# Multi-Level Video Representation with Application to Keyframe Extraction

Xiao-Dong Yu[1], Lei Wang[1], Qi Tian[2], Ping Xue[1]
[1] *Nanyang Technological University, School of Electrical and Electronic Engineering, Singapore*
[2] *Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore*
[1]*{exdyu,elwang,epxue}@ntu.edu.sg,* [2] *tian@i2r.a-star.edu.sg*

## Abstract

*Content-based video analysis calls for efficient video representation. In this paper, a novel multi-level representation of video is proposed based on the principle components derived from low-level visual features. It can characterize the video content from the coarse level to the fine level according to its intrinsic structure. This representation form provides a flexible scheme for video content analysis such as summarization, classification, and retrieval. A newly proposed subspace method, kernel based PCA, is explored to achieve this conveniently. The application in keyframe extraction is investigated to demonstrate the benefits of this representation.*

## 1. Introduction

For content analysis, video is often characterized by low-level visual features based on color, texture, shapes, motion, etc [1]. A video sequence can then be modeled as a set of points in the space spanned by these visual feature vectors. In this scenario, video representation turns to the characteristics of these points [2-4]. The technique discussed in this paper falls into this category.

Ideally, compact while informative video representation will facilitate fast and convenient browsing, indexing and retrieval. However, compact and informative are two conflicting demands: too compact representation faces the risk of losing some important information while too informative representation will result in redundancy. Finding a trade-off between compact and informative can be considered as a way, however, this is a tough problem in practice and the trade-off is not necessary to be consistent with the requirements of video analysis and the users. To achieve efficient and comprehensive content analysis, we argue that multi-level video representation (MLVR) is a practical alternative to avoid this problem. Under this scheme, video is represented from the coarse level to the fine level, according to its intrinsic structure, to satisfy different requirements on the properties of compact and informative. Video content analysis such as summarization, classification, and retrieval becomes flexible. The objective of this paper is to implement a MLVR scheme and demonstrate its application to keyframe extraction. Kernel-based Principal Component Analysis (KPCA) is explored to achieve this conveniently.

In literature, algorithms employing multi-level representation for video analysis have been studied for various applications. For video retrieval, Ngo et al [16] reported a two-level video clustering algorithm; Day et al [15] proposed a multi-level framework for abstraction and modeling in video databases. In both approaches, each level employs one or more distinct features and the number of levels is fixed. Thus such schemes are not applicable for applications that demand for flexible representation level. In applications of keyframe extraction, hierarchical structure has been employed by many researchers and such hierarchical keyframe can be applied to MLVR. For example, Uchihashi et al [17] reported a bottom-up iterative merging approach to build a tree-structured representation for video summary. Kobla et al [2] described a video as a curve in the feature space and applied curve simplification algorithm to this curve. The simplified curve can be represented as a tree structure and used to extract keyframes at different level of details. Nevertheless, a nontrivial weakness of these hierarchical algorithms is that only local video structure is taken into account while the resulting video representation is far from optimal from the global point of view.

Subspace method is an efficient way to capture the principal characteristic of the data in a high-dimensional space. Thus it can generate a globally optimized concise representation for the data. In video domain, several its variations, such as principal component analysis (PCA) and singular value decomposition (SVD), have been used as an attempt to capture the principal structure of a video [5-7]. However, these methods cannot give satisfactory multi-level representation due to the following fact. In both PCA and SVD, the transformation from the original space to the subspace is linear only. When the structures of the data present nonlinear property, PCA and SVD cannot effectively capture them therein. Consequently, it is found that in this case neither of them can offer sufficiently coarse and fine structures. For the video data represented by

the low-level visual feature, it may not keep a linear structure. Hence, a more powerful method has to be used instead.

Recently, kernel based PCA (KPCA) has been proposed as a new subspace method [8]. In KPCA, by employing the kernel trick, the given data are nonlinearly mapped from the feature space into a higher-dimensional kernel space, and linear PCA is performed there afterwards. It has been reported that, compared with linear PCA, KPCA has more powerful capability in representing nonlinear patterns [8,14]. This is expected to be helpful for exploring the nonlinear structure in video data. As a result, more efficient MLVR can be produced. In this paper, a novel MLVR scheme based on KPCA is proposed and realized. By using KPCA, the coarse and fine structures in the video are efficiently characterized by the components extracted from the feature space. The application to keyframe extraction is also investigated to demonstrate the benefit of this representation form.

