ACTION SYNTHESIS USING BODY SEGMENTATION

Jin-Hong Kim, Rae-Hong Park*,**

*Department of Electronic Engineering, Sogang University
**Interdisciplinary Program of Integrated Biotechnology, Sogang University
C.P.O. Box 1142, Seoul 100-611, Korea
{sosu02kjh, rhpark}@sogang.ac.kr

ABSTRACT

Recognizing action is an important part of a video surveillance and video retrieval system, in which motion information extracted from video is useful. This paper proposes an action synthesis method, in which Efros *et al.*'s action recognition method, Chen *et al.*'s body segmentation method, and Fujiyosi *et al.*'s method are combined. Chen *et al.*'s method and Fujiyosi *et al.*'s method are used as preprocessing of action synthesis. Experimental results with a number of test sequences show that the proposed method works efficiently for human action synthesis.

***Index Terms*—** Action synthesis, body segmentation, skeleton extraction, motion similarity

1. INTRODUCTION

Recognizing action is an important part of video surveillance and video retrieval systems, human-computer interaction, and human behavior understanding. The first step to action recognition is extraction of motion information. Extracted motion information is classified as preclassified actions in database by measuring a motion similarity between input video scenes and existing scenes in video database.

Action synthesis, as an application of action recognition, synthesizes new motions of a character from existing motion database. A user picks up portions of video and directs motion of a character. New motion is synthesized by computing the cost that is expressed as a sum of a motion smooth constraint term and a motion similarity term.

This paper proposes an action synthesis method, in which Efros *et al.*'s action recognition method [1] and Chen *et al.*'s body segmentation method [2] are combined. Efros *et al.*'s method extracts motion descriptors using optical flow and compares them with those in database of preclassified actions. Finally, it classifies action labels and extracts motions from a small-size figure in low-resolution video. Laptev and Lindeberg's method extracts local spatio-temporal interest points [3]. These interest points have the advantage of faithfully reflecting local motions and are robust to camera motion and scaling. Schuldt *et al.* [4] applied Laptev and Lindeberg's method to action recognition. Schuldt *et al.*'s method has a higher action recognition rate and can recognize action in various motion video. Chen *et al.*'s method is a body segmentation method that uses deformable triangulation for initialization and can classify even occluded body parts well. For example, with two legs having similar color and shadows, the segmentation result gives two legs even if one leg occludes the other. Chen *et al.*'s method solved the occlusion problem using a model-driven method.

The rest of the paper is organized as follows. Sections 2 and 3 describe Efros *et al.*'s action recognition method and Chen *et al.*'s body segmentation method, respectively. Section 4 proposes an action synthesis method for human action by combining the two methods. Section 5 gives experimental results and discussions. Finally, Section 6 concludes the paper.

2. ACTION RECOGNITION AND SYNTHESIS

Efros *et al.*'s method is the optical flow based method that can recognize action in a low-resolution video. For instance, in broadcast soccer video, moving human figures of small size are composed of almost 30 pixels, however such human figures have good movements for action recognition. Using optical flow based motion descriptors, Efros *et al.*'s method shows good recognition results for low-resolution video. Furthermore, motion descriptors also can be used for action synthesis.

Fig. 1 shows the block diagram of Efros *et al.*'s method. First, human figures in a video are tracked and stabilized. Stabilization is a preprocessing step for extracting motions from blurred video. Motion vectors in human figures are computed by the Lucas-Kanade optical flow method [5]. The motion descriptor is a set of spatio-temporal features that describe motion in video. Computed motion vectors F define motion descriptors and are decomposed into x and y motion vector components, F_x and F_y. Each of F_x and F_y consists of two channels, defined as

$F_x = F_x^+ + F_x^-$	(1)
$F_y = F_y^+ + F_y^-$	(2)

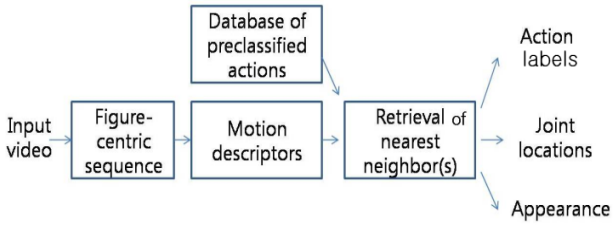

Fig. 1. Block diagram of Efros *et al.*'s method.

