ENST/UOB/LU@TRECVID2007
HIGH LEVEL FEATURE EXTRACTION USING 2-LEVEL PIECEWISE GMM

George Yazbek[1,2], Georges Kfoury[3], Gabriel El Alam[3], Chafic Mokbel[2], Gérard Chollet[1]

[1]École Nationale Supérieure des Télécommunications, [2]University of Balamand, [3]Lebanese University

ABSTRACT

We describe a high level feature extraction system for video. Video sequences are modeled using Gaussian Mixture Models. We have used those models in the past to segment video sequences into 2D+time objects. The segmentation result has been used with great success in a compression scheme. In the present work, the Gaussian components of the model are considered to completely model the corresponding objects in the video. Their parameters are used as low-level features for a high-level model used for the detection of a topic or high level feature. The system is not optimized for a particular feature and is thus scalable to any number of features. A threshold is manually selected for each feature after normalization.

The only difference between runs was normalization. We tested two runs: B_ENST_1 uses znorm per topic per video. B_ENST_2 use znorm per video. The second system provided better results.

This is an initial system that will be used to explore the effectiveness of the modeling of videos using GMMs.

1. INTRODUCTION

Several systems have been proposed in the past for the extraction of high level features from video sequences. Stochastic models are often proposed for this purpose. The system described in the present paper is also based on stochastic modeling. However, those models are applied in two levels. The basic idea is to first segment a video sequence into coherent 2D+time objects, then to use these objects to extract the high level features. Therefore, at a lower level a stochastic model, i.e. an extended piecewise Gaussian Mixture Model (GMM) [3][1], is used to perform this segmentation. The characteristics of the detected objects are used in order to extract, in another GMM model, the high-level features.

The paper is organized as follows. The next Section describes the system used for video segmentation, leading to the extraction of low level features. The Section 3 describes the high level model, i.e. the stochastic GMM model to detect the topics in the segmented video. Section 4 presents the annotation of the database used in our experiments. Once the models estimated the scoring for a new video and the corresponding normalization techniques are detailed in the Section 5. The experimental results are provided in the Section 6. Finally, the section 7 presents the conclusions and some perspectives for this work.

2. LOW LEVEL FEATURE EXTRACTION

As stated in the introduction, the extraction of low level features from raw video is done through the segmentation of the video into 2D+time objects. The detected objects are then characterized with a small set of parameters. In this work, low level feature extraction and segmentation is done in a single step. We have used our previous work described in [1] where a hierarchical piecewise Gaussian Mixture Model (GMM) is proposed to segment a video sequence into 2D+time objects. For an observable process $\underline{X}$ in a space of dimension p, a GMM λ with K Gaussian components is defined as the following probability density function:

$$p_\lambda(\underline{x}) = \sum_{k=1}^{K} w_k N_k(\underline{x}, \underline{\mu}_k, \underline{\underline{\Gamma}}_k)$$

where,

$$N_k(\underline{x}, \underline{\mu}_k, \underline{\underline{\Gamma}}_k) = (2\pi)^{-p/2} \left\| \underline{\underline{\Gamma}}_k \right\|^{-1/2} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_k)^{\#}\underline{\underline{\Gamma}}_k^{-1}(\underline{x}-\underline{\mu}_k)} \quad \text{(Eq. 1)}$$

The pixels in the video are considered to be the realization of such a process. Each pixel is represented by its geometrical coordinates (x, y), its time instant (t) and its colors (in YUV space in this work). This leads to a total of 6 parameters. A GMM is trained to estimate the distribution of those pixels. Afterwards, pixels are attributed, following the Maximum Likelihood criterion, to the Gaussian components of the GMM. Therefore, a Gaussian component of the GMM models a set of pixels in different video frames; since time is being including in the pixels' representation. In definitive a Gaussian component of the estimated GMM represents a 2D+time object. In [1] we have shown that this segmentation is very effective and efficient compression algorithms may be defined.

The GMM modeling of the video pixels provides more than the segmentation of these pixels into 2D+time objects. It offers models for those detected objects. Each Gaussian component represents a region. One can build a feature vector containing some of the Gaussian parameters. The following nine parameters have been selected from the GMM Gaussian component in order to represent the region:

$$\underline{v} = \left(\sigma_x, \sigma_y, \sigma_t, \mu_Y, \mu_U, \mu_V, \sigma_Y, \sigma_U, \sigma_V\right)$$

The space and time means have not been considered since the objects may be shifted in time and space between different videos. Therefore, only standard deviations of space and time parameters are taken into account. While for a specific object, the means and standard deviations of the pixels colors (here in YUV space) completely characterize the object.

In definitive, the low-level features for a video sequence are obtained after training a piecewise GMM on a sequence of frames leading to K objects, each object being represented by the standard deviations of its pixels positions and times and by its pixels colors means and standard deviations.

3. HIGH LEVEL MODEL

After the definition of the low level features, a high level model is to be determined in order to detect the different topics in a video sequence. In our system, a GMM is also used for this purpose. The low level features are modeled by a high level GMM with G Gaussian components. A video sequence at the entry of the system is first segmented and low-level features are extracted using the low-level GMM. A set of K feature vectors represent the video sequence at the output of the low level processing. These vectors are modeled by the high level GMMs; a GMM being defined for every topic of interest in the application.

