Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 219190, 11 pages
doi:10.1155/2010/219190

Research Article

Human Action Recognition Using Ordinal Measure of Accumulated Motion

Wonjun Kim, Jaeho Lee, Minjin Kim, Daeyoung Oh, and Changick Kim

Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 119 Munji Street, Yuseong-gu, Daejeon 305-714, South Korea

Correspondence should be addressed to Changick Kim, cikim@ee.kaist.ac.kr

Received 14 December 2009; Accepted 1 February 2010

Academic Editor: Jenq-Neng Hwang

This paper presents a method for recognizing human actions from a single query action video. We propose an action recognition scheme based on the ordinal measure of accumulated motion, which is robust to variations of appearances. To this end, we first define the accumulated motion image (AMI) using image differences. Then the AMI of the query action video is resized to a $N \times N$ subimage by intensity averaging and a rank matrix is generated by ordering the sample values in the sub-image. By computing the distances from the rank matrix of the query action video to the rank matrices of all local windows in the target video, local windows close to the query action are detected as candidates. To find the best match among the candidates, their energy histograms, which are obtained by projecting AMI values in horizontal and vertical directions, respectively, are compared with those of the query action video. The proposed method does not require any preprocessing task such as learning and segmentation. To justify the efficiency and robustness of our approach, the experiments are conducted on various datasets.

1. Introduction

Recognizing human actions has become critical with increasing demand of high-level scene understanding to analyze behaviors and interactions of humans in the scene. It can be widely applied for numerous applications, such as video surveillance, video indexing, and event detection [1]. For example, irregular actions in public places can be detected by using the action recognition systems [2]. However, such action recognition systems still suffer from problems depending on variations of appearance. For example, the different clothes and genders yield significant differentiation of appearance in conducting similar actions. Also, same actions may be misclassified as different actions due to objects carried by actors [3] (see Figure 1). In these situations, traditional template matching based algorithm may fail to detect a given query action. Thus, it is worth noting that building an efficient and robust action recognition system is a challenging task.

There are two types of human action recognition models: *learning-based models* and *template-based models*. In the former, reliable action dataset is essentially needed to build a classifier whereas the single template (i.e., training-free) is used to find the query action in target video sequences in the latter. Since it is hard to maintain the large dataset for real applications, the latest algorithms for human action recognition tend to be template-based. In this sense, we also propose a template-based action recognition method for static camera applications.

Main contributions of the proposed method are summarized as follows: first, the accumulated motion image (AMI) is defined by using image differences to represent the spatiotemporal features of occurring actions. It should be emphasized that only areas containing changes are meaningful for computing AMI instead of the whole silhouette of human body as in previous methods [4, 5]. Thus, the segmentation task such as background subtraction to obtain the silhouette of human body is not required in our method. Secondly, we propose to employ the ordinal measure of accumulated motion for detecting query actions in target video sequences. Our method is motivated by the earlier work using the ordinal measure for detecting image and video copies [6, 7], in which authors show that the ordinal measure is robust to various modifications of original images. Thus, it can be employed to cope with variations of appearance for the accurate action recognition. Finally, the energy histograms, which are obtained by projecting AMI values in horizontal and vertical directions, are used to determine the best match among local windows detected as candidates close to the query action.

The rest of this paper is organized as follows: the related work is briefly summarized in Section 2. The technical details about the steps outlined above are explained in Section 3. Various real videos are tested to justify the efficiency and robustness of our proposed method in Section 4 and followed by conclusion in Section 5.

2. Review of Related Work

Human action recognition has been widely studied for last several decades. Bobick and Davis [8] propose the temporal templates as models for actions. They construct two vector images, that is, motion energy image (MEI) and motion history image (MHI), which are designed to encode a variety of motion properties. In detail, an MEI is a cumulative motion image whereas an MHI denotes recent moving pixels. Finally, these view-specific templates are matched against the model of query actions. Schüldt et al. [9] use space-time interest points proposed in [10] to represent the motion patterns and integrate such representations with SVM classification schemes. Ikizler et al. [11] propose to use lines and optical flow histograms for human action recognition. In particular, they introduce a new shape descriptor based on the distribution of lines fitted to the silhouette of human body. In [12], authors define the integral video to efficiently calculate 3D spatiotemporal volumetric features and train cascaded classifiers to select features and recognize human actions. Hu et al. [13] use the MHI along with foreground image obtained by background subtraction and the histogram of oriented gradients (HOG) [14] to obtain discriminative features for action recognition. Then they build a multiple-instance learning framework to improve the performance. Authors of [15] propose to use the mixture particle filters and then cluster the particles using local nonparametric clustering. However, these approaches require supervised learning based on the large reliable dataset before recognizing human actions.

Yilmaz and Shah [17] encode both shape and motion features to represent the 3D action models. More specifically, they treat actions as 3D objects in (x, y, t) space and compute action descriptors by analyzing the differential geometrical properties of spatiotemporal volume. Gorelick et al. [18] also induce the silhouette in the space-time volume for human action recognition. Unlike [17], they use the blobs obtained by background subtraction instead of contours. However, these silhouette-based approaches require accurate background subtraction.

A recent trend in human action recognition has been toward the template-based models as mentioned. Shechtman and Irani [19] introduce a novel similarity measure based on the correlation of behavior. They use intensity values in a small space-time patch. In detail, a space-time video template for the query action consists of such small space-time patches. It is correlated against a larger target video sequence by checking its consistency with every video segment to find the best match with the given query action. Furthermore, they propose to measure similarity between actions based on matching internal self-similarities [20]. Ning et al. [21] propose the hierarchical space-time framework enabling efficient search for desirable actions. Similar to [19], they also use the correlation between the query action template and candidates in the target video. However, these approaches may be unstable under noisy environments. In [3], authors propose the space-time local steering kernels (LSK) to represent the volumetric features. They compare the 3D LSK features of the query action efficiently against those obtained from the target video sequences using a matrix generalization of the cosine similarity measure. Although the shape information is well defined in the LSK features, it is hard to apply it for real-time applications due to the high dimensionality.