The remainder of this paper is organized as follows. Section 2 describes the algorithm of Kernel PCA briefly. Section 3 proposes a MLVR scheme based on KPCA. Section 4 presents its application in extracting keyframe. Experimental results are shown in Section 5. Finally, we draw a conclusion and discuss the future work in Section 6.

## 2. Kernel Principal Components Analysis

KPCA performs linear PCA in a high-dimensional kennel space rather than the original feature space. Given a data set $\boldsymbol{X}=\{\boldsymbol{x_i} \in \boldsymbol{R^N} \mid i=1,\ldots, n\}$, KPCA maps $\boldsymbol{X}$ into a kernel space, $\boldsymbol{F}$, by an nonlinear mapping, $\Phi$, associated with a given kernel function, $k$, where $k(x_i, x_j) = (\Phi(x_i), \Phi(x_j))$ and ( , ) denotes the dot product.

$$\Phi : \boldsymbol{R^N} \rightarrow \boldsymbol{F}.$$

The linear PCA problem in $\boldsymbol{F}$ finds the eigenvectors, $\boldsymbol{V(V \in F)}$, and the corresponding eigenvalues, $\lambda$ ( $\lambda \geq 0$), that satisfy

$$\lambda \mathbf{V} = \hat{C}\mathbf{V} \tag{1}$$

where

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^T . \tag{2}$$

Scholkopf et al [8] reformulate the problem for (1) as

$$\mathrm{n}\lambda\boldsymbol{\alpha} = K\boldsymbol{\alpha} , \tag{3}$$

where $\boldsymbol{\alpha}$ denotes the column vector with entries $\alpha_1, \ldots \alpha_n$ such that

$$\mathbf{V} = \sum_{i=1}^{n} \alpha_i \Phi(\mathbf{x}_i) \tag{4}$$

and $K$ is a symmetric $n$ x $n$ kernel matrix where

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) . \tag{5}$$

By solving the eigen problem in (3), $\boldsymbol{\alpha}$ can be obtained, and the $k$-th principal component (PC) of a datum, $\boldsymbol{x}$, is

$$PC_k(x) = (\mathbf{V}^k \cdot \Phi(\mathbf{x})) = \sum_{i=1}^{n} \alpha_i^k k(\mathbf{x}_i, \mathbf{x}), \tag{6}$$

where $1 \leq k \leq n$. The obtained subspace is the one spanned by the eigenvectors corresponding to the larger eigenvalues. These eigenvectors provide an orthogonal basis for $\boldsymbol{X}$ and the eigenvalue represent the importance of the corresponding PC. Therefore, the obtained PCs can be used to represent the given data in a subspace of $\boldsymbol{F}$. Such representation is better than the one in the original feature space as the former contains explicit structure information that facilitates the later analysis. Besides the advantage in capturing nonlinear structure, KPCA can produce up to $n$ (the sample size) PCs to describe the given data. However, linear PCA can only offer $N$ (the dimensionality of visual feature space) components. Considering that $n$ is usually larger than $N$ in video data, KPCA is expected to give both coarser and finer structures than linear PCA.

## 3. Muti-Level Video Representation with KPCA

Considering the capability of KPCA aforementioned, we apply KPCA to the video domain to achieve the proposed MLVR scheme. The principle components obtained from the low-level visual feature space are used to build up this scheme. By changing either the kernel parameters or the dimensionality of the subspace, two ways to achieve MLVR can be obtained, and either has its own characteristic and advantage.

Kernel is the soul of KPCA and its parameters have significant impact on the result. Gaussian Radial Basis Function (RBF) kernel function is one of the commonly used kernels, and it is defined as

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\| \mathbf{x} - \mathbf{y} \|^2}{2\sigma^2}), \tag{7}$$

where $\sigma$ is a parameter known as the "width", and it has to be set beforehand [8]. The magnitude of $\sigma$ affects the nonlinearity of the mapping, $\Phi$, associated with the kernel, $k$. Commonly, a smaller $\sigma$ results in higher nonlinearity while a larger $\sigma$ corresponds to lower one. As mentioned before, the nonlinearity of the mapping affects the capability of KPCA on capturing the coarse structure of data if it is nonlinear. Hence, the coarseness

level of the structure in a fixed lower-dimensional subspace can be adjusted by changing the magnitude of $\sigma$. This is an elegant property that the structure of video can be clearly shown from coarse level to fine level in a fixed dimensional subspace. The potential advantage is that the analysis results in different levels can be obtained by performing video analysis and processing in the same dimensional subspaces. This is very helpful to the algorithms sensitive to the dimensionality of subspace. The following example illustrates the claimed property explicitly.

