Behavior Modeling and Recognition Based on Space-Time Image Features

Heping Li, Zhanyi Hu, Yihong Wu, Fuchao Wu

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, P.O. Box 2728, Beijing 100080, P.R.China
Email: {hpli, huzy, yhwu, fcwu} @nlpr.ia.ac.cn

Abstract

A novel method based on space-time image features is proposed for automatic behavior modeling and recognition. The method is composed of the following three steps: (1). Video sequences are converted into a space-time image, and the image is divided into equal length segments; (2). From these segments, an unsupervised technique based on Dynamic Time Warping is used to determine the groups of different behaviors; (3). The behavior templates are built according to these groups, and are used for the recognition of behaviors. The method does not need to track the body parts and can model the behaviors without any prior knowledge. Experiments also demonstrate the effectiveness of our new method.

1. Introduction

Behavior analysis and recognition is driven by a wide ranges of applications, such as visual surveillance, advanced user interface, virtual reality. Given a long video, it is desirable for these systems to know how to automatically model and recognize human behaviors.

Most existing works [1-4] mainly focused on modeling the behaviors with manually labeling. For example, Li et al. in [1] built a multi-scale model from time-varying contours and Gong et al. in [2] learned a Dynamically Multi-Linked Hidden Markov Model (D-ML-HMM). Manually labeling is laborious and error prone [6]. Moreover, in many real-world applications, it is impossible for us to know the predefined models of behaviors. Fortunately, in every day life, the normal behaviors often occur over a relatively big proportion of the whole time and considering this fact, we may build the models of these behaviors with unsupervised clustering method. In [5-6], this fact has been considered. They both used the spectral clustering method [7] to build the behavior models. Zelnik-Manor et al. in [5] used the multiple temporal scales

technology to solve the time-varying problem of human behavior. In fact, multiple temporal scales technology cannot well deal with it. And Xiang et al. in [6] used the Multi-Observation Hidden Markov model (MOHMM) to avoid this problem, where the initialization of MOHMM in the case of sparse data is required.

Figure 1. Overview of the proposed method.

In this paper, we use Dynamic Time Warping (DTW) [8] as the basic tool to model normal behaviors. DTW is often used to analyze temporal patterns in speech recognition and it has been adapted to the behavior analysis [9]. DTW can be used to measure the similarity of two time series whether their time lengths or periods are equal and to solve the time-varying problem of behavior. At the same time, we adopt novel behavior features called space-time image features. The features and the space-time features directly extracted from time series are fundamentally different. The methods based on the space-time features, such as Pixel Change History (PCH) [2], Motion Energy Image (MEI) and Motion History Image (MHI) [3], need to accurately determine the time intervals of the movements. But Space-time image features are no need. Moreover, space-time image features are simple,

insensitive to noise and they do not need for tracking the parts of human body.

The overview of our method is shown in Figure 1. It consists of two major parts. (1). Motion signature representation: including background modeling, foreground extraction, shape feature extraction and space-time image representation. (2). Modeling and recognition: to train the models with spectral clustering method [7] based on DTW and recognize the behaviors using the trained models.

The remainder of this paper is organized as follows: Motion signature representation is outlined in Section 2. Section 3 presents the modeling and classification. Experimental results are reported in section 4. And, conclusions and future research directions are listed in section 5.

2. Motion signature representation

Figure 2. Examples of foreground detection and shape feature extraction: (a) the recovered background; (b) an original image; (c) the detected foreground; (d) the shape feature.

Figure 3. An example of space-time image representation: (a) the original video sequences; (b) the space-time image about the behavior "hand clapping".

2.1. Shape feature extraction

Background subtraction has been widely used to detect foreground with a fixed camera. The critical step is to recover the background. Here, we use the Least Median of Squares (LMedS) method [10] to construct the background from pre-captured video sequences. Then the brightness change is obtained by subtracting the background image (figure 2(a)) from the current image (figure 2(b)). After removing noise, we get the binary image (figure 2(c)) by a threshold value.

