Periodic Motion Detection with ROI-based Similarity Measure and Extrema-based Reference-Frame Selection

Xintong Han[1], Gaojian Li[2], Weiyao Lin*[1], Xiaoqiong Su[1], Hongxiang Li[3], Hua Yang[1], Hui Wei[2]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University, China (*Corresponding author)
[2]School of Computer Science and Technology, Fudan University, China
[3]Department of Electrical and Computer Engineering, University of Louisville, USA

***Abstract*— This paper presents a new algorithm for detecting and analyzing the periodic motions in video sequences. Different from the previous methods which detect periodic motions from the entire frame, we propose a convex-hull-based process to automatically determine the regions of interest (ROI) of the motions and utilize an ROI-based similarity measure to detect the motion periods. Furthermore, we also propose an extrema-based method to select the optimal reference frame for further improving the periodic detection performance. Our proposed algorithm can not only effectively detect motion periods with both constant and variable period lengths, but also have obvious advantage when handling periodic motion with slight movements. Experimental results demonstrate the effectiveness of our proposed method.**

I. INTRODUCTION

Periodic motions happen frequently in our daily life. Some example periodic motions include waving hands, walking, dumbbell lifting, or ocean waves. Nowadays, detecting and analyzing these periodic motions is of great importance in many applications [1-4]. For example, extracting gait periods for gait recognition [4], counting the number of periods for automatic sport analysis [1], and detecting activities with irregular periods as abnormal activities [7].

Basically, the motion period detection methods can be divided into two classes: the transfer-based methods and the waveform-based methods.

The transfer-based methods first transfer the video signals into some transform domain and then perform period detection accordingly. For example, some frequency-domain-based methods transfer the video signal into the frequency domain and extract the largest peak other than the zero frequency as the estimated period length. Briassouli and Ahuja [6] project the pixel values of each video frame onto the x and y axes to get two signals over time and then utilize time-frequency analysis on these two signals for period estimation. However, since the transfer-based methods are based on the analysis in the transform domain, they can only deal with the motions with constant periods and will fail to handle motions with varying period lengths. Furthermore, transfer-based analysis normally requires large number of periods in order for estimating the accurate period length. Thus, they may have low accuracy when applied to the video sequence including few periods.

The waveform-based methods first extract a 1-dimension waveform from the video sequence which reflects the motion's periodic variation over time. Then the periods can be extracted by analyzing this waveform. Since the waveform-based methods are more flexible and have low requirements on the number of periods, they are more widely used for period detection. In this paper, we also focus on discussing this class of methods. Various waveform-based algorithms have been proposed. For example, Cutler and Davis [7] compute the self-similarity waveform based on the absolute differences between frames and then create a 2-dimension lattice structure for matching the self-similarity waveform to find the period. However, this method is still limited due to its low capability in handling varying-length periodic motions. Wang et al. [4] extract the width of the object's lower body as the waveform to extract gait periods. Although, this method can be effective in detecting the period of gaits, it is based on very specific assumptions such that it cannot be extended for detecting other motions.

Furthermore, most of the existing waveform-based methods have the following two major limitations:

(a) Many existing methods (also including the transfer-based methods) find the periods based on the entire object or the entire frame. However, in practice, many periodic motions are only reflected by parts of the objects (e.g., the hands in waving-hand motion). When including the entire object, the noisy movements of the other irrelevant parts may greatly affect the final results. These noisy effects will become extremely obvious for periodic motions with only slight movements. Although some algorithms [1, 2, 4] have tried to exclude the noisy movements by identifying or tracking the specific parts, most of them are quite ad-hoc which cannot be easily extendable. Therefore, a more general and automatic method is desired.

(b) Since the similarity between frames can well reflect the motion's periodic variation over time, the similarity can be a good feature when extracting the 1-dimension waveforms. However, in order to calculate this similarity value, a reference frame is required (i.e., the similarity is calculated between the current frame and this reference frame). Most existing similarity-based methods simply select the first frame or manually select a frame as reference [1, 5]. This less-optimal reference selection may also greatly affect the final performance. Thus, it is also important to develop new methods to find the optimal reference frame.

