Visualization and Clustering of Crowd Video Content in MPCA Subspace

Haiping Lu, How-Lung Eng, Myo Thida
Institute for Infocomm Research
Agency for Science, Technology and Research
1 Fusionopolis Way, Singapore
{hlu,hleng,mthida}@i2r.a-star.edu.sg

Konstantinos N. Plataniotis
Dept. of Electrical & Computer Engineering
University of Toronto
10 King's College Road, Toronto, ON, Canada
kostas@comm.utoronto.ca

ABSTRACT

This paper presents a novel approach for the visualization and clustering of crowd video contents by using multilinear principal component analysis (MPCA). In contrast to feature-point-based approach and frame-based dimensionality reduction approach, the proposed method maps each short video segment to a point in MPCA subspace to take temporal information into account naturally through tensorial representations. Specifically, MPCA projects each short segment of a video to a low-dimensional tensor first. A few MPCA features are then selected according to the variance captured as the final representation. Thus, a video is visualized as a trajectory in MPCA subspace. The trajectory generated enables visual interpretation of video content in a compact space as well as visual clustering of video events. The proposed method is evaluated on the PETS 2009 datasets through comparison with three existing methods for video visualization. The MPCA visualization shows superior performance in clustering segments of the same event as well as identifying the transitions between events.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms,Experimentation, Human Factors, Management

Keywords

Visualization, clustering, crowd, video analysis, MPCA

1. INTRODUCTION

Recent proliferation of surveillance cameras has led to a strong demand for automatic data processing tools for the enormous amount of video data generated in applications such as surveillance, healthcare, and ambient intelligence [8, 10, 4]. While earlier works focused on videos with single or just a few subjects [8, 7], the analysis of crowd video content starts to attract more and more attentions recently [4].

A popular approach to video analysis is to detect and track a set of feature points [7], which has been quite successful for videos with single or just a few subjects. However, this approach is sensitive to noise and the performance is heavily dependent on the detection and tracking modules. Furthermore, when there is a crowd in the scene, it will be extremely difficult to detect and track individuals. On the other hand, it might not be necessary to know the movement of each individual, instead, a characterization of the crowd movement as a whole would provide insights on the scene.

To avoid object detection and tracking, recent approach considers frames of a video as a set of images with each frame (image) as a basic video element. In [8], video is considered as a collection of unordered images and an image space is defined through Isomap [9]. Thus, video sequences specify a trajectory through that image space. Similarly in [10], input frames are represented in a low-dimensional space using Laplacian Eigenmaps [1], where a graph is defined based on similarity of frames. Since the video sequences used in [10] consist of video events well separated by the so-called no event, a rule-based temporal graph is further introduced to incorporate temporal information.

Instead of embedding the raw frames, another approach is to extract motion patterns through the calculation of optical flow. In [11], a video sequence is divided into video segments (or clips) and then each segment is split into several cuboids. A "video word" representation is obtained for each video segment by concatenating histograms of optical flow fields for all the cuboids in the segment. Diffusion map [3] is employed to embed the motion pattern information.

In this work, we consider a short video segment as a basic video element represented as a tensor (a multidimensional array) and propose to use multilinear principal component analysis (MPCA) [5] to embed each video segment into more compact manifolds for analysis. The proposed approach is less sensitive to noise and it operates on raw video sequences, without background subtraction, foreground segmentation, or silhouette extraction. Video content can be visualized as a trajectory in MPCA subspace and the visual cluster rendering scheme in [2] can further be adopted for interpretation. The proposed solution is evaluated on sequences from the Performance Evaluation of Tracking and Surveillance (PETS) 2009 [4] against three other methods to show its superior performance.

CIKM'10, October 25–29, 2010, Toronto, Ontario, Canada.

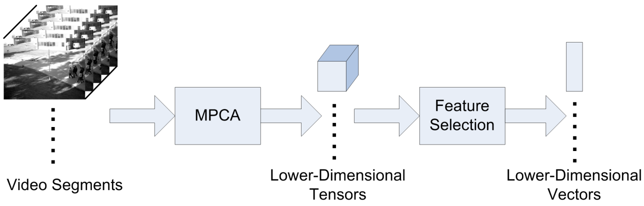


Figure 1: Illustration of the proposed MPCA-based approach for crowd video content analysis.