Figure 1 (a) to (c) show plots of the first two PCs of the video "news" produced by KPCA with $\sigma$=0.10, 1.00 and 16.0 respectively, in which the visual feature is color histogram extracted from the video as in (8). For comparison, the first two PCs produced by linear PCA are plotted in Figure 1 (d). In these plots, the frames with similar color histogram are represented by points close to each other and vice versa. The temporal relationship between the frames is represented by the lines connecting the points. The video "news" begins with an interview with Mr. Blair (shot 1:1-197), then an anchorperson shot (shot 2:198-583) followed by a series of Blair's activities: the Blairs with children (shot 3: 584-655, shot 4: 656-714, shot 5:715-783), the Downing Street (shot 6: 784-834), Blair came out of his office (shot 7: 835-985) and the meeting room (shot 8: 986-1034). The descriptions of the camera motion are summarized in Table 1. As expected, when $\sigma$ is very small, e.g. 0.10 in Figure 1 (a), the PCs given by KPCA show the coarser structure of this video with three clusters corresponding to shot 1, 2 and 3-8 respectively. With the increase of $\sigma$, finer structures are disclosed in Figure 1 (b) and Figure 1 (c) gradually: shots are separated first and then the internal activities in each shot. However, only finer structure in the video can be seen in the plot of PCs given by linear PCA (Figure 1 (d)). It is similar to that in Figure 1 (c), where $\sigma$ = 16.0 and the mapping is less nonlinear for the given data set.

The other way to achieving MLVR is to increase the dimensionality of the subspace while fixing the $\sigma$ value. It is known that in KPCA, the magnitude of an eigenvalue represents the importance of the corresponding component. The subspace spanned by the first $M$ PCs represents the video structure in a certain level of details. The larger the value of $M$, the finer the represented structure. The benefit of this way is that we only need perform KPCA once to achieve multi-level representation and the computational load is less. If we change the dimension of the linear subspace method, say, linear PCA, multi-level can also be displayed. However, due to the limitation of the linear transformation, the resolution is limited in scale, and the video structure obtained can neither coarser nor finer than those given by KPCA. The influence of $M$ can be demonstrated clearly in the following example.

Figure 2 demonstrates the influence of $M$. The video is the same as above. We set $\sigma$=0.10 and plot the first two and three PCs given by KPCA in Figure 2 (a) and (b) respectively. The cases from linear PCA are also plotted in Figure 2 (c) and (d) for comparison. It can be seen that the coarser structure of the video is shown with three clusters in the subspace produced by KPCA, where $M$ =2. When $M$ increases to 3, five clusters are obtained and the finer structure is shown. But the influence of $M$ on the linear PCA is not as significant as that on KPCA because the structure represented by PCs given by linear PCA has already been very fine. Note that although the cases of higher dimensionality are not plotted due to inconvenience, similar comparison results can be expected.

In summary, we proposed a MLVR scheme based on KPCA in this section. Thanks to the capability of KPCA, we achieve MLVR conveniently by changing either $\sigma$ or $M$. Such representation facilitates many video analysis applications, such as video summary, video indexing and retrieval, etc. A MLVR based keyframe extraction application is presented in the next section to demonstrate its benefits.

## 4. Muti-Level Keyframe Extraction

After obtaining MLVR with KPCA, the structure of the video has been described efficiently in multiple levels. In this section, we demonstrate a multi-level keyframe extraction algorithm, which can benefit directly from the proposed MLVR scheme. Figure 3 illustrates the overall diagram of the proposed algorithm. It is detailed as follows.

