A Dynamic Probabilistic Multimedia Retrieval Model

Tzvetanka I. Ianeva
Departamento de Informática,
Universidad de Valencia, Valencia, Spain
tzveta.ianeva@uv.es

Arjen P. de Vries, Thijs Westerveld
Centrum voor Wiskunde en Informatica,
Amsterdam, The Netherlands
arjen@acm.org, thijs.westerveld@cwi.nl

***Abstract*— We describe the application of a probabilistic multimedia model to video retrieval. From video shots, we compute Gaussian-mixture models that capture correlations in time and space, such as the appearance and disappearance of objects. These models improve the precision of "query by example/s" results in the TRECVID 2003 collection when compared to models that only make use of static visual information. Furthermore, integrated with information from automatic speech recognition (ASR) transcripts, they outperform ASR only results.**

I. Introduction

Video search systems can be compared by the open metrics based evaluation process at TRECVID 2003 funded by ARDA and NIST. Formatted descriptions of information needs (topics) are given to all participants and their video search systems are asked to return a list of up to 1000 shots which meet the need, based on likelihood of relevance. The video shot is considered as a basic unit of video data. In current video retrieval systems, there are two video representation schemes used for retrieval, which we classify as *static* and *spatio-temporal*. A wide range of video retrieval models is based on the static approach by representing the shot by its keyframe [1], [2], so information about the temporal evolution of the video is lost. But video is a temporal media, so a 'good' model will be one that solves the limitation of keyframe-based shot representation by exploiting the spatio-temporal information of the video. Spatio-temporal models attempt to extract backgrounds and independent objects in the dynamic scenes captured in the sequences. Spatio-temporal grouping techniques can be classified into two categories: (1) segmentation with spatial priority and tracking of regions from frame to frame, and (2) those based on joint spatial and temporal segmentation. Our approach belongs to the second category in which video is considered as a spatio-temporal block of pixels, by treating the spatial and temporal dimensions simultaneously. A good argument to follow this approach is that human vision finds salient structures jointly in space and time, as described by Gepshtein and Kubovy in [3].

A. Motivation

In [4], we demonstrated the limitations of our static model: the models represent only keyframes rather than shots, and this can hurt retrieval performance. Consequently, we built a new *dynamic* retrieval model which we used for TRECVID2003. Section II describes the dynamic retrieval model and presents our two hypotheses. Section III-C reports on experimental results. Conclusions and directions for future work are described in the final section.

B. Related work

DeMenton proposed spatio-temporal grouping in a vector space, using similarity clustering [5]. He uses seven-dimensional feature vectors composed of three color and four motion descriptors. Pixels with similar color and motion are close in feature space and grouped by hierarchical mean shift analysis over increasing large regions.

Spatio-temporal graph-based image segmentation methods are based on a graph whose nodes are the image features grouped using graph cut techniques. The edges connecting pixels in spatial as well temporal directions are weighted according to some measure of similarity (affinity) between nodes. Similarity based on motion profiles is used by Shi and Malik [6]. A motion profile represents the probability distribution of the motion vector at a given point. Fowleks et al. [7] define similarity based on spatio-temporal location (x, y, t), color (L, a, b), and the optical flow feature vector x_i attached to each pixel i of the sequence. To compute segmentations more efficiently they use the Nyström approximation of the normalized cut algorithm. Greenspan et al. [8] explore spatio-temporal grouping by fitting a mixture model for linear motion detection by using the Gaussian covariance coefficients between spatial and temporal dimensions. In [9], they extend the spatio-temporal video-representation scheme to a piecewise GMM framework in which a succession of GMMs are extracted for the video sequence, for the description of non-linear motion patterns.

Our approach is most similar to [8]. We also build mixture models to describe the shots in the collection. Our contribution is that for first time we applied these techniques in practice and show they are useful for video retrieval from a large generic heterogeneous collection.

II. Building Dynamic GMMs

A. Gaussian Mixture Models

The Dynamic retrieval model we use to rank video shots is a generative model built as an extension and optimization of the static retrieval model [1], inspired by the probabilistic approach to image retrieval [10] and the language modeling approach to information retrieval [11]. We present concisely the visual part of the dynamic model and the consideration that we take into account. Details regarding the other parts can be found in [1].