Basically, our approach belongs to the *template-based model*. Unlike previous methods, the ordinal measure employed in our method easily generalizes across appearance variations due to different clothes and body figures. Further technical details will be presented in the following section.

3. Proposed Method

The proposed method consists of three stages: AMI computation, candidate detection by using the ordinal measure of accumulated action, and determination of the best match based on the energy histograms. Overall procedure of the proposed method is shown in Figure 2.

3.1. Accumulated Motion Image (AMI). Since the accumulated motion is differentiable across various actions, it can be regarded as a discriminative feature for recognizing human actions. Based on this observation, we introduce a new feature, AMI, enabling efficient representation of the accumulated motion.

Our feature, AMI, is motivated by the gait energy image (GEI) popularly used for the individual recognition [22] and gender classification [23]. However, as compared to GEI, only areas including changes are used to compute AMI instead of requiring the whole silhouette of human body. To this end, the gray-level AMI is defined by using image differences as follows:

$$\mathrm{AMI}(x, y) = \frac{1}{T} \sum_{t=1}^{T} |D(x, y, t)|, \tag{1}$$

where $D(x, y, t) = I(x, y, t) - I(x, y, t-1)$ and T denotes the length of the query action video (i.e., total number of frames). We name it as accumulated motion image because: (1) AMI represents the time-normalized accumulative action energy and (2) pixels with higher intensity values in the AMI denote that motions occur more frequently at the positions. Although our AMI is related to MEI and MHI proposed

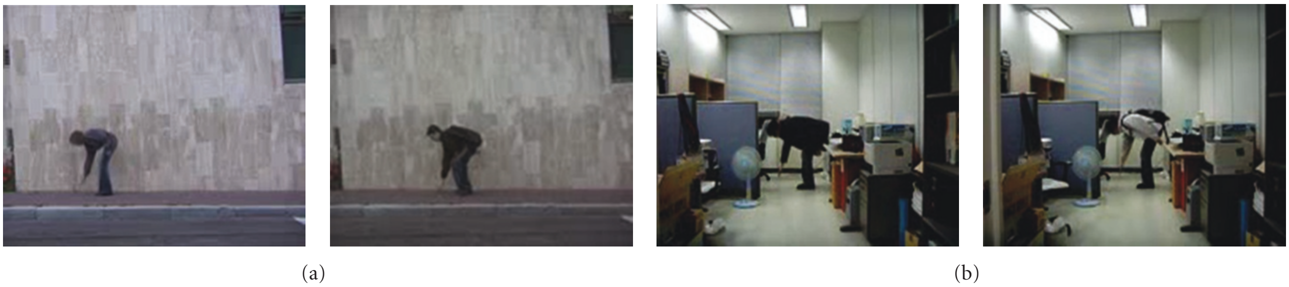


Figure 1: (a) Variations of appearance due to different clothes in the same action. (b) Different appearance in conducting the same action due to backpack.

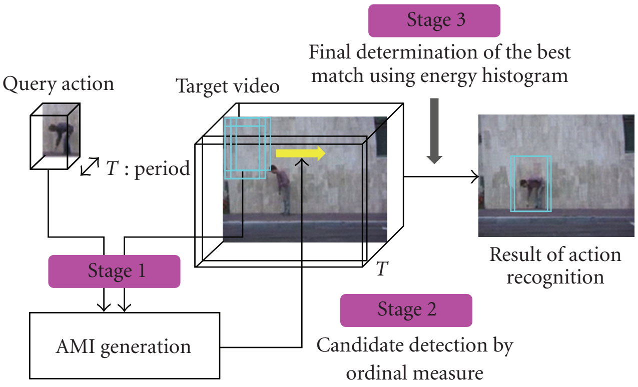

Stage 3
Final determination of the best match using energy histogram
Query action
Target video
T : period
T
Result of action recognition
Stage 1
Stage 2
AMI generation
Candidate detection by ordinal measure

Figure 2: Overall procedure of the proposed method.

by Bobick and Davis [8], there is a fundamental difference. More specifically, the equal weights for all change areas are given in MEI. The higher weights are assigned to new frames whereas low weights are assigned to older frames in MHI. Therefore, both of them are not suitable for representing the accumulated motion for our ordinal measure, which will be explained in the following subsection. As compared to MEI and MHI, AMI describes the accumulated motion by using the pixel intensity. The examples of AMI for some actions are shown in Figure 3.

3.2. Ordinal Measure for Detecting Candidates. Traditional template-based action recognition techniques have relied on the shape correspondence. The distances between the query action and all local windows in the target videos are computed based on the shape similarities of corresponding windows. However, most of them are apt to fail in tolerating variations of appearance due to the clothes and objects carried by actors, which is often observed in surveillance environments. To solve this problem, we employ the ordinal measure for computing the similarity between different actions, which is very robust to various signal modifications [7]. For example, two subimages of the same action obtained by resizing AMIs are shown in Figure 4, which have variations of appearance due to different clothes and backpack. The values of resized AMI are quite different between two subimages whereas the ordinal signatures between corresponding subimages are identical. Thus, we believe that the ordinal measure of accumulated motion can provide a more efficient way of recognizing human actions.