In this paper, we propose a new algorithm for periodic motion detection. The proposed algorithm has the following major contributions:

(a) We propose a convex-hull-based (CHB) process to automatically determine the regions of interest (ROIs) of the motions. By this way, only the ROIs are considered for periodic detection while the noisy movements of the other irrelevant parts can be effectively excluded.

(b) We propose an ROI-based similarity measure (ROI-SIM) to detect the motion periods. The ROI-based similarity measure incorporates the weighted sum of features over the extracted ROIs such that the relative importance of ROIs can be properly balanced.

(c) We also propose an extrema-based (EB) method to select the optimal reference frame for further improving the periodic detection performance.

The rest of the paper is organized as follows. In section II, the framework of the proposed algorithm is described. Section III describes the details of the algorithm, including the CHB process, the ROI-SIM, and the EB method. Section IV describes the two proposed metrics for evaluating the periodic motion detection performances. And Section V shows the experimental results.

II. FRAMEWORK

The framework of our proposed algorithm is shown in Fig. 1. In Fig. 1, for the input video sequence, the object foregrounds

are first extracted and aligned. Then the CHB process is used to determine the ROIs of the motion object. After that, features are extracted from the ROIs, and ROI-SIM is calculated between frame pairs based on the extracted features. The ROI-SIMs for different frame pairs are used to construct a similarity plot and our EB method is utilized to select the optimal reference frame from this similarity plot. Finally, with the optimal reference frame, a similarity waveform can be created and the motion periods can be estimated from this waveform.

In the following section, we will describe the steps in Fig. 1 in detail. It should be noted that the grey blocks in Fig. 1 (i.e., "Convex-hull-based process for ROI extraction", "ROI-SIM calculation between any two frames", and "use EB method for selecting the optimal reference frame") are the key contributions of our algorithm. Therefore, more attention will be put on these steps in the following section.

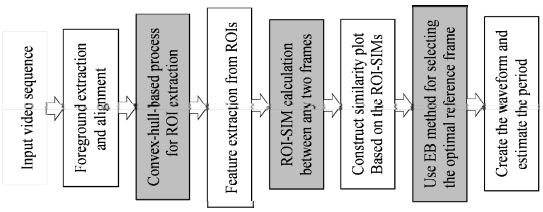


Fig. 1 The framework of our proposed algorithm.

III. Details of the Proposed Algorithm

A. Foreground Extraction and Alignment

In order to determine proper ROIs about the periodic motion, foregrounds of the motion objects need to be first extracted. There can be various ways to extract the object foregrounds such as background subtraction or frame difference [1]. In this paper, we use a Kinect camera with the depth information to extract the foreground [8]. Furthermore, for motions with location change (such as walking), we also align the foreground from different frames (i.e., align the foregrounds to the same location in the frame) [3, 4, 8]. Some examples of the extracted foreground is shown in Fig. 2 (a). It should be noted that in our algorithm, only the foreground information is used for periodic detection while the more sophisticated functions such as human parts detection and tracking are not utilized. This makes our algorithm general and easily extendable to detecting various periodic motions such as ocean waves.

B. CHB Process for ROI Extraction

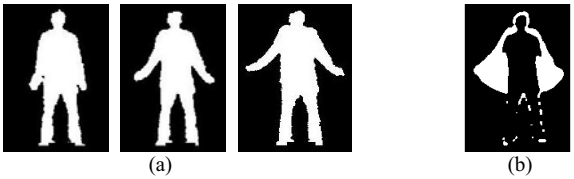
The CHB process includes two sub-steps: (a) calculate the motion regions, and (b) determine the ROIs by convex hull. They are described in the following, respectively.