Firstly, we extract the color histogram from the MPEG video as the feature and KPCA is performed thereafter. The color histogram in our algorithm is constructed from the DC-image, which is recovered from the MPEG stream by Yeo's method [9]. It is composed of 24 bins, in which 16 for Y component and 4 for Cb and Cr components respectively. For the $i$-th frame, the color histogram $H_i$ is

$$H_i = \{\ h_Y(m), m = 1-16,\ h_{Cb}(n), n = 1-4,\ h_{Cr}(l), l = 1-4\ \}. \quad (8)$$

The color histogram of a MPEG sequence form a data set $\boldsymbol{H}$ $\{\boldsymbol{H_i} \in \boldsymbol{R^{24}} \mid i=1,\ldots, n\}$ and $n$ is the number of frame in the sequence. KPCA is performed in $\boldsymbol{H}$ and the first $M$ components, $\boldsymbol{PC}$=$\{PC(x_i)_k \mid i=1,\ldots,n,\ k=1,\ldots, M\}$ are computed by (6).

Then, Fuzzy c-means clustering [10] is used to find the cluster structures in the subspace spanned by the first $M$ components. Since the cluster number $c$ is unknown *a priori*. We employ cluster validity analysis to find the optimal cluster number [11] among integers from 1 to $N$. The maximum cluster number $N$ for a video with length $S$ is defined as [12]

$$N = N(S) = 10 + round\left(\frac{S}{25}\right). \quad (9)$$

Finally, a keyframe is extracted from each cluster by finding the frame closest to the cluster centroid $c_i$

$$K_i = \min_{1 \le j \le n} \left\| PC(j \mid j \in i) - PC(c_i) \right\|. \quad (10)$$

Multi-level keyframe extraction can be achieved by changing $\sigma$ or $M$. The procedures of FCM, cluster validity analysis and keyframe extraction are performed in a multi-pass manner and each pass produces a keyframe set in a certain level with specified $\sigma$ and $M$.

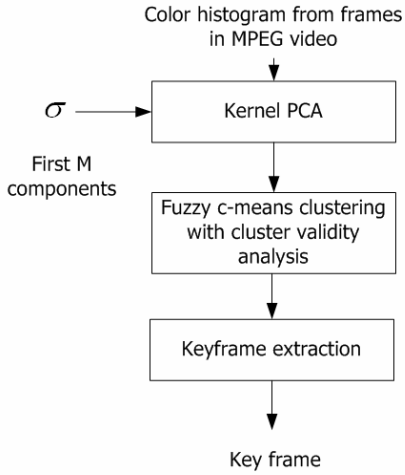



Figure 3. Multi-level keyframe extraction based on MLVR

## 5. Experimental Results

We conduct experiments on three video clips taken from MPEG-7 test video set. They are in MPEG-1 format with size 352 x 240, frame rate 29.97fps and CBR 1150kbps. Their characteristic descriptions are summarized in Table 1. In all experiments, we set m=2.0, $\varepsilon$=1.0e-5 and $L$ = 100 for the FCM clustering algorithm.

Firstly, we test the influence of $\sigma$ on performance of keyframe extraction. Figure 4 shows the keyframes extracted from the video "news" with KPCA, where $\sigma$=0.10, 1.00 and 16.00 respectively. Keyframes extracted with linear PCA are also shown for comparison. In all these experiments, we set $M$=2. By changing $\sigma$, we obtained keyframes which describe the video content from coarse to fine level while linear PCA can only provide fine level results. This is consistent with the discussion of the influence of $\sigma$ on MLVR in Section 3. Table 2 lists the results of the other two videos and they are similar to that of the video "news".

Then we investigate the influence of $M$ on extracting keyframes. Figure 5 shows the keyframes of video "news" with KPCA where $\sigma$=0.10 while $M$=2,3, and 8 respectively. Again, the cases for linear PCA are presented for comparison. In Section 3, we have plotted the PCs with $M$=2 and 3. Although we cannot conveniently plot the PCs when $M$>3, the keyframes with $M$>3 can be easily extracted for comparison. It can be seen that $M$ has more significant impact on keyframe extraction with KPCA than the on with linear PCA. The keyframes obtained with linear PCA almost remain unchanged for all $M$ while those with KPCA present multi-level structure. The test results on the other two sequences are also listed in Table 3 and similar conclusion can be drawn.

From above test results, we can see that multi-level keyframe extraction is greatly simplified when it is performed on the proposed MLVR scheme. The proposed MLVR has clearly represented the intrinsic structure of the video data, and the consequent video processing and analysis, e.g., extracting keyframe, in different levels becomes much easier.

## 6. Conclusion and future work

In this paper, a multi-level video representation scheme is proposed. KPCA is employed to achieve this representation and its benefits on keyframe extraction are demonstrated.

