Human Action Recognition using Temporal-State Shape Contexts

Pei-Chi Hsiao[1,2], Chu-Song Chen[1], and Long-Wen Chang[2]
[1]*Institue of Information Science, Academia Sinica, Taipei,Taiwan*
{*pchsiao, song*}*@iis.sinica.edu.tw*
[2]*Institue of Information Systems and Application, National Tsing Hua University, Hsinchu,Taiwan*
lchang@cs.nthu.edu.tw

Abstract

In this paper, we present a temporal-state shape context (TSSC) method that exploits space-time shape variations for human action recognition. In our method, the silhouettes of objects in a video clip are organized into three temporal states. These states are defined by fuzzy time intervals, which can lessen the degradation of recognition performance caused by time warping effects. The TSSC features capture local characteristics of the space-time shape induced by consecutive changes of silhouettes. Experimental results show that our method is effective for human action recognition, and is reliable when there are various kinds of deformations. Moreover, our method can identify spatially inconsistent parts between two shapes of the actions, which could be useful in action analysis applications.

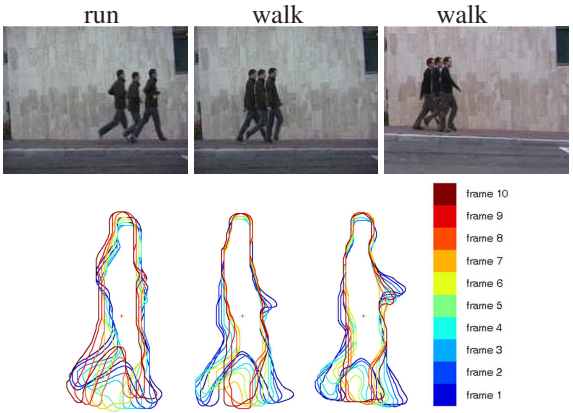


Figure 1. Three space-time shapes induced by consecutive subject silhouettes in three video sequences.

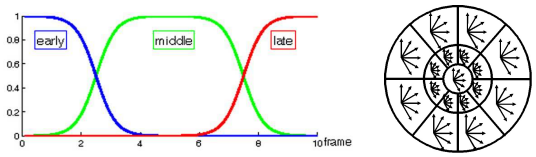
1. Introduction

Analyzing human actions plays an important role in many computer vision applications, such as gesture detection, activity recognition, and event analysis. Two kinds of approaches are usually used. The first employs local motion information such as optical flow [4, 5]. The second analyzes space-time shape variations to characterize motion kinematics [3, 2, 8]. In this paper, we present a method that uses space-time shape variations for human action recognition.

In this research track, Bobick and Davis [3] modeled human actions by motion-history images containing continuously-changed object silhouettes over time. Blank *et al.* [2] analyzed 3D shapes of object silhouettes in a space-time volume. Wang and Suter [8] considered a video sequence as a raw data vector and adopted a graph embedded method to learn the dynamic shape manifolds. The above methods all used global descriptors to represent human-action characteristics. The first two extract spatial or spatial-temporal moments, and the third uses raw data vectors of the video sequences for representation. However, the global descriptors may not effectively utilized local spatial information. This will degrade the discriminative performance, particularly when discriminative differences between two actions only exist in few regions, such as the identification of running and walking.

To resolve this problem, some researchers use local features to capture local space-time shape variations [6, 7]. Roh *et al.* [6] proposed Curvature Scale Space features, which are extracted from interesting points detected on 2D contours. Sminchisescu *et al.* [7] use local descriptors that combined shape contexts and pairwise edge features. In these methods, local features are extracted from each frame of a video sequence. Therefore, time warping effects have to be handled when matching two temporal sequences. Typically, a dynamic time warping algorithm or a Hidden Markov model can be adopted to compensate the time warping effect.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE



(a) (b)

Figure 2. Temporal-state shape contexts are constructed from (a) three temporal states defined by three membership functions and (b) edge orientation histograms computed from neighboring log-polar regions

In this paper, we propose a local feature called *temporal-state shape context* (TSSC) that can capture local space-time shape characteristics effectively. The time warping effects are compensated by segmenting the video sequence into several temporal states defined by soft fuzzy intervals. This paper is organized as follows. In Section 2, we introduce the TSSC method. In Section 3, we present some experimental results. A conclusion is given in Section 4.

