FAST VIDEO SHOT RETRIEVAL BASED ON A SEQUENCE TRACE IN THE 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, across different platforms, and over different communication channels. In this work we present a fast retrieval algorithm based on matched filtering utilizing the video sequence trace characteristics in the principal component space. Experimental 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 work for labeling the content and is 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 a video frame and a similarity metric based on these features is typically used. Visual features 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 temporal behavior therefore of the sequence is not addressed. The retrieval performance can also be negatively affected by the scale variance, noise, and 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) [5][8] space. 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 the problem of matching the geometry of the traces; when the dimensionality of the PC space is small, this can be done efficiently. We more specifically propose the use of the differential trace (an one-dimensional quantity irrespectively of the dimensionality of the PC space) for matching and retrieval. Implementation details are described to further speed up the retrieval but 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 principal component 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. Principal Component Space Projection

Let n denote the dimensionality of a frame, i.e., a video frame f_j belongs to R^n. The Principal Component Analysis (PCA) [8] finds an nxd transformation Q_d, with d orthogonal unit $nx1$ vectors, that maps the frames of the

video sequence f_j to a low *d*-dimensional ($d<n$) Principal Component space, that is,

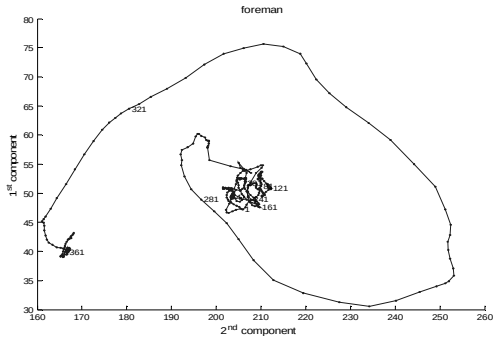
$$x_j^d = Q_d^T f_j \quad (1)$$

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

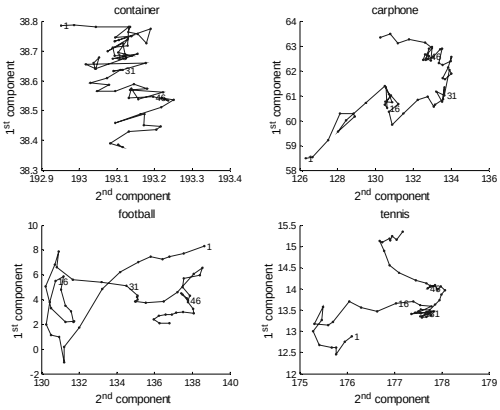
$$Q^* = \min_Q \sum_{j=1}^{N} \| (f_j - f_0) - QQ^T (f_j - f_0) \|^2 , \quad (2)$$

where f_0 is the average of the frames observed. Notice that *Q* is sample dependent and the accurate computation of *Q* requires large number of samples.

For *d=2*, the mapping of the video sequence frames into 2-dimensional plane points can be visualized. Examples of mappings for the "foreman" sequence is plotted in Fig. 1a, and certain "mixed" sequences consists of segments from the "football", "tennis", "container" and "car phone" sequences are illustrated in Fig. 1.



(a). "Foreman" sequence



(b). "Mixed" sequences

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

The trace of the sequence is obtained by connecting frames in time sequence. Notice that the traces of different clips occupy different areas in the 2D space, and have different trace geometry.

2.2. Matching Metrics

It is difficult to visualize higher dimensional traces. A scalar feature of traces, the differential trace step, can be of use. It is defined by

$$l_j = \begin{cases} 0, & if \quad j = 1 \\ | x_j^d - x_{j-1}^d |, & if \quad j > 1 \end{cases} \quad (3)$$

Let $L = [l_0, l_1, \cdots, l_{m-1}]$ denote the differential trace of an *m*-frame sequence. The differential trace for the "foreman" sequence is shown in Fig.2, for 2- and 4-dimensional cases. It appears that the differential trace is relatively invariant with respect to the dimensionality *d* of the space for *d* greater than 2. This is because most energy is captured in the first 2 dimensions.

