INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Gaussian Mixture Densities for Indexing of Localized Objects in a Video Sequence

Riad Hammoud and Roger Mohr

No 3905

Mars 2000

THÈME 3

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--3905--FR+ENG

Gaussian Mixture Densities for Indexing of Localized Objects in a Video Sequence

Riad Hammoud and Roger Mohr

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Movi

Rapport de recherche n° 3905 — Mars 2000 — 36 pages

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

Abstract: The appearance of non-rigid objects in a video stream is highly variable and therefore makes the identification of similar objects very complex. Furthermore, the indexing process of all detected objects is a very challenging problem when all appearances of an object would be stored: The database produced would become so large that searching would be intractable. In this paper we present a framework for object-based indexing which on one side increases the robustness of existing feature detectors used for object recognition and on the other side reduces the size of the database. The temporal variation of features of a tracked object in the video-shot is modeled by a mixture of Gaussians. Given a tracked object, this consists in separating the feature distribution into homogeneous clusters. Each cluster corresponds to a stable view of the tracked object. We put in competitions seven different Gaussian models and the number of Gaussian components varies up to four. The EM algorithm is applied to estimate the parameters of the mixture of Gaussians where the number of its components and the Gaussian model are a priori fixed. The choice of the best structure of the data (model and number of Gaussians) is realized by different criteria: BIC, ICL and NEC. The training of the system is done on a set of different tracked objects and the Gaussian mixture classifier is used to recognize new occurrences of objects. Experiments on a video base of twelve different objects are conducted and eight color features are tested. A comparison in the performance of the proposed system and the *temporal feature method* is analyzed and reported.

Key-words: Object recognition, Mixture of Gaussians, Video indexing, Video shot, Color descriptors

(Résumé : tsvp)

This work is supported by the Alcatel CRC

Densités de mélange gaussien pour l'indexation des objets localisés dans une séquence vidéo

Résumé : L'identification des objets similaires localisés dans une séquence vidéo est complexe à cause de la grande variation de leurs apparences dans le temps. En outre, un processus d'indexation de tous ces objets n'est pas faisable vue la taille énorme de la base d'index ainsi produite. Dans ce papier, nous présentons un système pour indexer les objets vidéo, qui permet d'une part de robustifier les descripteurs utilisées en reconnaissance d'objets et d'autre part de réduire la taille de la base d'index. La variation temporelle des descripteurs de tous les occurrences d'un objet suivi dans le plan-vidéo est modélisée par un mélange gaussien. Ceci consiste à partitionner la distribution de descripteurs en des clusters homogènes qui correspondent à plusieurs vues stables de l'objet suivi. Nous mettons en compétitions sept modèles gaussiens différents et un nombre de composantes gaussiennes allant jusqu'à quatre. L'algorithme EM est appliqué pour estimer les paramètres du mélange gaussien dont le nombre de ses composantes et le modèle gaussien sont fixés à priori. Le choix de la meilleur structure de données (modèle et nombre de gaussiennes) est réalisé à l'aide des critères BIC, ICL et NEC. L'apprentissage du système est effectuée sur un ensemble d'objet suivis et le classificateur du mélange gaussien est utilisé pour classer des nouvelles occurrences d'objets. Des expérimentations sur une base vidéo de douze objets différents sont menés et huit descripteurs de couleurs sont testés. Une comparaison en performance du système proposé et de la méthode de *descripteur temporel* est aussi analysée et rapportée.

Mots-clé : Reconnaissance d'objets, Mélange gaussien, Indexation de la vidéo, Plan vidéo, Descripteurs de couleurs

1 Introduction and Motivation

Video has a rich implicit temporal and spatial structure based on shots, camera motions, object motions and interactions, etc. To enable high level searching, browsing and navigation, this implicit structure needs to be made explicit [HC99]. For this purpose, cut detection [BG96] [Ua93] and object acquisition [GB97] are performed first. A classification strategy of these objects into homogeneous classes will create links in the video stream, allowing for instance to jump to the next shot where the same person appears. So, the navigation and search will become more powerful and less time consuming for the users in numerous domains [SBS97]. This motivates the research in this area and applications ranges from video surveillance to human-computer interaction.

The dynamic nature of video makes object-based recognition very difficult: low resolutions, large-scale changes, rotations, variable illumination and occasional partial occlusions. Therefore recognition with classical recognition methods [DH73] [CJ91] gives poor results. In order to overcome this difficulty, we propose an approach that models the intra-shot variations of tracked objects. This approach deals with the general problem of non-rigid object recognition and could be used to improve robustness of existing recognition methods.

Retrieval and browsing video require the source material to be first effectively *indexed* [HJZW95]. Most of previous research in video indexing have been text-based [Dav93] [RBE94]; Content based indexing of video using low-level visual image features is still a research problem and classifying the feature of non-rigid objects is a very challenging and important problem in computer vision. In addition to the problems of appearance-based recognition, the enormous volume of low-level image features within each shot is a significant problem for video indexing. For example, in a movie of 90 minutes projected at 24 frames per second, about 129600 objects would be extracted even if there was only one object of interest per frame. Linear searching into such objects database would inefficient, so indexing is the issue, as index would allow to select efficiently a subset of relevant template. However, indexing data with uncertain descriptors in large dimension is a hard problem which has yet only partial solutions (see for instance [WZ98]). So reducing the descriptor size and the number of data is a key issue.

One popular way to reduce the set of data is to index only objects which appear in "representative" key-frames of shots [O'C91]. This is reasonable in the case of still shots.

However, most video shots are moving and the representative key-frame technique can not easily handle the resulting intra-shot variability of features.