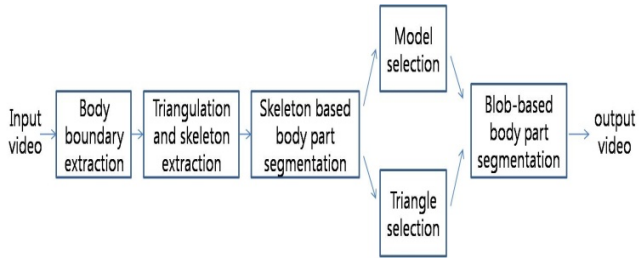

Fig. 2. Block diagram of Chen *et al.*'s method.

where F_x^+ and F_y^+ are positive half-wave rectified values of F_x and F_y, respectively. Similarly, F_x^- and F_y^- are negative half-wave rectified values of F_x and F_y, respectively. Finally, these four channels are blurred and normalized and then used as motion descriptors.

Similarity of motion descriptors $S(i,j)$ between frame i of sequence A and frame j of sequence B is expressed as

$$S(i,j) = \sum_{t \in T} \sum_{c=1}^{4} \sum_{x,y \in I} a_c^{i+t}(x,y) b_c^{j+t}(x,y) \quad (3)$$

where a_1^i, a_2^i, a_3^i, and a_4^i are motion descriptors of frame i in sequence A whereas b_1^j, b_2^j, b_3^j, and b_4^j are those of frame j in sequence B, T represents the temporal extent of the sequence, and I denotes the spatial extent of the sequence.

3. BODY SEGMENTATION

Chen *et al.*'s body segmentation method uses deformable triangulation for initialization. It is effective for segmentation of occluded body parts. Fig. 2 shows the block diagram of Chen *et al.*'s method.

First of all, moving human objects are extracted and then extracted objects are initialized using deformable triangulation. Control points are extracted from the boundary of an extracted human figure. Using extracted control points, Chew's divide-and-conquer algorithm [6] can obtain triangulation of a human object figure.

After initialization, the center of each triangulation is connected. This connectivity generates the skeleton of a human object, by which a human object is segmented roughly. This rough segmentation result using skeletons is realized by the expectation-maximization (EM) algorithm [7] that consists of the iterative E-step and M-step.

Finally, rough segmentation result and human model selection are used for fine-level segmentation. Using skeletons of rough segmentation result, we can choose a proper model for segmenting different body parts.

4. PROPOSED HUMAN ACTION SYNTHESIS

We use human body segmentation as preprocessing for human action synthesis. We use Chen *et al.*'s method and Fujiyosi *et al.*'s method [8]. Human body parts are decomposed into head, arm, body, and leg. Human body segmentation is used as a guide to extract a star skeleton of a human object, which can be obtained from human body segmentation. Using a star skeleton of a human object, several regions are obtained for computing motion descriptors. Computing motion descriptors between human body parts can consider movements of each human part.

The proposed method is composed of three steps and Fig. 3 shows its block diagram. It starts with extracting moving human objects. The second step is to segment human objects into body parts. Synthesizing action is the final step.

The proposed method is described as follows. First of all, we extract moving objects in video, in which Fujiyosi *et al.*'s moving object extraction method is used. Simple background subtraction is sensitive to changes of light, but their method can adapt to slow changes of light. This advantage results from using a statistical model, running average value and standard deviation for extracting foreground.

