International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:8, 2008

Graph-based High Level Motion Segmentation using Normalized Cuts

Sungju Yun, Anjin Park and Keechul Jung

***Abstract*—** Motion capture devices have been utilized in producing several contents, such as movies and video games. However, since motion capture devices are expensive and inconvenient to use, motions segmented from captured data was recycled and synthesized to utilize it in another contents, but the motions were generally segmented by contents producers in manual. Therefore, automatic motion segmentation is recently getting a lot of attentions. Previous approaches are divided into on-line and off-line, where on-line approaches segment motions based on similarities between neighboring frames and off-line approaches segment motions by capturing the global characteristics in feature space. In this paper, we propose a graph-based high-level motion segmentation method. Since high-level motions consist of several repeated frames within temporal distances, we consider all similarities among all frames within the temporal distance. This is achieved by constructing a graph, where each vertex represents a frame and the edges between the frames are weighted by their similarity. Then, normalized cuts algorithm is used to partition the constructed graph into several sub-graphs by globally finding minimum cuts. In the experiments, the results using the proposed method showed better performance than PCA-based method in on-line and GMM-based method in off-line, as the proposed method globally segment motions from the graph constructed based similarities between neighboring frames as well as similarities among all frames within temporal distances.

***Keywords*—**Capture Devices, High-Level Motion, Motion Segmentation, Normalized Cuts

I. Introduction

MOTION capture devices have been widely used to capture motion information of human natural behavior in detail, and have been utilized in producing several contents, such as animation, movies, and video games. However, motion capture devices are expensive, cumbersome, and inconvenient to use. To solve this problem, in general cases, captured sequential data is first segmented into distinct motions, then segmented motions are synthesized according to the given contents. However, the sequential data is usually segmented by contents producers in manual, and it requires many times and labors. Therefore, automatic motion segmentation has recently attracted a lot of attentions [1-8].

Previous approaches to segmenting input sequential data into distinct motions are divided into two categories: on-line and off-line [1]. Off-line approaches captured the global characteristics of features in feature space, and generally used clustering algorithm, such as *k*-means and GMMs (Gaussian mixture models). Barbic et al. [1] proposed a motion segmentation method using GMM. Kim et al. [2] proposed a novel scheme for synthesizing a new motion from segmented motions stored in database, and they used *k*-means to segment distinct motions from captured sequences. Lee and Elgammal [3] proposed motion segmentation framework for human motion synthesis from the motion sequences, and *k*-means algorithm was used to find distinct motion after dimensionality reduction using SOM. Sakamoto et al. [4] proposed motion map for constructing a graphical user interface, and the motion map was used for motion retrieval and segmentation by simply clicking key-frames. Above-mentioned off-line approaches showed good performance with captured sequences composed of salient motions in feature space, but it is difficult for natural sequences to have salient motions due to external factors, such as noises.

To tackle this problem, on-line approaches have been proposed to consider relationships between neighboring frames, such as similarities. Fod et al. [5] proposed a novel method to automatically segment arm movement, by analyzing the trajectory executed by the arm in neighboring frames. Kwon and Shin [6] proposed a novel scheme for motion segmentation and synthesis, and they segmented the captured sequences into distinct motions by exploiting the behavior of the center of mass trajectory of the performer. Yamasaki and Aizawa [7] proposed a method of motion segmentation and retrieval for 3D data captured by several cameras, and the local minima extracted in similarities between neighboring frames was verified whether they were truly segmentation boundaries or not by thresholding. Barbic et al. [1] proposed a new method to segment motion capture data into distinct behaviors, based on the standard deviation of each frame derived from the PCA. Bouchard and Bardler [8] proposed a motion segmentation method using neural networks trained with temporal variance in order to create a classifier that is more robust with regard to motion boundaries in motion capture sequences, and the actual segmentation was derived by analyzing output of the neural network in neighboring frames. Although on-line approaches are widely utilized recently, compared with off-line ones, they can get stuck in local minima, as only relationships between neighboring frames are considered to segment the distinct motion from captured sequential data.