The visual part of the dynamic model ranks video shots by their probability of generating the samples (pixel blocks) in query example(s). The model is smoothed using background probabilities, calculated by marginalization over the collection. So, a collection video shot ω_i is compared to an example document (video shot or image) $\boldsymbol{x}$ by computing its retrieval status value (RSV), defined as:

$$\text{RSV}(\omega_i) = \frac{1}{N} \sum_{j=1}^{N} \log\left[\kappa P(\boldsymbol{x_j}|\omega_i) + (1-\kappa) P(\boldsymbol{x_j})\right], \quad (1)$$

where κ is a mixing parameter. The example image consists of N samples ($\mathcal{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N})$), i.e., the example video shot consist of $29 \times N$ samples. Collection documents ω_i are modelled as mixtures of Gaussians with a fixed number of components:

$$P(\boldsymbol{x}|\omega_i) = \sum_{c=1}^{N_C} P(C_{i,c})\ \mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}_{i,c}, \boldsymbol{\Sigma}_{i,c}), \quad (2)$$

where N_C is the number of components in the mixture model, $C_{i,c}$ is component c of class model ω_i and $\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})},$$

where n is the dimensionality of the feature space and $(\boldsymbol{x} - \boldsymbol{\mu})^T$ is the matrix transpose of $(\boldsymbol{x} - \boldsymbol{\mu})$.

B. Selecting frames

Our approach to extending the static Gaussian-mixture model to a dynamic model has been to retain the computation of feature vectors for 2D images, augmented with a new temporal feature. Shots consist of a large number of frames (in our collection, shot sizes range from 24 to 66629 frames), so it was not feasible to consider all frames in a shot.

In selecting a sub-sequence of frames, we considered the following two approaches: (1) modeling the video shot as one-second video sequence around the keyframe or (2) modeling the entire video shot as sequence of frames sampled at regular intervals. The first one captures even small changes in time around the keyframe, but is very sensitive to the choice of the keyframe in long-duration shots. In the second approach, the dependency on the keyframe is completely removed, we will have information of what is going on at the beginning and at the end even in very long shots. On the other hand, we lose precision if the chosen frames are too diverse. Another disadvantage is that the computation cost scales with the length of the shot.

We determined experimentally the retrieval performance of the two approaches on the TRECVID2003 search development collection. We always obtained better results when applying the first approach. As a consequence, we model a one-second video sequence around the keyframe (29 frames) as a single entity. Next, we explain the retrieval process.

C. Features

The samples are 8×8 blocks described by discrete cosine transform (DCT) coefficients and spatio-temporal position. In a given shot, like the one in Figure 1), we obtain a set of feature vectors by the following process, as outlined in Figure 2: each frame of a shot is decomposed into blocks of size 8×8 pixels; for each of those blocks, we compute a 15-dimensional feature vector: the Y, Cb, Cr color channels of each block are transformed by the DCT. The first 10 DCT coefficients from the Y channel, the first coefficient from the Cb channel, and the first coefficient from the Cr channel make up the first 12 features. Three more features for each block describe the x and y coordinates of the block in the frame, and temporal position of the frame inside the shot; the time is normalized between 0 and 1. Using the EM algorithm, we compute the parameters of the GMM ($\boldsymbol{\mu}, \boldsymbol{\Sigma}$, and $P(C)$) for this set of feature vectors.

To handle the case when we have as query example a single image, i.e., we do not have the time dimension, we tried again two alternative approaches: (1) We artificially made a sequence of 29 images same as the query example where the time is normalized between 0 and 1. The intuitive explanation in this case is that we observe the same image all the time. (2) We extend the query example image's feature vector with a fixed temporal feature value of 0.5. The intuitive explanation in this case is that we do not know what happens before and after the central moment, when matching the keyframe. Experiments on the TRECVID2003 training data set gave slightly better results using the second approach, so the second approach has been implemented. It has the additional advantage of lower computational cost.