To this end, AMI is firstly resized to a $N \times N$ subimage by intensity averaging as shown in Figure 4. Let us define the $1 \times M$ rank matrix of resized AMI for the query action video V^q as $\mathbf{R}(V^q)$ where M equals to $N \times N$. It is set to 9 in our implementation. For example, the rank matrix of the query action can be represented as $\mathbf{R}(V^q) = [5, 1, 6, 4, 2, 3, 9, 7, 8]$ in Figure 4 and also each element of the rank matrix can be expressed as $\mathbf{R}^1(V^q) = 5, \mathbf{R}^2(V^q) = 1, \ldots, \mathbf{R}^9(V^q) = 8$. Thus, the accumulated motion of query video is effectively encoded in a single rank matrix.

Then the rank matrix of the query action video should be matched against the rank matrices of all local windows to detect candidates close to the query action. Here centers of local windows are positioned four pixels apart from each other in the target video frame and thus they are densely overlapped in horizontal and vertical directions, respectively (see Figure 2). For example, total 1681(= 41 × 41) comparisons need to be performed for the target video frame of 200 × 200 pixels with given local windows of 40 × 40 pixels. The ith frame of the target video can be represented as follows:

$$V^t[i] = \langle V_1^t[i], V_2^t[i], \ldots, V_P^t[i] \rangle, \quad i = 1, 2, \ldots, L, \tag{2}$$

where P and L denote the total number of local windows in the ith frame of the target video and the length of the target video, respectively. Thus, the rank matrix of resized AMI for the kth local window in the ith image frame of the target video can be defined as $\mathbf{R}(V_k^t[i])$. Then the distance between two rank matrices is expressed by using $L1$-norm as follows:

$$d_k[i] = \frac{1}{M} \sum_{j=1}^{M} \left\| \mathbf{R}^j(V^q) - \mathbf{R}^j\left(V_k^t[i]\right) \right\|, \tag{3}$$

where $k = 1, 2, \ldots, P$ and $i = T, T+1, \ldots, L$. T denotes the length of the query action video as mentioned. j denotes the index of the rank matrix. This $L1$-norm is known to be more robust to outliers than $L2$-norm [24] and also computed efficiently. The rank matrix of query action is consistently applied to compute the distance regardless of the frame and local window indexes of the target video as shown in (3). Finally, if the distance defined in (3) is smaller than the threshold, the corresponding local windows are detected

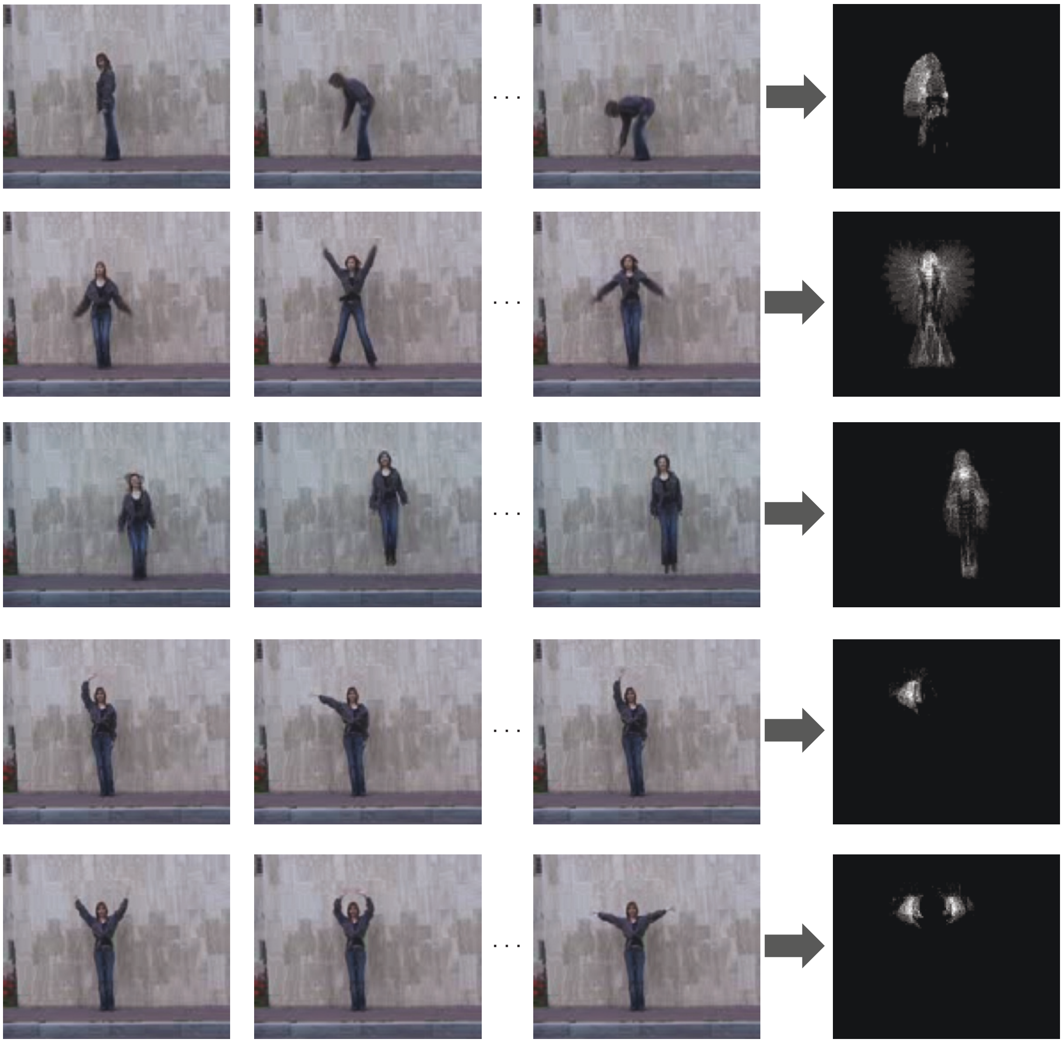


Figure 3: Examples of AMI for five actions from Weizmann dataset [16]: bend, jack, vertical jump, one-hand wave, two-hand wave (from top to bottom).

