea: penalize the model in some way by offsetting the increase in log-likelihood with a corresponding increase in the number of parameters, and seeking to minimize the combination.

A recent attempt to tackle the above problem was done by Biernacki [1]; a novel criterion called Integrated Classification Likelihood (ICL) was proposed (equation 3). The

advantage of this criterion is that it is more robust to high dimensional spaces and it works more better than the other criteria with non Gaussian data. This criterion is implemented by the proposed approach.

$$ICL(M) = -2L_M + Q_M ln(n) - 2\sum_{i=1}^{n}\sum_{j=1}^{J}\hat{c}_{ij}t_{ij}, \quad (3)$$

where L_M is the maximized log-likelihood of the model M, Q_M is its number of free parameters and $\hat{c}_{ij}$ represents the estimated partition deduced from t_{ij}.

For a reasonable range of the number of Gaussians ([1..3] in our experiments), the values of ICL criterion are computed using EM at each iteration and finally the minimum is picked which indicates the best number of Gaussians.

4 Hierarchical strategy for classifying density functions

The multiple features of a tracked object are reduced in an efficient way to a standard feature vector: the θ parameter. Based on these parameters the classification of tracked objects will be performed.

The mixture density model differs from standard parametric modeling and not all clustering algorithms can be performed. Just considering the set of parameters of these densities as a feature point makes no sense; for instance, if an object is not moving in one shot its variance will be close to zero, and this representation could correspond to just one image in a shot where the tracked object is seen in variable appearances. Therefore distance in this parameters space is meaningful.

An alternative approach to completely avoid modeling parameters (of mixture Gaussian classifier for instance) in parameter space is to define a distance measure between the *densities* themselves. Then, based on the resulting distance matrix, a Hierarchical Classification approach is applied. Two distances between Gaussian densities are implemented and interesting conclusions are extracted from the test results.

Distance similarity. As it was stated in section 3, a Gaussian mixture is used as a tracked object might have very different appearances. Therefore computing the distance between the global appearance distribution makes no sense, and the distance to be considered should be the distance between the two closest appearances. So a distance measure is computed for each pair of Gaussians of different mixtures, and the minimum is picked. In the following we describe the two implemented distance measures.

Kullback distance. This is a well-known measure of the divergence between two distributions p and q ([15] pages 3-6). It is based on information theoretic motivation and defined as the cross-entropy $\int p(\theta) log \frac{p(\theta)}{q(\theta)}$. Monte-Carlo procedures are used to efficiently evaluate it. As the Ascendant Hierarchical classification technique (see below) needs a symmetric proximity matrix the following form of the Kullback distance is implemented. This one is non-negative and is zero if and only if $p = q$. However, it is not termed a metric because the inequality property is not satisfied.

$$d_K(p,q) = \frac{1}{2}\left(\int p(\theta) log \frac{p(\theta)}{q(\theta)} d\theta + \int q(\theta) log \frac{q(\theta)}{p(\theta)} d\theta\right). \quad (4)$$

Another implemented measure between two Gaussian distributions is the Bhattacharyya distance ([7], page 99) defined as

$$d_B(p,q) = \underbrace{\frac{1}{8}(\mu_p - \mu_q)^T \left(\frac{\Sigma_p + \Sigma_q}{2}\right)^{-1} (\mu_p - \mu_q)}_{B1} + \underbrace{\frac{1}{2} log \left(\frac{\left|\frac{\Sigma_p + \Sigma_q}{2}\right|}{\sqrt{|\Sigma_p||\Sigma_q|}}\right)}_{B2} \quad (5)$$

where $|\ .\ |$ is the determinant. The first term in this expression, $B1$, is similar to the well-known Mahalanobis distance and measures the distance between the two populations caused by the mean shift. The second term $B2$ gives class separability due to the covariance-difference.