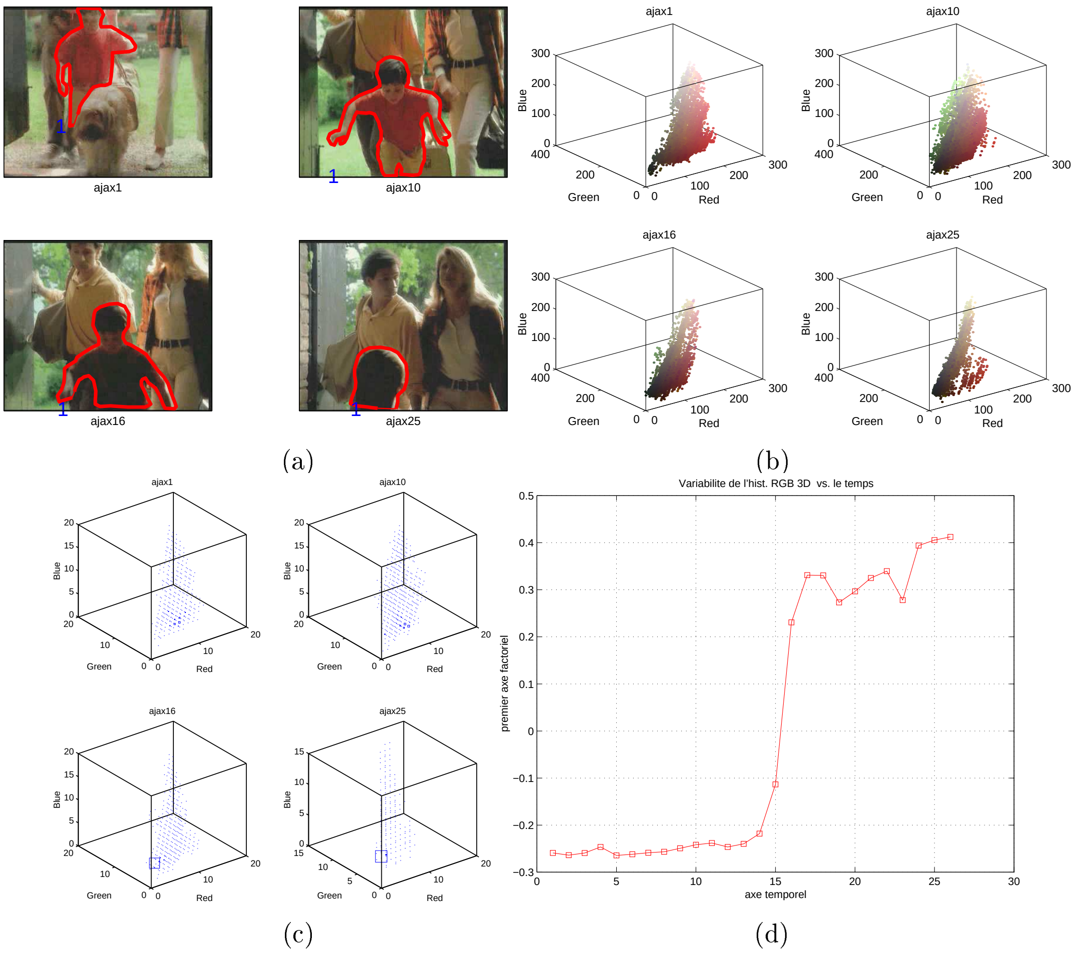

Figure 1: *Example of the temporal intra-shot variability: (a) appearance-based segmentation of a running child under different light conditions and occlusions (frames 1, 10, 16 and 25 of 26); (b) the corresponding RGB histograms; (c) the corresponding histograms where each color (R, G, B) is represented by a square whose size is proportional to the color's frequency; (d) the variation of the first principal components of the histograms over the shot (26 frames).*

Figure 1 illustrates the variability of a tracked object; Figure 1.a shows four different occurrences of a child running from sunlight into shade in a shot of 26 frames. At the beginning, the child progressively appears and at the end of the shot he disappears. Figure 1.b represents the RGB color histograms of each view of the child. This gives an idea of the intra-shot variations of objects. Figure 1.c shows these histograms where each color (R, G, B) is represented by a square whose size is proportional to the color's frequency. Figure 1.d illustrates the very significant evolution of the first principal components of these histograms over time. This makes it clear that a flexible object-based indexing process should not be limited to representative key-frames, but should take into consideration the temporal variations of features during shots. For example, as above, changes from sunlight into shade produce a significantly bimodal distribution with two different mean colors, one for each lighting condition.

In this paper we propose the use of mixture densities to model intra-shot changes of feature appearance. Such a model not only allows to capture variability but it allows also a more compact model which speeds up the retrieval process. Based on this, we then associate novel feature occurrence of an object in the video with known tracked objects. We assume that the video stream has already been segmented into shots and objects, and that some of these tracked objects have been selected and used as "labels" defining `tracked object models`. Section 3.1 contains more details of this process.

The first part of the approach consists of modeling the intra-shot changes of features within each tracked object model using a mixture of Gaussians. The goal is to separate feature distributions given by the tracked object class into homogeneous clusters. Each cluster being modeled by a single Gaussian distribution. For this purpose, we use the EM algorithm [DLR77] [MK97]. Seven different variants of Gaussian mixtures are used to describe the feature distributions. To identify the best fitting model for recognition and to choose the optimal number of components (clusters) for each feature distribution we use the Bayes information criterion (BIC) [Sch78]. Also, two alternative criteria ICL (Integrated Classification Likelihood) and NEC (Normalized Entropy Classification) are used to complete the comparative study of choosing a mixture model.

The second part of the approach consists of identifying the class of a novel occurrence of an object in the video. For that, we build the complete Gaussian mixture of all observed feature vectors of tracked object models. Occurrences are assigned to the tracked object class most likely to contain them using the maximum a posteriori probability (MAP) rule [McL92].

The organization of this paper is as follows: Section 2 introduces previous work done in the area and relevant approaches to the present work, section 3 gives a brief description of the proposed framework, section 4 details our approach, and section 5 exhibits its experimental behavior on a decompressed `MPEG` sequence extracted from the "Avengers" TV movie of "Institut National de l'Audiovisuel en France". Color histograms computed on standard, *essentially different*, colors spaces related to intensity or normalized with respect to intensity, were used as standard features widely experimented to perform evaluation of a such approach. In section 6 an alternative method to index video objects, the method of temporal mean feature proposed by Zhang [Zha96], is approached and experimented. The performance of the proposed approach is discussed in section 7 and in section 8 we give some conclusions and perspectives.

2 Related work

When dealing with the face recognition problem, [SP98] [MGR98], systems have also to deal with many different appearances: with beard or not, glasses or not, etc. To model such variability, Sung and Poggio [SP98] learn the face descriptors as a collection of six clusters, each being a Gaussian distribution; these clusters are built using a k-means clustering algorithm. In a second step near miss data are also introduced to refine the process.

For the same purpose, McKenna [MGR98] use the Gaussian color mixture to track and model face classes in natural scenes (video). This work is the closest to the contribution presented in this paper; it differs mainly by the input data which are tracked objects in our case, and in technical details like Gaussian models and the related criterion.

Mixture of Gaussians distribution is becoming more popular in the vision community. For the problem of motion recognition, Rosales [Ros98] evaluates the performance of different classification approaches, K-nearest neighbor, Gaussian, and Gaussian mixture, using a view-based approach for motion representation. According to the results of his experiments on eight human actions, a mixture of Gaussians could be a good model for the data distribution, its performance is almost as good as not assuming the form of the data distribution. In fact his work was an extension of the work of Davis [DB97], where the same descriptors representing human motion were used: Motion History Images (MHI) and Motion Energy

Images (MEI). However, in [DB97], matching was just performed using a nearest neighbor approach and they are improved using the mixture of Gaussians by [Ros98].

3 Description of the framework

3.1 Basic segmentation

In order to build up a frame-to-frame links between video objects, two visual tasks are required before running the classification process: *video-shot segmentation* and *object acquisition*.

1. *Video-shot segmentation:* This segments the sequence into temporal slices. Each slice is a set of continuous frames representing a continuous action in time or space. Most existing algorithms detect discontinuities in some chosen video parameter using inter-frame differences [Ua93] [BG96]. Parameters used include intensity, RGB or hue color, motion vectors and 3D hints. Our system implements the method of [BG96] which is based on the detection of the dominant motion in the successive images.

2. *Object acquisition:* This is done by detecting and tracking moving objects through the frames within a single shot. For this purpose, dominant motion and cross-correlation are widely used. In our system, the method of [GB97] have been implemented which uses the dominant motion to detect and track independently moving objects; otherwise, static objects are manually selected and tracked.

3.2 Feature extraction

Many different types of features could be computed on the segmented objects. In general, features are combinations of measurements that attempts to summarize the considered appearances in an efficient way. They define a multidimensional space of dimension d: the *feature space*. Thus, each object occurrence is represented by a single point in this feature space, whose coordinates are the values of its features.

One standard example of features is the color histogram. It represents the distribution of discrete color feature values in a n-dimensional color feature space ($n = 3$ for RGB). The color histogram provides a computationally efficient yet effective method which is robust under rotations, translations, scale changes and partial occlusions [SB91]. Also, different techniques of intensity normalization increase the importance of the histogram by making it invariant under illumination changes. Generally, the dimension of histogram (number of bins) is determined empirically; it has a little influence on the retrieval accuracy when it ranges from 32 to 256 [GS96] [GS98b]. A complete description of the experimented feature database is gived in section 5.2.

3.3 Classification strategy

We now give an overview of the classification strategy of the approach. A set of different tracked objects are labeled by a user using the registration system designed for this work. For each object in the video, one tracked object is selected. It is called "tracked object model" in the rest of this paper. Data from different tracked object models are collected and labeled.

For each tracked object model, we use a Gaussian mixture densities to model intra-shot variability of features. The high dimensional feature space is reduced in a statistically optimal way using the Principal Component Analysis (PCA). However, this reduced feature space is still of high dimension with respect of the number of occurrences of tracked objects in a short video shots (duration less than one second). Therefore the fitting of arbitrary Gaussian mixtures is often highly under-constrained due to limited data and the "curse of dimensionality". To make the fit stabler, we introduce in section 4.2 a selection of constrained Gaussian models with constraints on the form (linear, quadratic, ...) and on the number of estimated parameters, especially on the covariance matrix.