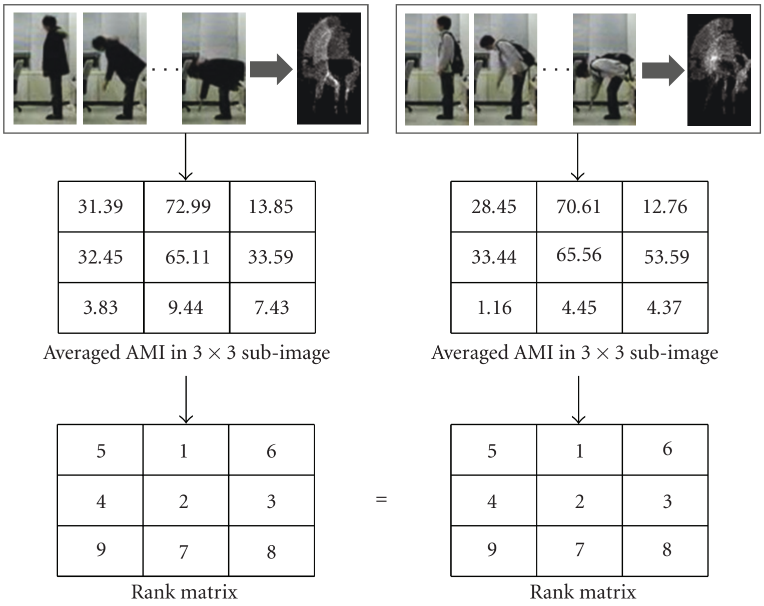
as candidates close to the query action. It is important to note that a comparison between the rank matrices of the query action video and local windows is conducted after initial T frames in (3). It is because that the length of query action video is required at least to generate the reliable AMI of each local window for the accurate comparison. Thus, the latest T frames of the target video need to be stored. However, It should be emphasized that computing (3) with all local windows in each target video frame is very fast since $1 \times M$ rank matrices are only used as our features for the similarity measure instead of full 3D feature vectors (i.e., spatiotemporal cubes shown in [3, 19]).

3.3. Determination of the Best Match Using Energy Histograms. To determine the best match among candidates efficiently, we define the energy histograms by projecting AMI values in horizontal and vertical directions, espectively, as shown in Figure 5. First, the horizontal projection is performed to accumulate all the AMI values in each row of the candidate window. The projection is also conducted in the vertical direction. To be invariant to the size of the local window, accumulated AMI values of each bin are normalized by the maximum value among AMI values belonging to the corresponding bin. Our energy histogram for each direction is defined as follows:

$$\mathrm{EH}_h(i) = \sum_{j=0}^{W-1} \frac{\mathrm{AMI}(i,j)}{\mathrm{max\_AMI}\ (i)}, \quad i = 0,\ldots,H-1, \tag{4}$$

$$\mathrm{EH}_v(j) = \sum_{i=0}^{H-1} \frac{\mathrm{AMI}(i,j)}{\mathrm{max\_AMI}(j)}, \quad j = 0,\ldots,W-1, \tag{5}$$

where H and W denote the height and width of the local window, respectively. max_AMI $(\cdot)$ denotes the maximum value among AMI values belonging to the ith or jth bin in each energy histogram. The two energy histograms of the candidates, EH_h^c and EH_v^c, are compared with those of the query action video, EH_h^q and EH_v^q, to determine



31.39	72.99	13.85
32.45	65.11	33.59
3.83	9.44	7.43

Averaged AMI in 3 × 3 sub-image

28.45	70.61	12.76
33.44	65.56	53.59
1.16	4.45	4.37

Averaged AMI in 3 × 3 sub-image

5	1	6
4	2	3
9	7	8

Rank matrix

=

5	1	6
4	2	3
9	7	8

Rank matrix

Figure 4: Two different 3 × 3 subimages (i.e., $M = 9$) of the same action having identical ordinal signature.

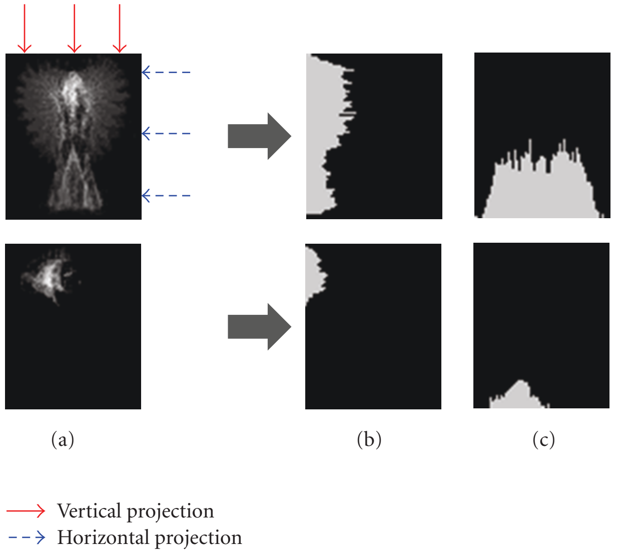


Figure 5: (a) AMIs for jack and one-hand wave from top to bottom. (b) Horizontal energy histogram. (c) Vertical energy histogram.