D. Motion

To reduce the complexity of our model, the Gaussians are axis-aligned, i.e., their covariance matrix ($\boldsymbol{\Sigma}$) is diagonal. Therefore, we can decompose a 15-dimensional Gaussian into 15 *independent* 1-dimensional models for each component. In particular, this means that a single Gaussian in our 15-dimensional model cannot capture correlations, e.g., between time and space. However, with multiple Gaussians, motion does give rise to different models under certain circumstances: a moving object is one that is visible in location x_1, y_1 at time t_1 and visible at x_1, y_2 at time t_2 (with $t_1 < t_2$). With two Gaussians with means (x_1, y_1, t_1) and (x_2, y_2, t_2), respectively, we can distinguish linear motion in one direction from linear motion in the opposite direction. This is impossible with static models. On the other hand, many shots without linear motion, e.g., with "jumping" objects, also fit such GMMs, so arguably we do not capture motion in an intrinsic way.

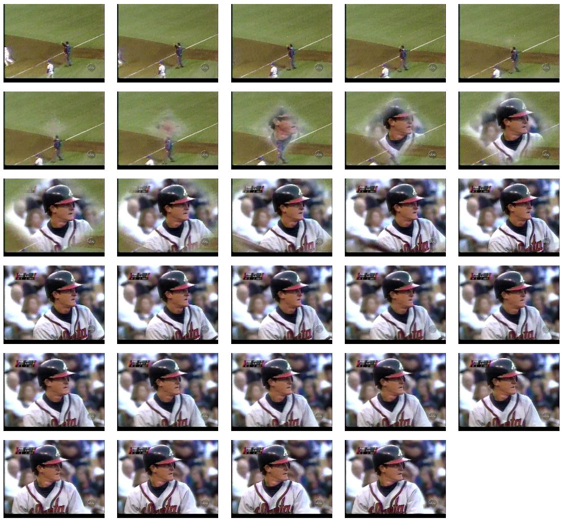
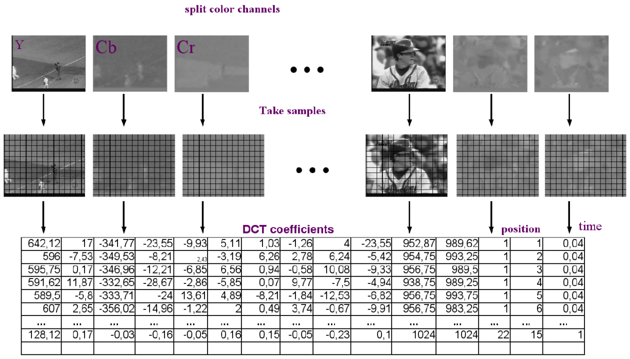


Fig. 1. A shot represented by 29 frames around the keyframe



642,12	17	-341,77	-23,55	-9,93	5,11	1,03	-1,26	4	-23,55	952,87	989,62	1	1	0,04
596	-7,53	-349,53	-8,21	2,43	-3,19	6,26	2,78	6,24	-5,42	954,75	993,25	1	2	0,04
595,75	0,17	-346,96	-12,21	-6,85	6,56	0,94	-0,58	10,08	-9,33	956,75	989,5	1	3	0,04
591,62	11,87	-332,65	-28,67	-2,86	-5,85	0,07	9,77	-7,5	-4,94	938,75	989,25	1	4	0,04
589,5	-5,8	-333,71	-24	13,61	4,89	-8,21	-1,84	-12,53	-6,82	956,75	993,75	1	5	0,04
607	2,65	-356,02	-14,96	-1,22	2	0,49	3,74	-0,67	-9,91	956,75	983,25	1	6	0,04
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
128,12	0,17	-0,03	-0,16	-0,05	0,16	0,15	-0,05	-0,23	0,1	1024	1024	22	15	1

Fig. 2. The process of obtaining feature vectors.

In any case, spatio-temporal correlations are captured in the 3D GMMs; we capture the appearance and disappearance of objects. An example of resulting Gaussians for the shot given in Figure 1 is given in Figure 3;[1] in the visualization, the 15-dimensional Gaussians are reduced to 3 spatial dimensions (for the location and time), and their mean color and texture are visualized. It is possible to correlate Gaussians with objects in the shot, which appear and disappear at given times: e.g., the grass is only visible at the beginning of the sequence and the corresponding component, the green dynamic space-time blob, also disappears at about $t = 0.4$. In other words, the GMM captures the disappearing of the grass. This effect cannot be captured in the static model where time is not taken into account.