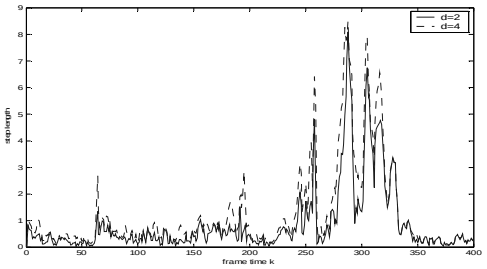


Figure 2. "foreman" seq trace step length plots

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,

$$d(L^a, L^b) = \sqrt{\sum_{j=1}^{m} (l_j^a - l_j^b)^2} \quad (4)$$

The differential trace can now be used 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

$$k^* = \arg\min_k d(L^q - L_k^b) \quad (5)$$

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*. The operation in (5) can be implemented efficiently with a matched filter-like structure, with L^q convolving with L^b and detecting the spike in the output at location k^*. This now becomes a very fast retrieval algorithm, since no visual features matching operation is needed (assuming the features have been extracted), but a scalar operation instead.

The search time can be further reduced if we assume that a video shot segmentation [6] is performed first for the sequences in the database. The trace of each shot then can be bounded by a hyper-parallelepiped. If the bounding parallelepipeds of the traces of the querying

sequence and the shot do not intersect, then no matching of the corresponding differential traces (5) is attempted. Clearly such an approach is independent of the dimensionality of the PC space.

Because of the relatively low dimensionality of the PCA trace representation of the video shots, effective indexing schemes like R^*-Tree [13] can be utilized to improve the retrieval efficiency.

Other methods like direct matching of the points in the PCA space, projection of the querying video clip onto the parameterized curve representation of the database clips can also be utilized as retrieval solution.

3. IMPLEMENTATION CONSIDERATIONS

As mentioned in section 1, the querying video clip very often is of different spatial and temporal resolutions than the clips in the database. To address the spatial resolution incompatibility (plus additive noise issues) we low-pass filter and down-sample both the querying and the database sequences. That is, they are all brought down to a common $w_1 x w_2$ resolution before applying PCA,

$$f_j' = D_{w1xw2}(LP(f_j)) \qquad (6)$$

Typical common resolutions used for (6) are 8x8, 12x12, or 16x16. This down-sampling 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. A related issue to the differences in temporal resolution is when there are random frame drops, due, for example, to transmission errors. The missing frames need to be interpolated in this case before the differential trace is computed. We perform linear interpolation in the PC space.

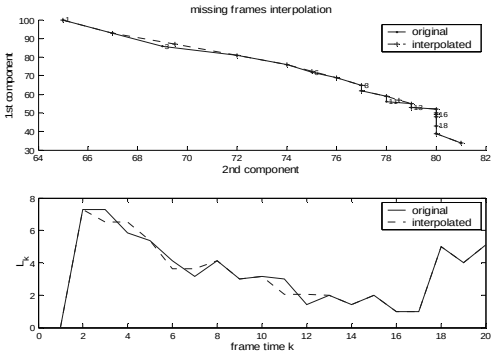


Figure 3. Interpolation of missing frames

Although more sophisticated interpolation methods could be employed, we experimentally found out that linear interpolation is adequate for retrieval purposes. An interpolation example for the first 20 frames of the "foreman" sequence with missing frames 3, 6, 11, 16 is shown in Fig.3. The upper plot in Fig.3 depicts the trace interpolation result in the 2D space, while the lower plot shows the corresponding interpolated differential trace L_k.

4. SIMULATION RESULTS

In our simulations we low-pass filter and average down the frames to 8x8 thumbnail images before the PCA process. The eigen-values and the 1st and 2nd component basis vectors of the "foreman" sequence are plotted in Fig. 4.