2. VIDEO ANALYSIS IN MPCA SUBSPACE

Figure 1 illustrates the proposed MPCA-based approach for crowd video content analysis. Each segment is represented naturally as a third-order tensor with three modes (row, column and time). Input video segments are first mapped to low-dimensional tensors by MPCA. Then, a few MPCA features are selected to obtain low-dimensional vectors for final representation. Since we treat a video segment as a basic video element, we analyze the input video in overlapping segments by using an observation window of L frames. This window is shifted by S frames in each step. Thus, for a video with H frames in total, the number of overlapping segments will be

$$M = \lfloor (H - L + S)/S \rfloor, \tag{1}$$

where $\lfloor \cdot \rfloor$ denotes the floor operation.

2.1 Notations

In this paper, vectors are denoted by lowercase boldface letters, e.g., $\mathbf{x}$; matrices by uppercase boldface, e.g., $\mathbf{U}$; and tensors by calligraphic letters, e.g., $\mathcal{A}$. Their elements are denoted with indices in brackets. Indices are denoted by lowercase letters and span the range from 1 to the uppercase letter of the index, e.g., $n = 1, ..., N$. An N^{th}-order tensor is denoted as $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$. It is addressed by N indices i_n, $n = 1, ..., N$, and each i_n addresses the n-mode of $\mathcal{A}$. The n-mode product of a tensor $\mathcal{A}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is a tensor with entries [5]:

$$(\mathcal{A} \times_n \mathbf{U})(i_1, ..., i_{n-1}, j_n, i_{n+1}, ..., i_N) = \sum_{i_n} \mathcal{A}(i_1, ..., i_N) \cdot \mathbf{U}(j_n, i_n). \tag{2}$$

A rank-1 tensor $\mathcal{A}$ equals to the outer product of N vectors:

$$\mathcal{A} = \mathbf{u}^{(1)} \circ ... \circ \mathbf{u}^{(N)}, \tag{3}$$

which means that $\mathcal{A}(i_1, ..., i_N) = \mathbf{u}^{(1)}(i_1) \cdot ... \cdot \mathbf{u}^{(N)}(i_N)$ [5].

2.2 MPCA feature extraction and selection

MPCA [5] is a multilinear subspace learning method that extracts features directly from tensorial representation of multi-dimensional objects. In [5, 6], MPCA is proposed for gait recognition by representing each half cycle of gait silhouette sequences as a third-order tensor. In this paper, we apply MPCA to crowd video analysis by extracting MPCA features from raw video sequences.

As described in the beginning of this section, M overlapping video segments are obtained first from the video to be analyzed. They are represented as M third-order tensors $\{\mathcal{X}_1, ..., \mathcal{X}_M \in \mathbb{R}^{I_1 \times I_2 \times I_3}\}$ ($I_3 = L$), as the input to MPCA. The MPCA algorithm solves for a multilinear projection

$$\{\tilde{\mathbf{U}}^{(n)} \in \mathbb{R}^{I_n \times P_n}, n = 1, 2, 3\}, \tag{4}$$

where $P_n < I_n$ for $n = 1, 2, 3$, that maps the original video tensor space $\mathbb{R}^{I_1} \bigotimes \mathbb{R}^{I_2} \bigotimes \mathbb{R}^{I_3}$ into a lower-dimensional tensor subspace $\mathbb{R}^{P_1} \bigotimes \mathbb{R}^{P_2} \bigotimes \mathbb{R}^{P_3}$:

$$\mathcal{Y}_m = \mathcal{X}_m \times_1 \tilde{\mathbf{U}}^{(1)^T} \times_2 \tilde{\mathbf{U}}^{(2)^T} \times_3 \tilde{\mathbf{U}}^{(3)^T}, m = 1, ..., M, \tag{5}$$

such that the total tensor scatter $\Psi_{\mathcal{Y}} = \sum_{m=1}^{M} \parallel \mathcal{Y}_m - \bar{\mathcal{Y}} \parallel_F^2$, is maximized, where $\bar{\mathcal{Y}} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{Y}_m$ is the projection of the mean sample (i.e., the average of the M tensorial samples). This MPCA problem is solved through an iterative alternating projection method in [5].