In section 4 we will discuss in detail the first part of the proposed approach. Based on the estimated probability densities of Gaussian mixtures for all components of the all tracked object models, a complete mixture density is built. Indeed, using the maximum a posteriori probability we classify novel occurrences according to the most probable component among all the tracked object models. More details about this second part of the approach are exposed in section 4.3.

4 Gaussian mixture for classifying video objects

Let L be the set of different tracked object models labeled by the user. Each tracked object has different image occurrences, and each occurrence belongs to one and only one of L

tracked objects. This means that each occurrence has a class label. Let x_i be the feature vector of dimension d that characterizes the occurrence i.

During the tracking within a shot, x_i is variable due to all conditions listed in the introduction. Let X to be the set of feature vectors data collected during this tracking. The distribution of X is modeled as a joint probability density function, $f(x \mid X, \theta)$ where θ is the set of parameters for the model f. We assume that f can be approximated as a mixture of Gaussians density functions; this point is revised in the next section.

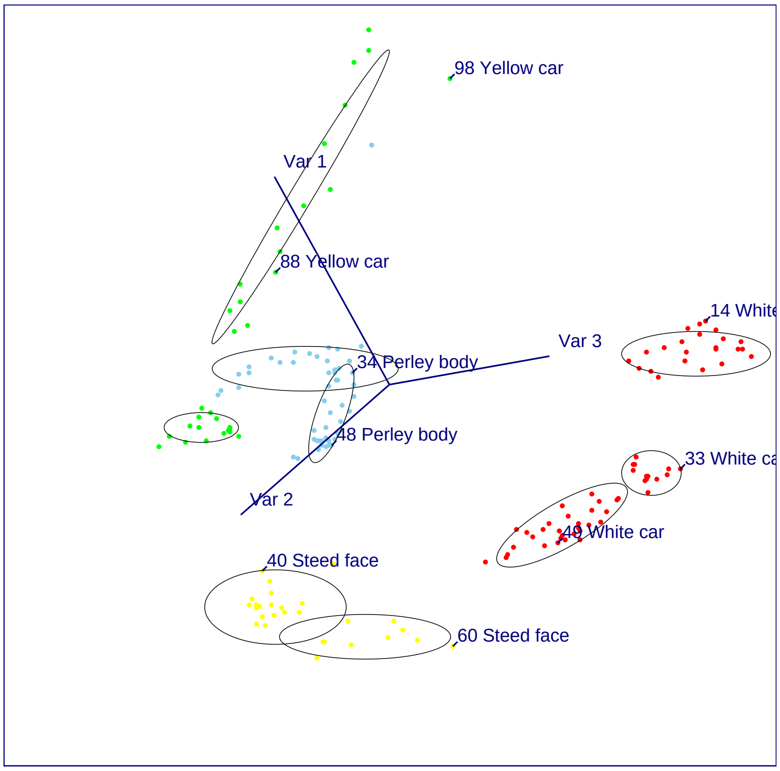

Figure 2: *Modeling the variability of the first* 4 *tracked objet models of figure 3 in the three principal components of the RGB histogram space, using mixture of several Gaussians. The covariance ellipses of each mixture component are shown.*

It remains to estimate the parameter θ for each tracked object model. We use standard density estimation techniques such as EM [MK97] (see section 4.1.2). To determine the optimal number of components for an object class (tracked object model) we use the Bayesian criteria [Sch78] [BCG98] (see section 4.2.2).

As the object is tracked, x_i varies in a continuous way. However, this continuous track in the feature space is unpredictable due to various conditions, for instance partial occlusion. Figure 2 illustrates the distribution of the first 4 tracked objects of figure 3, in the three principal components of the RGB histogram space. Each of them has to be modeled by a mixture of several Gaussian distributions; for instance the covariance ellipses of these components are shown.

4.1 Interpretation of a Gaussian mixture model

In this section we formulate the Gaussian mixture estimation procedure.

4.1.1 Definition

A J-component mixture in $\mathbb{R}^d$ is defined as

$$f(y|\theta) = \sum_{j=1}^{J} p_j \varphi(y|\alpha_j) \tag{1}$$

where the p_j's ($0 < p_j < 1$ and $\sum_{j=1}^{J} p_j = 1$) are the mixing proportions and where $\varphi(y|\alpha)$ is a density function parameterized by α. The vector of parameters to be estimated is $\theta = (p_1, \ldots, p_J, \alpha_1, \ldots, \alpha_J)$.

Following our assumptions on the data, we assume that $\varphi(y|\mu, \Sigma)$ denotes the density of a Gaussian distribution with mean μ and variance matrix Σ (positive defined symmetric). The parameters to be estimated are therefore:

$$\theta = (p_1, \ldots, p_J, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J).$$

In the following, we denote $\theta_j = (p_j, \mu_j, \Sigma_j)$, for $j = 1, \ldots, J$.

The d-dimensional Gaussian density for the component j takes the following form:

$$\varphi(y \mid \mu_j, \mathbf{\Sigma}_j) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}_j|^{1/2}} exp\left(-\frac{1}{2}(y - \mu_j)' \mathbf{\Sigma}_j^{-1} (y - \mu_j) \right) \tag{2}$$

4.1.2 Mixture density estimation

Mixture density estimation is a missing data estimation problem to which the EM algorithm [DLR77] can be applied. The type of Gaussian mixture model to be used (see next section) has to be fixed and also the number of components in the mixture. If the number of components is one the estimation procedure is a standard computation (step M), otherwise the expectation (E) and maximization (M) steps are executed alternately until the log-likelihood of θ stabilizes or the maximum number of iterations is reached.

Let $\mathbf{y} = \left\{y_i;\ 1 \leq i \leq n \ and \ y_i \in I\!R^d\right\}$ be the observed sample from the mixture distribution $f(y|\theta)$. We assume that the component from which each y_i arises is unknown, so that the missing data are the labels c_i $(i = 1, \ldots, n)$. We have $c_i = j$ if and only if j is the mixture component from which y_i arises. Let $\mathbf{c} = (c_1, \ldots, c_n)$ denote the missing data, $\mathbf{c} \in B^n$, where $B = \{1, \ldots, J\}$. The complete sample is $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_i = (y_i, c_i)$. The complete log-likelihood is

$$L(\theta, \mathbf{x}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{J} p_j \varphi(x_i|\mu_j, \Sigma_j) \right\}. \qquad (3)$$

The body of the loop of the EM algorithm at iteration "m" has two steps:

Step-E: For $i = 1, \ldots, n$ and $j = 1, \ldots, J$ compute the conditional probability, given $\mathbf{y}$, that y_i arises from the mixture component with density $\varphi(.|\mu_j^m, \Sigma_j^m)$ and mixing proportion p_j^m

$$t_{ij}(\theta^m) = \frac{p_j^m \varphi(x_i, \mu_j^m, \Sigma_j^m)}{\sum_{\ell=1}^{J} p_\ell^m \varphi(x_i|\mu_\ell^m, \Sigma_\ell^m)}. \qquad (4)$$

Step-M: Maximize the log-likelihood conditionally on t_{ij}^m. Indeed, in the case of a general Gaussian model we get for θ^{m+1}

$$p_j^{m+1} = \frac{1}{n} \sum_{i=1}^{n} t_{ij}(\theta^m) \ ; \ \mu_j^{m+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^m)\ y_i}{\sum_{i=1}^{n} t_{ij}(\theta^m)} \qquad (5)$$

$$\Sigma_j^{m+1} = \frac{\sum_{i=1}^{n} t_{ij}(\theta^m)(y_i - \mu_j^{m+1})(y_i - \mu_j^{m+1})^\mathsf{T}}{\sum_{i=1}^{n} t_{ij}(\theta^m)}. \tag{6}$$

At each iteration, the following properties hold.
For $i = 1, \ldots, n$

$$\sum_{j=1}^{J} t_{ij}(\theta^m) = 1 \tag{7}$$

and

$$\sum_{j=1}^{J} p_j^m = 1. \tag{8}$$

Initialization of the clusters is done randomly. In order to limit dependence on the initial position, the algorithm is run several times (10 times in our experiments) and the best solution is kept.