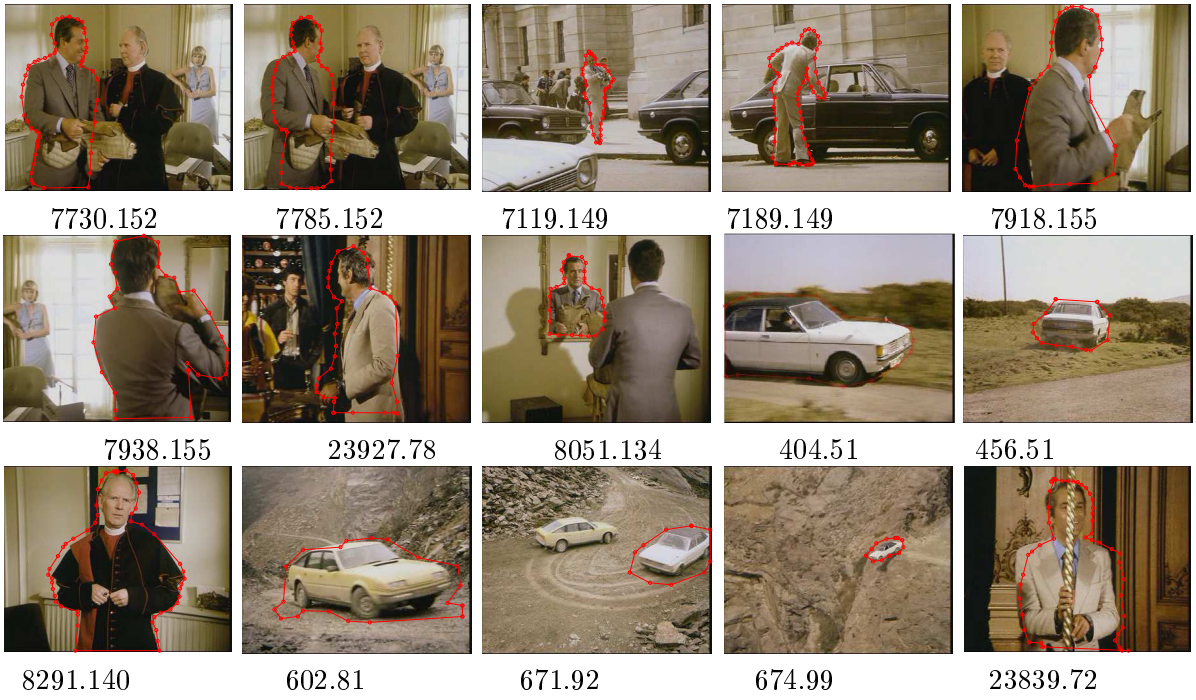

Figure 2: *Subset of the "Avengers" object database*

Hierarchical classification algorithm. In order to identity clusters of tracked objects, the Ascendant Hierarchical classification (AHC) is used by the proposed framework. In contrast to the partitional methods (mixture of Gaussians, k-means, ...), the AHC does not require any a priori knowledge (the number of clusters) and it represents data as a nested sequence of partitions [14]. However, the crucial point for the AHC is the order in which the clusters are formed when different distance between clusters (hierarchical clustering methods) are used. The choice of a suitable distance between clusters is an

important matter in applications, but theory provides few guidelines for optimizing the choice. Commonly, the single-link and complete-link methods (the minimal and maximal distances between two clusters) are used; test results with these two methods are described in the next section.

The cutting point in the hierarchy is selected by the user as there is no completely satisfactory method for determining the number of data clusters for the AHC technique [14]. This point is discussed in more details in section 5.2.