1) Calculating the Motion Regions

Let F_i be the i-th foreground frame of the input sequence ($F_i(x, y)$=1 if (x, y) is a foreground pixel and $F_i(x, y)$=0 otherwise). In order to calculate the motion regions, we first calculate the absolute difference between neighboring frame pairs F_i and F_{i+1}. After that, all the absolute differences are summed up to form a binary image, we call this image the binary change image (BCI) and it can be calculated by:

$$BCI(x,y)=\begin{cases}1 & if \sum_{i=0}^{N-1}\left|F_{i+1}(x,y)-F_i(x,y)\right|>0\\ 0 & otherwise\end{cases} \quad (1)$$

where N is the length of the sequence. Obviously, non-zero pixels in BCI represent the motion regions for the input sequences. Some examples of the input foreground frames and the resulting BCI is shown in Fig. 2. It should be noted that in order to further exclude the noise in the foreground frames, median filtering is applied on the BCI.



(a) (b)

Fig. 2 (a) Foreground frames of a periodic motion video (a person shrugging and holding out his hands). (b) The binary motion change image (BCI) of the video.

2) Determine the ROIs by Convex Hull

From Fig. 2 (b), we can see that although many of the non-moving areas have been excluded in BCI, there are still many noisy regions remaining. These regions may still greatly affect the final performance. Therefore, further filtering is required for extracting the accurate ROIs. In this paper, we propose to use a convex-hull-based method for extracting the ROIs. It is described in the following.

We first divide the BCI into several connected regions. And then, the convex hull of each region is calculated (i.e., the smallest convex polygon containing the region, see Fig. (a)). Here, we use R_i to denote the i-th connected region, and C_i to denote the convex hull of R_i.

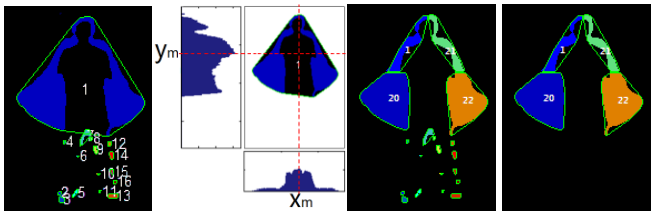
Then, for each convex hull C_i, we calculate its solidity SD_i, i.e., the portion of non-zero pixels in the convex hull: $SD_i=A(R_i)/A(C_i)$, where $A(R_i)$ is the number of pixels for R_i, and $A(C_i)$ is the number of pixels in the convex hull.

The solidity indicates whether the convex hull is "fully" filled with the connected region. If a region has low solidity value (i.e., below a threshold Th_s), it means that the convex hull is not well suited for the connected region and further partition is needed. The partition of the convex hull and the connected region can be described as follows:

(a) We project the zero-value pixels (i.e., $BCI(x, y)$=0) in the convex hull C_j onto x and y axes to form two histograms. And then find the peaks x_m and y_m on these two histograms.

(b) Then the connected region R_j is separated into smaller connected regions by the two lines: $x=x_m$ and $y=y_m$.

(c) Finally, the solidity of these smaller convex hulls are re-calculated to decide whether to further partition.



(a) (b) (c) (d)

Fig. 3 The example of determining regions of interest by convex hull. (a) Calculate the connected regions and their convex hulls. (b) Partition of R_1 which has low solidity value by projecting pixels on x and y axes. (c) R_1 is partitioned into 4 smaller regions and their convex hulls are re-calculated. (d) After filtering, the ROIs C_1, C_{20}, C_{21}, C_{22} are determined.

Fig. 3 (b) is an example to show the partition process of a connected region. This partition process continues until the solidities of all connected regions exceed Th_s (Fig. 3 (c)).

Now we regard all the regions in the convex hulls as candidate ROIs and try to filter out the noisy ones. The filtering is based on the following rules:

(a) Since small convex hulls are more likely to contain noise and non-periodic parts which will influence the detection of the period, we apply a threshold to eliminate smaller convex hulls.