III. Retrieval using Dynamic GMMs

A. The TRECVID 2003 collection

The TRECVID2003 video search collection contained 113 broadcast news videos, from which 32318 shots have been extracted. TRECVID 2003 provides 25 topics described by video shots and image examples. Given the topics the different video search systems were asked to return a list of up to 1000 shots based on likelihood of relevance with the topic. Returned shots are pooled, judged for relevance, and, performance of a search run is measured by recall, precision, and mean average precision. As part of the Lowlands Group, we participated in the TRECVID 2003 search task with runs based on the dynamic models and one static run for comparison purposes.

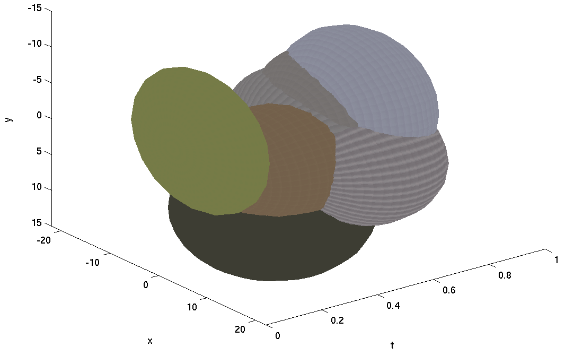


Fig. 3. Visualization of the Gaussian model for the shot depicted in Figure 1

B. Experimental setup

The video sequences in the TRECVID search collection are already segmented into shots. For each shot, we build a static model, a dynamic model (see Section II) and a language model. Details about the language models can be found in [12]

For the TRECVID search task, topic descriptions provide images and shots as query examples. The selection of images and videos to be used as query examples for ranking is the only manual action. From there on, the retrieval process is fully automatic; we compute the retrieval status value for each document model using equation 1 from Subsection II-A. To combine models with ASR we (unrealistically) assume textual and visual information are independent. Under this assumption we easily compute the joint probability of observing query text and visual example by the sum of the individual log probabilities [13].

C. Experimental Results

Table I shows the results for static models, dynamic models and language models (based on ASR transcripts provided by LIMSI [14]). We report here mean average precision (MAP) and precision at 10 (P@10). Although static and dynamic models have the same MAP, dynamic models appear to be more useful because they have higher initial precision (see Table I). An example of a dynamic query with higher initial precision than the static variant is shown in Figure 4, where we search for anchor persons.

Initial precision is important, not only because users are often interested in finding just a few good shots, but also

[1] A version of this paper with color figures is available from `http://www.cwi.nl/projects/trecvid/trecvid2003.html`

TABLE I

PERFORMANCE OF THE STATIC/DYNAMIC MODELS

Description	MAP	P@10
ASR only	.130	.268
static only	.022	.076
static + ASR	.105	.220
dynamic only	.022	.096
dynamic + ASR	.132	.272

Mean Average Precision (MAP) and precision at 10 (P@10) for ASR only, visual only and combined runs (static and dynamic models).

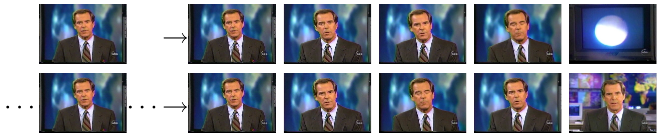


Fig. 4. Top 5 results for static and dynamic runs; only keyframes are shown.

because it allows for a useful combination with ASR results. In previous experiments [12], we found a combination of textual and visual results is useful if both modalities have useful results. For the dynamic models run, this seems to be the case. Apart from the fact that the dynamic model outperforms the static model in a visual-only setting, perhaps even more important is the fact that it can improve on the ASR-only results (see Table I). For most topics, the ASR run is better than the visual run, but for some topics, this is reversed. Apparently, the combined run is able to automatically select the best modality per topic without letting the inferior modality disturbing the results too much. This lifts the burden of selecting the appropriate modality from the user, and ensures robustness against choosing the wrong modality.