5 Experiments

Video base. The experimentations were done on a real video sequence extracted from the standard *Avengers* TV movie given by the "National Institute of Audiovisuel" in France (INA). This video sequence of 1938 frames was segmented into 2749 individual objects which correspond to 29 tracked objects of seven different classes (classes of *J. Steed*, *Ford Car*, *Licorne*, ...). Figure 2 displays some views of 12 tracked objects which correspond to 5 different classes. In this figure, different views of a tracked object within a shot have the same label, for example images 7918 and 7938 show two views of the tracked *white car* with the label 51. This data set of non-rigid objects is very variable (partial occlusions, 3D-rotations, changes of views, ...). The presence of partial occlusions up to 50%, the unknown model of illuminations and the large changes of views together in the same time (for example, see *J. Steed* in images 7730, 7918 and 8051) make the clustering of objects into homogeneous groups very difficult.

Feature base. As presented in section 2 color distribution is the feature space going to be considered for these experiments. The color distribution is approximated by an histogram in the d-dimensional space (named the real space in the following table); it is computed in the RGB,HSV and H (hue) color spaces where this later one is invariant to illumination changes. Then, the Principal Component Analysis (PCA) is used as a linear method to reduce the dimension of the feature space. For this small size of data (few individuals within shots) this dimensionality reduction allows to model more accurately the distribution.

The preprocessed data set of each tracked object is modeled as a mixture of J-Gaussians. The maximum number of Gaussians permitted is fixed to 3. The ICL criterion determines for each modeled tracked object the best value of J ($J \in [1..3]$).

5.1 Evaluation and test results

Applying the AHC on the density set of the whole tracked objects of the experimented video, a hierarchy of 29 levels is constructed (the down level with 29 clusters and the top level with 1 cluster). In order to evaluate the approach, the hierarchy is cut at level 22 to obtain seven clusters. The total percentage of correctly classified objects is computed. A tracked object is correctly classified in a cluster if the majority of cluster elements are similar to this object. Table 1 summarizes the test results at this level of the hierarchy. The approach is evaluated either when the content variation of tracked objects is modeled

as a mixture of J-Gaussians (see above to know how the J is determined) or by only one Gaussian.

Features	d-space		Distance	**Total** %			
	real	PCA		*single-link*	*complete-link*	*single-link*	*complete-link*
				J-Gaussians		**one-Gaussian**	
h_{RGB}	64	10	d_B	44.83	62.02	44.83	55.17
h_{RGB}	64	10	d_K	**51.72**	**79.31**	44.83	62.07
h_{HSV}	64	10	d_B	44.83	65.52	51.72	62.07
h_{HSV}	64	10	d_K	51.72	68.97	51.72	62.07
h_{hue}	32	5	d_B	55.17	55.17	51.72	58.02
h_{hue}	32	5	d_K	58.62	62.07	55.17	62.07

Table 1: *Test results with the single and complete link hierarchy methods, in the cases of J-Gaussians and One-Gaussian modeling of the intra-shot variability; Total percentage of correctly clustered objects.*

5.2 Comparative analysis and discussion

A guess from the table 1 that the approach gives the better clustering ratio (79.30%) when the content variation in the RGB histogram space is modeled as a mixture of J-Gaussians, the Kullback distance is computed between Gaussian components and when the complete-link hierarchical clustering method is applied. In general, the results are the best when the modeling is done in the RGB space. This is related to our assumption on the form of the data to be Gaussian. This could be valid for RGB but it is not for example the case for H feature where it is cylindrical.

In the clustering process, it is quite hard to make a decision between which criterion to select for measuring cluster distance. It is well known that the criteria of the distance of the closest representative produces clusters as strings or contours which might be far from compact. Running the algorithm with the distance of the two closest representative (single-link) and the two farthest ones (complete-link) led to the conclusion that the later is always better (form 10% up to 20%).

The mixture approach is obviously very advantageous. The experiments that a maximum of one or three Gaussian components were allowed in the distribution. The three components case always provides the best result, except for H feature where the results are similar. Of course if objects would always be static (like object 8291.140), without occlusion or illumination changes, this should be reconsidered.