(b)Furthermore, since noisy convex hulls have some common patterns, we also select some noisy convex hulls as the training data and pre-calculate their x and y axes histograms. And convex hulls whose histograms are similar to these training ones will also be excluded.

After filtering the noisy convex hulls, the remaining convex

hulls will be the determined ROIs (see Fig. 3 (d)).

C. Feature Extraction for ROIs

After ROIs are determined, we need to extract the features from these ROIs in each frame to capture the motion variations. Note that our proposed framework is general and it allows various feature extraction methods either on the original color frame or on the foreground frame. In this paper, for an input frame, we first calculate the SURF feature vectors for the interest points [10] in each ROI and then sum them up for representing this ROI [10]. Note that the locations and scales of the SURF interest points in each ROI are fixed for different frames in order to make the feature vectors to have the same length over different frames. We denote $\boldsymbol{h_{i,j}}$ as the summed-up feature vector of the j-th ROI I_j in the frame F_i.

D. ROI-based Similarity between Frame Pairs

After extracting the features for each ROI in the current frame F_i, these features are weighted and concatenated in order as a long vector $\boldsymbol{H_i}$:

$$\boldsymbol{H}_i = [A(I_1)\boldsymbol{h}_{i,1}, A(I_2)\boldsymbol{h}_{i,2}, ..., A(I_M)\boldsymbol{h}_{i,M}] \quad (2)$$

where $\boldsymbol{h_{i,j}}$ is the feather vector for ROI I_j in F_i, and $A(I_j)$ is the weight equal to the number of non-zero pixels in I_j. After normalizing $\boldsymbol{H_i}$, the resulting normalized $\boldsymbol{H_i^*}$ will be the ROI-based feature vector for representing the current frame F_i. The process is illustrated in Fig. 4.

Then the ROI-based similarity (ROI-SIM) $S_{i,j}$ between any frame pairs F_i and F_j can be calculated by:

$$S_{i,j} = IH\left(\boldsymbol{H}_i^*, \boldsymbol{H}_j^*\right) \quad (3)$$

where $IH(\cdot)$ is the operation to calculate the intersection area of the two vectors [11]. Normally, bigger $S_{i,j}$ indicates larger similarity between frame pairs.

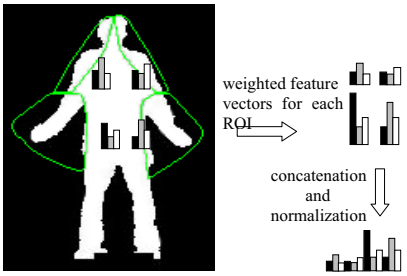


Fig. 4 Example of computing the ROI-based feature vector of a frame. (a) in each ROI, SURF features are calculated and summed up as the feature vectors for each ROI, (b) upper: weight the feature vectors for each ROI by the ROI's size, down: the weighted feature vectors for different ROIs are concatenated and normalized as the final descriptor of the current frame.

E. Constructing Similarity Plot

With the ROI-SIM between any frame pairs, we can construct a similarity plot: $\boldsymbol{M}=[S_{i,j}]_{N\times N}$ where $\boldsymbol{M}$ is a $N\times N$ matrix and its i-th row j-th column element is the ROI-SIM between frame pairs F_i and F_j. Fig. 5 (a) shows one example similarity plot for the shrugging-shoulder motion in Fig. 2.

In Fig. 5 (a), the bright regions show larger similarity. We find that there is a bright line on the main diagonal of the similarity plot. This is because the image is always similar to itself. Also note that for a motion with constant period, since every frame is similar to the frame after a constant time, the bright lines in the similarity plot should be straight and has a slope of 45° or 135°. However, since the motion in Fig. 2 has varying period length, it is obvious that the bright lines in Fig. 5 (a) are not perfectly straight.

