FAST VIDEO SHOT RETRIEVAL BY TRACE GEOMETRY MATCHING IN PRINCIPAL COMPONENT SPACE

[+*]*Zhu Li*, [*]*Aggelos K. Katsaggelos, and* [+]*Bhavan Gandhi*

[+]Multimedia Communication Research Lab (MCRL), Motorola Labs, Schaumburg
[*]Department of Electrical & Computer Engineering, Northwestern University, Evanston

ABSTRACT

Content-based video retrieval technology holds the key to the efficient management and sharing of video content from different sources, in different scales, across different platforms, and shared over different communication channels. In this work we present a fast retrieval algorithm based on matching the geometry of video sequence traces in the principal component space. Techniques to address scale (spatial and temporal) issues, as well as, noise and other possible distortions, such as frame dropping, are discussed. Experimental results demonstrate the effectiveness of the proposed approach.

1. INTRODUCTION

With the proliferation of digital video capturing, storage and communication devices, the amount of information in video form is growing rapidly in personal entertainment, security, and military applications. To effectively share and manage video content presents a technical challenge to the existing information management systems. Semantic features based management systems require substantial amount of manual labeling of the content and are therefore in general not practical.

Consider the following example representing an application addressed by this work. A mobile phone user has just watched a low visual quality (e.g., QCIF size, 10fps), short (e.g., 5 sec) segment of a soccer game from some unknown source. S/he wants to now watch the complete game in SDTV format from her personal soccer game video collection, or some content provider's collections. The system will therefore need to search a database based on this 5-sec segment and return the locations of the full size program, if it exists. The semantic information is clearly not present in the querying segment. The matching has thus to be "content-based". In addition, the variance in temporal and spatial scale, as well as, the noise and distortion incurred during the communication must also be addressed.

Content-based retrieval approaches have been investigated extensively by many researchers [1]-[4], [7], [9]-[12]. Such approaches are typically based on the visual features of video frames and a similarity metric defined on these features. Visual features are typically high dimensional and the commonly used are color, shape, texture, and motion. Drawbacks of such approaches are the computational expense associated with the extraction and matching of visual features, and the fact that the video sequence is treated as a collection of images and the collective temporal behavior of the sequence is typically not addressed. The retrieval performance can also be negatively affected by the scale variance, noise, and quantization distortion of the video content.

In the proposed approach video sequences are viewed as temporal traces in some high dimensional space. Each video frame is reduced to a point in its Principal Component (PC) space [5][8] of much lower dimension. The trace over time of a video sequence in this space should provide sufficient information to differentiate it from other sequences. In the PC space, the matching of sequences becomes a problem of matching the geometry of the traces; when the dimensionality of the PC space is small, this matching as well as indexing can be done efficiently. Implementation details are described to further speed up the retrieval and also address spatial and temporal resolution differences between the query and the database video.

The paper is organized into the following sections. In section 2 we present the method for computing the trace of a video sequence in its PC space and the matching method. In section 3 we discuss implementation issues, in section 4 we present simulation results, and in section 5 we draw conclusions and outline our future work.

2. PRINCIPAL COMPONENT SPACE TRACE AND MATCHING METRICS

2.1. Scaling and Principal Component Space Projection

Let n denote the dimensionality of a frame, i.e., a video frame f_j belongs to R^n. Principal Component Analysis (PCA) [8] finds an nxd transformation V_d, with d

orthogonal unit *n*x*1* vectors, that maps the frames of the video sequence f_j to a low *d*-dimensional ($d<n$) Principal Component space, that is,

$$x_j^d = V_d^T f_j \qquad (1)$$

where x_j^d is an *d*x1 vector. V_d is found according to ,

$$V_d = \min_V \sum_{j=1}^{N} \| (f_j - f_0) - VV^T (f_j - f_0) \|^2 , \qquad (2)$$

where f_0 is the average of the frames observed. Notice that *V* is sample dependent and its accurate computation requires correct modeling of the covariance of the sample frames with a large number of samples.

For a sequences with frame width *W* and height *H*, the original data dimension $n=W*H$, can be quite high, and the amount of available data is typically not adequate in accurately performing PCA. Therefore we would like to reduce the frame size to a desirable scale *w* and *h* first, with minimum information loss possible, before the PCA. This is typically done via the wavelet scaling.

For $w=8$, $h=6$ and $d=2$, the PCA basis vectors and eigen-values are shown in Fig. 1. Notice that the first 4 components captured most energy of the sample frames.