After extracting moving objects, we segment them into several human body parts. In our proposed method, we assume that moving objects are humans. Chen *et al.*'s body segmentation method is used as a preprocessing step of Fujiyosi *et al.*'s skeleton extraction to segment occluded human body parts. Fujiyosi *et al.*'s skeleton extraction method is simple and can extract features of articulated objects. Using body segmentation before skeleton extraction, features of articulated objects can be extracted more accurately. We use body segmentation as a guide to skeleton extraction as well as to video synthesis.

Because Fujiyosi *et al.*'s method uses silhouettes of human figures for skeleton extraction, some skeleton features are missing when some body parts are occluded. For instance, this method cannot extract some skeleton features, for example, when the left (right) leg occludes the right (left) leg in a human walking scene. Using Chen *et al.*'s body segmentation method, the proposed algorithm can obtain more detailed human body information. After occluded body parts are segmented, Fujiyosi *et al.*'s method

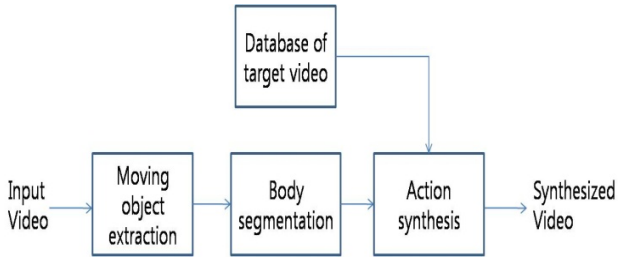

Fig. 3. Block diagram of proposed method

can extract skeleton features from each of segmented body parts. Their star skeleton extraction method is simple and can extract features of articulated objects. Each star skeleton feature is connected to the centroid of a human figure for overall star skeleton. A human object is segmented into several regions by star skeleton. Finally, Efros *et al.*'s method is applied to each of segmented regions for synthesizing action. Synthesizing action is performed by using the motion similarity in (3). Synthesized frames are made of the best matched frames which have the largest motion descriptor similarity. In computing motion descriptors for each segmented region, movements of each human part (head, legs, and arms) are considered in detail.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

We apply the proposed method to a number of test sequences. Database contains 16 different videos (720×480), with database videos classified into eight classes. Each class has two motions (upward and downward motions), with each motion consisting of 30 frames containing different motions. Fig. 4 shows database videos used in experiments, with 16 different motions indicated by arrows.

Fig. 5 shows skeletonization results of the input frame. Fujiyosi *et al.*'s method is applied to input video for image skeletonization. Fig. 5(a) shows a moving object in the input video and Fig. 5(b) shows a segmented human object. Fig. 5(c) shows the result of border extraction, in which the white point inside the object represents the centroid of a moving object. The white cross symbol on the boundary is the starting point for computing the distance between the centroid of a moving object and pixels on the object boundary. Fig. 5(d) shows the final result of the moving object extracted from an input video. Using distances between the centroid and boundary points, each feature point is extracted and connected to the centroid of a moving object.

Fig. 6 shows result of computation of the motion similarity between database and input sequences. We use input sequences with the background different from that in database videos. Fig. 6(a) shows an input video which has upward right arm and leg motion. Fig. 6(b) shows motion similarity results, in which some spurious peaks occur at other database sequences. Note that the highest peak occurs at the correct sequence in database and is about twice larger than the second peak.

Fig. 7 shows final result of synthesized video using the proposed method. Fig. 7(a) shows an input video which has upward right arm motion and right leg motion. Fig. 7(b) shows the result of synthesized video. Although input and database video have different background and human characters, the proposed method can synthesize videos well.

6. CONCLUSIONS

We propose an action synthesis method using body segmentation, in which body segmentation is used as a guide to star skeleton extraction as well as to video synthesis. Using star skeleton and body segmentation results, Efros *et al.*'s method is applied to each of segmented regions. Experimental results with a number of test sequences show that the proposed method can synthesize videos well. The future work will focus on action synthesis of video sequences containing various multiple objects.