2 Human action recognition using local space-time features

We present in this section the preprocessing procedures, definition of the TSSC features, and the matching procedure using the TSSC features for human action recognition.

2.1 Preprocessing

Given a video clip of a person performing a certain action, we assume that a sequence of his/her silhouettes can be obtained either by a background subtractor or by a contour tracker. The global translation is compensated by aligning the centroid of silhouette of each frame. The scale differences among different people are alleviated by dividing the median points distance [1]. Figure 1 shows the aligned and scaled silhouettes for three video sequences. These consecutive silhouettes consist of the space-time shapes caused by human actions. We analyze the space-time shape variations among different actions for recognition.

2.2 Temporal-state shape contexts

We segment the video sequence into three temporal states: early, middle, and late states. These states are defined by fuzzy intervals. Bell-shaped membership functions are chosen to represent the soft temporal intervals, as illustrated in Fig. 2(a), and all of the membership functions sum to unity at any frame t:

$$w^{early}(t) + w^{middle}(t) + w^{late}(t) = 1. \quad (1)$$

With these membership functions, the temporal information can be exploited and performance degradation caused by time warping effects can also be reduced.

Now we describe the definition of the TSSC features. To capture local characteristics of the space-time shape, we extend the shape context method proposed by Belongie *et al.* [1] as follows.

Assume there are N contour points on the subject silhouettes:

$$p_i = [x_i, y_i, t_i, \theta_i], \quad i = 1 \dots N, \quad (2)$$

where x_i, y_i are the pixel coordinates, t_i is the frame number, and θ_i is the edge orientation computed at position (x_i, y_i) from the silhouette image of frame t_i. Given a center point $p_c = (x_c, y_c)$ at which we want to extract the TSSC feature, the neighborhood of p_c is partitioned into R regions by a log-polar coordinate system, as shown in Fig. 2(b). In each region r, three histograms of edge orientations of contour points p_i are computed for the three temporal states, respectively:

$$h(r) = [\ h^{early} \quad h^{middle} \quad h^{late}\ ]^T, \quad (3)$$

$$h^{state}(k) = \sum_{\substack{(x_i,y_i)\in r,\\ \theta_i \in bin(k)}} w^{state}(t_i)\ , \quad (4)$$

where $state \in \{early, middle, late\}$.

The TSSC feature at the center point p_c is then obtained by concatenating the three histograms of each neighboring region of p_c:

$$TSSC(p_c) = [h(1), h(2), \dots, h(R)]^T. \quad (5)$$

The TSSC feature captures local characteristics of the space-time shape centered at p_c, and the temporal states accounts for the reservation of temporal information.

In our experiment setup, we used 17 log-polar regions and 6 bins for each edge orientation histogram of the temporal states. Therefore, the TSSC feature is a $17 \times 6 \times 3 = 306$ dimensional descriptor.

2.3 Human action recognition using TSSC

The previous description of TSSC features ignores one important remaining issue to decide where the TSSC features should be computed, that is, where the center points p_c should be placed. There are several

strategies to choose center point positions in previous works [6, 4, 5]. One popular strategy is relying some salient point detectors to locate salient points on the 2D contours of subject silhouettes [6]. Contour points with high curvature are likely to be selected as candidate center points. This strategy is not suitable for human action recognition, since the shape variations in low curvature regions is also important. Therefore, constructing local descriptors only at high curvature points is not appropriate.

Another possible strategy is to place center points in regions where there is prominent subject motion [4, 5]. Even though describing shape variations in moving regions is certainly necessary, the fact the there is no shape variation in static regions is also an important clue for action recognition. Therefore, extracting local descriptors in static regions is also needed.