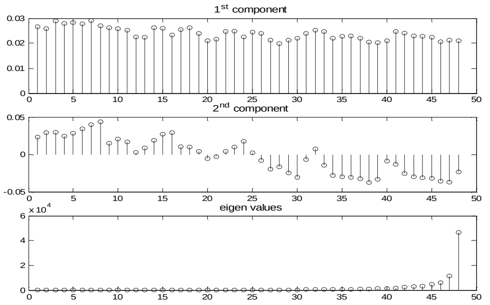


Figure 1. Principal component basis vectors and eigen-values

With *d*=2, frames are reduced to feature points in the 2D PC space and the trace of video sequences can be visualized. Examples of such trace for the “foreman” and a certain “mixed” sequences are illustrated in Fig. 2. The “mixed” sequence consists of 50 various video clips of 60 frames each. This is a very compact representation compared with other image features like color, shape and texture features.

2.2. Matching Metrics

The traces of different video clips occupy different areas in the 2D space, and have different trace geometry, as shown in Fig.2.

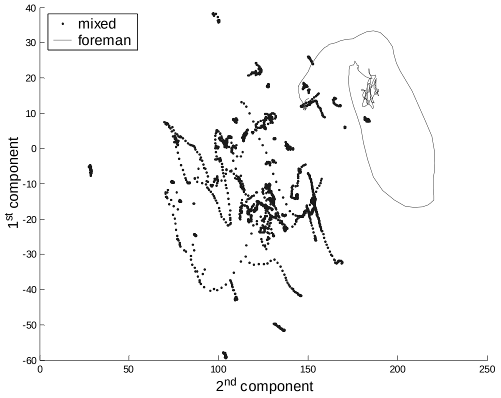


Figure 2. Sequence traces in the 1st-2nd principal component space

For a *q*-frame querying video clip with PC feature representation $Q=\{x_1, x_2,\ldots, x_q\}$, and a *p*-frame video collection with feature representation $D=\{y_1, y_2,\ldots, y_p\}$, with $p>q$, let the distance between *Q* and sub segment of *D* starting at *k* be,

$$d_Q^k = (\sum_{j=1}^{q} (x_j - y_{k+j-1})^2)^{1/2} \qquad (3)$$

Let the minimum distance between *Q* and *D* be $d^* = \min_{k\in[1,p-q+1]} d_Q^k$. If the video collection contains the query clip, then d^*=0. However due to different frame sizes, quantization scales when dealing with compression, and errors/noise in a communication application, we declare the query clip exists if the minimum distance is below certain threshold d_0, and the location is determined by,

$$k^* = \arg \min_{k\in[1,p-q+1]} d_Q^k \qquad (4)$$

Alternatively, a scalar feature of a trace, the differential trace step, can be used in matching. It is defined by

$$l_j = \begin{Bmatrix} 0, & if \;\; j=1 \\ |\, x_j - x_{j-1} \,|, & if \;\; j>1 \end{Bmatrix} \qquad (5)$$

Let $L = \{ l_1, l_2,\ldots,l_m, \}$ denote the differential trace of an *m*-frame sequence. As an example, the differential trace for the “foreman” sequence is shown in Fig.3, for *d*=2 and *d*=4 cases. Based on Fig. 3 and additional data, it appears that the differential trace is relatively invariant with respect to the dimensionality of PCA for $d > 2$. This is because most energy is captured by the first 2 dimensions.

Let us denote by $d(L^a, L^b)$ the distance between two differential traces L^a and L^b of length *m*. For example, if the L_2 norm is used, is computed as,

$$d(L^a, L^b) = \sqrt{\sum_{j=1}^{m} (l_j^a - l_j^b)^2} \qquad (6)$$

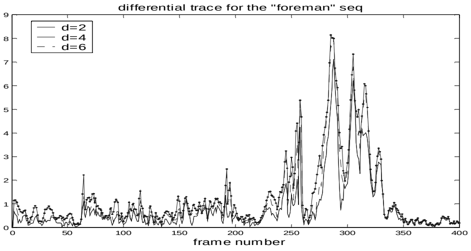


Figure 3. "foreman" seq trace step length plots

The differential trace can now be used as a matching metric for retrieval. That is, for an *m*-frame querying video clip, with differential trace L^q, its best match is found in a database according to

$$k^* = \arg\min_k d(L^q - L_k^b) \qquad (7)$$

where $L_k^b = [l_k^b, l_{k+1}^b, \cdots, l_{k+m-1}^b]$ is the partial differential trace of a database sequence of length *m* starting at time instance *k*. Compared with the retrieval method in (3), this differential trace based method places more emphasis on the temporal behavior of the sequence, than the location and geometry. It seems to be more appropriate to use this method when the query clips are long.