The MPCA projection matrices $\{\tilde{\mathbf{U}}^{(n)}, n = 1, 2, 3\}$ can be viewed as $\prod_{n=1}^{3} P_n$ EigenTensors [5]:

$$\tilde{\mathcal{U}}_{p_1 p_2 p_3} = \tilde{\mathbf{u}}_{p_1}^{(1)} \circ \tilde{\mathbf{u}}_{p_2}^{(2)} \circ \tilde{\mathbf{u}}_{p_3}^{(3)}, \tag{6}$$

where $\tilde{\mathbf{u}}_{p_n}^{(n)}$ is the p_n^{th} column of $\tilde{\mathbf{U}}^{(n)}$. Since our objective is to perform unsupervised analysis on crowd video content, in particular visualization, only the first a few most important EigenTensors are needed. Therefore, we further perform a feature selection based on an importance score $\Upsilon_{p_1 p_2 p_3}$ calculated from the variation captured in each EigenTensor. $\Upsilon_{p_1 p_2 p_3}$ for the eigentensor $\tilde{\mathcal{U}}_{p_1 p_2 p_3}$ is defined as

$$\Upsilon_{p_1 p_2 p_3} = \sum_{m=1}^{M} \left[\mathcal{Y}_m(p_1, p_2, p_3) - \bar{\mathcal{Y}}(p_1, p_2, p_3) \right]^2 \tag{7}$$

where $\mathcal{Y}_m$ is the projection of $\mathcal{X}_m$ in MPCA subspace, and $\bar{\mathcal{Y}}$ is the mean feature tensor defined above.

For the EigenTensor selection, the entries in $\mathcal{Y}_m$ are arranged into a feature vector $\mathbf{y}_m$ according to $\Upsilon_{p_1 p_2 p_3}$ in descending order. Only the first D entries of $\mathbf{y}_m$ are kept for subsequent analysis.

2.3 MPCA-based visualization and clustering

The selected D features allow us to embed each video segment to a D-dimensional MPCA subspace, resulting a trajectory for the video. Since the raw video segments are projected, the relative positions between data points in MPCA subspace will be closely related to the visual changes between the corresponding video segments. Small changes will result in closely spaced points while abrupt changes will lead to points far apart. In the case of crowd video content, the spacing between points in MPCA subspace tends to be affected mainly by two factors: crowd size and activity speed. Generally, small crowd with slow activity results in very closely spaced data points while big crowd with fast activity results in large spacing between data points. Therefore, we expect the MPCA subspace embedding to provide a good visualization of the video content. Two key benefits of the proposed approach are summarized below:

1. The raw video segments are taken as the input so there is no need for object detection, silhouette extraction, and object tracking. In addition, the method is robust against noise, occlusion or shadows.

2. Through representing video segments in their inherent 3D form using tensors, the temporal information is naturally taken into consideration in MPCA subspace. The MPCA projection maps a spatial-temporal volume to a point in a subspace. In contrast, in the frame-based approach [8, 10], each frame is mapped to a point based on image similarity so the distance between points reflects only the similarity between two

frames, while temporal information is not being considered in such mapping.

Benefiting from human cognitive abilities, visualization of patterns is a practical way to gain insight into large databases [2]. As MPCA-based visualization can provide us cues on the characteristics of the video content in a compact form, human subjects can perform various tasks by observing and interpreting the visualization of a video rather than the video itself, such as anomaly detection, video summarization and clustering of video events. In the next section, experiments on real-world crowd video sequences are carried out to demonstrate the superiority of the MPCA-based approach in visualization of crowd video content for summarization and visual clustering of events.

3. EXPERIMENTAL EVALUATION

In this section, several experiments on real-world sequences from surveillance cameras are performed to show the effectiveness of our proposed method in video content visualization and event clustering. Different from commercial movies, video sequences from surveillance cameras usually contain a smooth change of motion in the frames rather than frequent scene cuts and rapid changes [10]. Therefore, we expect a smooth manifold of motion over time for these sequences.

3.1 Experimental data and design

The proposed method is evaluated on the dataset S3 "Event Recognition" of the PETS 2009 database[1] with dense crowd and subjective difficulty of L3 (the most difficult) [4]. There are four sequences with time stamps 14-16, 14-27, 14-31 and 14-33. Due to space limitation, only results on sequence 14-27 from view 001 are reported in this paper. There are 334 frames in this sequence ($H = 334$). In simulations, the original sequence is down-sampled for lower computational cost. The original color sequences are converted to gray-level and each frame is resized to 192 × 144 pixels. The sequences are recorded at 7 frames per second. Figure 2(d) shows four frames from this sequence. For performance evaluation, we compare the visualization of sequence Time1427 from PETS 2009 in a two-dimensional (2D) space using four methods with the following settings.

Isomap-based method [8] and **Laplacian-Eigenmap-based method** [10]: the distances between each frame pair of a sequence are calculated and for each frame, the distance for the k nearest neighbors are kept. Four values of $k = 4, 6, 8, 10$ are tested and the best results are reported.