For the above reasons and for simplicity, we place center points in uniform grid points around the centroid of silhouettes. Then a video sequence is represented as a set of TSSC features, each extracted from a grid point. The distance (dissimilarity) between two video sequences, V_i and V_j, can be measured by the summation of χ^2 distance between each pair of TSSC features extracted from the corresponding position:

$$d(V_i, V_j) = \sum_{p_c} \chi^2(TSSC_i(p_c),\ TSSC_j(p_c)), \quad (6)$$

where p_c is the center point position, $TSSC_i(p_c)$ and $TSSC_j(p_c)$ represent the TSSC features extracted at p_c from video V_i and V_j respectively, and χ^2 computes the χ^2 distance between the pair of TSSC features. In our experiments, a 3×5 grid is used, that is, totally 15 local TSSC features are extracted at 15 grid points.

3 Experiments

We use the human action database provided by Blank *et al* [2]. This database contains 90 videos (180×144, 25 fps), where nine people perform ten different actions. The subject silhouettes, which are not perfect but suffice for our method, are extracted by [2] and [8].

3.1 Action classification

For comparison with [2], we also crop each video into multiple clips, each containing 10 frames, where there are 5 overlapped frames with the next clip. Totally, 945 clips are obtained. Note that classifying these clips is difficult because repeated periodic information is not available in such a limited number of frames. Leave-one-out recognition experiment is conducted. For each test clip, the clips obtained from the same video sequence are excluded from the gallery set in advance. By using nearest neighbor with the distance measure defined in Section 2.3, only 26 clips (2.75%) are misclassified. The confusion matrix is shown in Table 1.

Table 1. Leave-one-out test on clips

	wave2	wave1	walk	skip	side	run	pjump	jump	jack	bend
wave2	99.1	0.9	0	0	0	0	0	0	0	0
wave1	0	93.9	0	0	0	0	6.1	0	0	0
walk	0	0	99.1	0	0	0.9	0	0	0	0
skip	0	0	0	87.3	0	7.0	0	5.6	0	0
side	0	0	0	0	100	0	0	0	0	0
run	0	0	1.8	0	0	98.2	0	0	0	0
pjump	0	0	0	0	0	0	100	0	0	0
jump	0	0	0	1.3	0	0	0	98.7	0	0
jack	0	0	0	0	0	0	0	0	100	0
bend	0	5.4	0	0	0	0	0	0	0	94.6

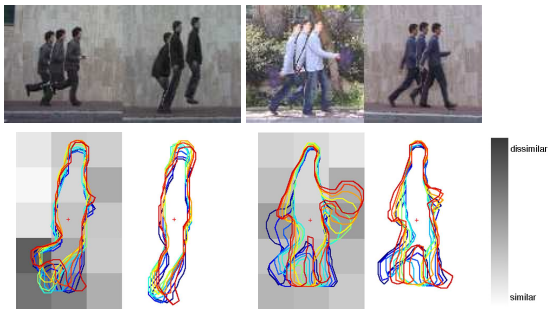
Table 2. Leave-one-out test on videos

	wave2	wave1	walk	skip	side	run	pjump	jump	jack	bend
wave2	100	0	0	0	0	0	0	0	0	0
wave1	0	100	0	0	0	0	0	0	0	0
walk	0	0	100	0	0	0	0	0	0	0
skip	0	0	0	77.8	0	11.1	0	11.1	0	0
side	0	0	0	0	100	0	0	0	0	0
run	0	0	11.1	0	0	88.9	0	0	0	0
pjump	0	0	0	0	0	0	100	0	0	0
jump	0	0	0	0	0	0	0	100	0	0
jack	0	0	0	0	0	0	0	0	100	0
bend	0	0	0	0	0	0	0	0	0	100

To verify the effectiveness of the TSSC features against time warping effects, we also perform a frame-by-frame matching experiment. Each frame in a clip is represented as a grid of the original shape contexts [1] and the distance between two clips is defined as the summation of shape context distances in all frames. With this frame-by-frame matching, the number of misclassified clips increases to 36 (3.8%). This result shows that time warping effects can be alleviated by the temporal states in our TSSC features.