F. Use the Extrema-based (EB) Method to Select the Optimal Reference Frame

After achieving the similarity plot, we need to find a reference frame such that the similarity waveform can be achieved by calculating the ROI-SIM between each frame and this reference frame: $\boldsymbol{W}_k=[S_{1,k}, S_{2,k}, S_{3,k}, ... S_{N,k}]$, where $\boldsymbol{W_k}$ is the resulting similarity waveform when the k-th frame F_k is selected as the reference frame. For periodic motions, the ideal waveform should show up-and-down shapes while peaks can be viewed as the starting point for each period (i.e., peak frames are the times when the object turns back to the posture as the reference one).

However, as mentioned, the estimated period results are sensitive to the selection of reference frames. This is because: (a) The period lengths for many motions may be varying. (b) The similarity may create "fake" peaks (i.e., peaks not for the period starting points) for non-optimal reference frames. For example, in the arm-waving motion, if the frame where the arm is waved in the middle is selected as the reference frame, a fake peak will appear when the arm first turns back to the middle position. Therefore, it is also important to select the proper reference frame for creating less noisy waveforms. Thus, we also propose an extrema-based method to select the optimal reference frame.

It is obvious that $\boldsymbol{W_k}$ actually corresponds to the k-th row in the similarity plot $\boldsymbol{M}$. Therefore, the problem of selecting the optimal reference frame is the same as selecting a suitable row in the similarity plot to best reflect the periodic motion variations. Based on the observation that peaks for the true periods are less affected by the reference frame while "fake" peaks are sensitive to reference frames, our proposed EB method tries to find the reference frame F_r with the smallest number of significant peaks as the optimal reference frame:

$$F_r = \underset{k}{arg\,min}\left(NP\left(\boldsymbol{W}_k\right)\right) \quad (4)$$

where $NP(\boldsymbol{W_k})$ is the total number of significant peaks in $\boldsymbol{W_k}$. A frame F_i is a significant peak frame for $\boldsymbol{W_k}$ if:

$$\begin{cases} S_{i,k} > max\left(\left\{S_{i-q,k} \mid q = -L, -L+1, ..., L, q \neq 0\right\}\right) \\ S_{i,k} > \mu_{W_k} + \sigma_{W_k}/2 \end{cases} \quad (5)$$

where L is the neighborhood area around F_i. μ_{Wk} and σ_{Wk} are the mean and variance of the waveform $\boldsymbol{W_k}$.

Fig. 5 shows an example to explain the effect of different reference frames, the significant peaks are marked out by red circles. If the reference frame is not properly selected (as in Fig. 5 (b)), many fake peaks will be included. By using our EB method to select a suitable reference frame (as in Fig. 5 (c)), the fake peaks can be effectively avoided.

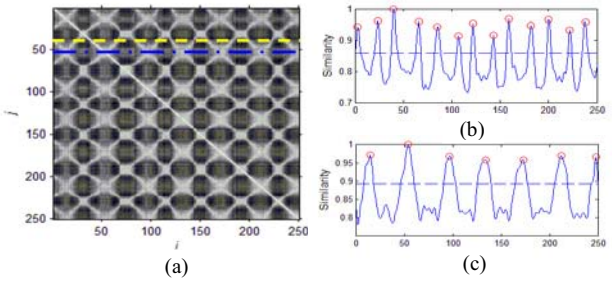


Fig. 5 (a) The similarity plot of the periodic motion in Fig. 2. (b) Similarity waveform when the reference frame is F_{40} (correspond to the dashed line in (a)). (c) Similarity waveform for reference frame F_{54} (the dot-dash line in (a)).

G. Estimate the period from the waveform

After achieving the optimal reference frame, the corresponding similarity waveform can be calculated. Finally, the periods of the periodic motions can be detected by finding the significant peaks in (6).

IV. Metrics for Evaluating Periodic Motion Detection

In this section, we propose two metrics for evaluating the periodic motion detection performances: the average period length difference e_P and the average starting point deviation e_T. They are defined in (6) and described in detail in the following.