S. Yun is with the Department of Digital Media, Soongsil University, Seoul, 156-743, Korea (e-mail: yuyu04@ssu.ac.kr).

A. Park is with the Department of Digital Media, Soongsil University, Seoul, 156-743, Korea (e-mail: anjin@ssu.ac.kr).

K. Jung is with the Department of Digital Media, Soongsil University, Seoul, 156-743, Korea (corresponding author to provide phone: 82-2-828-7250; fax: 82-2-822-3622; e-mail: kcjung@ssu.ac.kr).

In this paper, we propose graph-based motion segmentation, which integrates on-line approach into off-line one, to segment captured sequential data into distinct high-level motions. The high-level motion means meaningful distinct behavior, such as walking, running, and punching, and consists of many primitives of human behaviors, i.e. each frame in captured data [1]. The high-level motions have important characteristic that the primitives are repeated in one high-level motion. We exploit this fact and consider similarities between neighboring frames as well as all similarities among all frames to segment the high-level motions. However, common primitives, e.g. standing, should be in several high-level motions. Therefore, similarities within temporal distances are considered to segment the high-level motions, not all the similarities among all frames. In this paper, we transform the motion segmentation into a graph partitioning problem, and the procedure for the graph partitioning comprises of building a graph, called motion similarity graph (MSG), in which each partition defines one of all possible motions. The graph is built with vertices corresponding to frames and the edges between the frames are weighted by their similarity. The MSG is then partitioned into subgraphs by applying the normalized cuts [9] to globally maximize intra-subgraphs similarities and minimize inter-subgraph similarities. Therefore, the proposed method considers global similarities of frames to segment distinct motions rather than local similarities, and is robust to any local mismatch between two frames that belong to one high-level motion although we use similarities between neighboring frames. In the experiments, the proposed method showed better performance than the GMM-based method in off-line approaches and the PCA-based method in on-line approaches, by globally considering the similarities between neighboring frames as well as temporal proximity of frames.

The remainder of this paper is organized as follows. The features used in this paper and MSG are described in section 2, then normalized cuts for graph partitioning is described in section 3. Some experimental results are presented in section 4, and the final conclusions are given in section 5.

II. Motion Similarity Graph for Motion Segmentation

A. Features

Captured sequential data used in the experiments consists of 120 frames per seconds, and each frame is represented by 114 dimensional data that is 3 coordinates of 38 joints. To extract features invariant to the translation and orientation of motions, we use the set of distances from Hip joint centered on human body to all the joints, and the feature of *k*-th frame is represented as follows:

$$\mathbf{F}^k = \{P_1, P_2, \ldots P_n \ldots, P_{37}\},$$

where P_n denotes Euclidean distance from Hip joint to *n*-th one, and the maximum number of *n* is 37 that is the number of joints except for the Hip joint. Consequently, features used in this paper are represented by 37 dimensional data per each frame.

B. MSG with Temporal Proximity of Motions

Fig. 1 shows a part of captured sequential data used in present experiments, and consists of **Walk** (behavior 1) and **Kick** (behavior 2). The high-level motions are composed of several primitives indicated by alphabets, and the primitives deem to be frames in this paper. As shown in Fig. 1, the high-level motions have important characteristic that several primitives are repeated in one high-level motion, e.g. A, B, C, and D are repeated in high-level motion **Walk**.

Therefore, we consider similarities between neighboring frames as well as all similarities among all frames to segment input sequential data into high-level motions, and transform the motion segmentation into a graph partitioning problem. The procedure for the graph partitioning comprises of building a graph, in which each partition defines one of all possible motions. The graph $G = \langle v, \varepsilon \rangle$ is first built with vertices corrpsonding to the frames and the edges between the frames are weighted by their similarity.

