FAST VIDEO SHOT RETRIEVAL WITH 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 of the video sequence trace characteristics in the principal component space. Techniques to combat scale variance, noises and distortions are also investigated, resulted into a robust and fast content-based video shot retrieval solution.

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 these video content presents a technical challenge to the existing information management system. Semantic features based management system requires substantial amount of human work in the labeling of content and is generally not practical.

For example, if a mobile phone user watched a low visual quality, 5-sec segment of soccer game in QCIF size and 10 fps from some unknown source, and he wants to find out the full game in SDTV format from his personal soccer game video collections, or some content provider's collections, the system will need to search based on this 5-sec segment and return the full size program location if it exists. The semantic information is not present in the querying segment. The matching has to be "content-based", and the variance in temporal and spatial scale, as well as noise and distortion incurred during the communication must also be addressed.

The content-based approaches [1]-[4][7][9]-[12] have been investigated extensively by many researchers. These approaches are typically video frames based, and the retrieval is done via a metric function based on the visual features of the frames. Visual features used are color, shape, texture and motion. However a drawback is that the visual features extraction and matching are expensive in computation, and the visual feature based approach treat the video sequence as a collection of images and the temporal behavior of the sequence is not well addressed. The retrieval performance can also be negatively affected by the scale variance, noise and distortion in the video content.

In our approach, video sequences are viewed as temporal traces in some high dimensional space generated by some stochastic process. The video frames are reduced to points in its Principal Component (PC) [5][8] space, and the observed trace of a video sequence in the should give us sufficient information to differentiate it from other sequences. In the PC space, the matching of the sequences becomes the problem of matching the geometry of the traces, and when the dimension of PC space is small, this can be done efficiently. Under this framework, the problem of spatial scale variance and noise can be addressed by the filtering and transform processes. The temporal variance and distortion can be addressed by interpolating.

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 the method to address the issues of scale variance and noise/distortion. In section 4 we present the simulation results. In section 5 we draw conclusions and outline our future work.

2. PRINCIPAL COMPONENT SPACE TRACE AND MATCHING METRICS

The spatial dimension of the video frames is large. For QCIF sized intensity image sequences, the dimension is $R^{n=176x144}$. In reality the space populated by the video sequences is much smaller in dimension, due to the image formation constraints. The Principal Component Analysis (PCA) [5] finds an nxd transform Q_d, with d orthogonal unit nx1 vectors, that maps the frames of the video sequence f_j to a low d-dimensional Principal Component space Points (PCP) representation,

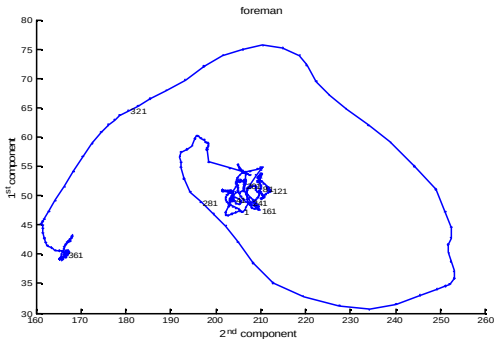
$$x_j^d = Q_d^T f_j \tag{1}$$

while preserving most energy or information of the sequence, that is

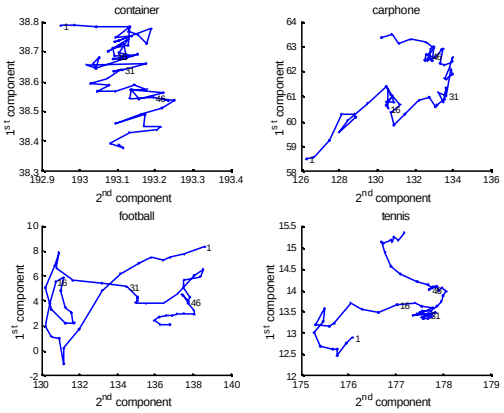
$$Q^* = \min_Q \sum_{j=1}^{N} \| (f_j - f_0) - QQ^T (f_j - f_0) \|^2 \tag{2}$$

in which f_0 is the average of the frames observed. QQ^T transform the rank d reconstructions back into the original n-dimensional space. 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 mapping for the "foreman" sequence and the mixed 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.

It is difficult to visualize higher dimensional traces, but there is a scalar feature of traces in the form of frame-by-frame step length vector that can be useful. Let the trace step length at frame time j in d-dimensional space be,

$$L_j = \begin{cases} 0, & if \quad j = 1 \\ |\, x_j^d - x_{j-1}^d \,|, & if \quad j > 1 \end{cases} \qquad (4)$$

The trace step length for the "foreman" sequence is illustrated in Fig.2, for 2 and 4 dimension cases. Notice that the step length vector is relatively invariant with respect to the dimension of the space, and 2 dimensional PC space is adequate in retrieval performance for most cases.