For an input video sequence, after the reference frame is determined, we can manually achieve the ground-truth starting point frames for each period in this sequence as: $P=\{t_1,t_2,...,t_K\}$. Also, we can get an estimated starting point set P' from the periodic motion detection algorithms: $P=\{t_1',t_2',...,t_{K'}'\}$. Then, the two evaluation metrics can be calculated by:

$$e_P = \frac{\left|\frac{t_{K'}'-t_1'}{K'-1} - \frac{t_K-t_1}{K-1}\right|}{(t_K-t_1)/(K-1)}, e_T = \frac{\frac{1}{K}\sum_{i=1}^{K}|t_i^*-t_i|}{(t_K-t_1)/(K-1)} \quad (6)$$

where $t_i^* = arg_{t_j'} min\left|t_j'-t_i\right|$. From (6), we can see that the average period length difference e_P more reflects the total number of periods counted by the algorithm. When a period is missed or over-counted, e_P will become large. On the other hand, the average starting point deviation e_T more reflects the algorithm's accuracy in determining the starting points of the periods. If the detected period starting points have larger distances to those of the ground truth, e_T will become large.

V. Experimental Results

In this section, we show experimental results of our proposed algorithm. Due to the limited space, only parts of the results are shown in this paper. Fig. 7 (a) shows an example periodic motion where the person is moving his hand left and right, and Fig. 7 (b) shows the extracted binary motion change image (BCI) as well as the ROI. Fig. 6 compares the similarity waveforms of the action in Fig. 7 (a) for the following methods:

(a) Use the sum of absolute difference (SAD) between frames for creating the similarity waveform [5]. And the reference frame is randomly selected as the first frame (SAD+RF$_{random}$).

(b) Use the SAD as the similarity metric, but the reference frame is selected by (5) (SAD+RF$_{opt}$).

(c) Do not extract ROIs. Directly extract SURF features [10] on the entire frame and calculate the similarity by (3) while the reference frame is selected by (4). (non-ROI+RF$_{opt}$).

(d) Use our proposed ROI-SIM for waveform and (4) for reference frame selection (ROI-SIM+RF$_{opt}$).

From Fig. 6, we can see that if the reference frame is not properly selected, the resulting waveform will be extremely noisy with numerous fake peaks, as in (a). Comparatively, by selecting a proper reference frame, the periodic variations can be more obviously captured in (b)-(d). Further comparing (b)-(c) with (d), we can see that since the entire frame information is used for similarity calculation, the waveforms in (b) and (c) are still noisy, which will greatly affect their period detection performance (these noisy effects are extremely severe for some sequences). However, by using our proposed ROI-SIM, the waveforms are obviously smoother and the periodic variations are more precisely represented.

Furthermore, Fig. 7 (c) compares the period detection results of the four methods: (a) SAD+RF$_{opt}$ [5] (note that we detect significant valley instead of peak for this method), (b) non-ROI+RF$_{opt}$, (c) the time-frequency-analysis method (TFAM) [7], and (d) ROI-SIM+RF$_{opt}$. In Fig. 7 (c), the red dashed lines are the ground truth period starting points while the blue solid lines with different markers are the detected period starting points. From Fig. 7 (c), we can see that since the waveform of SAD+RF$_{opt}$ is quite noisy, there are many over-counted period points. Although these over-counting problem is reduced for the TFAM and non-ROI+RF$_{opt}$ methods, their detected period starting points have lower accuracy (i.e., the points have larger distance to the ground-truth points. Compared to these methods, the periods detected by our proposed ROI-SIM+RF$_{opt}$ algorithm not only have low over-counting rates, but are also close to the ground truth points.