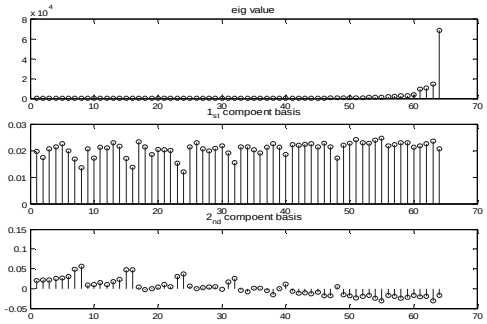


Figure 4. Principal component basis vector

For other thumbnail frame sizes like 4x4, 16x16 and 32x32, the eigen-values demonstrate the same tendency of capturing most energy in 2-4 dimensions. To learn the principal component space with 2-4 dimensions for all sequences, we collected data from more than 3000 frames and find the optimal projection Q from this data set. For different thumbnail frame sizes, the eigen-vectors are different but the sequence traces have similar geometry, as illustrated in Fig. 5. for the "foreman" sequence.

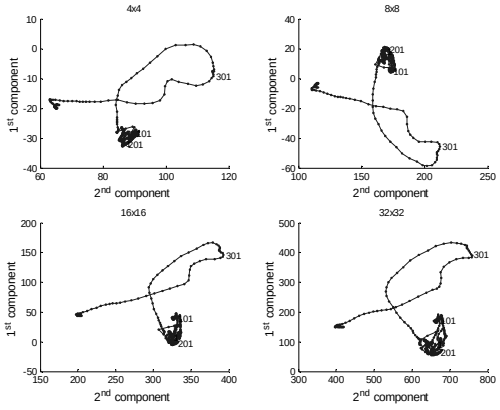


Figure 5. "foreman" sequence PCA trace with different thumbnail image size

To demonstrate the effectiveness of the proposed approach we set up a small video database of 1600 frames by manually mixing together different sequences. We set up 4 queries from QCIF sized clips of the "container", "tennis", "carphone" and "football" sequences, each of 20 frames in length. Their correct matching locations in the database are frames (i.e., time instances) 160, 380, 1030, and 1330.

Retrieval results are illustrated in Fig. 6. Results with a noise-free query are shown in the upper plot, while results with queries corrupted by additive Gaussian noise and dropped frames are shown in the lower plot (the amount of distortion is shown in Table 1. The relevance values, i.e., the distance values in (4) normalized to [0,1] by an exponential function *exp(-ad)* are shown, with *a=0.05*. Notice that in both cases the correct matching locations in the database are found, although in the noisy case the retrieval relevance values are less than 1.0. This increases the probability of false matches.

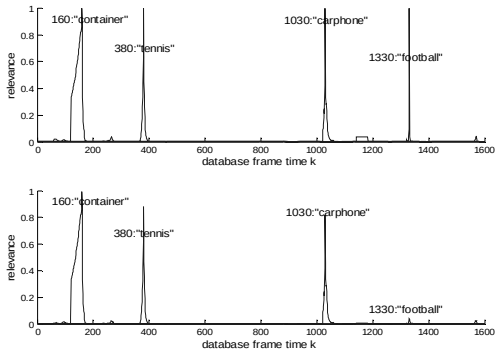


Figure 6. Retrieval results

Sequence	SNR (dB)	Drop Rate	Relevance
"Container"	13.13	0%	0.99
"Tennis"	0	25%	0.88
"Carphone"	16.98	10%	0.81
"Football"	13.13	25%	0.11

Table 1. Noisy retrieval results

The performance values of the experiment with the noisy queries are shown in Table 1. Notice that the SNR values represent the noise level in the PC space representation, not the original noise level in the frames. The retrieval performance seems to be more sensitive to the percentage of the dropped frames than the spatial noise.

5. CONCLUSION AND FUTURE WORKS

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

Work is underway to find out under what conditions the operations in (6) coupled with the PCA preserves the trace geometry. We are also investigating efficient indexing by utilizing the geometric properties of the trace of video clips in a very low dimensional space, and the template matching based solution for missing frames when time stamp information is not available.

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.