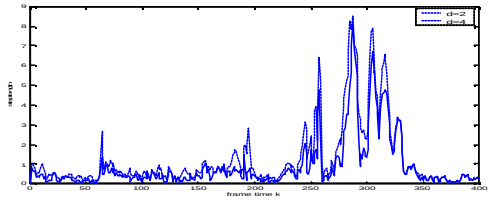


Figure 2. "foreman" seq trace step length plots

The distance between two video clips step vectors L^a and L^b of length n is computed as,

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

Notice that in addition to L2 norm used in (4), other norms can also be employed.

Therefore instead of direct trace geometry matching, we find out a faster alternative is by matching the trace's frame-by-frame step vector L as in (4). For an m-frame querying example video clip, with step vector L^q, its best match in the n-frame database clip with step length vector L^b is found by,

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

where $L_k^b = [l_k^b, l_{k+1}^b, \cdots, l_{k+m-1}^b]$, the k* in (5) is the time offset found in the database clip. The operation of (5) can be implemented efficiently with a matched filter like structure, with L^q convolves with L^b and detect the spike in the output.

Notice that this is a very fast algorithm. Each frame comparison is reduced to a scalar operation, while other solutions require a visual feature (extraction) matching operation.

Video database can contain many video clips. Assuming the video shot segmentation [6] is performed and each video shot is represented by the trace, then instead of matching every clip in the database with (5), we can further reduce the candidate clips by looking at the distances of the querying clip PCPs to the database clip PCPs. Let the PCP range spanned by the querying clip q be represented by $S_q=[x_{min}^q, x_{max}^q]$, where the components of x_{min}^q and x_{max}^q define the lower and upper boundaries in each dimension. Similarly the PCP range spanned by the database video clip b be represented by $S_b=[x_{min}^b, x_{max}^b]$, The database clip is rejected if,

$$S_q \cap S_b = NULL \qquad (6)$$

This means that if two video clips does not have overlaps in the PC space, then reject it. It can be explained by Fig 1.b as well. If the querying clip is some segment of the "coast guard" sequence as shown, then the candidate sequences "calendar" and "fish" will be rejected without operations in

(5). A tighter restriction than (6) is also feasible. For example, place a constraint on the volume ratio of,

$$\frac{S_q \cap S_b}{S_q} \le q \tag{7}$$

The combination of (5), (6) and (7) gives us a very fast video shot retrieval solution. Experimental results showed that it is very robust in performance. However, when the querying clip has different frame rate, and contains frame distortion and frame drops due to the communication and storage constraints, additional processing is needed to maintain the performance.

3. TEMPORAL-SPATIAL SCALE INVARIANCE AND DISTORTION REDUCTION

As mentioned in the section 1, the querying video clip could be produced in different frame size and frame rate, QCIF and 10 fps for example, with lower visual SNR quality and dropped frames due to the communication process, while the database clips are stored in SDTV size and 30 fps with high visual qualities. A direct PCA mapping of the same querying clip and database clip to the principal component space will result in different traces which degrades the retrieval performance. To address this problem, we have a two-step solution.

To combat the spatial scale variance and noise/distortion, we apply low-pass filtering and down averaging operations on the frames to a lower common *wxw* resolution before applying PCA,

$$f_j' = D_{wxw}(LP(f_j)) \tag{8}$$

These processes mitigate the problem of spatial distortion and scale variance. The size of common resolutions can be 8x8, 12x12 or 16x16. This process can also reduce the dimension of the original frame space and can improve the accuracy of PCA process (2) with limited samples.

The problem of different frame rate can be solved by pre-computing multiple step length vectors L for the database clips at different frame rates like 10fps, 15fps, 20fps, 25fps and 30fps. This is reasonable because that covers most typical frame rates, and when the querying frame rate is low, matching at lower rates reduces the computational cost. The other solution is to compute the database clips' L on the fly before feeding into the matched filter for match detection.

When the random frame drops occur due to the communication loss, the missing frames need to be interpolated and the step length adjusted. If frame k is missing from the clip, its PCP vector x'_k is interpolated as,

$$x'_k = \frac{x_{k-1} + x_{k+1}}{2} \tag{9}$$

More complex interpolation methods can also be employed, but we found (9) to be adequate for retrieval purpose. An interpolation example for the first 20 frames of the "foreman" sequence with missing frames 3, 6, 11, 16 are shown in Fig.3.