the best match. For the similarity measure between energy histograms in each direction, we employ the histogram intersection to attain simple computation, which is defined as follows:

$$S_k\left(\mathrm{EH}_k^T, \mathrm{EH}_k^C\right) = \frac{\sum_{i=0}^{l} \min\left\{\mathrm{EH}_k^q(i), \mathrm{EH}_k^c(i)\right\}}{\max\left\{\sum_{i=0}^{l} \mathrm{EH}_k^q(i), \sum_{i=0}^{l} \mathrm{EH}_k^c(i)\right\}}, \quad (6)$$

where $k = \{h, v\}$ and corresponding $l = \{H - 1, W - 1\}$. Finally, the best match is determined based on the combination of S_h and S_v as follows:

$$S_{\mathrm{val}} = \alpha \cdot S_h + (1 - \alpha) \cdot S_v, \quad (7)$$

where α denotes the weight, which is set to 0.5 in our implementation. If the similarity value defined in (7) is smaller than the threshold, the corresponding candidates are removed. It is worth noting that since our energy histograms express the shape information of AMIs correctly using one-dimensional histograms, falsely detected candidates in the target video can be effectively removed and thus the reliability of the proposed method increases. The example of the false positives elimination is shown in Figure 6. We can see that falsely detected windows in the two-hand wave video are effectively removed by using the energy histograms.

For the sake of completeness, the overall procedure of our proposed method is summarized in Algorithm 1.

4. Experimental Results

In this section, we divide the experiments into three phases. First of all, we test our proposed method in the Weizmann dataset [16] to evaluate the robustness and discriminability. The performance for the query action recognition among multiple actions is also evaluated in the second phase. Finally, the performance of our method for real applications such as surveillance scenarios and event retrieval is evaluated.

4.1. Robustness and Discriminability. The robustness determines the reliability of the system which can be represented by the accuracy of the query action detection before false detections begin to occur whereas the discriminability is concerned with its ability to reject irrelevant actions such that false detections do not occur. To evaluate the robustness and discriminability of our proposed method, we employ the Weizmann human action dataset [16], which is one of the most widely used standard datasets. In this dataset, total 10 actions conducted by nine people (i.e., 90 videos) are contained, which can be divided into two categories: global

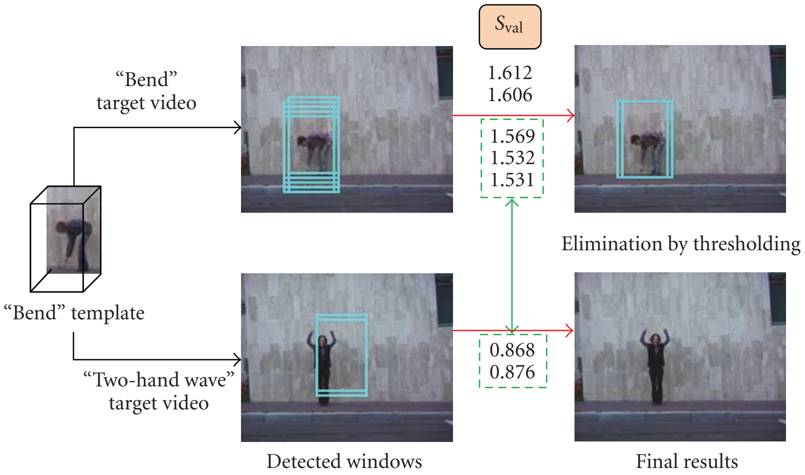


Figure 6: Verification procedure using energy histograms.

```
**Stage 1.** Compute AMI of the query action and local windows on the target video.
**Stage 2.** Ordinal measure for the query action recognition
    (1) Generate the rank matrix based on resized AMI.
    (2) Compute the distance between rank matrices of the query action and local windows from the target video.
        $d_k[i] = \frac{1}{M}\sum_{j=1}^{M} \|R^j(V^q) - R^j(V_k^t[i])\|$
**Stage 3.** Determination of the best match using energy histograms
        $S_{val} = \alpha \cdot S_h + (1 - \alpha) \cdot S_v$
```

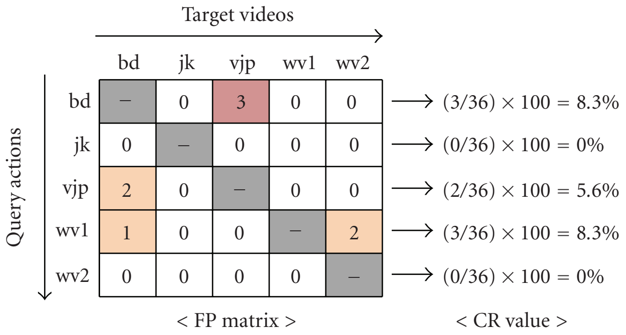
Algorithm 1: Human action recognition using ordinal measure of accumulated motion.

actions (like run, forward jump, side jump, skip, walk) and local actions (like bend (bd), jack (jk), vertical jump (vjp), one-hand wave (wv1), two-hand wave (wv2)). Since most events observed in static camera applications are related to local actions, we thus focus on the five local actions in the Weizmann dataset (see Figure 3).