However, common primitives should be in several high-level motions, e.g. As are repeated in both two motions in Fig. 1. Therefore, similarities within temporal distances are considered to segment the high-level motions, not all the similarities among all frames, and then the weights assigned to the edges are computed as follows:

$$W(i,j) = w(i,j) \times t(i,j),$$

where $w(i,j)$ denotes similaity between frames *i* and *j*. The notation $t(i,j)$ is used to represent temporal distance and is assigned lower costs if distance between two frames *i* and *j* are longer, computed as follows:

$$t(i,j) = e^{-\left\|\frac{i-j}{\sigma}\right\|},$$

where σ reflects length of one high-level motion, and we set this value in manual. Fig. 2 shows the value of *t* for repeated

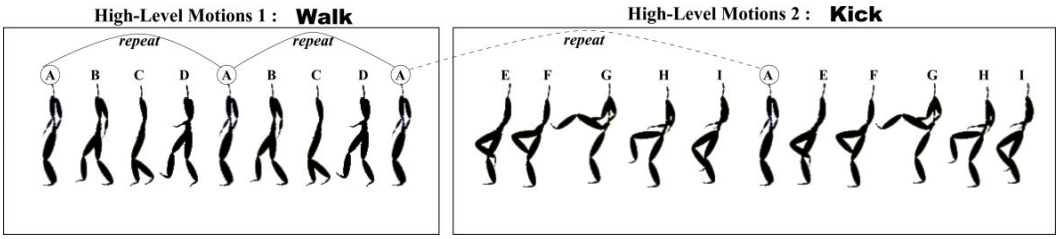


Fig 1. Motion data.

primitives A according to the σ, where σ is assigned 150 and x-axis represent frames. Consequently, the notation $W(i,j)$ reflects the likelihood that two frames i and j belong to one high-level motion based on temporal distance.

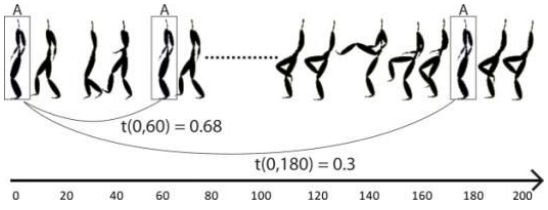


Fig 2. Temporal proximity.

III. Normalized Cuts

We employ the graph partitioning technique proposed by Shi and Malik[10] called normalized cuts. Starting with an initial MSG, $G = (V,E)$, we seek to partition into two disjoint sub-graphs, $G' = (V',E')$ and $G'' = (V'',E'')$ such that $V' \cup V'' = V$ and $V' \cap V'' = \emptyset$. Such a partition is achieved by removing the edges connecting subgraphs G' and G''. In graph theory literature, the summation of weights associated with the edges being removed is called a cut and it reflects the degree of dissimilarity between the two parts, that is

$$\mathrm{cut}(V',V'') = \sum_{i\in V',j\in V''} W(i,j) \qquad (1)$$

A partition that minimizes the cut value is considered an optimal bipartitioning of the graph, However, minimizing the cut value does not always provide a globally optimal solution as it favors cutting small sets of isolated nodes in the graph, To overcome this shortcoming, also considered the degree of association of the partitions being created with regard to the parent graph. Thus, the association of the graph is computed as follows:

$$\mathrm{assoc}(V',V) = \sum_{i\in V',j\in V} W(i,j)$$

where assoc(X,V) is the association measure of the graph and equals the total connection from all nodes in X to all nodes in V. The new degree of disassociation is the cut cost as a fraction of the total edge connections to all the nodes in the graph and called normalized cuts or N-cut

$$\mathrm{N-cut}(V',V'') = \frac{\mathrm{cut}(V',V'')}{\mathrm{assoc}(V',V)} + \frac{\mathrm{cut}(V',V'')}{\mathrm{assoc}(V'',V)}.$$

In the same spirit, a measure for total normalized association within groups can be defined for a given partition, as follows:

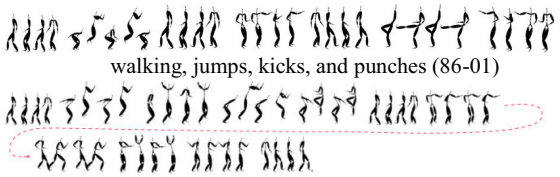
$$\mathrm{Nassoc}(V',V'') = \left(\frac{\mathrm{assoc}(V',V')}{\mathrm{assoc}(V',V)} + \frac{\mathrm{assoc}(V,V)}{\mathrm{assoc}(V'',V)}\right), \qquad (2)$$

where, $\mathrm{assoc}(V',V')$ and $\mathrm{assoc}(V,V)$ are total weights of edges connecting vertices within V' and V. We can see Nassoc reflect how tightly on average vertices within the group are connected on each other.