4.2 Different Gaussian mixture models

Mixture of Gaussians distribution is becoming more popular in the vision community [MGR98] [SP98] [Ros98] following to many important characteristics of this method:

- Density estimation is performed in a semi-parametric way so that the number of components scales with the complexity of the data and not with the size of the data set.
- Mixtures are sufficiently general to model arbitrarily complex, non-linear distribution accurately given enough data.
- Given limited data, the method should be constrained to provide better conditioning for the estimation.

As mentioned in the previous sections, the number of samples of tracked objects in shots is relatively small compared to the dimensions of the feature spaces. In order to estimate accurately the parameters in such a case, next section introduces the technique used in the proposed framework to constraint the method. In addition, the intra-shot variation of feature vectors could be different (more or less important) from one tracked object model to another one. This implies that the number of components of mixtures has not to be the same in all modeled tracked objects. To approach this problem, section 4.2.2 details some criteria used to select the appropriate number of components for each modeled tracked object.

4.2.1 Gaussian models

The various possible constraints on the parameters of a Gaussian mixture (e.g. all classes have same proportions, the same covariance matrix, an identity covariance matrix, ...), define 28 different models [CG95] [BG97]. Since the estimation of the covariance matrix is complex especially when the data is limited, a spectral decomposition into eigenvalues and eigenvectors is done. So, the covariance matrix for the component j is written as

$$\Sigma_j = \lambda_j D_j A_j D'_j$$

where $\lambda_j = |\Sigma_j|^{1/d}$, D_j the matrix of eigenvectors of Σ_j and A_j a diagonal matrix, such that $|A_j| = 1$, with the normalized eigenvalues of Σ_j.

The parameter λ_j determines the volume of the jth component, D_j its orientation and A_j its shape. Indeed, eight models are obtained by assuming equal or different volumes, shapes or orientations. For example the notation $[\lambda_j D_j A D'_j]$ means that the components have different volumes and orientations, and all have the same shape A. In addition, two other families of covariance matrix exists: four models assume diagonal variance matrices ($[\lambda_j B_j]$) and two models assume spherical shapes ($[\lambda_j I]$).

Of these 28 models we have implemented the following seven models derived from the three general families of covariance forms (M_1 and M_7 from ($[\lambda_j I]$); M_2 and M_3 from $[\lambda_j B_j]$; M_4, M_5 and M_6 from the general family). In the following we present a brief description of them where we consider in all cases that the proportion parameter is free for all mixture components. The degree of freedom, dof, i.e the estimated number of independent parameters, is given for each case. Let $\alpha = Jd + J - 1$ and $\beta = \frac{d(d+1)}{2}$.

- M_1: $\Sigma_j = \lambda_j I$ where $\lambda_j = \sigma_j^2$ is unknown and I is the identity matrix ($dof = \alpha + d$).
- M_2: $\Sigma_j = \lambda_j \mathrm{Diag}(a_1, \ldots, a_d)$ where $\lambda_j = \sigma_j^2$ is unknown, $|\mathrm{Diag}(a_1, \ldots, a_d)| = 1$ with unknown $a_1, \ldots, a_d$ and $\mathrm{Diag}(a_1, \ldots, a_d)$ denotes a diagonal matrix with diagonal vector $a_1, \ldots, a_d$ ($dof = \alpha + d + J - 1$).

- M_3: $\Sigma_j = \lambda_j \mathrm{Diag}(a_1^j, \ldots, a_d^j)$ where $\lambda_j = \sigma_j^2$ is unknown and $|\mathrm{Diag}(a_1^j, \ldots, a_d^j)| = 1$ with unknown $a_1^j, \ldots, a_d^j$ ($dof = \alpha + Jd$).
- M_4: $\Sigma_j = \lambda_j C_j$ where $\lambda_j = \sigma_j^2$ and $C = DA_jD'$ with $|C| = 1$; in this case we assume that all components have the same orientation and identical ellipsoidal shapes ($dof = \alpha + \beta + J - 1$).
- M_5: $\Sigma_j = \lambda_j C_j$; no restriction is placed on the covariance matrices $\Sigma_1, \ldots, \Sigma_J$. It is the most complex model ($dof = \alpha + J - 1 + J\beta$).
- M_6: $\Sigma_j = \lambda C_j$; all covariance matrices have the same volume ($dof = \alpha + J\beta - (J-1)$).
- M_7: $\Sigma_j = \lambda I$ where I is the identity matrix. This is the simplest model where only the means must be estimated ($dof = \alpha + 1$).

For each of these seven Gaussian models, the maximization step of the EM algorithm has its own specific form. Details are given by Celeux and Govaert [CG95].

4.2.2 Choosing models and mixture components' number

The major problem is to define how many Gaussians should be used, and what should be the Gaussian model. There will be an increase in the log-likelihood of the data with respect to the model whenever we increase the number of Gaussians, until the degenerate case of non invertible covariance matrices makes it infinite. But there is a penalty to be paid in terms of the number of parameters needed to specify the increasing number of Gaussians.

Many criteria are proposed in the literature [Sch78] [Bry91] for selecting the best number of Gaussians with a known Gaussian model (see previous section); 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.

The Bayes Information Criterion (BIC) due to Schwartz uses Bayesian arguments [Sch78]. It minimizes the following criterion:

$$BIC(M) = -2L_M + Q_M ln(n) \tag{9}$$

where L_M is the maximized log-likelihood of the model M and Q_M is its number of free parameters.

A recent attempt to tackle the above problem was done by Biernacki [BCG98]; a novel criterion called Integrated Classification Likelihood (ICL) was proposed (equation 10). 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.

$$ICL(M) = -2L_M + Q_M ln(n) - 2\sum_{i=1}^{n}\sum_{j=1}^{J}\hat{c}_{ij}t_{ij}, \tag{10}$$

where $\hat{c}_{ij}$ represents the estimated partition deduced from t_{ij}.

Another well-known criterion to select the best structure of the data (number of Gaussians and Gaussian model) is the Normalized Entropy Criterion (NEC) [CS96]. This criterion, to minimize, is gived by:

$$NEC(M) = \frac{E(M)}{L(M) - L_1(M)} \tag{11}$$

where

$$E(M) = -\sum_{i=1}^{n}\sum_{j=1}^{J} t_{ij}\log t_{ij} \geq 0,$$

represents the entropy witch can be regarded as a measure of the ability of the J-component mixture model to provide a relevant partition of the data, $L_1(M)$ denotes the maximized likelihood for a single Gaussian distribution.

For a reasonable range of the number of Gaussians (see section 5.5) and for the different Gaussian models (M_1 .. M_7), the values of a selected criterion are computed using EM to maximize the log-likelihood in each iteration and finally the minimum is picked witch indicates the best pair of Gaussian model and number of Gaussians.

4.3 Probabilistic recognition process

As mentioned in the introduction, the goal of this work is to build a similarity measure between video objects. A set of L tracked object models has been selected by the user, and each one has been modeled by a Gaussian mixture. In order to be able to classify a new occurrence (region) in the video, to one of these learned classes, we collect their corresponding L Gaussian mixtures to a single global Gaussian mixture of $\mathbf{W}$ components, where $\mathbf{W} = \sum_{l=1}^{\mathbf{L}} J_l$ and J_l is the number of components of the lth Gaussian mixture. For that, the mean and covariance of each component of the global Gaussian mixture being build are held fixed and only the proportions parameters are recomputed. As consequences, the feature space is divided into W regions. Then, for a novel point y_i corresponding to a requested occurrence in the video, the a posteriori probability, t_{ij} (formula 4) that y_i arises from the mixture component j ($j \in 1..W$) is computed. Then the probability, t_{il}, of membership of y_i in the lth tracked object model ($l \in 1..L$) is computed by summing the probabilities t_{ik} where $k \in 1..J_l$. In this case, all components are assumed independent. Finally, the y_i is assigned to the most probable tracked object model.