Since the proposed method does not determine the type of action performed in the target video but localizes windows including the query action in the target video, the confusion matrix, which is widely used in the *learning-based models*, cannot be applied for evaluating robustness and discriminability of our method. Instead, we define our metric, confusion rate (CR) as follows:

$$\mathrm{CR}(i) = \frac{\sum_{j=1}^{5} \mathrm{FP}(i,j)}{\mathrm{Card}(D)}, \quad \text{where } i, j = 1, 2, \ldots, 5. \qquad (8)$$

Here five local motions (i.e., bd, jk, vjp, wv1, wv2) are mapping to the number from 1 to 5 in turn. FP (i, j) denotes the number of videos containing falsely detected windows with a given query action where i and j denote indexes of the query actions and actions included in target videos, respectively (see Figure 7). D denotes a set of videos excluding videos related to the query action. For example, if false detections occur in the one of "bd" target videos and the two of "wv2" target videos when the "wv1" is given as the query action, we can represent FP(4, 1) = 1 and FP(4, 5) = 2.



Target videos

Query actions

	bd	jk	vjp	wv1	wv2	
bd	–	0	3	0	0	→ (3/36) × 100 = 8.3%
jk	0	–	0	0	0	→ (0/36) × 100 = 0%
vjp	2	0	–	0	0	→ (2/36) × 100 = 5.6%
wv1	1	0	0	–	2	→ (3/36) × 100 = 8.3%
wv2	0	0	0	0	–	→ (0/36) × 100 = 0%

< FP matrix > < CR value >

Figure 7: Confusion rate for five local actions from Weizmann dataset.

Furthermore, the CR can be computed as follows: CR(4) = $\{(1 + 2)/(45 - 9)\} \times 100 = 8.3\%$. The CR values for five local actions are shown in Figure 7. Note that the CR is evaluated only at the level where the query action is perfectly recognized in the videos including the actual query action.

The total classification rate of the proposed method can be defined as follows [3]:

$$C = \frac{(N - \#\text{ of misclassification})}{N} \times 100, \qquad (9)$$

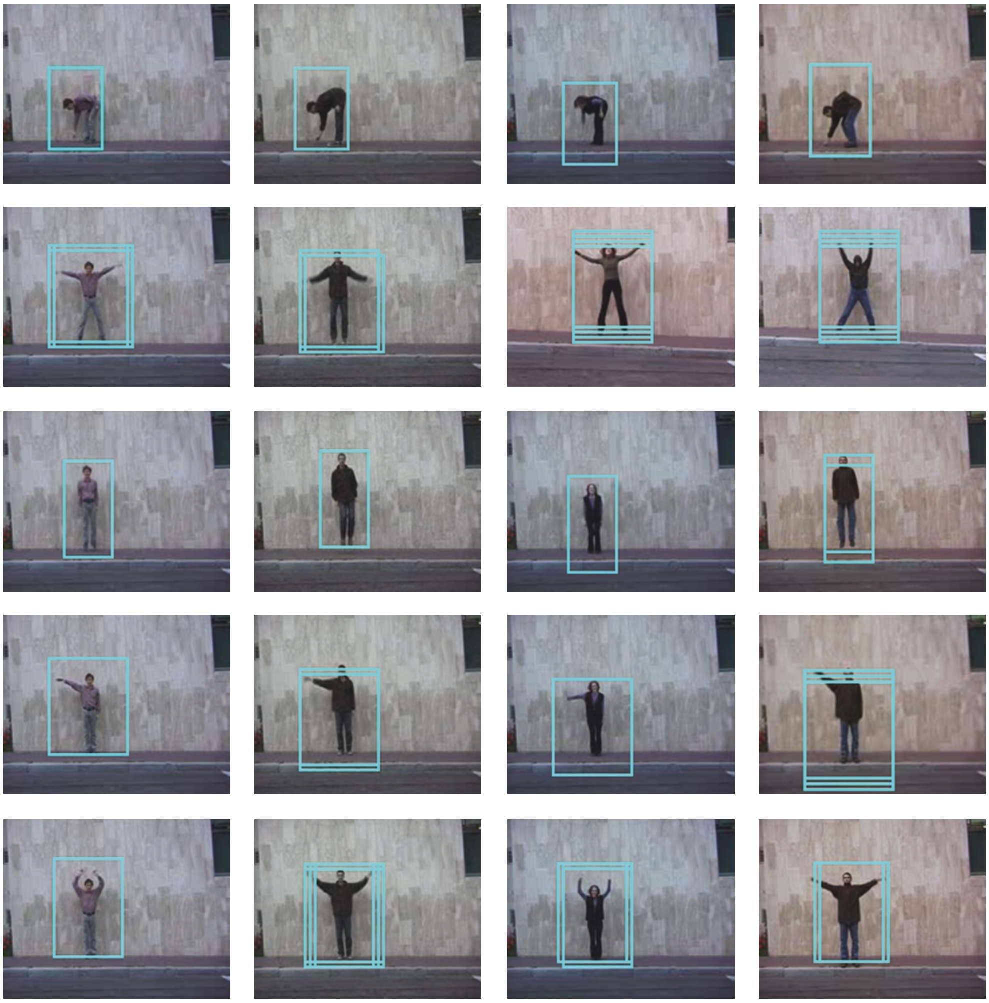


Figure 8: Examples of the query action localization using the proposed method (from the top to bottom: bd, jk, vjp, wv1, wv2).

where N denotes the total number of videos used for comparison. The total classification rate can be computed based on our results (see Figure 7) as follows: $C = [\{(9 \times 5 \times 5) - 8\}/(9 \times 5 \times 5)] \times 100 = 96.4\%$, which is comparable to the classification rates of other methods such as [3, 19]. The results of the query action localization in target videos are also shown in Figure 8.