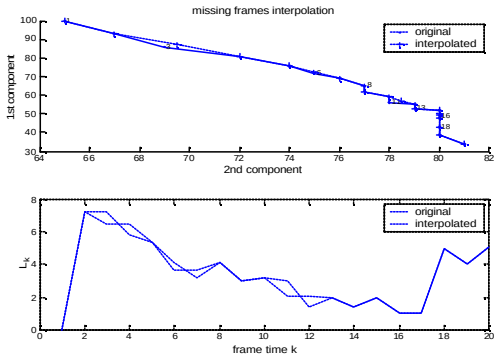


Figure 3. missing frames interpolation

The upper plot of Fig.3 is the trace interpolation results in 2D space, the lower plot is the interpolated step length function L_k.

4. SIMULATION RESULTS

In our simulation, we low-pass filter and average down (8) the frames to an 8x8 icon images before the PCA process (2). The eigen values and the 1st and 2nd component basis vectors are plotted in Fig. 4. Notice that most energies are captured by the 4 largest eigen vectors.

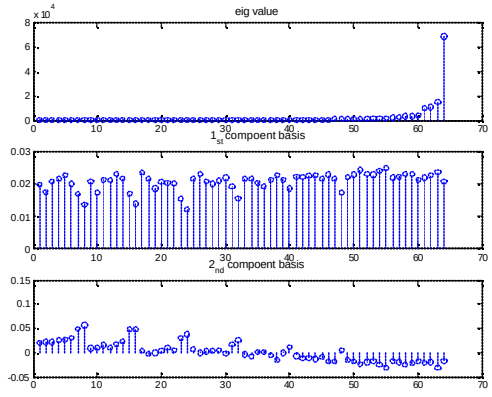
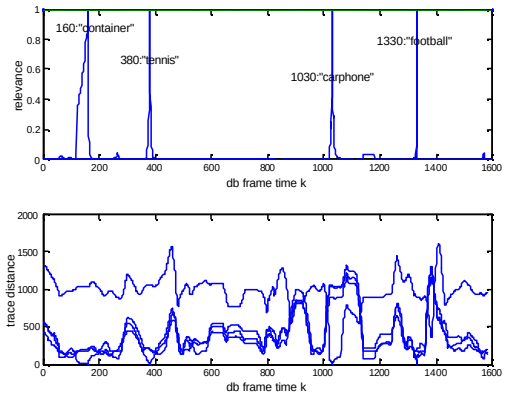


Figure 4. Principal component basis vector

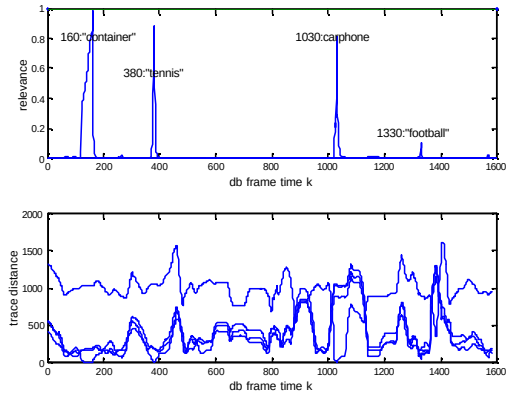
To demonstrate our retrieval solution we set up a small video database of 1600 frames by manually mixing together different sequences. We set up 4 queries from the QCIF sized clips of the "container", "tennis", "carphone" and "football" sequences, with 20 frames in length. Their correct matching locations in the database are 160, 380, 1030, and 1330.

The retrieval results are illustrated in Fig.5. The noiseless case is plotted in Fig.4a, the lower plot is the distance (4) at database frame time, the upper plot is the

retrieval relevance values normalized from the distance metric (4) into [0,1]. Notice that in noiseless case, correct matching with 1.0 relevance are found for all 4 query cases. In Fig.4b, the query clips are corrupted by the noises, even though they all find the correct matching locations in the database, the retrieval relevance values are not 1.0. The noise performance is summarized below in Table 1.



(a) noise free case



(b) Spatial-temporal noise case

Figure 5. retrieval results

Query Sequence	Spatial SNR (dB)	FrameDrop Rate (%)	Max Retrieval 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

Notice that the SNR in Tab. 1 is the noise level in the PCP representation, not the original noise level in the frames. The retrieval performance is more sensitive to the frame drop distortion 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 sequence trace geometry. Our solution is fast in performance, and robust to noise and distortions, as well as scales variance in both temporal and spatial domain. This solution can be useful in a wide range of practical applications that requires real time response to video queries.

Work is underway to investigate efficient indexing method that is made possible by the trace geometric representation of video clips in a very low dimensional space.

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.

.