In addition to creating topic-based or feature-based GMMs, a universal model λ_u is also created per topic. It contains all the Gaussian components corresponding to the all the frames independently of their topic/feature content. This universal model is used as a reference for log-likelihood ratio computation and comparisons. Actually, the log-likelihood ratio is considered as the basis for statistical decision in the topic detection problem. If λ_t is the high level GMM for the topic t and λ_u the universal high level GMM corresponding to this topic, the decision score for a low level feature vector $\underline{v}$ is computed as the log-likelihood ratio and is compared to a threshold γ_t:

$$score = \log \frac{p_{\lambda_t}(\underline{v})}{p_{\lambda_u}(\underline{v})} \underset{>}{<} \gamma_t \qquad \text{(Eq. 2)}$$

The estimation of the GMM parameters is performed using the Estimation-Maximization (EM) algorithm [4]. This algorithm allows Maximum Likelihood estimation based on some training data. The selection of the training data is crucial. The training of the low level GMM is straightforward. Actually, a new GMM is trained for every video sequence to allow sequence specific segmentation. In comparison, the training of the high level GMMs is not specific to a video sequence but to a topic. For this purpose, a training set must be available where all the topics are present. The low-level features extracted for every topic sequence may be used to train the topic-specific GMM models. However, not all the objects of the video containing a topic are relevant to this topic. Therefore, a special care must be taken in order to provide the training algorithm with relevant training data. The selection of the training data is described in the next Section.

4. ANNOTATION

In order to cope with the object specific data for topic/feature based GMM training, we have used two annotation methods, a frame-base annotation and a region based annotation.

In the frame-based annotation, a few key frames from each shot are selected and manually annotated using a web-based tool. One such system is the TRECVID collaborative annotation system based on active learning [2]. The basic premise of active learning is that by having an iterative process where "interesting" frames are annotated and the system refined to select the next most probable frames.

In region-based annotation, the results of low-level feature extraction results are manually annotated. This is a labor-intensive approach that cannot be reused by other systems. The annotation results however point to the video objects that correspond to the particular feature.

Once annotated the video frames are used for both training and evaluation of the system. The frame-based annotation provides low level feature vectors corresponding to the key frame. These vectors are used to train the high level topic GMMs if the corresponding topic is indicated to exist in the key-frame. For region-based annotation, we add the low level feature vectors corresponding to the region if the region is labeled with the topic of interest. As the region-based components are much more accurate than the frame-based components, the corresponding feature vectors are considered in the training process with much higher weight.

Finally, all the feature vectors in the training set are used to estimate the parameters of the Universal GMM.

5. SCORING AND NORMALIZATION

The log-likelihood ratio (Eq. 2) between the high level topic GMM and the high level universal GMM is considered as the score. While it is possible for the training to select the region-specific feature vector for training, this is not possible in the detection phase because we do not know the region corresponding to the topic. All the low-level feature

for a test video sequence are passed to the detection GMMs. Therefore, variation of the scores exists between the different video for the same topic.

To overcome the existing variations in the scores, we propose the use of normalization techniques. Two normalization approaches have been experimented in this work. The idea is to normalize the scores in order to ensure that the distribution of the scores has zero average and one for standard deviation. For a given video sequence, if μ and σ are the average and standard deviation of the scores, the normalized score s^n may be obtained from the original score s using the following formula:

$$s^n = \frac{s - \mu}{\sigma} \qquad \text{(Eq. 3)}$$

Two variants are considered for the calculation of the scores first and second order statistics. The first variant considers only the scores obtained for the topic of interest in the different sequences of the same video in order to calculate μ_t and σ_t. This normalization is called the Z-norm per topic and per video. The second variant uses the scores obtained for all topics for a given video in order to compute the μ_v and σ_v. This variant is called the Z-norm per video.

6. RESULTS

The two variations of normalization were tested with similar results in terms of mean average precision (0.006). This is an initial system to test the efficiency of the representation of GMM modeling to video retrieval.

7. CONCLUSIONS

In this paper a new topic or high level feature detection system has been proposed. This system is based on two-level GMM stochastic modeling. At a low level, a piecewise GMM allows the segmentation of a video sequence into a set of 2D+time objects, each object being associated with a Gaussian component of the GMM. The Gaussian component parameters are used to calculate a vector of feature parameters that better represents, at low level, the associated object. Those low level feature vectors are then passed to a high level GMM used for the classification and detection of topics and high level features. For each high level feature a classification GMM is computed. A Universal GMM is also estimated. The log-likelihood ratio, between the topic GMM and the universal GMM, permits to detect the presence of a high level feature or topic in a video sequence.

Since the log-likelihood scores vary largely between the different videos, two normalization techniques have been proposed: a normalization per topic per video and a normalization per video.

Several perspectives exist to this work. One would like to estimate to only consider the best score among all the feature vectors obtained for a sequence of video frames. Actually, this would correspond to a 2D+time object that generally indicates the presence of the high level feature or topic in the video. Another point is to consider the correlation that may exist between different objects in order to detect the presence of a topic in the video.

REFERENCES

[1] G. Yazbek, C. Mokbel, and G. Chollet, "Video Segmentation and Compression using Hierarchies of Gaussian Mixture Models," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, USA, pp. I-1009-I-1012, April 15-20 2007.

[2] S. Ayache and G. Quénot, "Evaluation of active learning strategies for video indexing", in *Fifth International Workshop Content-Based Multimedia Indexing (CBMI'07)*, Bordeaux, France, June 25-27, 2007.

[3] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic Space-Time Video Modeling via Piecewise GMM", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 384-396, 2004

[4] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data Using the EM Algorithm", *Journal of the Royal Statistical Society, Series B,* vol. 39(1), pp. 1-38, 1977

[5] http://tsi.enst.fr/becars