As the occurrences object corresponding to y_i is tracked within its shot, the classification probabilities can be integrated over time as detailed in the next section.

5 Experiments

Experiments have been conducted on a video clip of 1016 frames extracted from the *MPEG* "Avengers" TV movie of "`Institut National` de l'`Audiovisuel`" (INA). The *Shot segmentation* and *Object acquisition* tasks (see section 3.1) produce 1391 occurrences of 52 different tracked objects.

5.1 Object models

Occurrences of all tracked objects correspond to 12 different classes (white Ford, blue Mercedes, Volvo, Steed actor, Perley actress, etc). So, 12 different tracked objects were selected (randomly) and labeled in shots using the registration system designed for this work. Thus, 448 different views (occurrence images) of them were collected to estimate parameters of the Gaussian mixture densities. Figure 3 displays 4 different occurrences of each one of the 6 learned tracked object models.

5.2 Feature database

Most of the work on object recognition is based on matching sets of geometric image features (e.g. edges, lines and corners) to 3D objects models. For the complex case of image as displayed in Fig. 3 and Fig. 4, such tools would not solve the matching in particular as geometric features may be absent. In the contrast, color is a powerful information for object recognition [GS96] since colors of objects correlate strongly with object identity. A simple and effective recognition scheme is to represent and match objects on the basis of color-

Figure 3: *Subset of learned tracked object models*

metric histograms as proposed by Swain and Ballard [SB91]. Part of the appeal of Swain and Ballard's method is that a color histogram is independent of many imaging conditions: e.g. the orientation of a scene, the relative position of particular scene elements and the absence (or occlusion) of some of the colors. The work presented here is based on this technique.

A histogram $\{h_i\}$ is a mapping from a set of d-dimensional integer vectors i to the set of nonnegative integers. These vectors typically represents bins while partitioning the relevant region of the underlying space, and the associated integers are a measure of the mass of the distribution that falls into the corresponding bin. The histogram approach is well known as an attractive method for object recognition because of its simplicity, speed and robustness.

Now, multiple well-known color spaces include: *RGB, Lab, HSV, xyz, rgb,* $l_1 l_2 l_3$, [GS98a] could be used for representing color features in images or objects. However, a number of these color features are related to intensity I (L, V), or they are linear combinations of RGB (XYZ) or normalized with respect to intensity *rgb* ($l_1 l_2 l_3$, *xyz*). In this paper, experiments are conducted on the following standard, *essentially different*, color features: *RGB, I, rgb, H, HS, HSV* and $l_1 l_2 l_3$.

Color Definitions. Let R, G and B, obtained from a color camera, represents the tri-stimulus components defining a mapping from image space to a 3-D sensor space:

$$\int_{\lambda} E(\lambda) U_C(\lambda) d(\lambda) \tag{12}$$

for tri-stimulus values $C \in (R, G, B)$, where $E(\lambda)$ is the radiance spectrum and U_C are the three filter transmission functions. To represent the RGB-color space, a cube can be defined on the R, G and B axes. The axis connecting the black and white corners defines the intensity (or value V in HSV space):

$$I(R, G, B) = \frac{R + G + B}{3} \tag{13}$$

The projection of RGB points on the chromaticity triangle is defined by:

$$r(R, G, B) = \frac{R}{R + G + B},\ g(R, G, B) = \frac{G}{R + G + B},\ b(R, G, B) = \frac{B}{R + G + B} \tag{14}$$

yielding the rgb color space. This normalization of the length of each RGB (making $R+G+B=1$) is an effective way of removing this intensity dependence [FT99].

The transformation from RGB to HSV used here is given by:

$$H(R,G,B) = \arctan(\sqrt{\frac{\sqrt{3}(G-B)}{(R-G)+(R-B)}}) \tag{15}$$

representing the *hue*: tint or tone the most significant characteristic of the color and

$$S(R,G,B) = 1 - \frac{min(R,G,B)}{I(R,G,B)} \tag{16}$$

representing the saturation S that measures the relative white content of a color (i.e S for a white color is equal to zero).

The $l_1l_2l_3$ color space, proposed by [GS98a], is given by equations 17, 18 and 19:

$$l_1(R,G,B) = \frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \tag{17}$$

$$l_2(R,G,B) = \frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \tag{18}$$

$$l_3(R,G,B) = \frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2} \tag{19}$$

In this way, all color features can be calculated from the original R, G and B values from the corresponding red, green and blue images provided by the color camera. Now, opposed to RGB, HSV, HS and I other color features H, S, rgb and $l_1l_2l_3$ are invariant under a given body (diffuse) and surface reflection model, witch are assumed to be Lambertian [GS96].

Alternatively to normalized color space there is the constant color one (i.e diagonal matrix model). According to the results in [FT99], Finlayson demonstrates that it exists an equivalence between them; both H, S, rgb and $l_1l_2l_3$ are invariant with uniform intensity scaling.

5.3 Efficient considerations

Color space quantization. Since the RGB color space (or other derived spaces) is continuous, or possibly discrete with a large number of values (i.e., digital photographs are typically represented in RGB color space discretized to 256 levels per axis, which gives

Figure 4: *Subset of query objects*

over 1.67 millions distinct color points), the color space must be quantized or partitioned into a smaller number of colors. For this, histogram axes of different color feature spaces are partitioned uniformly with fixed intervals. The resolution on the axis is related to the computational efficiency considerations of histograms and furthermore of the estimation process of parameters of the mixture of Gaussians. To satisfy these conditions, the number of bins must be small but discriminant. Therefore, in the present experimentations, all the 3-dimensional (*RGB, HSV, rgb and* $l_1 l_2 l_3$), 2-dimensional (*HS*) and 1-dimensional (*H, S and I*) color feature spaces are quantized into 64, 49 and 32 colors respectively.

Dimensionality reduction. An histogram constructed on the basis of a color space is represented by a d-dimensional point in the feature space. Now, given a limited set of points those corresponding to occurrences of a tracked object model (in shot), estimating the Gaussian mixture density function from these samples could be inaccurate since the dimension of the space is still very high which leads to empty-spaces (for example, in the 64-dimensional space, 24 occurrences of a tracked object in a shot of one second, are not sufficient to accurately estimate the general Gaussian mixture model of one component). To approach this problem, the Principal Components Analysis (PCA) [Hot33] is used to reduce the data dimension by a linear projection on a subspace of the original data space that best preserves the variance in the data. The PCA is a standard method in data analysis; it consists to solve the well known eigenvalue decomposition problem:

$$\lambda = \nabla^t \Sigma \nabla \tag{20}$$

where ∇ is the eigenvector matrix of the covariance of the data and λ is the corresponding matrix of eigenvalues. Only p eigenvectors are kept corresponding to the p largest eigenvalues. The dimension of the new feature space is not fixed a priori but it's determined by the representation quality of the data, Q_E, in the PCA space E ($d_E \leq d$). Q_E is gived by the equation 21 which describes the ratio of the variance of the data in the E space.

$$Q_E = \frac{\sum\limits_{j \in E} \lambda_j}{\sum\limits_{i=1}^{d} \lambda_i} \times 100 \tag{21}$$

In the present experiments the Q_E is fixed to be greater than 95%. Following this, the features RGB, HSV, rgb, $l_1l_2l_3$, HS, H, S and I are projected in the $10, 10, 3, 10, 8, 5, 8$ and 8-dimensional spaces respectively.

Again, PCA is an important step beyond the overcoming of the curse of dimensionality since the number of samples (occurrences) of a tracked object model is not in general sufficient to fit an optimal Gaussian mixture model.

