

Mixture densities for video objects recognition

Riad Hammoud and Roger 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

Abstract

The appearance of non-rigid objects detected and tracked in video streams is highly variable and therefore makes the identification of similar objects very complex. Furthermore, indexing and searching of them represent a very challenging problem in computer vision. This paper presents a framework for object-based matching that increases the robustness of existing feature detectors used for object recognition. The Gaussian mixture densities are used to model intra-shot variations of observed features of tracked objects. This process is achieved by the EM algorithm which separates feature distributions given by a tracked object into homogeneous clusters. We use seven different variants of Gaussian mixtures and the Bayes information criterion to identify the best structure of the data (model and parameters). Experiments are conducted on a video sequence of fifteen different tracked objects and comparison in the performance of the mixture approach and the two key-frame methods is analyzed and reported.

1. Introduction and Motivation

Video has a rich implicit temporal and spatial structure based on shots, camera and object motions, etc. To enable high level searching, browsing and navigation, this implicit structure needs to be made explicit [6]. For this purpose, cut detection [1] and object acquisition [4] 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 that motivate the research in this area: video surveillance, human-computer interaction, etc.

Content based indexing of non-rigid video objects using low-level visual image features is still a challenging research problem in computer vision for the two following reasons. (1) The nature of video is dynamic: rotations, occasional occlusions, variable illuminations, etc. Therefore recognition with classical methods [3] gives poor results.

(2) The number of occurrences of all detected objects in shots is enormous; for example, 129600 objects can be localized in an MPEG video of 1h30 (assuming one object per image).

One popular way in video indexing consists of representing shots by "representative" key-frames [8]. 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 illustrates the variability of a tracked object in a shot of 26 frames; the left side shows four different occurrences of a child running from sunlight into shade. At the beginning, the child progressively appears and at the end of the shot he disappears. The right side illustrates the very significant evolution of the first principal components of their RGB histograms over time. This makes clear that a flexible video object recognition process should not be limited to representative key-frames, but should take into consideration the temporal intra-shot variations of features. For example, as above, changes from sunlight into shade produce a significantly bimodal distribution with two different mean colors, one for each lighting condition.

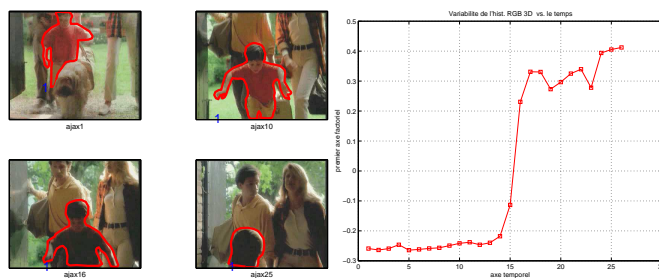


Figure 1. Intra-shot appearance of the running child (frames 1,10,16 and 25 of 26) (left); significant variability of their RGB histograms over time (right)

In order to overcome these difficulties, we propose in this paper the use of mixture densities to capture the intra-shot variability of features of "tracked object models" and



then to identify the class of a novel occurrence image in the video. We use the EM algorithm [2] to find homogeneous clusters in the feature space of a “tracked object model” and seven different variants of Gaussian models to describe the feature distributions. In order to identify the best fitting model for recognition (among these seven) and to choose the optimal number of components (clusters) for each feature distribution we use the Bayes Information Criterion (BIC) [10].

Our approach deals with the general problem of non-rigid object recognition and consists of improving robustness of existing recognition methods. Also, it allows a more compact model which speeds up the retrieval process. The organization of this paper is as follows: Sec. 2 introduces previous work done in the area and relevant approaches to the present work, Sec. 3 gives a brief description of the framework, Sec. 4 details our approach, and sec. 5 exhibit its experimental behavior. Finally, sec. 6 discusses the performance of the proposed approach and section 7 gives some conclusions and perspectives.

2. Related work

Mixture of Gaussian distribution is becoming more popular in the vision community. For the problem of motion recognition, Rosales [9] 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. McKenna [7] 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.

3. Description of the framework

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*. The *Video-shot segmentation* segments the sequence into temporal slices. In our system we have implemented the method of [1] which is based on the detection of the dominant motion in the successive images. The *Object acquisition* task is done by detecting and tracking moving objects through the frames of a single shot. For this purpose, dominant motion and cross-correlation are widely used. Our system implements the method of [4] which uses the dominant motion to detect and track independently moving objects; otherwise, static objects are manually selected and tracked.