The two threshold values used for candidate detection and determination of the best match are empirically set. The size of local windows is set to be equal to the image size of the query action video. Note that the spatial and temporal scale changes up to ±20% can be handled in our method. The framework for evaluating performance has been implemented by using Visual Studio 2005 (C++) under FFMpeg library, which has been utilized for MPEG and Xvid decoding. The experiments are performed on the low-end PC (Core2Duo 1.8 GHz). The test videos in the Weizmann dataset are encoded with the image size of 180 × 144 pixels. The query action video for each local motion is cropped from one of nine videos related to the corresponding action in our experiment. Since the processing speed of our algorithm achieves about 45 fps for the test videos, it can be sufficiently applied for real-time applications.

4.2. Recognition Performance in Multiple Actions. In this subsection, we demonstrate the recognition accuracy of the proposed method by using our videos captured in different environments (i.e., indoor and outdoor) with the image size of 192 × 144 pixels. In particular, the performance for the query action recognition among multiple actions is evaluated.

First, two people conduct different actions in consecutive sequences shown in Figure 9. More specifically, one person waves a one hand consistently in the indoor environment while the other one performs continuously different actions shown in Figures 9(a) and 9(b). We can see that the query action “wv2” and “jk” are correctly detected. In Figure 9(c), the query action “vjp” is detected. Especially, a case that “vjp” is conducted by different two actors at the same time is also successfully detected. Furthermore, our method captures invariably the query action although the color of

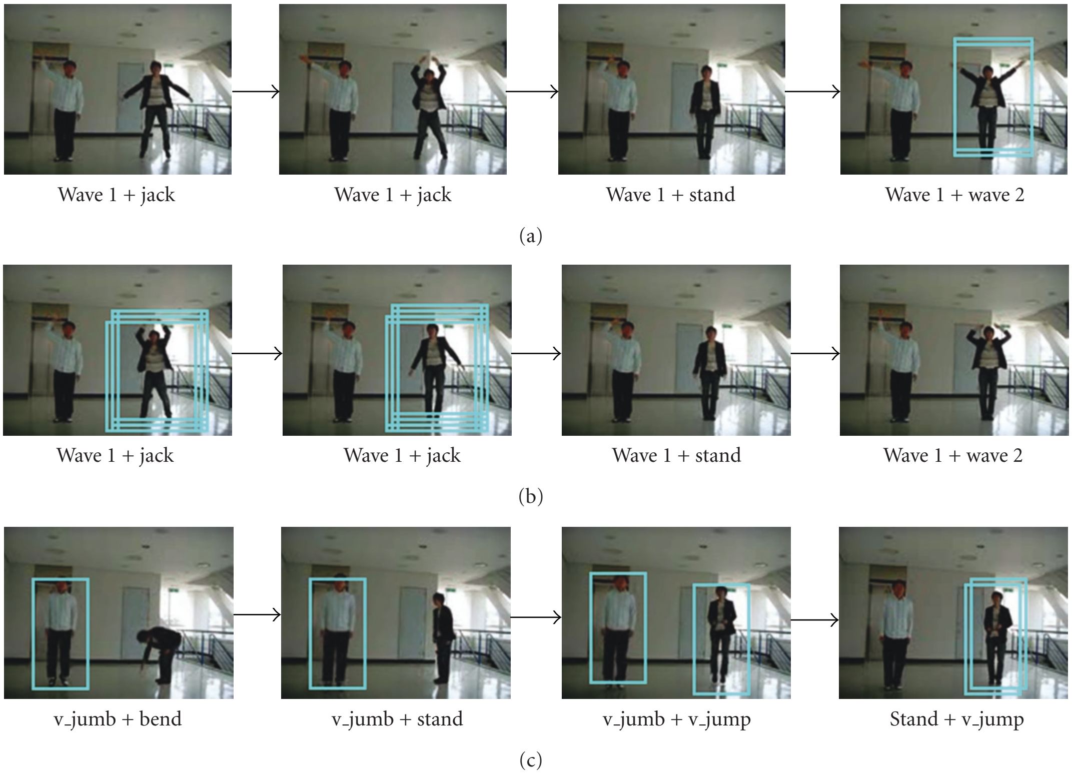


FIGURE 9: Query action recognition among multiple actions in the indoor environment. (a) Two-hand wave. (b) Jack. (c) Vertical jump.