IV. CONCLUSION

In general, from a larger number of frames much more solid evidence about the visual content of the shot can be accumulated than from a single keyframe. Another advantage of spatio-temporal modeling is the reduced dependency on choosing an appropriate keyframe: sometimes the keyframe is not well chosen, e.g., it is completely black, but the immediate vicinity of the keyframe still contains valuable information in the one-second sequence around it.

The resulting models capture the appearance and disappearance of objects and simple linear motion, but cannot describe complicated non-linear temporal events like movement of water. The extracted space-time blobs are not appropriate for representing non-convex regions in the 3D space-time domain due their Gaussian nature. In our current extension of the model we are looking for an approximation of non-linear motion appropriate for video retrieval. A balance between the computational processing cost and the complexity of the captured motion has to be found to be able to handle generic and large search collections as provided by TRECVID. According the official TRECVID2003 results, the main advantage of our dynamic model has been that the combination of dynamic shot models with language models of the ASR transcripts outperforms both individual runs. The combination with the static models performed however worse than the ASR only run. Thus, we conclude that the dynamic models are useful for the integration of different modalities. Currently, we are investigating the integration of audio in the model.

A number of open questions remain to be addressed by future research:

(1) We already optimized the process of building GMMs for a shot in three ways: by selecting a subset of the frames in the shot, by operating on 8×8 blocks of pixels, and by only considering Gaussians that are axis-aligned. However, we are looking for further ways to speed up the process of generating GMMs without sacrificing precision.

(2) We found out that choosing a dense interval of frames around the keyframe gives better results than a sparse regular sample of frames from the entire shot. This suggests that modeling spatio-temporal correlations is more important than to model all objects in a shot. Therefore extending our model with better modeling of motion may be a way to improve precision significantly. How this can be done within reasonable constraints on computing time remains to be seen. One possibility would be to consider Gaussians that are axis-aligned except for the spatial and temporal axes.

ACKNOWLEDGMENT

The first author was supported by grants of Generalitat Valenciana FPIB01_362 and CTESRR/2003/43.

REFERENCES

[1] T. Westerveld, A. P. de Vries, A. van Ballegooij, F. M. G. de Jong, and D. Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processing*, 2003.

[2] W. H. Adams *et al.*, "IBM research TREC-2002 video retrieval system," in *Proc. Text Retrieval Conference (TREC)*, 2002.

[3] S. Gepshtein and M. Kubovy, "The emergence of visual objects in space-time," in *Proceedings of the National academy of Science*, vol. 97, no. 14, USA, 2000, pp. 8186–8191.

[4] T. Westerveld and A. P. de Vries, "Experimental result analysis for a generative probabilistic image retrieval model," in *SIGIR 2003*, Toronto, Canada, 2003.

[5] D. DeMenton, "Spatio-temporal segmentation of video by hierarchical mean shift analysis," in *Statistical Methods in Video Processing Workshop*, 2002.

[6] J. Shi and J. Malik, "Motion segmentation and tracking using normalized cuts," in *IEEE Conference on Computer Vision*, 2002, pp. 1151–1160.

[7] C. Fowlkes, S.Belongie, and J. Malik, "Efficient spatiotemporal grouping using the nystrom method," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 231–238.

[8] H. Greenspan, J. Goldberger, and A. Mayer, "A probabilistic framework for spatio-temporal video representation and indexing," in *European Conference on Computer Vision*, vol. 4, 2002, pp. 461–475.

[9] ——, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, 2004.

[10] N. Vasconcelos, "Bayesian models for visual information retrieval," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[11] D. Hiemstra, "Using language models for information retrieval," Ph.D. dissertation, Centre for Telematics and Information Technology,University of Twente, 2001.

[12] T. Westerveld, A. P. de Vries, and A. Ballegooij, "CWI at TREC-2002 video track," in *The Eleventh Text Retrieval Conference (TREC-2002)*, 2003.

[13] T. Westerveld, T. Ianeva, L. Boldareva, A. de Vries, and D. Hiemstra, "Combining information sources for video retrieval," in *TRECVID 2003 Workshop*, 2004.

[14] J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1–2, pp. 89–108, 2002.