Nevertheless, there are still several open issues in the proposed algorithm. Firstly, KPCA solves the eigen problems of the kernel matrix whose size is directly proportional to the square of $n$, which corresponds to the number of frames in the given video sequence in this paper. This limits the application of KPCA in longer video sequence. Several ways have the potential to remove this obstacle. For example, the number of frames used in KPCA can be reduced by subsampling or setting up MLVR at shot level [13]. Computation complexity can be reduced significantly if we compute only the top several eigenvectors [8]. They will be further verified in our future work. Secondly, tuning of $\sigma$ and $M$ for a given application should be investigated in the future.

We demonstrated an application of the proposed MLVR scheme to keyframe extraction in this paper. It can be extended to more applications, such as video index and retrieval and we have included them in our future work.

## Acknowledgements

This work is supported by Agency for Science, Technology and Research, Singapore.

## References

[1] P.Aigrain, "Content-based Representation and Retrieval of Visual Media: A State-of-the-art Review", *Multimedia Tools and Applications*, Vol.3, No.3, pp179-202, 1996.

[2] V.Kobla, D. Doermann, and C. Faloutsos, "Developing High-Level Representations of Video Clips using *VideoTrails*", *Proc. of the SPIE Conference on Storage and Retrieval for Still Image and Video Databases VI*, Vol. 3312, pp. 81--92, 1998.

[3] H.S.Chang, S.S.Sull, and S.U.Lee, "Efficient Video Indexing Scheme for Content-Based Retrieval", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.9, No.8, 1999.

[4] L. Zhao, W. Qi, S.Z.Li, S.Q.Yang, and H.J.Zhang, "Keyframe Extraction and Shot Retrieval Using Nearest Feature Line", *International Workshop on Multimedia Information Retrieval*, 2000.

[5] E. Sahouria and A. Zakhor, "Content Analysis of Video Using Principal Components", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 9, No.8, December 1999.

[6] C.Y.Chang, A.A.Maciejewski, and V.Balakrishnan, "Eigendecomposition-Based Analysis of Video Images", *Proc. Storage and Retrieval for Image and Video Database*, VII, pp186-191, 1999.

[7] Y.Gong and X.Liu, "Video Summarization Using Singular Value Decomposition", *Computer Vision and Pattern Recognition*, Vol.2 , June 13 - 15, 2000.

[8] B.Scholkopf, A.J.Smola, K.R.Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", *Neural Computation*, Vol. 10, pp1299-1319, 1998.

[9] Boon-Lock Yeo, "On Fast Microscopic Browsing of MPEG-compressed Video", *Multimedia Systems*, Vol.7, pp 269-281, 1999.

[10] James C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press 1981.

[11] Anil K. Jain, *Algorithms for Clustering Data*, Englewood Cliffs, N.J. : Prentice Hall , 1988.

[12] A. Hanjalic and H.J.Zhang, "An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.9, No.8, pp1280-1289, December 1999.

[13] A.M.Ferman, A.M.Tekalp, "Two-Stage Hierarchical Video Summary Extraction to Match Low-Level User Browsing Preferences", *IEEE Transactions on Multimedia*, June 2003.

[14] Kwang In Kim, Keechul Jung, and Hang Joon Kim, "Face Recognition Using Kernel Principal Component Analysis", *IEEE Signal Processing Letters*,Vol.9, No.2, pp40-42, February 2002

[15] Young Francis Day, Ashfaq Khokhar, Serhan Dagtas, Arif Ghafoor, "A Multi-Level Abstraction And Modeling In Video Databases", *Multimedia Systems*, Vol.7, pp409–423,

[16] 1999C.W.Ngo, T.C.Pong and H.J.Zhang, "On Clustering and Retrieval of Video Shots Through Temporal Slices Analysis", *IEEE Transactions on Circuits and Systems for Video Technology*, pp446 – 458, Vol.4, No.4, December 2002

[17] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries", *Proceedings ACM Multimedia*, Orlando, FL, pp.383-392, 1999. .

