A Nfl-based and Feature Extraction Supported Shot Retrieval Approach

Zhen Lei, Yujun Liu, Wenge Zhang
Academy of Armored Force Engineering
Beijing 100072,China
williamjohnmail@126.com

Xuelin Liu
Beijing Airspace Control Center
Beijing,100720, China

***Abstract*—Nearest Feature Line (NFL) is a convenient and effective way to search in video database, but the Framework does not address the feature extraction for dimension reduction. In this paper, a novel method is proposed for content-based shot retrieval. Karhunen Loeve Transform (KLT) is used to reduce dimensionality of feature spaces. In addition, we present a new take-maximum-from-minima (TMFM) based key frame extraction algorithm, and key-frames extraction is combined with the NFL method to achieve a better performance. Experimental results have shown that our combined method not only achieves superior performance than the NFL and Equal interval (EI), and other classification methods such as Nearest Neighbor (NN) and EI and Nearest Center (NC) and EI but also increases the retrieval speed and reduces the memory significantly without sacrificing the retrieval accuracy.**

Keywords-Nearest Feature Line, Feature Extraction, shot retrieval, key frame extraction

I. INTRODUCTION

Retrieval in video database is a challenging task because of the huge data volume and rich content of video material. Traditional Content-based Retrieval(CBR) systems is usually performance based on comparison of low level features, such as color, texture or shape features, extracted from the images themselves. But lots of noise information which is unrelated to retrieval contained in the original features, which will affect the precision of retrieval. High dimension also increases operation burden and higher data memory demand. Thus, dimension reduction should be performed before indexing the feature vectors with a multi-dimensional indexing technique.

Dimension reduction is typically obtained using either the KLT or clustering techniques. The extracted features are grouped into some suitable data structure or mathematical construct, and suitable metrics are used to measure the similarity between an image and any other image. At this stage, the main challenges are the high dimensionality of the feature vectors. Solutions to the high dimensional indexing problem include reducing the dimension of the feature vectors and the use of efficient multi-dimensional indexing techniques. Our algorithm utilizes the KLT to remove redundancy in feature space and further improve the retrieval performance. KLT is a data-dependent transform that achieves optimal decorrelation and energy compaction.

Moreover, we propose a new TMFM based key frame extraction method. Key frames provide a suitable abstraction for video indexing, browsing and retrieval [1]. Users can quickly browse over the video by viewing only a few highlighted key frames. The use of key frames greatly reduces the amount of data required in video indexing and browsing. Key-frames extraction combined with the NFL method can achieve a better retrieval performance. In our scenario, each frame of a shot is considered as a point in the feature space. We use the lines between feature points of key-frames as an approximation of the whole trajectory of a shot. In order to make it more appropriate for video retrieval, we utilizes the improved nearest feature line method proposed by Zhao et. al[2].

The time and memory complexities in online retrieval are linear with the total feature dimension. Result shows that when only about $\psi = 35\%$ of the original total dimension is used, which is the case in our system, only 35% of the memory is needed and the speed of the proposed algorithm is about 3 times of that of the algorithm without the KLT dimension reduction.

This paper is organized as follows. Section II introduces the KLT process. Section III describes our TMFM based key frame extraction algorithm. In section IV, we focus on the description of improved shot retrieval approach. The experimental results are shown in Section V. Concluding remarks are provided in section VI.

II. FEATURE EXTRACTION

Feature extraction is to choose representative components (extractive features) from feature vectors (coarse features), namely to turn high-dimensional features in Pattern space to low-dimensional features in feature space. The coarse features become extractive features after reducing dimension may reduce workload greatly for clustering or training of samples. Studying feature extraction is still not only to reduce workload and that for an idiographic pattern recognition issue, recognizing precision doesn't improving along with increasing of feature number.