By (1) and (2),$\mathrm{N-cut}(V',V'') = 2 - \mathrm{Nassoc}(V',V'')$, thus the normalized cuts satisfies both criteria: minimization of the disassociation between the groups and maximization of the association within the groups. Consequently, the normalized cuts help to maximize intra-motions similarities and minimize inter-motion similarities.

IV. Experimental Results

In present experiments, quantitative evaluation was tested with motion capture data from the Graphics Lab of Carnegie Mellon University [1] to verify the effectiveness of the proposed method, and motion capture data consists of several motions, such as running, walking, and sitting. To reduce the computational times, we reduced the number of frames per second from 120 to 30, and compared the proposed method with GMM-based method in off-line approaches and PCA-based method in on-line approaches.



walking, jumps, kicks, and punches (86-01)

walking, jumping, jumping jacks, jumping on one foot, punching, chopping (86-05)

walking, swinging arms, stretching, jumping on one leg, and jumping (86-07)

walking, sitting, looking, stand up (86-09)

walk around, sit, stand up, and running (86-10)

walking, stretching, walking and turning (86-11)

Fig 3. Motion sequence

Fig. 3 shows sequential images sampled from captured sequential data used in the experiments, and the captions describe the set of high-level motions and index used in the Graphics Lab of Carnegie Mellon University [1]. The parameters used in the proposed method were σ that reflects the temporal proximity of frames and K that is the number of distinct high-level motions. One way of selecting σ is to obtain a reasonable tradeoff between recall and precision values. Segmentation with large number motions was likely to have higher recall and relatively poor precision and vice versa. Fig. 4 shows precision and recall rates according to the value of σ, and the precision and recall rates were computed as follows:

$$Precision = \frac{\text{Reported Correct Cuts}}{\text{Total Number of Reported Cuts}} \times 100,$$

$$Recall = \frac{\text{Reported Correct Cuts}}{\text{Total Number of Correct Cuts}} \times 100.$$

In the case of Figs. 4(a,b,c,f), since the length of each distinct motion is short, the smaller value of σ showed the higher precision rates, while in the case of Figs. 4(d,e), since the length of each distinct motion is relatively long, the larger value of σ showed the higher precision rates.

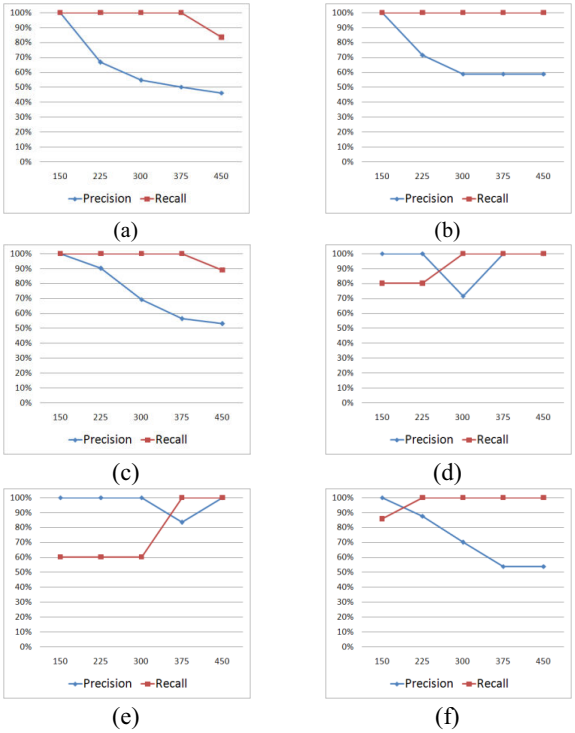


Fig 4. precision and recall rates according to the value of σ: (a) 86-01(x-axis indicates temporal proximity σ, y-axis indicates percentage), (b) 86-05, (c) 86-07, (d) 86-09, (e) 86-10, (f) 86-11.

Fig. 5 shows MSG overlapped with the results extracted by proposed method automatically and by human manually, where the circles indicate results by human and the rectangles indicate results by the proposed method. The value of σ was manually chosen to give the best performance for each motion capture data, and is written in the caption. The value of K is the number of distinct motions assigned in the Graphics Lab of Carnegie Mellon University [1].