Acknowledgment This work was supported by the Second Brain Korea 21 Project.

7. REFERENCES

[1] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. Int. Conf. Computer Vision*, vol. 2, pp. 726–733, Nice, France, Oct. 2003.

[2] C.-C. Chen, J.-W. Hsieh, Y.-T. Hsu, and C.-Y. Huang, "Segmentation of human body parts using deformable triangulation," in *Proc. Int. Conf. Pattern Recognition*, vol. 1, pp. 355–358, Hong Kong, China, Aug. 2006.

[3] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. Int. Conf. Computer Vision*, vol. 2, pp. 432–439, Nice, France, Oct. 2003.

[4] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. Int. Conf. Pattern Recognition*, vol. 3, pp. 32–36, Cambridge, UK, Aug. 2004.

[5] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artificial Intelligence*, pp. 674–679, Vancouver, Canada, Apr. 1981.

[6] L. P. Chew, "Constrained Delaunay triangulations," *Algorithmica*, vol. 4, no. 1, pp. 97–108, May 1989.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2000.

[8] H. Fujiyosi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 1, pp. 113–120, Jan. 2004.

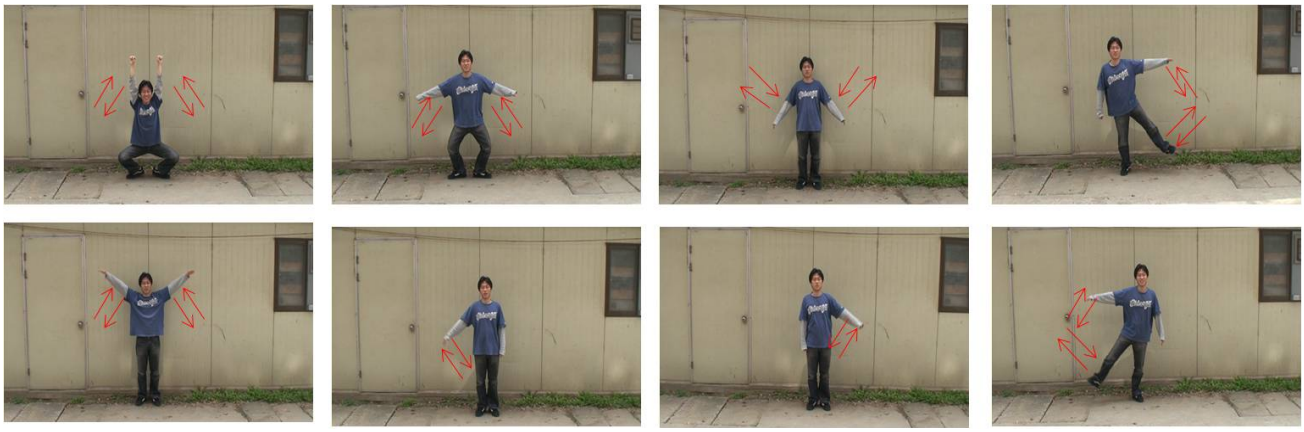

Fig. 4. Database videos used in experiments.

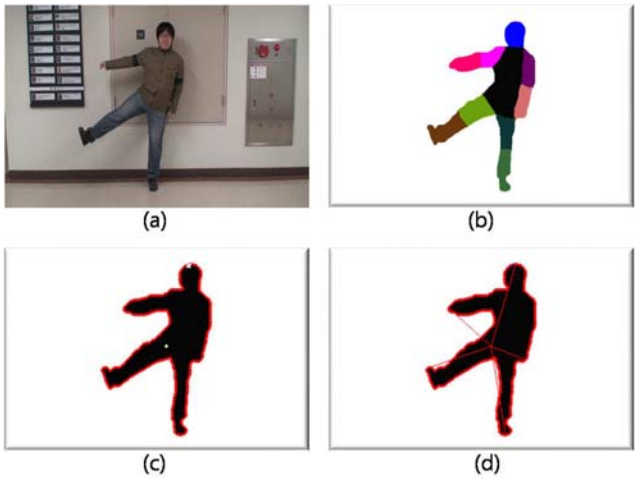

Fig. 5. Star skeletonization. (a) input frame. (b) segmented human body. (c) border extraction. (d) star skeletonization.