5.4 Queries

Each occurrence image of tracked objects can be considered as a potential query. The class of such query is obtained using the maximum a posteriori probability rule. However, we can make use of the tracking within a shot to make more robust decision, particularly when some images are heavily distributed due to occlusion or other unpredictable events. For such a case a robust decision is implemented through a majority decision rule: the final class is the class for which most of individuals belongs to it.

The query set consists of 1391 individual queries which corresponding to 52 different sequences of tracked objects. Each query has a unique correct answer (one of the learned tracked object model). However, queries represent various challenging situations like different views of the object, large changes in appearance, illumination changes, partial occlusions up to 50%, etc. Figure 4 shown some views of these queries to be classified by the system.

5.5 Results

The performance of the proposed approach is measured by computing the percentage of total queries correctly classified. First time, the total individual percentage, *ind.%*, is computed where each query is classified independently from the others. Next time, the total tracking percentage, *track.%*, is computed where each query is classified based upon a *majority vote*; all occurences of a tracked object are classified to the same class.

The test results for descriptors those are related to intensity and normalized with respect to intensity are shown in tables 1 and 2 respectively. Also, for each criterion used to select the best structure of the data the results are shown. The maximum number of Gaussian components, $MaxNbC$, used in the current experimentations was ranged from 1 to 4.

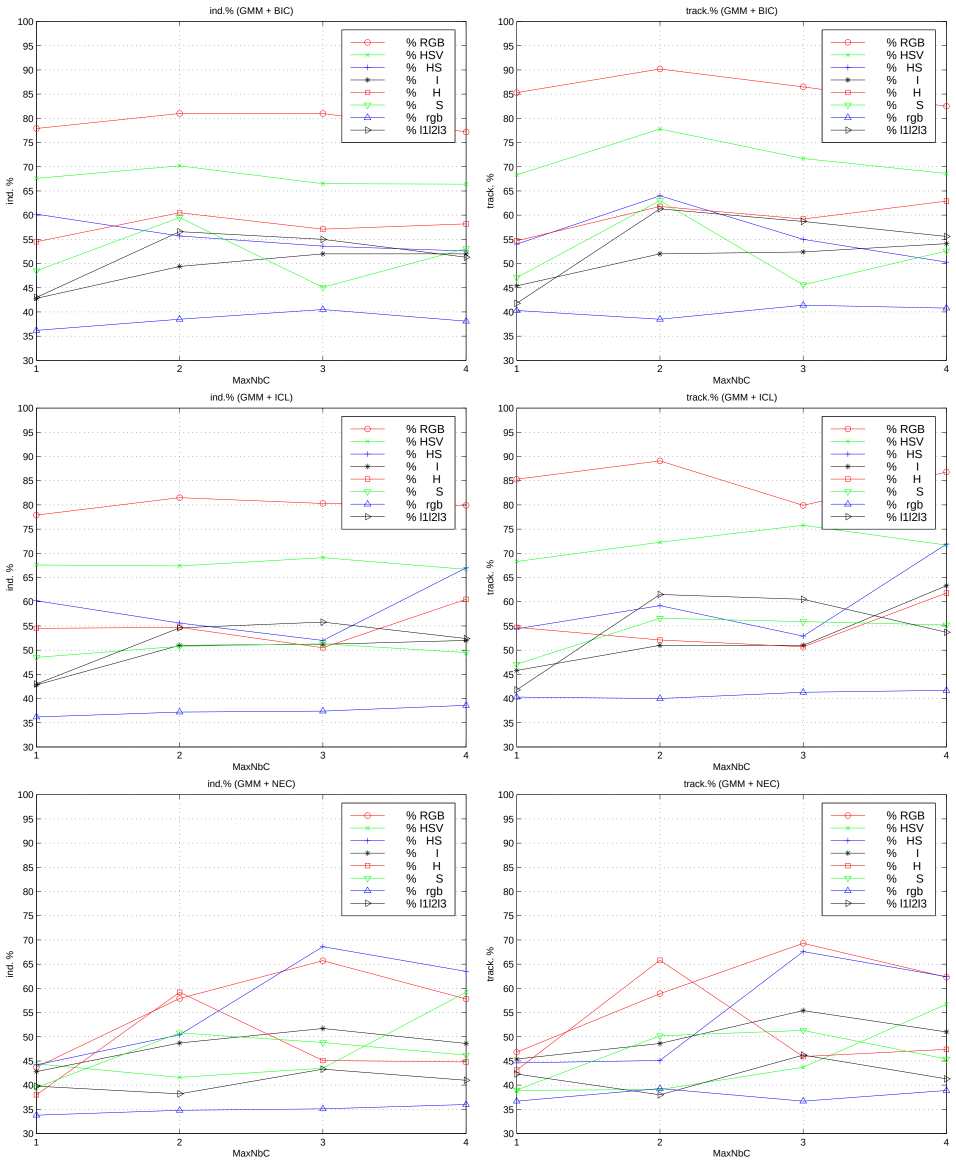

Figure 5: Test results with the Gaussian mixture method using BIC (first raw) ICL (second raw) and NEC (third raw) criteria. Total percentage of correctly classified queries: *ind.%* (left column) *track.%* (right column).

From these tables (1 and 2) one can draw the following conclusions: The RGB histogram gives the best ratio of queries classification (90.2%) with the mixture method of 2-Gaussians components maximum. It's followed by the HSV (77.8%), HS (71.9%), I (63.3%), S (63.0%), H (62.9%), $l_1l_2l_3$ (61.5%) and rgb (40.3%) hsitograms. The rgb gives the poor results.

The method gives relatively the same results using either of BIC or ICL criteria. These criteria are used to automatically determine whether a single Gaussian or a several- components Gaussian mixture provides the best fit with a selected Gaussian model among the

Features	d	d_E	*MaxNbC*	**Total ind.** %			**Total track.** %		
				BIC	ICL	NEC	BIC	ICL	NEC
RGB	64	10	1	77.9	77.9	43.7	85.3	85.3	46.8
RGB	-	-	2	81.0	81.5	57.9	90.2	89.1	58.9
RGB	-	-	3	81.0	80.3	65.7	86.5	79.9	69.3
RGB	-	-	4	77.2	79.9	57.8	82.5	86.8	62.3
HSV	64	10	1	67.6	67.6	44.4	68.3	68.3	38.9
HSV	-	-	2	70.2	67.4	41.6	77.8	72.3	39.1
HSV	-	-	3	66.5	69.1	43.5	71.7	75.8	43.7
HSV	-	-	4	66.4	66.7	59.2	68.6	71.7	56.7
HS	49	8	1	60.2	60.2	44.2	54.0	54.4	44.6
HS	-	-	2	55.7	55.6	50.4	64.0	59.2	45.1
HS	-	-	3	53.6	52.0	68.6	55.0	52.9	67.6
HS	-	-	4	52.6	67.0	63.5	50.3	71.9	62.4
I	32	8	1	42.8	42.8	42.8	45.4	45.8	45.4
I	-	-	2	49.4	51.0	48.7	52.0	51.0	48.6
I	-	-	3	52.0	51.2	51.7	52.4	51.0	55.4
I	-	-	4	52.0	52.0	48.6	54.1	63.3	51.0

Table 1: *Test results with the Gaussian mixture method for RGB, HSV, HS and I histograms; Total percentage of correctly classified queries:* **ind.** % *and* **track** %*; Different results using BIC, ICL and NEC criteria are displayed.*

seven implemented in the current framework. However, the method gives a poor results with the entropy criterion (NEC).