Fig. 6 shows results of motion segmentation using GMM-based, PCA-based and the proposed methods, where x-axis indicates frames, y-axis indicates three methods, and dotted lines indicate results manually segmented by human. GMM-based approach showed differentiation of the performance according to the features expressed in feature space. The case of 86-01, one of the capture motion data used in the experiments, showed good performance, as the distinct motions were expressed in salient clusters. However, in the most cases, the capture data was segmented into much or less number of motions than ground truth. The PCA-based method in on-line approach had the main problem that the number of segmented motions is much than the ground truth, as they can get stuck in local minima, and this is the general problem occurred in the on-line approaches. On the other hand, the proposed method showed stable segmentation results among all the motion data used in the experiments, and better performance than two approaches: GMM-based and PCA-based, as shown in Fig. 6. However, the case of motion capture data including longer motions than σ showed the length of segmented motions was not longer than σ, thus the segmented boundaries were not correct. Fig. 7 shows the performance evaluation of three methods: GMM-based, PCA-based, and the proposed method, using above-mentioned precision and recall rates.

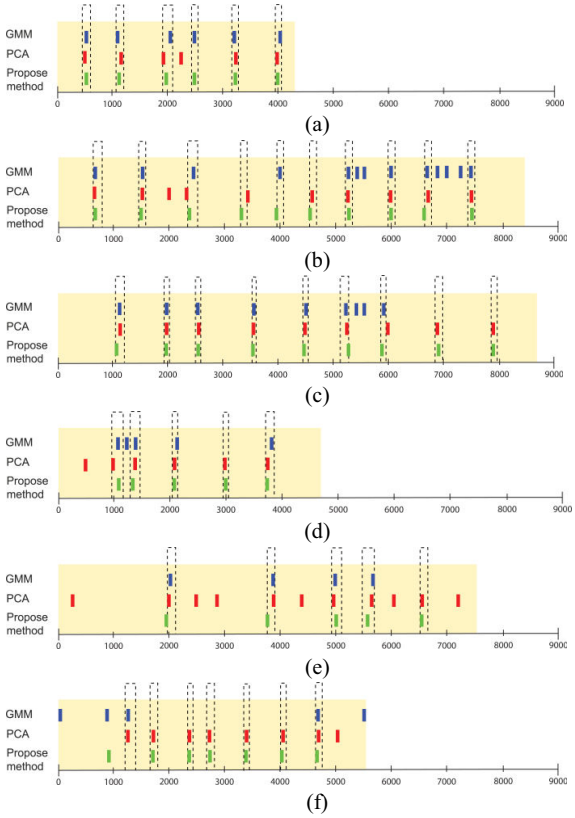


Fig 6. The result of motion segment of GMM-based, PCA-based and propose method: (a) 86-01, (b) 86-05, (c) 86-07,(d) 86-09, (e) 86-10, (f) 86-11.

Fig. 8 shows application of motion segmentation developed using the proposed method. The symbol A represents a graph of broken line based on similarities between neighboring frames and segmented sub-graphs expressed by colors. The symbol B represents MSG, and the symbol C represents window displaying segmented motions. Lastly, the symbol D represents input space for the parameters used in the proposed methods, σ and K.