Diffusion-map-based method [11]: a video is divided into overlapping segments, each with $L = 5$ frames. Two successive segments are overlapped by 4 frames ($S = 1$) so that almost every frame will have a corresponding segment for easy comparison with the first two methods. $8 \times 8 \times 5$ cuboids are obtained for each segment and 4-bin histograms of optical flow are computed for each cuboid, which are concatenated to form "video words". The frequency of each video word in different segments is normalized to obtain probability, with 10% of the video words with the highest and lowest conditional entropy are discarded. Four sets of diffusion map parameters (t, σ) are tested: $(2, 8)$, $(2, 10)$, $(4, 6)$ and $(8, 5)$. The best results are reported.

The proposed **MPCA-based method**: overlapping segments are first obtained with $L = 5$ and $S = 1$ as in the Diffusion-map-based method for easy comparison. They are then embedded in a 2D MPCA subspace ($D = 2$) following the process described in Section 2 with the same setting for MPCA in [5].

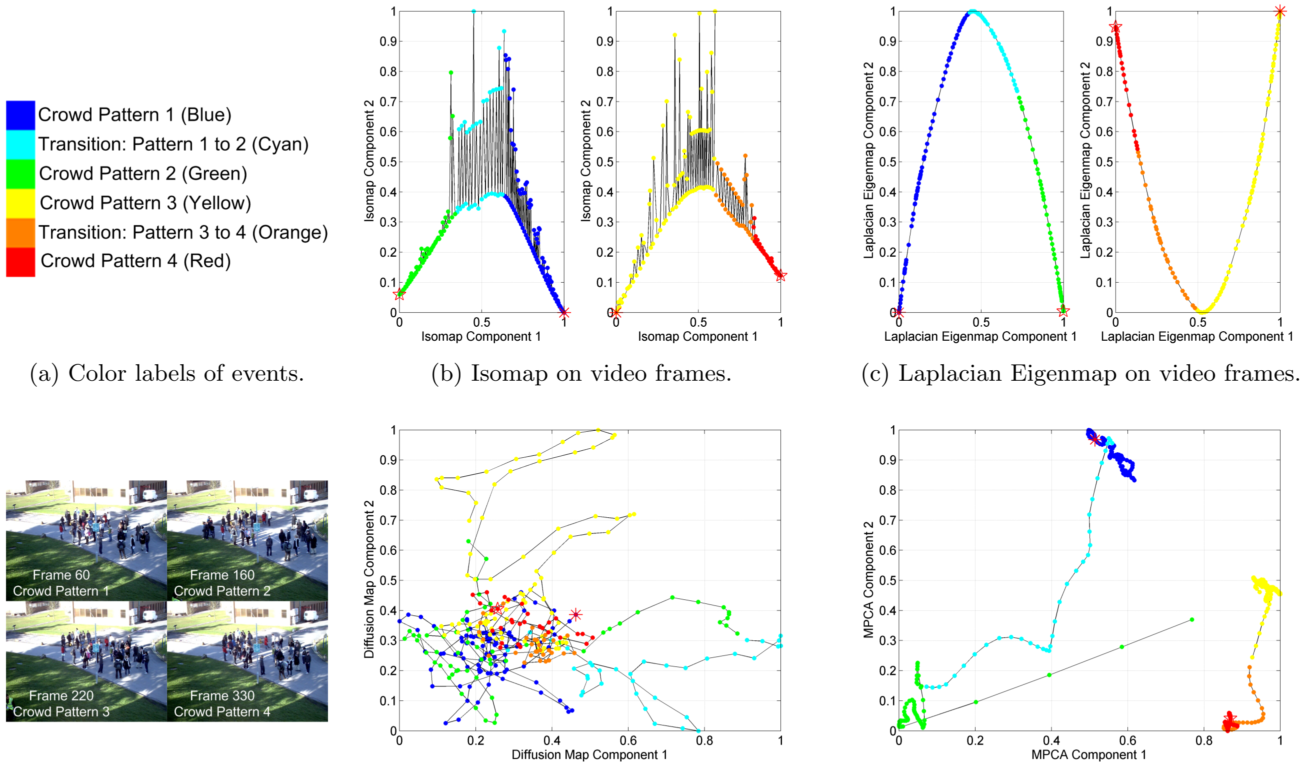
3.2 Experimental results and discussions

The visualizations obtained from the four methods enable visual cluster rendering of the video sequences [2]. For the convenience of evaluation, we use different colors to code different events. The color code for each event is depicted in Fig. 2(a). The labeling of events is produced by human subjects. The visualizations of the Time1427 sequence are reported in Figs. 2(b), 2(c), 2(e) and 2(f), each showing the 2D embedding by a method with each point corresponding to a frame or a video segment. There is a scene change in the sequence so Isomap-based method and Laplacian-Eigenmap-based method detect two separately connected components. Thus, the respective visualizations are shown separately for each connected component detected for fair comparison. For each visualization, a red asterisk marks the beginning of the trajectory and a red pentagram marks the end of the trajectory. For evaluation, we examine whether the visualizations provide clear clues about the events in the sequences and whether we can visually identify the start and end of the events as well as the transitions between them from the visualizations.