KLT and its variations have been used in dimension reduction in many areas such as features for facial recognition, eigen-images, and principal component analysis. Its theory is as follows:

A nonperiodic random process $x(t)$ may be expanded into a series of orthogonal function $\phi_n(t)$

$$x(t) = \sum_{n=1}^{\infty} \gamma_n x_n \phi_n(t), \quad a \le t \le b \tag{1}$$

The confirming of orthogonal function $\phi_n(t)$ and coefficient γ_n is as below:

For t and s in the interval $[a,b]$, the autocorrelation function $R(t,s)$ is given by:

$$R(t,s) = E\{x(t)x(s)\} = \sum_n |\gamma_n|^2 \phi_n(t)\widetilde{\phi}_n(t) \tag{2}$$

where $\widetilde{\phi}_n(t)$ is plural-conjugate of $\phi_n(t)$, thus,

$$R(t,s)\phi_n(s)ds = \int_a^b \sum_n |\gamma_n|^2 \phi_n(t)\int_a^b \phi_n(s)\widetilde{\phi}_n(s)ds \tag{3}$$

Consider the orthogonal relation

$$\int_a^b R(t,s)\phi_n(s)ds = |\gamma_n|^2 \phi_n(t) \tag{4}$$

We assume $\omega_1, \omega_2, ...\omega_m$ is M species of samples and consider them to be all real number or continuous random function, each class can be expressed as

$$x_i(t), \quad T_1 \le t \le T_2 \quad i = 1,2,..M \tag{5}$$

Then $x_i(t)$ can be expanded as below,

$$x_i(t) = \sum_{j=1}^{\infty} c_{ij}\phi_j(t) \quad T_1 \le t \le T_2 \quad i = 1,2,...M \tag{6}$$

where c_{ij} is coefficient, $E\{c_{ij}\} = 0$

If the observation for function $x_i(t)$ is carried through for n times continuously and equably in $[T_1, T_2]$, then observation vector appears

$$x_i = \begin{pmatrix} x_i(t_1) \\ x_i(t_2) \\ \vdots \\ .x_i(t_n) \end{pmatrix} \tag{7}$$

Therefore, (6) will become finite form

$$x_i = \sum_{j=1}^{n} c_{ij}\phi_j \tag{8}$$

where $E\{c_{ij}\} = 0$ and $\Phi_j = \begin{pmatrix} \phi_j(t_1) \\ \phi_j(t_2) \\ \vdots \\ \phi_j(t_n) \end{pmatrix}$.

(8)may take the form:

$$x_i = \Phi c_i \tag{9}$$

where matrix $\Phi = (\Phi_1\ \Phi_2 \cdots \Phi_n)$, and $c_i = (c_{i1}, c_{i2}, \cdots c_{in})^T$

If discrete form is adopted, then the autocorrelation matrix R is given by

$$R = \sum_{i=1}^{m} p(\omega_i)E\{x_i x_i^T\} \tag{10}$$

Substituting (9) into (10),

$$R = \sum_{i=1}^{m} p(\omega_i)E\{\Phi c_i c_i^T \Phi^T\} = \Phi\left[\sum_{i=1}^{m} p(\omega_i)E\{c_i c_i^T\}\right]\Phi^T = \Phi D_\lambda \Phi^T \tag{11}$$

From the orthonormality of Φ, we have

$$R\Phi = \Phi D_\lambda \tag{12}$$

and (8)may also take the form:

$$c_i = \Phi^T x_i \tag{13}$$

Moreover, Minimum error is given by

$$\varepsilon_{\min}^2 = \sum_{i=m+1}^{n} E\left[(c_i - E\{c_i\})^2\right] = \sum_{i=m+1}^{n} \Phi_i^T E[(x_i - E\{x_i\})(x_i - E\{x_i\})^T]\Phi_i = \sum_{i=m+1}^{n} \Phi_i^T C_x \Phi_i \tag{14}$$

where C_x is the covariance matrix of x, equation (14) can also be expressed as

$$\varepsilon_{\min}^2 = \sum_{i=m+1}^{n} \lambda_{Ci} \tag{15}$$

where λ_C is the eigen-value calculated from covariance matrix C_x, extend it to the autocorrelation matrix R of M classes of samples, we have

$$\varepsilon_{\min}^2 = \sum_{i=m+1}^{n} \lambda_{Ri} \tag{16}$$

where λ_R is the eigenvalue calculated from autocorrelation matrix R, now we can bring forward the principle of dimension reduction i.e. eigenvector of lesser eigenvalue correspond to R should be discarded to ensure minimum error.