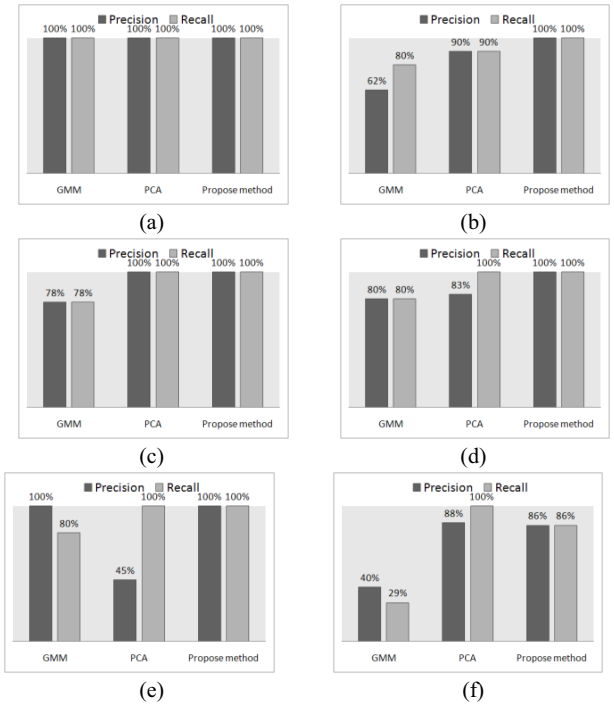


Fig 7. Precision and Recall rates of GMM-based, PCA-based, Propose method: (a) 86-01(x-axis indicates algorithm, y-axis indicates percentage), (b) 86-05, (c) 86-07,(d) 86-09, (e) 86-10, (f) 86-11.

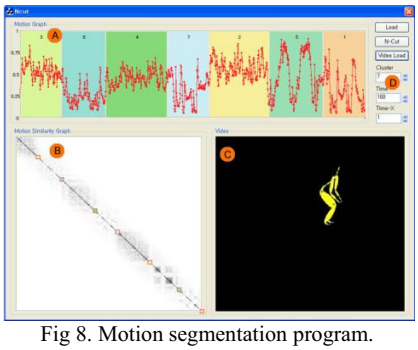


Fig 8. Motion segmentation program.

V. Conclusion

In this paper, we proposed a graph-based high-level motion segmentation method. Since high-level motions have repeated frames within temporal distances, we consider similarities between neighboring frames as well as all similarities among all frames within the temporal distance. This was achieved by constructing a graph, where each vertex represents a frame and the edges between the frames are weighted by their similarity. Then, normalized cuts algorithm was used to partition the constructed graph into several sub-graphs by globally finding minimum cuts.

The problem of the proposed method is σ that reflects length of one high-level motion was manually set and the case of motion capture data including longer motions than σ showed the length of segmented motions was not longer than σ, thus the segmented boundaries were not correct. Therefore, we will investigate to solve these problems.

Acknowledgment

This work was supported by grant No. (R01-2006-000-11214-0) from the Basic Research Program of the Korea Science.

References

[1] J. Barbic, A. Safonova, J.-Y. Pan, C. Faloutsos, J.K. Hodgins, and N.S. Pollard, "Segmenting Motion Capture Data into Distinct Behaviors", Proceedings of ACM International Conference on Graphics Interface, Vol. 62, pp. 185-194, 2004.

[2] T. Kim, S. Park, and S. Shin, "Rhythmic-Motion Synthesis based on Motion-Beat Analysis", ACM Transactions on Graphics, Vol. 22, 2003, pp. 392-401.

[3] C. Lee and A. Elgammal, "Human Motion Synthesis by Motion manifold Learning and Motion Primitive Segmentation", Proceedings of International Conference on Articulated Motion and Deformable Objects, Lecture Notes in Computer Science, Vol. 4069, 2006, pp. 259-266.

[4] Y. Sakamoto, S. Kuriyama, and T. Kaneko, "Motion Map: Image-based Retrieval and Segmentation of Motion Data", Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 259-266, 2004.

[5] A. Fod, M. J. Mataric, and O. Jenkins, "Automated Derivation of Primitives for Movement Classification", Autonomous Robots, Vol. 12, No. 1, 2002, pp. 39-54.

[6] T. Kwon and S. Shin, "Motion Modeling for On-line Locomotion Synthesis", Proceedings of ACM SIGGRAPH/ Eurographics Symposium on Computer Animation, pp. 29-38, 2005.

[7] T. Yamasaki and K. Aizawa, "Motion Segmentation and Retrieval for 3D Video based on Modified Shape Distribution", EURASIP Journal on Advances in Signal Processing, Vol. 2007, No. 2, 2007, pp. 1-11.

[8] D. Bouchard and N. Badler, "Semantic Segmentation of Motion Capture using Laban Movement Analysis", Proceedings of International Conference on Intelligent Virtual Agents, Lecture Notes in Computer Science, Vol. 4722, pp. 37-44, 2007.

[9] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, 2000, pp. 888-905.