6 Alternative methods for video objects recognition

In a related work in video representation, o'connor [O'C91] summarizes each video shot by a selected key-frame: the middle image of the shot in general. However, key frames utilize only spatial information. Zhang [HJZW95] integrates in his system of video structuring the

Features	d	d_E	*MaxNbC*	**Total ind.** %			**Total track.** %		
				BIC	ICL	NEC	BIC	ICL	NEC
H	32	5	1	54.5	54.5	38.0	54.7	54.7	43.1
H	-	-	2	60.5	54.7	59.2	61.8	52.1	65.8
H	-	-	3	57.1	50.5	45.1	59.2	50.7	45.9
H	-	-	4	58.2	60.5	44.8	62.9	61.8	47.4
S	32	8	1	48.5	48.5	39.5	47.1	47.1	39.0
S	-	-	2	59.5	50.8	50.8	63.0	56.6	50.2
S	-	-	3	45.1	51.3	48.8	45.6	55.9	51.3
S	-	-	4	53.1	49.5	46.2	52.6	55.2	45.4
rgb	64	3	1	36.2	36.2	33.8	40.3	40.3	36.7
rgb	-	-	2	38.5	37.2	34.8	38.5	40.0	39.3
rgb	-	-	3	40.5	37.4	35.1	41.4	41.3	36.7
rgb	-	-	4	38.1	38.6	36.0	40.8	41.7	38.9
$l_1l_2l_3$	64	10	1	43.0	43.0	39.8	41.8	41.8	42.3
$l_1l_2l_3$	-	-	2	56.6	54.6	38.2	61.3	61.5	38.0
$l_1l_2l_3$	-	-	3	55.0	55.8	43.3	58.7	60.5	46.2
$l_1l_2l_3$	-	-	4	51.3	52.4	41.0	55.6	53.7	41.3

Table 2: *Test results with the Gaussian mixture method for H, S, rgb and $l_1l_2l_3$ histograms; Total percentage of correctly classified queries:* **ind.** % *and* **track** %*; Different results using BIC, ICL and NEC criteria are displayed.*

temporal characteristics at shot level and proposes some temporal features like the mean of average brightness and a few dominant colors calculated over all frames in a shot.

Features	d_E	**Total ind.** %	**Total track.** %
RGB	10	56.6	55.3
HSV	10	53.7	60.3
HS	8	48.2	48.4
I	8	36.6	41.1
H	5	42.4	48.7
S	8	39.5	39.5
rgb	3	35.4	41.3
$l_1l_2l_3$	10	42.4	42.7

Table 3: *Test results with the temporal mean feature method in the PCA feature space (d_E)*

Using this "temporal mean feature" method, each tracked object model is represented by only one d-dimensional point in the feature space: the centroid of the distribution μ. To identify the class of object queries the first nearst neighboor method is used with the Euclidien distance as a dissimilarity metric.

Table 3 shows the test results with the temporal mean feature method in the PCA feature space.

7 Comparative analysis and discussion

Performance of the proposed method. The proposed method to classify localized objects in the video sequence, consists to modelise the temoporal variation of descriptors of all occurence images of tracked object models, using a mixture of several Gaussians. Such a method is efficient for non-rigid video objects recognition where their degree of variability is high. In the other hand, the method of temporal key-frame is simple and efficient when the object is static but most of our experimental video objects are moving and deforming.

For the set of tested features, the *track.*% measure of correctly classified queries is increased considerably, by the using of Gaussian mixture method, rather than the temporal mean feature method, for 10% to 35%. Figure 6 illustrates a comparision in the performamce of these methods.

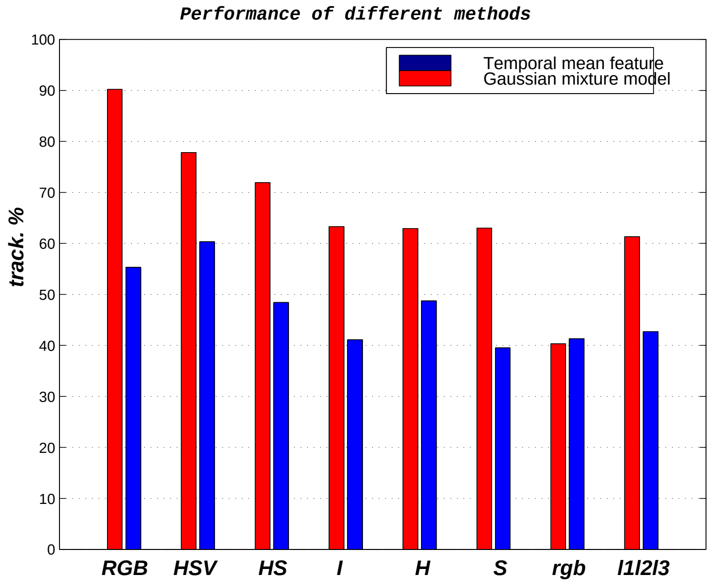

Figure 6: Total *track*.% of the two methods: Gaussian mixture model and temporal mean feature.

Form of the distributions and our approach. As illustrated in figure 2, some tracked object models were represented by a compact set of points in the feature space but others are more scattered. This is the consequence of the degree of variability of each tracked object within its shot. As an evidence, the key-frame method and the temporal mean feature one will give poor results. This was confirmed by the experimental results given in the previous section. The variability modeling approach using Gaussian mixture densities performs best.

On the other hand, the results obtained by the Gaussian mixture approach indicates that the distributions are not unimodal in general. A reliable estimate can be obtained by the BIC or ICL criteria. The NEC criterion is shown that is not a good criterion to select both the best number of Gaussians and the optimal Gaussian model. That the maximum number of Gaussians in the mixture that best represents the underlying distributions was

2 for RGB, HSV, $l_1l_2l_3$ and rgb histograms and 4 for the I, H, S, and HS, indicates the better suitability of multi-modal distributions to describe the data. We could see that the more multi-modal the estimated distribution, the better the classification results. However, the risk of over-fitting the data is also higher. This case is shown for the RGB data when *MaxNbC* is greater to 2. The explanation we have yet reached is that the amount of data within each cluster (one Gaussian component) becomes rather small, typically 6; in such a high dimensional space, this leads to unstable Gaussian distribution estimation.

The Gaussian mixture method with the ICL criterion [BCG98] gives a better results when the $MaxNbC$ is equal to 4 for the one and two-dimensional features (excepted the H color feature). However, for the three-dimensional features the ICL gives relatively similar results as the BIC where the entropy criterion NEC gives a poor results for the most of features.

Assumption on the form of the data. The basic assumption that has been made on the distributions of the experimented features was its Gaussian form. This could be valid for RGB and I for examples but it is not the case for the H feature where it is cylindrical. Another kind of mixtures could be used (Gamma distribution) to model the variability in such a case but it is hard to implement.

Classifying of single and tracked objects. Classifying tracked objects with a robust technique based on the *majority vote* increases the success rate up to 10% for the most tested features.

Low-level object representation and performance. In computer vision, the recognition rate is limited by the similarities in the class descriptions given by the feature vector. This is the main source of error where color histograms are not invariant to partial occlusions of moving video objects (see figures 3 and 4). Moreover, color histograms makes alike two different objects that have similar global color distributions. So, the Gaussian mixture classifier had the tendency to concentrate their error in misclassifying of some tracked object tests with tracked object models.

In the other hand, the color histograms that are invariant under illumination changes, H, S, rgb, and $l_1l_2l_3$ [GS98a] [FT99] did not improve the recognition rate on the current experimented objects database. The normalized rgb histograms gives the poorest results using the temporal mean feature method and for our method. These results have not been expected according to the results obtained by Finlayson [FT99]. Our explanation to this phenomena is that the normalization process compacts the histogram form and it loses information especially when the dimension of the histogram is small (64 for example). Then the normalization reduces again the discriminative characteristic of the data. This could be resolved by computing of histograms of higher dimensions (4096 bins for example). However, such dimensions represents a challenge problem for video indexing systems and the proposed method. It's important to notice that illumination changes in our objects database are not artificially done, in other words the model of illumination variation is not known a priori; in many images of figures 3 and 4 the illumination changes correspond to the partial presence or absence of the sun where on the other experimented databases of [FT99] and [GS98a] this is not the case.