III. KEY FRAME EXTRACTION

In this paper, we propose a new TMFM (take maximum from minima) based key frame extraction algorithm, The following describes how the TMFM algorithm is performed:

Step1. The total frame number of shot C is set to N, and initialize key frame number M=0;

Step2. Let frame f_0 be the first key frame Z_0;

Step3. Calculate the distances between other frames $f_k(k=1,2,3,...N-1)$ and f_0, if f_j satisfies condition:

$$Dist(f_0,f_j)=MAX(\|f_0-f_j\|),\ (j=1,2,3,..N-1) \quad (17)$$

then f_j is a new key frame Z_1, and $M=M+1$;

Step4. Calculate the distances between other frame f_m $(m=1,2,3,...N-1, m\neq j)$ and Z_0, Z_1 respectively: $Dist(f_m,Z_0)$ and $Dist(f_m,Z_1)$;

Step5. The minima of each pair of distances are reserved.

$$MindistPerpair_m = MIN(Dist(f_m,Z_0),Dist(f_m,Z_1)) \quad (18)$$
$$(m=1,2,3,...N-1, m\neq j)$$

Step6. The maximum of minimum distances is taken out:

$$MaxInMin = MAX(MindistPerpair_m) \quad (19)$$
$$(m=1,2,3,...N-1, m\neq j)$$

Step7. If

$$MaxInMin \geq C_1 \cdot Dist(Z_0,Z_1) \quad (20)$$

where $(C_1 \geq \frac{1}{2})$, corresponding frame is selected as new key frame Z_2, otherwise end approach;

Step8. The distances between residual frames and aforementioned three key frames are calculated respectively: $Dist(f_n,Z_0)$, $Dist(f_n,Z_1)$, $Dist(f_n,Z_2)$, and the minima of three are reserved., then the maximum of each minimum is selected, if the maximum is considerable part of above maximum, then corresponding frame is selected as key frame Z_3,otherwise end approach. Other key frames are worked out in the same principle.

IV. SHOT RETRIEVAL APPROACH

In feature space, two key frames I_i and I_j corresponding to two feature points f_i and f_j. We define:

$$f_k=\{f_{1k},f_{2k},\cdots f_{mk}\} \quad (21)$$

where m is the dimension of feature space, and the feature line connecting f_i and f_j can be expressed as $\overline{f_if_j}$, p_x is the projection of f_x on the feature line $\overline{f_if_j}$.

In order to compare the accuracy after KLT reduction with that obtained in the original feature space, we use KLT to reduce dimension of feature, for each feature point, do the following:

Step1. The autocorrelation matrix R of feature point: $f_k=\{f_{1k},f_{2k},\cdots f_{mk}\}$ is calculated, where m is the original dimension of feature space.

Step2. The eigenvalues of the autocorrelation matrix R and the corresponding eigenvectors are calculated to get eigenvectors matrix Φ.

Step3. Sort the eigenvectors in the order of descending eigenvalues, and the eigenvector of lesser eigenvalue correspond to R is discarded.

Step4. Transformation is carried out to reduce dimension of feature.

We say that the feature space after reducing the dimension of the feature vectors is KLT space. The query image F_x corresponds to point f_x in KLT space, then we decide similarity between F_x and each shot in video database by calculating the distance between f_x and all shots in KLT space. The distance between f_x and a shot is the minimum of distances between query point f_x and all feature lines. Then sort distances between query point and all shots in the order of ascending distances. And the shot corresponding to minimum distance is just the top similar shot.

V. EXPERIMENT RESULTS

To evaluate the performance of the proposed improved NFL+TMFM method, we build a video database of 120 shots. Among them, 12 shots can be unambiguous classed and are regard as test shots. These shots are taken from 30 minute news of CCTV. Three types of color features, two types of shape features and two types of texture features are used in our system, the total dimension is 475. We use accuracy and weight score[3] to measure the performance.