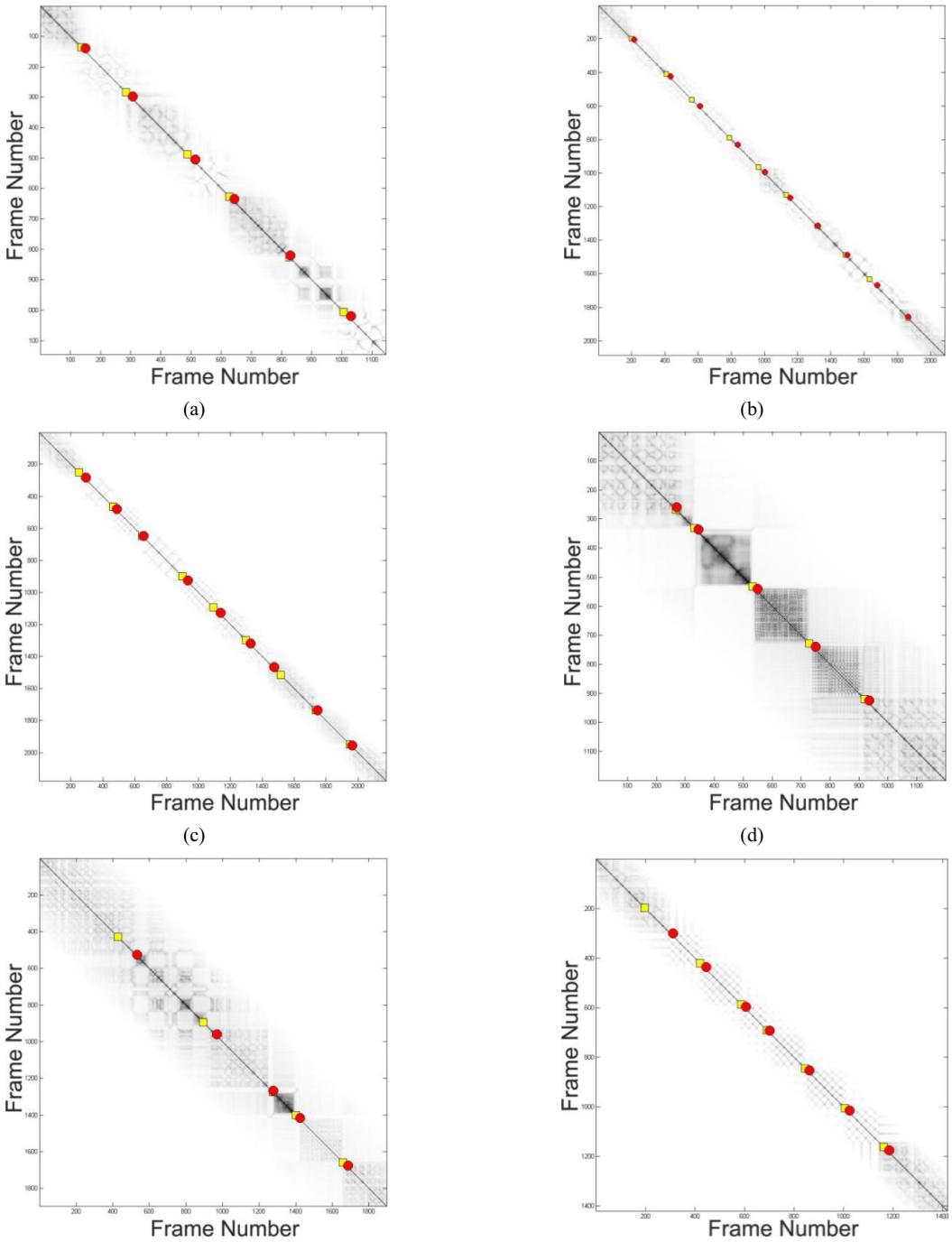


Fig 5. The result of Graph segmented using propose method: (a) 86-01(number of cluster K=7 and temporal proximity of frames σ=150), (b) 86-05(K=11, σ=150), (c) 86-07(K=10, σ=150),(d) 86-09(K=6, σ=450), (e) 86-10(K=6, σ=450), (f) 86-11(K=8, σ=150).