The search performance can be further improved by efficient indexing. Because of the relatively low dimensionality of the feature space, an R* tree [13] like indexing structure can be employed to store the features for each video shot.

3. IMPLEMENTATION CONSIDERATIONS

As mentioned in section 1, the querying video clip very often is of different spatial and temporal resolution than the clips in the database. To address the spatial resolution incompatibility (plus additive noise and quantization error issues) we scale the frames to a common spatial resolution by low-pass filtering and down-sampling both the querying and the database sequences. Typical common resolutions used are 8x6, 12x9 and 16x12. This scaling process can also improve the accuracy of the PCA process (2) with limited samples, since it reduced the data space dimension.

The differences in temporal resolution between querying and database sequences can be addressed by pre-computing or computing on the fly the traces and differential traces of the sequences in the database at different frame rates, like for example, 10fps, 15fps, 20fps, 25fps and 30fps. For a given frame rate in the query clip, we just match the database features that have the same timestamp offset with the querying clips.

A related issue to the differences in temporal resolution is when there are random frame drops, due, for example, to transmission errors. Assuming the frame timestamp is present, then for the matching method in (3), we simply ignore the frames that are missing. The distance is now defined as

$$d_Q^k = (\sum_{j=1}^{q} h_j (x_j - y_{k+j-1})^2)^{1/2} \qquad (8)$$

where h_j is zero if frame *j* is missing in the query clip, otherwise h_j=1.

For the differential trace based matching, the missing frames need to be interpolated. We perform linear interpolation in the PC space. Although more sophisticated interpolation methods could be employed, we experimentally found out that linear interpolation is adequate for retrieval purposes.

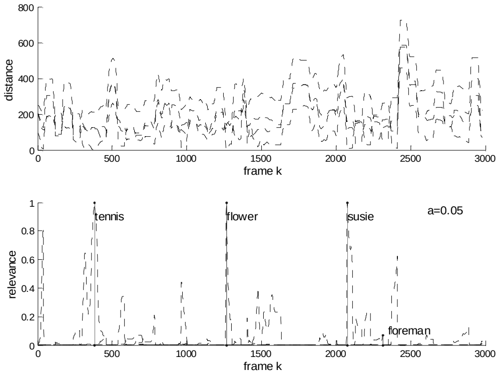
4. SIMULATION RESULTS

In our simulations the PCA is done based on 3200 randomly selected video clips from various collections. We experimented with scales 4x3, 8x6, 16x12, 32x24, and found out that the scale of 8x6 offers the best compromise between energy compaction and amount of information loss.

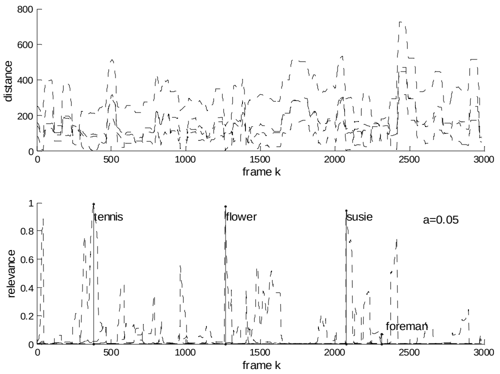
To demonstrate the effectiveness of the proposed method, we set up a video collection of 3000 frames from 50 different clips. We then set up 2500 positive queries with different clip length and locations from the collections. We also set up 500 negative queries from clips that do not exist in collections.

When there is no noise and no random frame drops, the retrieval achieves 100% accuracy for all 2500 positive queries and 500 negative queries, with query lengths of 8, 12, 16 and 20 frames for both geometry matching and differential trace matching methods.

Selected noise free retrieval results are illustrated in Fig. 5a. We have four 20-frame queries created from "tennis", "flower", "susie" and "foreman" sequences. Among them only the "foreman" query does not exist in the collections.



(a) Noise-free queries



(b) Noisy queries

Figure 4. Retrieval results

The upper plot is the distance d_Q^k in (3) for all four cases. From this we compute the relevance values, i.e., the distance values in (3) normalized to [0,1] by an exponential function $exp(-ad_Q^k)$ as shown in the lower plot, with $a=0.05$. A threshold of 0.9 is applied to the relevance values to determine if the query clip exists in the collections. In this case, we correctly determined that queries "tennis", "flower", and "susie" exist and found their correct locations. The threshold of 0.9 eliminated several false detections including the "foreman" query .