Another source of errors is a direct consequence of the dominant motion detection and tracking mechanism we use [GB97]. Dominant motion is hard to track when the object is moving and its variation between successive frames is high. Figure 4, illustrates many views of objects where the background of the underlying images represents a significant part of them (for example occurrence #602.81). A more sophisticated moving object detection and tracking techniques, that combine motion and color for example, would increase robustness.

Alternative to linear projection. Given limited data which are represented in a relatively high dimensional space, the PCA is applied in order to overcome this curse of dimensionality. This leads to a more accurate estimation of the Gaussian mixture models. However, the PCA method cannot take into account non linear structures of the data, structures consisting of arbitrarily shaped clusters or curved manifolds since it describes the data in terms of a linear subspace. Another, non linear projection method, multidimensional scaling could be used, but its major problem is that it is computationally very intensive for large data set.

Also, the PCA method is widely used in image retrieval systems (eigen images) in order to accelerate the search process. However, it reduces the accuracy of the recognition. This is shown in the present experiments by comparing the results shown in tables 3 and 4; the accuracy of object retrieval is relatively higher in the real feature space than in the PCA one. This experiments was not conducted with the mixture of Gaussians approach: the estimation of Gaussian mixture model is very unstable is such a high dimension space given limited data set.

In the real feature the distance used is the normalized χ^2 metric (equation 22).

$$\chi^2(Q,V) = \sum_i \frac{(q_i - v_i)^2}{(q_i + v_i)} \tag{22}$$

where q_i and v_i represent the frequencies of the ith bin of the Q and V histograms respectively.

Features	d	**Total ind.** %	**Total track.** %
RGB	64	61.4	59.3
HSV	64	59.0	60.0
HS	49	52.9	56.6
I	32	37.8	43.2
H	32	53.7	56.1
S	32	42.3	40.8
rgb	64	30.4	25.6
$l_1l_2l_3$	64	45.5	48.3

Table 4: *Test results with the temporal mean feature method in the real feature space (d)*

8 Conclusion and perspectives

In this paper we have presented a methodology for increasing the robustness of existing features used in video object recognition. Temporal variation embedded in the video, given by the segmentation in shots, is used in the recognition process. The variability of appearance-based objects is modeled by Gaussian mixtures where the number of components and the Gaussian model are chosen automatically. Each component of the mixture corresponds to a stable appearance of the object in the shot. The video shot is represented by a multi-modal probability distribution, rather than by a simple point (key or mean frame) in the feature space. As shown in the experimental study, on a very variable video objects database, such modeling improves the recognition rate considerably when compared to the classical temporal mean feature approach used in video indexing.

Future experimentation will be performed on other features like correlograms [HKM+97] and local invariants [SM97]. However, the color histogram is a good example to demonstrate the behavior of the approach. The dimension of the feature space is a critical point for the density estimation process. Thus, when the number of occurrences of a tracked object model is limited, and the variability of feature distribution is large, the estimation process becomes more difficult, and so suboptimal Gaussian models are selected (among seven) to prevent over-fitting. Artificial object images could be generated, from existing once, to solve this point.

The use of tracked objects yielded better recognition performance than the use of single occurrence images. The principle is based on the majority vote. One direct extension of this consists in estimating the Gaussian mixture of the tracked object to be classified and then to compute a metric distance (*Kullback-Leibler* for instance) between Gaussian components of it and the learned tracked object models.

Another extension of the approach would be made by eliminating the outliers points in order to estimate the Gaussian densities robustly. Finally, a future research direction we intend to explore is to render the classification process of video objects completely automatic, so no tracked object models need to be specified by the user. Such work would be very useful but it will be quite challenging. It fits the general unsupervised clustering problem in a context where individuals are mixtures of distributions estimated differently.

Acknowledgments

We would like to acknowledge `Alcaltel CRC` for its support of this work. We are grateful to `C. Biernacki` for his helpful comments on Gaussian clustering models. Finally, we would like to acknowledge the "Institut National de l'Audiovisuel en France" for giving permission to use their video material.

References

[BCG98] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated classification likelihood. Rapport de recherche, INRIA, 1998.

[BG96] 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.

[BG97] C. Biernacki and G. Govaert. Choosing Gaussian Models in Discriminant Analysis. In *IV International Meeting of Multidimentional Data Analysis*, Bilbao, Spain, September 10-12 1997.

[Bry91] P.G. Bryant. Large-Sample Results for Optimization Based Clustering Methods. *Journal of Classification*, 8:31–44, 1991.

[CG95] G. Celeux and G. Govaert. Gaussian Parsimonious Models. *Pattern Recognition*, 28(5):781–783, 1995.

[CJ91] D.T. Clemens and D.W. Jacobs. Space and time bounds on indexing 3d models from 2d images. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1007–1017, October 1991.

[CS96] G. Celeux and G. Soromenho. An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. *Journal of Classification*, 13(2):195–212, 1996.

[Dav93] M. Davis. Media streams : An iconoc visual language for video annotation. *Proc. Symposium on Visual Languages*, 1993.

[DB97] J. Davis and A. F. Bobick. The representation and recognition of human mouvement using temporal templates. In *Computer vision and pattern recognition (CVPR)*, 1997.

[DH73] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* Wiley-Interscience, 1973.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[FT99] G. D. Finlayson and G. Y. Tian. Color normalization for color object recognition. *International Journal of Pattern Recognition and Atrificial Intelligence*, 13(8):1271–1285, 1999.

[GB97] 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.

[GS96] T. Gevers and A. W. M. Smeulders. A comparitve study of several color models for color image invariant retrieval. In *Proceedings of the First International Workshop, IDB-MMS*, pages 17–26, Amsterdam, The Netherlands, August 1996.

[GS98a] T. Gevers and A. W. M. Smeulders. Image indexing using composite color and shape invariant features. In *ICCV*, pages 576–581, Bombay, India, 4-7 January 1998.

[GS98b] T. Gevers and A.W.M. Smeulders. Image indexing using composite color and shape invariant features. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pages 576–581, 1998.

[HC99] R. Hammoud and L. Chen. A spatiotemporal approach for semantic video macro-segmentation. In *European Workshop on Content-Based Multimedia Indexing*, pages 195–201, IRIT-Toulouse FRANCE, Octobre 1999.

[HJZW95] S.W. Smoliar H. J. Zhang, C. Y. Low and J.H. Wu. Video parsing, retrieval and browsing: An integrated and content-based solution. *ACM Multimedia*, pages 15–24, 1995.

[HKM+97] 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.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24:417–441, 1933.

[McL92] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[MGR98] S. J. Mckenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern recognition*, 31(12):1883–1892, 1998.

[MK97] G. L. McLachlan and T. Krishnan. *The EM algorithm and Extensions.* John Wiley and Sons, New York, 1997.

[O'C91] B.C. O'Connor. Selecting key frames of moving image documents : A digital environment for analysis and navigation. *Microcomputers for Information Management*, 8(2):119–133, 1991.

[RBE94] L.A. Rowe, J.S. Boreczky, and C.A. Eads. Indexes for user access to large video databases. *Proc. IS.T SPIE Conf. on Storage and Retrieval for Image and Video Databases II*, pages 150–161, 1994.

[Ros98] 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.

[SB91] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[SBS97] N. Sawheny, D. Balcom, and I. Smith. Authoring and navigating video in space and time. *IEEE Multimedia Magazine*, 4(4):30–39, 1997.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[SM97] 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.

[SP98] 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.

[Ua93] H. Ueda and al. Automatic structure visualization for video editing. In *Proc. InterCHI 93, ACM press, New York*, pages 137–141, 1993.

[WZ98] R. Weber and P. Zezula. A quantitative analysis of performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24th VLDB Conf.*, 1998.

[Zha96] Z. Zhang. On the epipolar geometry between two images with lens distortion. In *Proceedings of the 13th International Conference on Pattern Recognition,* *Vienna, Austria*, volume I, pages 407–411. IEEE Computer Society Press, August 1996.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399