Table 1. Characteristic description of the testing videos

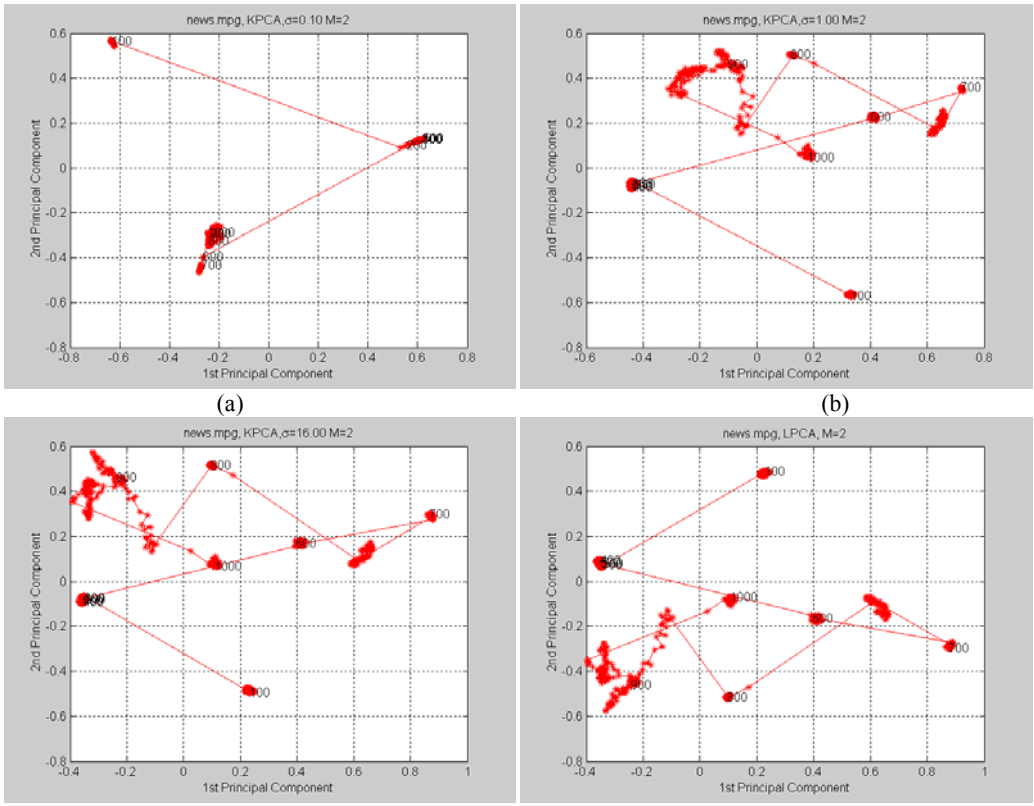
| Name | #Frame | Duration | #Shot | Camera Motion |
|---|---|---|---|---|
| News.mpg | 1034 | 41s | 8 | 1-197, 198-583 :medium shot, static<br>584-655:long shot, static<br>656-714, 715-783:medium shot, static<br>784-834:long shot, static<br>835-985: medium shot, following<br>986-1034, long shot, static |
| Golf.mpg | 951 | 32s | 2 | 1-141: long shot, static<br>142-311:closeup<br>312-485:zoom out<br>486-887:medium shot, static<br>888-951:zoom in |
| Basketball.mpg | 1046 | 35s | 4 | 1-155: medium shot, following<br>156-884:long shot, panning<br>885-951, 952-1046 : medium shot, following |

Table 2. Keyframe number obtained with changing of $\sigma$

| Name | #shot | KPCA, $\sigma$= 0.10, $M$=2 | KPCA, $\sigma$= 1.00, $M$ =2 | KPCA,$\sigma$= 16.00, $M$ =2 | LPCA, $M$ =2 |
|---|---|---|---|---|---|
| News.mpg | 8 | 3 | 8 | 13 | 11 |
| Golf.mpg | 2 | 3 | 6 | 12 | 5 |
| Basketball.mpg | 4 | 3 | 6 | 13 | 7 |