We also perform another experiment for video matching. Given a test video, the middle clip is extracted by cropping the beginning and ending 1/3 from it. Then this middle clip represents this test video. When matching with a target video, multiple target clips are extracted from the target video, each having the same length as the test middle clip, with 5 overlapped frames again. The distance between the test and target video is defined as the minimum distance between the middle clip and any of the target clips.

For all 90 videos, 87 of them are correctly classified, and only 3 videos are misclassified, two for skipping and one for running action. The recognition accura-



(a) (b)

Figure 3. Spatial consistency map between (a) skip and jump, (b) swinging with a bag and walk.

cies are shown in Table 2. As according to the results of previous works, videos of the two actions, skip and jump, are very likely confused in most methods since they represent jumping forward on one leg or on two legs respectively.

Besides estimating the distance measure between two actions, our method can also provide spatial consistency information as shown in Fig. 3, where dark regions indicate dissimilar parts between two actions, while white regions indicate similar parts.

3.2 Robustness

Since our method is primarily based on extracted subject silhouettes, here we evaluate the reliability of our method with respect to various deformations that may occur in real world walking videos. Three datasets provided by the authors of [2, 8] are used in this experiment. Dataset A contains 10 walking videos of different walking styles, carrying objects, and non-rigid deformations. Dataset B contains 10 walking videos taken from different viewpoints, varying from 0°to 81°. Dataset C contains 17 walking videos with partial or serious occlusions by fences, bench, branches, or poles.

Each video is split into clips and then the smallest *Median Hausdorff distance* between the clip and each action is computed [2]. Table 3 shows the first and the second best matched actions for each video sequence, among which only six videos are misclassified, and four misclassifications (B7–B10) are because of large viewpoint variations (larger than 45°). This is reasonable since currently our method does not account for large viewpoint variations. The other two misclassified videos (C6 and C16) are due to severe occlusions by branches and nonstandard walking style (walking upstairs) respectively. Generally speaking, our method performs well under different kinds of deformations in real world videos.

Table 3. Robustness experiments

Type[2]	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
1st	walk	walk	walk	walk	walk	walk	walk	walk	walk	walk
2nd	side	skip	side	run	side	skip	side	side	side	run
Type[2]	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
1st	walk	walk	walk	walk	walk	walk	**side**	**pjump**	**pjump**	**pjump**
2nd	side	side	side	side	side	side	pjump	side	side	side
Type[8]	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1st	walk	walk	walk	walk	walk	**run**	walk	walk	walk	walk
2nd	skip	skip	jump	side	skip	walk	jack	run	run	skip
Type[8]	C11	C12	C13	C14	C15	C16	C17			
1st	walk	walk	walk	walk	walk	**skip**	walk			
2nd	run	run	run	run	run	walk	side			

A1:Swinging a bag A2:Carrying a briefcase A3:Knees up A4:Limping A5: Sleepwalking A6:Occluded legs A7:Normal walk A8:Occluded by a pole A9:Walking in a skirt A10:Walking with a dog
For details about video sequences in Dataset B and C, please refer to [2, 8].

4 Conclusions

We introduce a local space-time feature, called TSSC, for human action recognition. This feature exploits both temporal states for compensating time warping effects and shape contexts for space-time shape variations. Encouraging results are obtained in our experiments, even with deformations in real world videos. Spatial consistency information between two actions can also be provided by our method, which is useful for action analysis applications.

ACKNOWLEDGEMENT: This work was supported in part by Grants NSC 96-3113-H-001-011 and NSC 95-2221-E-001-028-MY3.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002. 2, 3

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 1, 3, 4

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001. 1

[4] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, 2003. 1, 3

[5] I. Laptev and P. Perez. Retrieving actions in movies. *ICCV*, 2007. 1, 3

[6] M.-C. Roh, B. Christmas, J. Kittler, and S.-W. Lee. Gesture spotting for low-resolution sports video annotation. *Pattern Recogn.*, 41(3):1124–1137, 2008. 1, 3

[7] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *ICCV*, 2005. 1

[8] L. Wang and D. Suter. Learning and matching of dynamic shape manifolds for human action recognition. *IEEE TIP*, 16(6):1646–1661, 2007. 1, 3, 4