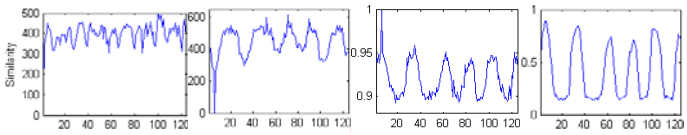


Fig. 6 The similarity waveforms for different methods: (a) SAD+RF$_{random}$ (b) SAD+RF$_{opt}$ (c) non-ROI+RF$_{opt}$ (d) ROI-SIM+ RF$_{opt}$

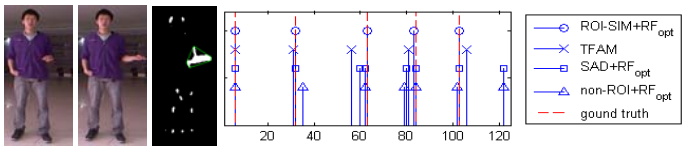


Fig. 7 Moving hand sequence: (a) Frame 24, 33 (b) The white part is the BCI and extracted ROI is within the green polygon. (c) The estimated period starting points for different methods and the red dotted lines are the ground truth.

Finally, Table I compares the e_T and e_P metrics for different methods on two datasets: our created dataset and Weizmann dataset [9]. Note that our dataset was captured by Kinect [8] and it includes 50 different periodic motions with each sequence contains 200-300 frames. And the Weizmann dataset includes 9 periodic motions performed by 9 different people [9]. From Table I, it is clear that some methods (such as SAD+RF$_{opt}$) have high e_T rate since they will easily miss or over-count periods. And some methods (such as TFAM and non-ROI+RF$_{opt}$) have high e_P rate since their detected period starting points have larger distance to the ground truth. Comparatively, our proposed algorithm (ROI-SIM+RF$_{opt}$) has the lowest rates on both metrics.

TABLE I Performance of Different Methods

	Our dataset		Weizmann dataset [9]	
	e_T	e_P	e_T	e_P
SAD+RF$_{opt}$	22.05%	5.88%	6.98%	6.51%
non-ROI+RF$_{opt}$	5.44%	9.71%	1.72%	3.74%
TFAM	4.41%	12.50%	3.36%	8.51%
ROI-SIM+ RF$_{opt}$	**1.79%**	**4.95%**	**0.47%**	**2.47%**

Acknowledgements

This work was supported in part by the following grants: National Science Foundation of China grants (61001146, 61025005, 61103124), Chinese national 973 project grants (2010CB731401), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), the SMC grant of SJTU, Shanghai Pujiang Program (12PJ1404300), and China National Key Technology R&D Program (2012BAH07B01).

References

[1] Y. Ren, B. Fan, W. Lin, X. Yang, H. Li, W. Li, D. Liu, "An efficient framework for analyzing periodical activities in sports videos," *IEEE Int'l Cong. Image and Signal Processing* (*CISP*), pp. 502-506, 2011.

[2] A. B. Albu, R. Bergevin, and S. Quirion, "Generic temporal segmentation of cyclic human motion," *Pattern Recognition*, vol. 41, pp. 6–21, 2008.

[3] I. Laptev, S.J Belongie, P. Perez, and J. Wills, "Periodic motion detection and segmentation via approximate sequence alignment," *ICCV*, 2005.

[4] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-Gait Image: A Novel Temporal Template for Gait Recognition," *ECCV*, 2010.

[5] R. Cutler, L. Davis, "View-based detection and analysis of periodic motion," *Int'l Conf. Pattern Recognition* (*ICPR*), pp. 495-500, 1998.

[6] A. Briassouli and N. Ahuja, "Extraction and analysis of multiple periodic motions in video sequences," *TPAMI*, vol. 29, pp. 1244-1261, 2007.

[7] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis and applications," *TPAMI*, vol. 22, pp. 781-796, Aug. 2000.

[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *CVPR*, pp. 1297-1304, 2011.

[9] Weizmann set: www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions

[10] H. Bay, A. Ess, T. Tuytelaars, L. Gool, "SURF: speeded up robust features", *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, 2008.

[11] M. Swain and D. Ballard, "Color indexing." *Int'l Journal on Computer Vision*, vol.1, no. 7, pp. 11-32, 1991.