Registration system for the classification. A set of different tracked objects are labeled by the user using the

registration system designed for this work. For each different object in the video, one tracked object is selected. It is called “tracked object model” in the rest of this paper. Data (features) from different tracked object models are collected and labeled. For each tracked object model, we use a Gaussian mixture density to model the intra-shot variability of its collected features. The high dimensional feature space is reduced in a statistically optimal way using the Principal Component Analysis (PCA). However, as we will see in sec. 5 and 6, this reduced feature space is still of high dimension with respect of the number of occurrences of tracked objects in short video shots (duration $\leq 1s$). 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 the next section a selection of constrained Gaussian models with constraints on the form (linear, spherical, ...), the number of estimated parameters and especially the covariance matrix.

4. Gaussian mixture for classifying 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 y_i be the feature vector of dimension d that characterizes the occurrence i . During the tracking process, y_i is variable due to all conditions listed in the introduction. Let Y to be the set of feature vectors data collected during this tracking. The distribution of Y is modeled as a joint probability density function, $f(y | Y, \theta)$ where θ is the set of parameters for the model f . We assume that f can be approximated as a J -component mixture of Gaussians: $f(y|\theta) = \sum_{j=1}^J p_j \varphi(y|\alpha)$ where the p_j 's are the mixing proportions and φ is a density function parameterized by the center and the covariance matrix, $\alpha = (\mu, \Sigma)$. In the following, we denote $\theta_j = (p_j, \mu_j, \Sigma_j)$, for $j = 1, \dots, J$ the parameters to be estimated.

As the moving object is tracked, y_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 illustrated the distribution of the 4 tracked object models 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. The covariance ellipses of these components are shown.

Parameters Estimation. Mixture density estimation is a missing data estimation problem to which the EM algorithm [2] can be applied. The type of Gaussian mixture model to be used (see next paragraph) 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), oth-

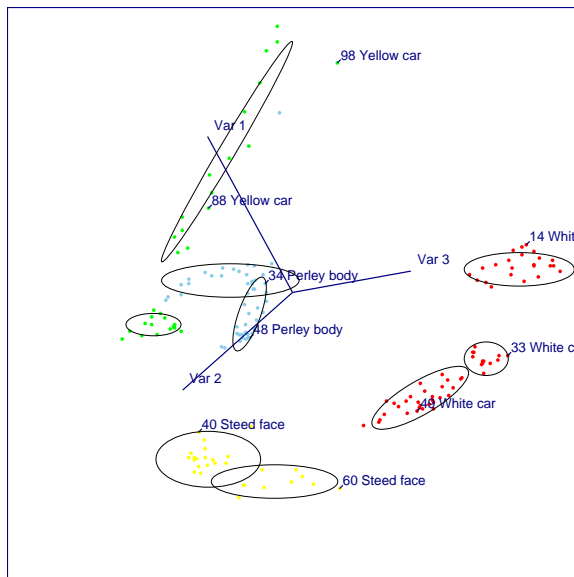


Figure 2. Modeling the variability in the RGB space of the 4 object models of figure 3

erwise 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 $\mathbf{y} = \{y_i; 1 \leq i \leq n \text{ and } y_i \in \mathbb{R}^d\}$ 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 ($i = 1, \dots, n$). We have $c_i = j$ if and only if j is the mixture component from which y_i arises. Let $\mathbf{c} = (c_1, \dots, c_n)$ denote the missing data, $\mathbf{c} \in B^n$, where $B = \{1, \dots, J\}$. The complete sample is $\mathbf{x} = (x_1, \dots, x_n)$ with $x_i = (y_i, c_i)$. The complete log-likelihood is $L(\theta, \mathbf{x}) = \sum_{i=1}^n \log \left\{ \sum_{j=1}^J p_j \varphi(x_i | \mu_j, \Sigma_j) \right\}$. More details on the EM algorithm could be found in [2]. 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.