The size of the foreground region varies with the distance of object to camera, camera parameters and the size of object. We therefore need to normalize the

foreground region. Firstly, we equidistantly divided the bounded rectangle of the foreground into 10×7 non-overlapping sub-blocks (figure 2(d)). Then, the normalized value of each sub-block is calculated as follows:

$$d_i = \left\lfloor \frac{\text{sub}(i)}{\max} \times 255 \right\rfloor, \quad i = 1, 2, \dots, k, \quad (1)$$

where $k = 10 \times 7$ is the number of sub-blocks; $\text{sub}(i)$ is the number of the foreground pixel in the i^{th} sub-block; $\max$ is the maximum value of $\text{sub}(i)$ ($i = 1, 2, \dots, k$). The resulting shape descriptor D_t at image frame t is given as:

$$D_t = [d_1, d_2, d_3, \dots, d_k]. \quad (2)$$

2.2. Motion signature representation

To represent the motion signature of a sequence with t frames, a matrix I of size $t \times k$ is constructed as follows:

$$I = [D_1, D_2, D_3, \dots, D_t]_{t \times k}^T. \quad (3)$$

We call I space-time image (Figure 3(a) shows an original video sequences about the behavior "hand clapping"; Figure 3(b) represents the corresponding space-time image).

3. Motion modeling and Recognition

3.1. Motion signature segmentation

Given the space-time image I produced from a long video, we first slice it into N short segments along the time axis: $I = \{I_1, I_2, \dots, I_n, \dots, I_N\}$. Ideally, each segment contains a single behavior pattern. Broadly speaking, there are three methods for the partition, namely, non-activity gaps, fixed time duration with overlapping window and the points of abrupt change. Here, we adopt the second one with a temporal window of size $T=64$ frames. So the size of each short space-time image is 64×70 .

3.2. Space-time image features extraction

Figure 3(b) shows that the space-time image can provide abundant texture information. We can explore the representations of behaviors from the different texture characters of different space-time images. We extract the features by Gabor filtering as follows:

$$F_n(x, y) = \iint I_n(x_1, y_1) G(x - x_1, y - y_1, f) dx_1 dy_1 \quad (4)$$

$$G(x, y, f) = \frac{1}{2\pi\delta_x\delta_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\delta_x^2} + \frac{y^2}{\delta_y^2}\right)\right] M(x, y, f) \quad (5)$$

$$M(x, y, f) = \cos[2\pi f(x \cos \theta + y \sin \theta)] \quad (6)$$

where $I_n(x, y)$ is the n^{th} space-time image; $F_n(x, y)$ is the filtered space-time image; δ_x and δ_y are the space constants of the Gaussian envelope along the x and y axis, respectively; f is the frequency of the sinusoidal function; θ denotes the orientation of Gabor filter.

To reduce the computation complexity and characterize local texture information, we extract the statistical features in each 4×5 block of each short space-time image. The feature value is the mean:

$$m = \sum_B |F_n(x, y)| / K, \quad (7)$$

where B is a 4×5 block in the short space-time image; K is the number of pixels in the block B . So each short space-time image V_i ($1 \leq i \leq N$) can be represented as a 16×14 feature vector:

$$V_i = [M_1, M_2, \dots, M_n, \dots, M_{16}]^T, \quad (8)$$

where $M_n = (m_{n1}, m_{n2}, \dots, m_{ni}, \dots, m_{n14})$.

3.3. Affinity matrix and spectral clustering

Dynamic Time Warping can align two signals and get the warping path. The path can be found very efficiently by using dynamic programming. Given two motion signatures V_m and V_n , the best warping path will minimize the cumulative distance:

$$\gamma(i, j) = d(i, j) + \min \{ \gamma(i-1, j), \gamma(i, j-1), \gamma(i-1, j-1) \}, \quad (9)$$

where $1 \leq i, j \leq 16$;

$$\begin{aligned} \gamma(1, 1) &= d(1, 1); \\ \gamma(i, 1) &= d(i, 1) + \gamma(i-1, 1); \\ \gamma(1, j) &= d(1, j) + \gamma(1, j-1). \end{aligned} \quad (10)$$

In Eq.(9), $\gamma(i, j)$ is the partial cost, and $d(i, j)$ is defined as follows:

$$d(i, j) = 1 - M_i M_j / \|M_i\| \|M_j\|. \quad (11)$$

Where M_i is the i^{th} element of motion signature V_m , M_j is the j^{th} element of V_n .

Considering that the best warping paths may have different lengths, we use the normalized distance to measure the similarity of two motion signatures:

$$S(m, n) = \exp[-\gamma_{mn} / k\sigma], \quad (12)$$

where $S(m, n)$ is the similarity distance of two motion signature V_m and V_n ; γ_{mn} is the cumulative distance; k is the length of the warping path; σ is a constant factor.

Consequently, an $N \times N$ affinity matrix $S = [S(m, n)]$ with $1 \leq m, n \leq N$ is obtained. This matrix provides a new representation for the long training video series. We then use the spectral clustering method, i.e. the normalized-cut approach [7], to cluster the data.