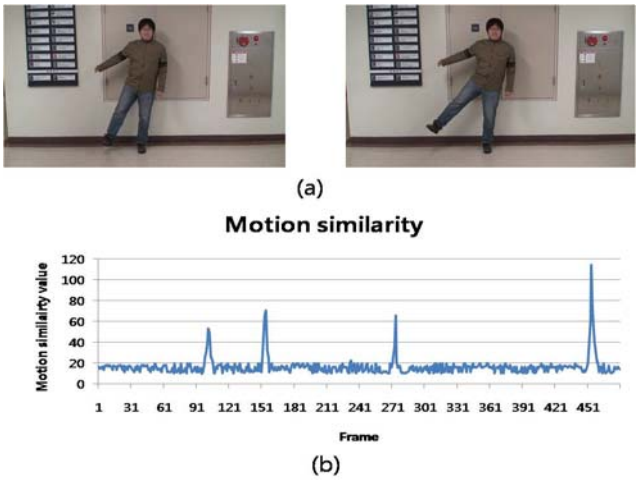

Fig. 6. Computation of the motion similarity. (a) input sequences (two frames). (b) motion similarity.

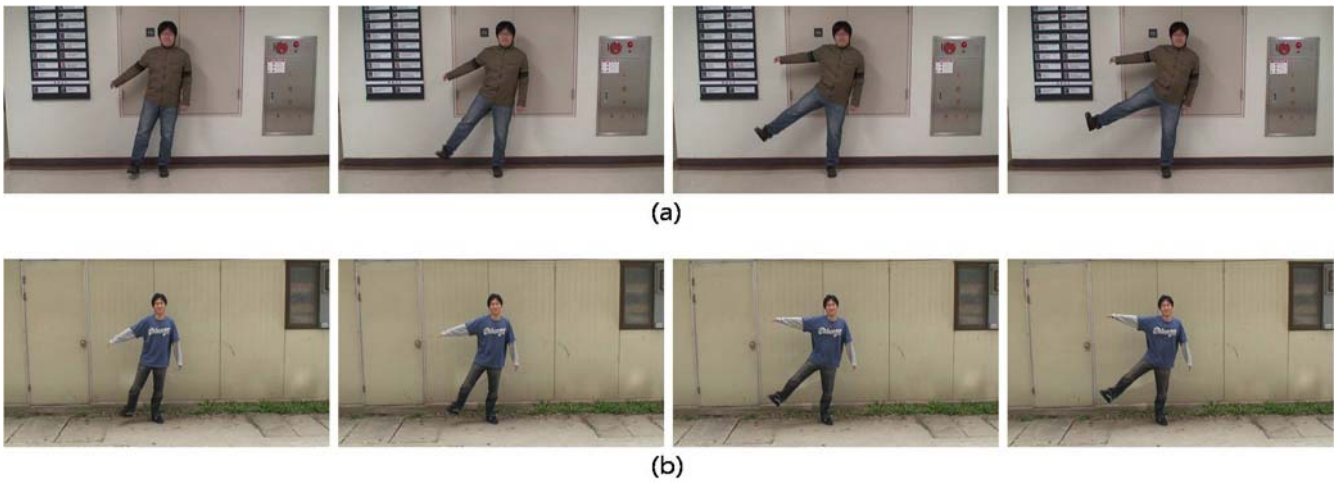

Fig. 7. Action synthesis. (a) input video. (b) synthesized video.