Gaussian models. Gaussian mixtures are sufficiently general to model arbitrarily complex, non-linear distribution accurately given enough data [7]. When the data is limited, the method should be constrained to provide better conditioning for the estimation. The various possible constraints on the covariance parameters of a Gaussian mixture (e.g. all classes have the same covariance matrix, an identity covariance matrix, ..), defines 14 models. We have implemented the following seven models derived from the three general families of covariance forms: $M_1 = \sigma_j^2 I$ and $M_7 = \sigma^2 I$ the simplest model from the spherical family (I is the identity matrix); $M_2 = \sigma_j^2 \text{Diag}(a_1, \dots, a_d)$

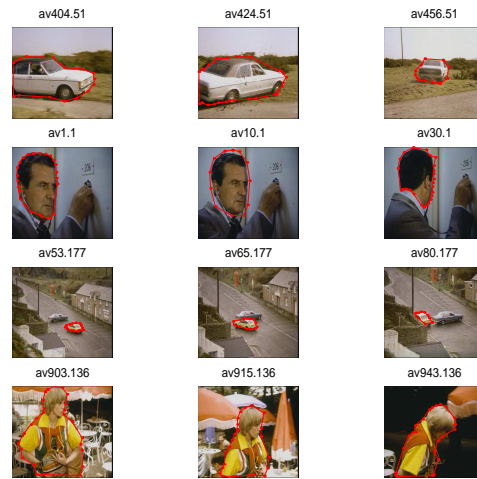


Figure 3. Subset of tracked object models

and $M_3 = \sigma_j^2 \text{Diag}(a_1^j, \dots, a_d^j)$ from the diagonal family where $|\text{Diag}(a_1^j, \dots, a_d^j)| = 1$ with unknown a_1^j, \dots, a_d^j ; M_4 from the general family which assumes that all components have the same orientation and identical ellipsoidal shapes; M_6 from the general family which assumes that all covariance matrices have the same volume; Finally, M_5 is the most complex model with no restrictions. See [2] for more details about their complexity and their maximization step of the EM algorithm.

Model choice criterion. To avoid a hand-picked number of modes of the Gaussian mixture, the Bayes Information Criterion (BIC) [10] is used to determine the best probability density representation (appropriate Gaussian model and number of components). It is an approach based on a measure that determines the best balance between the number of parameters used and the performance achieved in classification. It minimizes the following criterion: $BIC(M) = -2L_M + Q_M \ln(n)$ where L_M is the maximized log-likelihood of the model M and Q_M is its number of free parameters.

Probabilistic recognition process. 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 only a global Gaussian mixture of \mathbf{W} components, where $\mathbf{W} = \sum_{l=1}^L J_l$ and J_l is the number of components of the l th Gaussian mixture. For that, the mean and covariance of each component of the global Gaussian mixture being build are held fixed and only the proportion parameters are recomputed. Using the maximum a posteriori probability we classify novel occurrences in the video.

5. Experiments

Object models. Experiments have been performed on a decompressed MPEG sequence of 1016 frames, extracted

from the *Avengers* movie (INA). Using our system it was segmented into 1370 object occurrences corresponding to 51 tracked objects. 15 different tracked objects were selected (randomly) and labeled in shots using the registration system designed for this work. Thus, 480 different views of them were collected to estimate parameters of the Gaussian mixture density. Figure 3 displays 3 different occurrences of each one of the 4 learned tracked object models.

Features. The approach is evaluated using gray and color global features resumed by histograms. The histogram approach is well known as an attractive method for object recognition because of its simplicity, speed and robustness. The RGB space is quantized into 64 colors, and the gray-scale one into 16 intensities. Then, the PCA was applied on the entire set of initial data of the each feature space in order to reduce their dimensionality (d_E in table 1). This is an important step to overcome the curse of dimensionality where the number of samples of an object model is not in general sufficient to fit an optimal Gaussian mixture model.



Figure 4. Subset of tracked object tests

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 occlusions or other unpredictable events. For such case, a robust decision is implemented through a majority decision rule: the final class is the class for which most of each individual belongs to; a threshold of 50% consistent answers is required, otherwise the classification process rejects the query. The query set consists of 1370 individual queries which corresponding to 51 different sequences of tracked objects.

Results. To measure the performance of methods we compute the percentage of total correctly classified

Meth.	feat. hist.	d_E	MaxNbC or dist.	Total	
				indi. %	trac. %
mix.	gray	5	1	45.80	45.69
mix.	gray	5	2	52.10	51.00
mix.	gray	5	3	49.42	45.47
mix.	gray	5	4	56.57	54.53
key	gray	16	χ^2	31.12	32.10
key	gray	5	d_e	32.95	33.72
mean	gray	16	χ^2	30.09	31.56
mean	gray	5	d_e	32.00	33.72
mix.	RGB	10	1	73.65	82.99
mix.	RGB	10	2	79.10	86.10
mix.	RGB	10	3	81.50	86.30
mix.	RGB	10	4	71.09	73.87
key	RGB	64	χ^2	45.16	45.91
key	RGB	10	d_e	42.90	44.50
mean	RGB	64	χ^2	43.09	45.00
mean	RGB	10	d_e	47.90	45.50