Table 3. Keyframe number obtained with changing of $M$

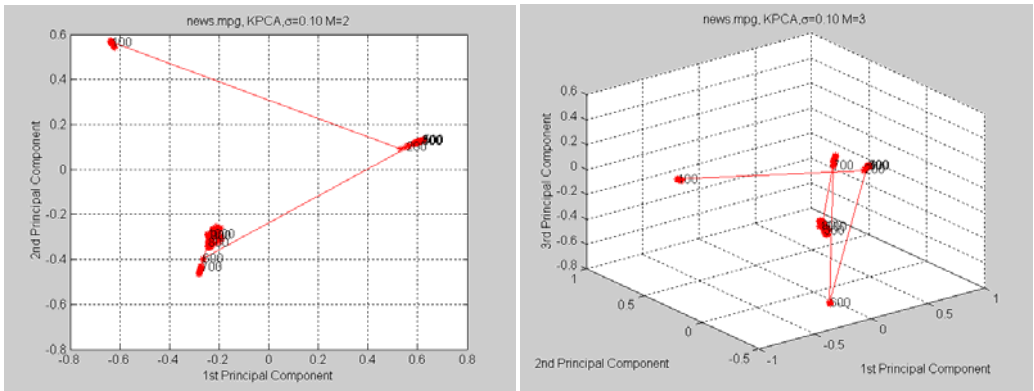
| Name | #shot | KPCA, $\sigma$= 0.10, $M$ =2 | KPCA, $\sigma$= 0.10, $M$ =3 | KPCA, $\sigma$= 0.10, $M$ =8 | LPCA, $M$ =2 | LPCA, $M$ =3 | LPCA, $M$ =8 |
|---|---|---|---|---|---|---|---|
| News.mpg | 8 | 3 | 5 | 10 | 11 | 12 | 13 |
| Golf.mpg | 2 | 3 | 4 | 8 | 5 | 6 | 6 |
| Basketball.mpg | 4 | 3 | 4 | 12 | 7 | 7 | 7 |



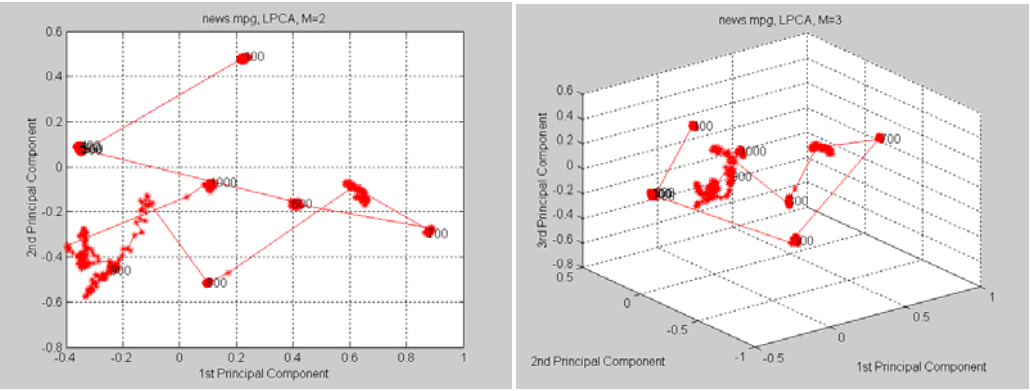

(a) (b)

(c) (d)

Figure 1. Plots of the first 2 PCs for news.mpg. (a) – (c) plot the PCs produced by KPCA, where in (a) $\sigma$=0.10, in (b) $\sigma$=1.00 and in (c) $\sigma$=16.00. (d) plots the PCs produced by linear PCA. All components have been normalized into (-1, 1).




(a) (b)




(c) (d)

Figure 2. The plot of the PCs of news.mpg. (a) and (b) plot the PCs produced by KPCA, where $\sigma$=0.10 , in (a) $M$=2 and in (b) $M$=3. (c) and (d) plot the PCs produced by linear PCA, where in (c) $M$=2 and in (d) $M$=3.

| | |
|---|---|
| KPCA<br>$\sigma$ = 0.10<br>$M$=2 |  |
| KPCA<br>$\sigma$ = 1.00<br>$M$=2 |  |
| KPCA<br>$\sigma$ = 16.00<br>$M$=2 |  |
| LPCA<br>$M$=2 |  |

Figure 4. The influence of $\sigma$ on keyframe extraction. All keyframes are produced by FCM with cluster validity analysis in the 2-dimension subspace, i.e., $M$=2. The first three rows are those extracted with KPCA, where $\sigma$ = 0.10, 1.00 and 16.00 respectively and the last row is those produced with linear PCA.


Figure 5. The influence of $M$, dimensionality of subspace, on keyframe extraction. The first three rows are keyframes produced with KPCA , where $\sigma$=0.10 and $M$=2, 3, 8 respectively and the last three rows are those with linear PCA, where $M$=2, 3, 8 respectively.