Firstly, our improved NFL + TMFM keyframe extraction scheme is compared with previous approaches, including the NFL+ Breakpoint (BP) keyframe extraction method, the NFL+Equal interval(EI) keyframe extraction method, the Nearest Center (NC)[4]+EI keyframe extraction method and the Nearest Neighbor(NN)+EI keyframe extraction method in the original feature space without dimension reduction. Fig.2 show that the NFL + TMFM keyframe extraction scheme is close to the NFL+ Breakpoint(BP) keyframe extraction method, the two methods produce the best performance, and the NFL+ EI method is close to the NN+EI method, they produce better performance, the Nearest Center (NC)+EI keyframe extraction method produces the worst performance compared with aforementioned methods.

Finally, the accuracy after KLT reduction is compared with that obtained in the original feature space. The results in Fig.3 show that the accuracy can be maintained when the retained total dimension is 35% of the original feature space; the benefit brought about by this reduction is that the retrieval speed is almost tripled and that the memory requirement is about 1/3 of the original.

Figure 1. An example of retrieval result

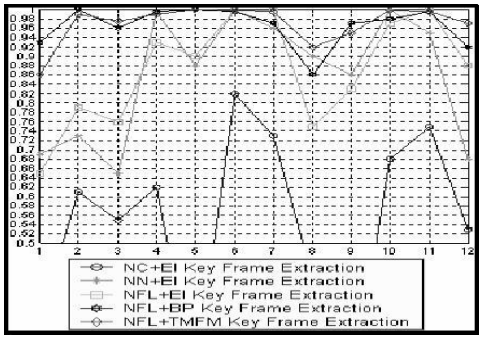

Figure 2. The score of different methods.

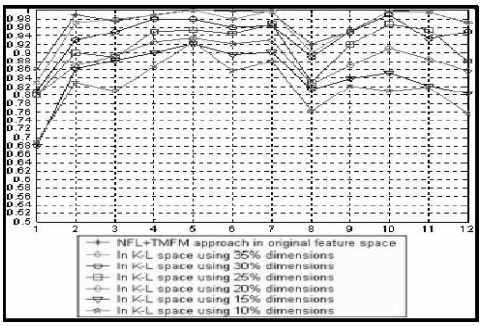

Figure 3. comparing the retrieval accuracy in KLT space between the dimension updating and no dimension updating

VI. CONCLUSION

In this paper, we have presented a new approach to content-based representation of video shots and the application in example-based shot retrieval. KLT is incorporated into improved NFL to extract representative components. Moreover we have proposed a novel TMFM based key frame extraction approach, with this approach, the number of key frames and the location of the key frames in a given video are determined automatically.

The experimental results have shown that the proposed NFL+TMFM method achieves high performance not only better than the traditional methods for classification such as NN+EI and NC+EI, but also the NFL+EI method in original feature space, moreover, it can increase the retrieval speed and reduce the memory significantly without sacrificing the retrieval accuracy. In principle, KLT can be incorporated to any other content-based retrieval methods to save memory and to speed-up computation. In the future, we will incorporate relevance feedback(RF) to achieve better performance.

REFERENCES

[1] J. Peng and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," IEEE Multimedia, 2010, 17(2), PP: 64–73.

[2] Zhao, L., Qi, W., Li, S.Z., Yang, S.Q., Zhang, H.J.(2001). Content-based Retrieval of Video Shot Using the Improved Nearest Feature Line Method. In Proceedings of ICASSP. Salt Lake City, 2001, PP: 1625 – 1628.

[3] Li, S. Z.. Content-based Classification and Retrieval of Audio Using the Nearest Feature Line Method. IEEE Transactions on Speech and Audio Processing, 2000, 8(5), PP: 619-625.

[4] Zhang, H.J., Zhong, D., and Smoliar, S.W. An Integrated System for Content-Based Video Retrieval and Browsing. Pattern Recognition, 1997,30(4), PP:643-658.