Modeling the variability as a Gaussian mixture, the Kullback distance provides better results than the Bhattacharyya (from 3% up to 17%). More expensive experimentations are required in order to search a more definitive experimental conclusion. At this experimented level, these results can only be considered as indicative trends.

Practically one cannot expect to get perfect clustering of the data under these difficult

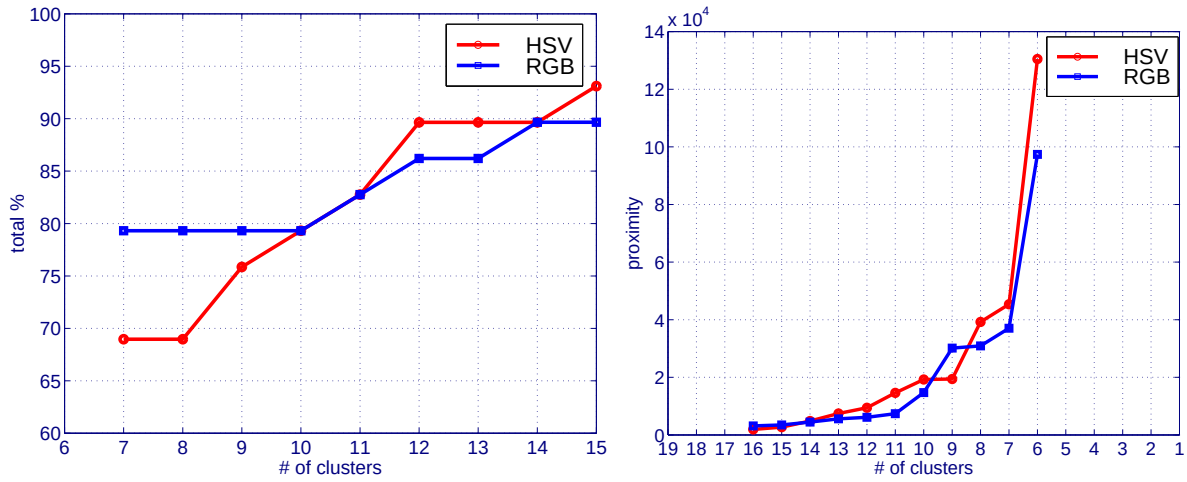

Figure 3: *Graphical representation for the total % of correctly classified objects versus the number of clusters (left) and the proximity values versus the number of clusters(right); cases of RGB and HSV features using J-Gaussians models with the Kullback distance and the complete-link method)*

but realistic conditions. As the human operator should remain in the loop, one way to proceed could be the following:
(**1**) Run the AHC algorithm with a number of classes that is approximately twice the real number of objects expected; our experiments show that the number of correct clusters obtained is close to perfect, but of course with over-clustering. Figure 3 (left) illustrates the total percentage of correctly classified objects versus the number of clusters (from 7 to 15) in the RGB and HSV color histogram spaces.
(**2**) Manually group clusters with same objects together.
If m tracked objects appear n times giving raise to $n \times m$ tracked individuals, this would reduce the checking to the $2m$ clusters instead of handling the $n \times m$ tracked individuals.

In order to simplify the selection of the number of clusters it is always a good idea to provide a graphical illustration of the number of clusters versus the distance between formed groups at the different levels of the hierarchy. The user can select the number of clusters at the curve points. Figure 3 (right) illustrates the hierarchical indices (proximity values) versus the number of clusters. For example, levels 9 and 7 can be selected as cut points in the hierarchy.

6 Conclusion and perspectives

Recently, the **M**oving **P**icture **E**xpert **G**roup, MPEG-7, is working on a standard description of the video in order to provide a large accessibility to this media. One challenging problem addressed by this standardization is the content classification into clusters.