There are six events (with two transitions) for the sequence Time1427, with a scene change in the middle. The Isomap-based method performs poorly for most of the trajectory, mapping points of the same event to points far apart, except for events 3 (green) and 6 (red). The Laplacian-Eigenmap-based method is able to map two different crowd patterns to different parts of the trajectory in Fig. 2(c). However, the trajectory does not truly reflect the actual characteristics of the event. E.g., the event crowd pattern 1 (blue) consists of small local movement of the same pattern, but it is mapped to a long trajectory in Fig. 2(c). The Diffusion-map-based method fails in visualizing this sequence, with points of different events heavily mixed. The MPCA-based method has visualized this sequence particularly well. Events 1 (blue), 3 (green), 4 (yellow) and 6 (red) all have fixed crowd patterns and their MPCA visualizations form small clusters in Fig. 2(f), except a few green points due to the scene change. The transition (cyan) between patterns 1 (blue) and 2 (green) is more significant and it is clearly observed in Fig. 2(f). The transition (orange) between patterns 3 (yellow) and 4 (red) is more subtle so it is less distinguishable from pattern 4 (red) in Fig. 2(f).

In summary, the Isomap visualization tends to be noisy while the Laplacian Eigenmap visualization produces smooth curves for events of very different characteristics so it is difficult to interpret events. The Diffusion map visualization cannot provide insights to the corresponding video content at all. In contrast, the proposed MPCA visualization is more meaningful and gives a better interpretation of the video content. In particular, small local movements for a particular crowd pattern nicely form clusters in MPCA subspace. Furthermore, its performance is consistent over the other PETS2009 sequences with various characteristics, which are not reported here due to space constraint. Therefore, the proposed MPCA-based method for video visualization and clustering provides a powerful tool to video analysis.

[1] http : //www.cvg.rdg.ac.uk/PETS2009/a.html



(a) Color labels of events. (b) Isomap on video frames. (c) Laplacian Eigenmap on video frames.

(d) Example frames. (e) Diffusion map on video motion patterns. (f) MPCA on video segments.

Figure 2: Visualization results for sequence Time1427 (best viewed on screen or in color print).

4. CONCLUSIONS

Analysis of crowd video content is becoming an important topic, where detection and tracking of individual subjects have become extremely difficult due to the large number of subjects in the scene. In this paper, we use MPCA, a recent multilinear statistical method, to analyze crowd activities and events, with no need for object detection and tracking. We consider a video segment as a basic video element and map it to a point in MPCA subspace. The MPCA visualization characterizes the entire video sequence with an abstract description of the events and it provides a valuable tool in analyzing video content for video summarization, anomaly detection, and behavior understanding. In particular, the visualization enables visual clustering of events. Experiments show that the proposed MPCA-based method gives much better visualization of challenging PETS 2009 crowd video sequences and produces more visible clusters of events than three existing methods.

5. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

[2] K. Chen and L. Liu. Clustermap: labeling clusters in large datasets via visualization. In *Proceedings of CIKM*, pages 285–293, Nov. 2004.

[3] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, July 2006.

[4] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *Proc. 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–6, Dec. 2009.

[5] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19(1):18–39, Jan. 2008.

[6] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos. Boosting discriminant learners for gait recognition using MPCA features. *EURASIP Journal on Image and Video Processing*, 2009.

[7] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[8] R. Pless. Image spaces and video trajectories: Using isomap to explore video sequences. In *Proceedings of ICCV*, volume 2, pages 1433–1440, 2003.

[9] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, Dec. 2000.

[10] I. Tziakos, A. Cavallaro, and L.-Q. Xu. Video event segmentation and visualisation in non-linear subspace. *Pattern Recognition Letters*, 30(2):123–131, Jan. 2009.

[11] Y. Yang, J. Liu, and M. Shah. Video scene understanding using multi-scale analysis. In *Proceedings of ICCV*, pages 1669–1676, Nov. 2009.