Table 1. Test results with Gaussian mixture (mix.), Key-frame and mean-histogram methods.

queries: individual occurrences (*indi.%*) and tracked objects (*trac.%*). The maximum number of permitted Gaussian components, *MaxNbC*, was ranged from 1 to 4. The BIC is used to chose the appropriate number of mixture components and the best fitting Gaussian model (among seven models). Table 1 shows the test results of the Gaussian mixture method (mix.) and the two others compared methods (next paragraph). Column 1 indicates the method used and column 2 indicates the feature type. For the mixture method column 4 indicates the maximal number of components; otherwise, it indicates the type of the metric distance computed.

Comparative analysis. Related work in video indexing represents each video shot by a key-frame [8]. Generally, the key-frame is chosen to be the middle frame of the shot. Only localized objects in the key-frames are indexed, and a median database is generated for the 15 object classes for each kind of feature being tested. To match request object occurrences to indexed tracked object classes in the median database, both the χ^2 -test in the real feature space and the Eucliden metric, d_e , in the PCA space were considered. On the other hand, we can compute the *mean* histogram (μ) of each tracked object model.

Comparing the results shown in table 1 of the three methods, we see that the *indi.%* and *trac.%* of correctly answers are increased considerably, by the using of Gaussian mixture method, rather than the two key-frame methods, by at least 40% and 20% for color and gray histograms respec-

tively. However, the two key-frame methods are simple and efficient when the object is static but the most of our experimental video objects are moving and deforming.

6. Discussion

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 disperse. This is the consequence of the degree of variability of each tracked object within its shot. As an evidence, the two key-frame methods will give poor results. This was confirmed by the experimental results given above. The variability modeling approach proposed in this paper performs best.

On the other hand, the results obtained by the Gaussian mixture approach indicate that the distributions are not unimodal in general. A reliable estimate can be obtained by the BIC criterion. Using BIC, we found that the maximum number of modes that best represents the underlying distributions was 3 for the RGB histograms and 4 for the Gray histograms, 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 equal to 4. The only 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. This is strongly related to the follow: for the RGB histogram, the Gaussian model, M_5 (most general), was only chosen twice but the model M_4 , which assumes that all mixture components have the same orientation and identical ellipsoidal shapes, was chosen 13 times. Compared to the gray-scale histogram, M_5 was selected 18 times and M_4 only 2 times. Again, the number of data is too limited with respect to the RGB space dimension ($d_E = 10$).

Representation and performance. In computer vision, the recognition rate is limited by the similarities in the class descriptions given by the feature vector. So using invariant color features [5] will hopefully decrease the misclassifying of the Gaussian mixture. Whatever, methods will be used this will not overcome tracking errors which occurs very often in such a complex case, for instance shadow being a significant part of the tracked yellow car, 30, in image 260 (av260.30) of figure 4.

7. Conclusion and perspectives

This paper presents a methodology for increasing robustness of existing features used in video object recognition. The tracked object is represented by a multi-modal probability distribution, rather than by a simple point (key-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 key-frame approach used in video indexing. The use of tracked objects yielded better recognition performance than the use of single occurrence images. Tracked object tests were classified by a majority vote. One direct extension of this consists in estimating their Gaussian mixtures. Then, to compute a metric distance (*Kullback-Leibler* for instance) between their components and those of learned object models. Finally, a future research direction we intend to explore is to completely automate the classification process of video objects, so that 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 in different ways (Gaussian models).

Acknowledgments. We would like to acknowledge *Alcatel CRC* for its support of this work and *C. Biernacki* for his helpful comments on Gaussian mixture models.

References

- [1] 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.
- [2] G. Celeux and G. Govaert. Gaussian Parsimonious Models. *Pattern Recognition*, 28(5):781–783, 1995.
- [3] D. Clemens and D. 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.
- [4] M. Gelgon and P. Bouthemy. A region-level graph labeling approach to motion-based segmentation. In *CVPR*, pages 514–519, Puerto Rico, June 17-19 1997.
- [5] T. Gevers and A. 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.
- [6] 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.
- [7] S. J. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern recognition*, 31(12):1883–1892, 1998.
- [8] B. 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.
- [9] 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.
- [10] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.