In this work we have mainly presented a method for identifying clusters of tracked objects in a real movie film. Here the problem differs from conventional classification in that

the underlying measurements are not feature vectors but density estimates. The Gaussian mixtures are used to model the multi-modal distributions of tracked objects in the color feature space. Such a modeling consists in capturing the content variation of an object during tracking in an efficient way; This variation in the feature space is due to the high changes of appearance of moving objects in time. Then the matching process is performed on the density estimates since a distance between two mixture densities is considered as the minimal value of all distances between pairs of Gaussians. Based on the computed matrix of proximity, a sequence of nested clusters of objects are identified using the Hierarchical Classification technique. The experimental results on the *Avengers* TV movie demonstrate the efficiency of the proposed method. The use of Gaussian mixtures is more adequate for modeling the distribution than a single Gaussian. In this paper, only color histograms are tested as a low-level feature image, but future experiments will consider other descriptors such as correlograms [12]. invariants [20]. The experiments showed that the data are not always Gaussian (for example the H feature). In this case more appropriate statistical approximation functions like Gamma model should be considered. Also, the dimension of the feature space is a critical point for the density estimation process when the number of individual occurrences of tracked objects is limited. One issue to prevent over-fitting is to use suboptimal Gaussian models with constrained covariance matrix for example [4].

Acknowledgments

We would like to acknowledge `Alcaltel CRC` for its support of this work, and the "`Institut National de l'Audiovisuel en France`", dept of Innovation, for providing the video used in this paper.

References

[1] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated classification likelihood. Rapport de recherche RR-3521, INRIA, October 1998.

[2] P. Bouthemy and F. Ganansia. Video partitionning and camera motion characterisation for content-based video indexing. *Proc. 3rd IEEE Int. Conf. Image Processing.*, september 1996.

[3] I.V. Cadez, C.E. McLaren, P. Smyth, and G. J. McLachlan. Hierarchical models for screening of iron deficiency anemia. In *The Sixteenth International Conference on Machine Learning*, Bled, Slovaine, 27-30 June 1999.

[4] G. Celeux and G. Govaert. Gaussian Parsimonious Models. *Pattern Recognition*, 28(5):781–783, 1995.

[5] P. Chang and J. Krumm. Object recogntion with color cooccurence histograms. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA*, pages 498–504, 1999.

[6] C. Colombo, A. D. Bimbo, and P. Pala. Semantics in visual information retrieval. *IEEE MultiMedia*, 6(3):38–53, July 1999.

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[8] M. Gelgon and P. Bouthemy. A region-level graph labeling approch to motion-based segmentation. In *CVPR*, pages 514–519, Puerto Rico, June 17-19 1997.

[9] R. Hammoud and R. Mohr. Gaussian mixture densities for indexing of localized objects in a video sequence. Technical report, INRIA, March 2000. http://www.inria.fr/RRRT/RR-3905.html.

[10] R. Hammoud and R. Mohr. Interactive tools for constructing and browsing structures for movie films. In *ACM Multimedia*, pages 497–498, Los Angeles, California, USA, October 30 - November 3 2000. (demo session).

[11] R. Hammoud and R. Mohr. Mixture densities for video objects recognition. In *International Conference on Pattern Recognition*, volume 2, pages 71–75, Barcelona, Spain, 3-8 September 2000.

[12] J. Huang, S. Ravi Kumar, M. Mitra, W.J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA*, pages 762–768, June 1997.

[13] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(6):577–589, June 1998.

[14] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[15] S. Kullback. *Information theory and Statistics*. Dover, New York, NY, 1968.

[16] S.J. Mckenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.

[17] G.L. McLachlan and T. Krishnan. *The EM algorithm and Extensions*. John Wiley and Sons, New York, 1997.

[18] F. Nack and A.T. Lindsay. Everything you wanted to know about mpeg-7. *IEEE Multimedia*, pages 65–77, July-September 1999.

[19] R. Rosales. Recognition of human action using moment-based features. Technical Report Report BU 98-020, Boston University Computer Science, Boston, MA 02215, November 1998.

[20] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, May 1997.

[21] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, March 1978.

[22] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[23] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.