When the spatial noise is present, the retrieval performance is degraded, as expected. The retrieval error rates in percentile are summarized in Table 1. The columns are the different query lengths, while the rows are the different levels of noise added in PSNR. Notice that the retrieval accuracy holds well for queries with number of frames >15 and PSNR >24dB.

%	10	15	20	25	30	35	40
36dB	0	0	0	0	0	0	0
32dB	4	2	2	2	2	2	0
28dB	6	2	4	0	0	0	0
24dB	10	4	4	2	2	2	2
20dB	14	14	10	8	8	4	4

Table 1. Noisy retrieval error rates

The effect of random frame drops on the retrieval performance is negligible for query clips with length greater than 30 and drop rate below 30%. This is the expected operating range for clip retrieval.

Sequence	PSNR	Drop Rate	Relevance
"tennis"	36 dB	12.5%	0.994
"flower"	28 dB	25%	0.987
"susie"	20 dB	50%	0.958
"foreman"	36 dB	0%	0.068

Table 2. Noisy retrieval results

Selected retrieval results with noise and frame drops are shown in Fig 5.b. The amount of noise and frame drops are summarized in Table 2.

5. CONCLUSION AND FUTURE WORKS

In this paper we presented a new content-based video shot retrieval solution. The video frames are reduced to points in low (2~4) dimensional space and the retrieval is based on matching the location and geometry of the sequence trace. Our solution is fast and robust to noise, distortions and differences in spatial and temporal resolutions between querying and database sequences. The low dimensionality of the proposed feature also makes efficient indexing possible. The proposed solution can be useful in a wide range of practical applications that require real time response to video queries.

6. REFERENCES

[1] Calic, J. and Izquierdo, E., "A multiresolution technique for video indexing and retrieval", *Proceedings of Int'l Conference on Image Processing*, September, 2002, Rochester, NY.

[2] S.-F. Chang, Chen, W., Meng, H.J., Sundaram, H., Di Zhong, "A fully automated content-based video search engine supporting spatiotemporal queryies", *IEEE Trans. on Circuits and System for Video Technology*, vol.8, No.5, September 1998.

[3] Chiou-Ting Hsu, and Shang-Ju Teng "Motion trajectory based video indexing and retrieval", *Proceedings of Int'l Conference on Image Processing*, September, 2002, Rochester, NY.

[4] Dagtas, S., Al-Khatib, W., Ghafoor, A. and Kashyap, R.L.; "Models for motion-based video indexing and retrieval ", *IEEE Trans. on Image Processing*, Vol. 9 No. 1, Jan. 2000.

[5] Forsyth, D., and Ponce, J., *Computer Vision A Modern Approach*, pp.507-509, Prentice Hall, New Jersey, 2003.

[6]Hanjalic, A., "Shot-boundary detection: unraveled and resolved? ", *IEEE Trans. on Circuits and System for Video Technology*, vol.12, No.2, Feb. 2002.

[7] Hanjalic, A., Lagendijk, R.L., and Biemond, J., "Automated high-level movie segmentation for advanced video-retrieval ", *IEEE Trans. on Circuits and System for Video Technology*, vol.12, No.2, Feb. 2002.

[8] Hastie, H., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, Chapter 14, Springer Series in Statistics, 2001.

[9] Kim, Sang Hun; Park, Rae-Hong, "An efficient algorithm for video sequence matching using the modified Hausdorff distance and the directed divergence", *IEEE Trans. on Circuits and System for Video Technology*, vol.12, No.7, July 2002.

[10] Muneesawang, P., Guan, L., "Automatic relevance feedback for video retrieval", *Proceedings of Int'l Conference on Multimedia and Expo*, July 2003, Baltimore, MD.

[11] Smith, J.R., Basu, S., Ching-Yung Lin, Naphade, M. and Tseng, B., "Interactive content-based retrieval of video", *Proceedings of Int'l Conference on Image Processing*, September, 2002, Rochester, NY.

[12] Wei Zeng, Wen Gao, and Debin Zhao, "Video indexing by motion activity maps", *Proceedings of Int'l Conference on Image Processing*, September, 2002, Rochester, NY.

[13] N. Beckmann, H.-P. Kreigel, R. Schenider, and B. Seeger, "The R*-tree: An efficient and robust access method for points and rectangles", *Proceedings of ACM SIGMOD ICMD*, 1990.