3.4. Motion modeling

After spectral clustering, we obtain the space-time image segments that belong to the same class. Based on the result, we select a prototype for each cluster. Generally, there are two methods to choose the prototype. One is to select the segment minimizing the distance to all other segments, and the other is to select the average of all the segments in the cluster. Different from these two methods, in this paper, we combine them with DTW. Firstly, we select the prototype minimizing the distance to all other segments. To reduce the computation complexity, we use Euclidean distance to measure the similarity between two segments in this step. Then, according to the best warping path, we warp all other segments into the prototype. Finally, we take the average value as the final prototype of the cluster. At the same time, the minimum distance of all the segments to the prototype in each cluster is calculated by Eq.(12) and will be used as a threshold value in the following recognition process.

3.5. Recognition

Similar to Li et al. [1], the recognition process is carried out in two steps. The first step is to eliminate the markedly different behavior models according to the threshold value. Moreover, DTW is used to warp the test behavior pattern into remainder models. In the second step, space-time image features are extracted with the method mentioned above and the similarity distances are calculated as:

$$f_p = \exp\left(-\sqrt{\sum_{k=1}^K \sum_{l=1}^L [V_i(k, l) - V_p(k, l)]^2 / KL}\right). \quad (13)$$

where V_i is the test behavior vector; V_p is the vector of the p^{th} behavior prototype; $K = 16$, $L = 14$; f_p is the similarity distance between the test behavior and the p^{th} behavior prototype. So the test behavior belongs to the class with maximum similarity.

4. Experiments and Results

We test our method on a total of four gestures performed by six subjects. Images are of size of 320×240 pixels and frame rate is of 10 frames/sec. The video totals 13004 frames. Gesture No.1 to 4, namely, “waving right hand”, “hand clapping”, “kicking left foot”, “kicking right foot” are shown in figure 4. Except the four gestures, the video contains other behaviors that occur over a relatively small proportion of the total time. We slice this video into 1617 segments with the fix time duration $T = 64$

frames and the step length is 8 frames. The 1000 segments selected at random are classified into 10 classes (figure 5). We select the first 4 clusters corresponding to the most frequently occurring gestures. The four clusters are “ waving right hand ” (165 segments), “ hand clapping ”(262 segments), “ kicking left foot ” (186 segments) and “ kicking right foot ” (123 segments), respectively.

Figure 4. Several frames from gestures “waving right hand”, “hand clapping”, “kicking left foot”, “kicking right foot”.

Figure 5. The normalized affinity matrix ordered according to the result of clustering.

Table 1. Experimental results

Gesture No	1	2	3	4	Total
#Training	165	262	186	123	736
#Testing	119	136	124	85	464
#Correct	105	117	110	72	404

The remainders (617 segments) with manual labeling are used for testing. The final recognition results are shown in Table 1. The recognition rate on the testing data is 87.1%.

5. Conclusions

A novel space-time image features based method is proposed for behavior modeling and recognition without manual labeling. Experiments show that our method is effective. In our method, it doesn't need to

track the body parts and the behaviors can be modeled without any prior knowledge. In the future, behavior modeling from multiple views will be studied to increase the correct recognition rate and behavior modeling in the case of sparse data will be considered.

Acknowledgements

This work was supported by the National High Technology R@D Program of China under the grant No: 2005AA114130.

References

- [1] H. Li and M. Greenspan. Multi-scale gesture recognition from time-Varying contours. In *IEEE Int'l Conf. On Computer Vision*, pp.236-243, October 2005.
- [2] S.G.Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *IEEE Int'l Conf. On Computer Vision*, pp.742-749, October 2003.
- [3] A.F.Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(3): 257-267, March 2001.
- [4] I.Laptev and T. Linderberg. Space-time interest points. In *IEEE Int'l Conf. On Computer Vision*, pp.432-439, October 2003.
- [5] L. Zelnik-Manor and M. Irani, Event-based analysis of video. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pp.123-130, December 2001.
- [6] T. Xiang and S. G. Gong. Video behaviour profiling and abnormality detection without manual labeling. In *IEEE Int'l Conf. On Computer Vision*, pp.1238-1245, October 2005.
- [7] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *IEEE Int'l Conf. On Computer Vision*, pp.975-982, September 1999.
- [8] S.Chu, E.Keogh, D.Hart, M.Pazzani, Iterative deepening dynamic time warping for time series, in: *International Conference on Data Mining, SDM*, 2002.
- [9] A. Corradini. Dynamic time warping for off-line recognition of a small gesture vocabulary. In *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 82-89, 2001.
- [10] Y. Yang and M. Levine. The background primal sketch: an approach for tracking moving objects. *Machine Vision and Application*, pp. 17-34, 1992.