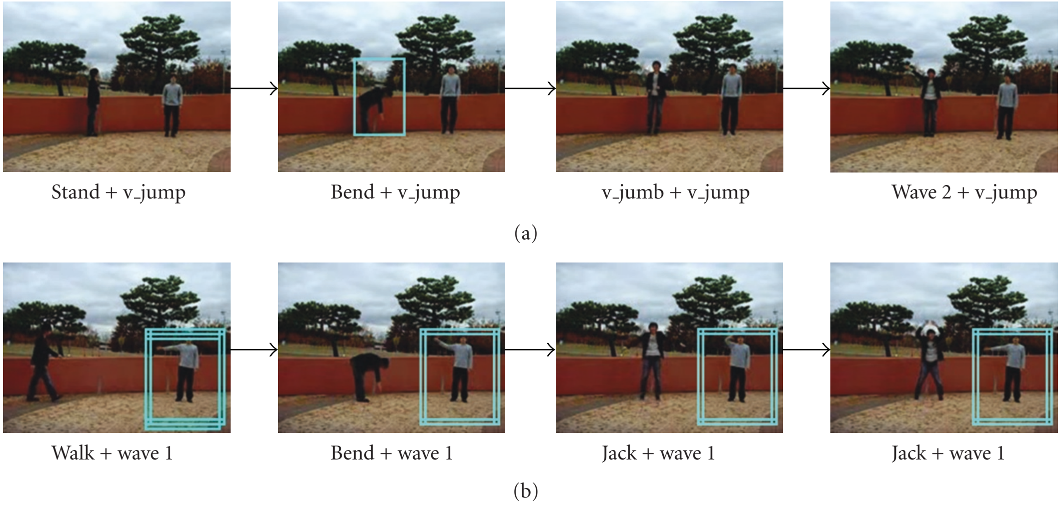
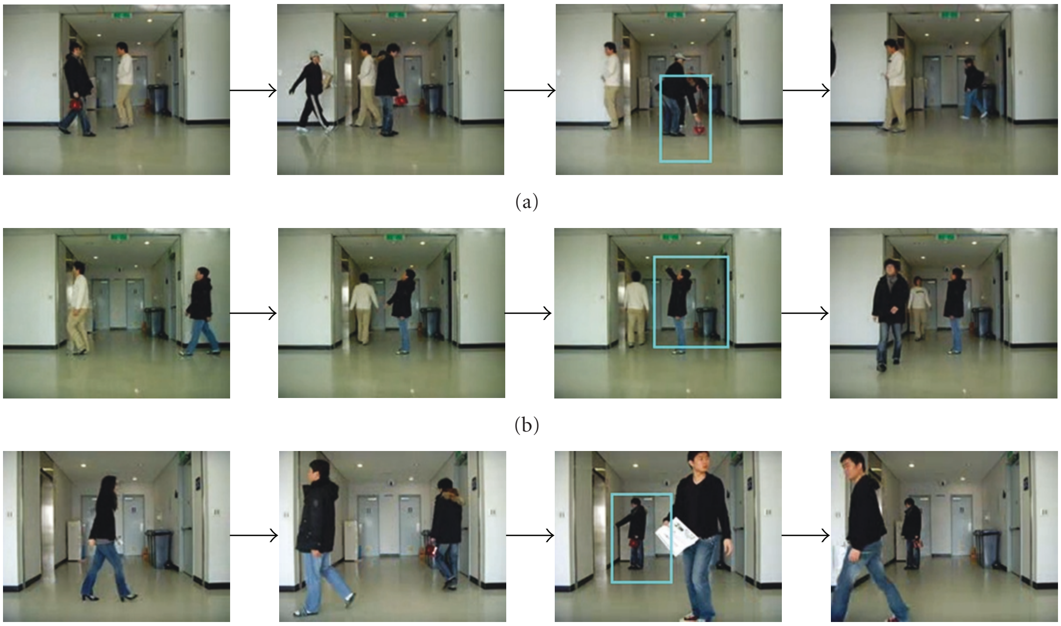


FIGURE 10: Query action recognition among multiple actions in the outdoor environment. (a) Bend. (b) One-hand wave.

background is similar with that of actors (see Figure 9(c)). We also demonstrate the performance of our method in the outdoor environment. The query action "bd" is correctly detected among various actions conducted by one person as shown in Figure 10(a). In Figure 10(b), the query action "wv1" is successfully detected even if there is global motion (i.e., walk) in the target video. Note that the all templates for query actions are obtained from the Weizmann dataset. Based on these results, it is shown that the query action can be robustly recognized among various multiple actions by our proposed method.

4.3. Recognition Performance for Real Applications. Since most standard action dataset including the Weizmann dataset is captured in well-controlled environments while actions in the real world often occur in much more complex scenes, there exists a considerable gap between these samples and real world scenarios.

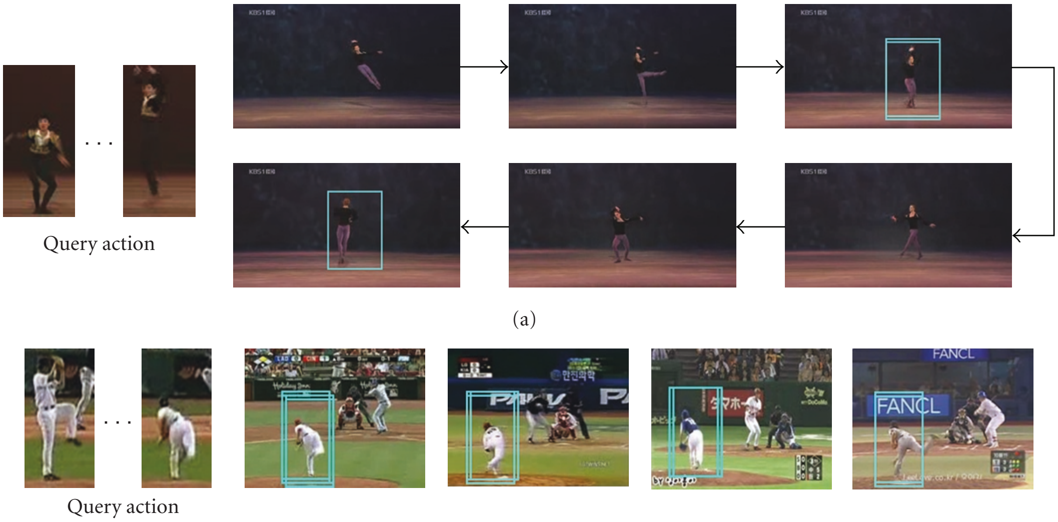


(a)

(b)

(c)

Figure 11: Performance in the selected videos of the surveillance system. Each video is composed of 1070, 800, 850 frames, respectively. (a) Put-objects. (b) Call-people. (c) Push-button.



(a)

(b)

Figure 12: Query action recognition for event retrieval. (a) Turn-jump in ballet. (b) Examples of recognizing pitching action.

First of all, to show the robustness and efficiency of the proposed method for the surveillance systems, we try to recognize three specific actions, which are often observed in surveillance scenarios: put-objects, call-people, push-button. Figure 11 shows the recognition results of our method in each surveillance video with the image size of 192 × 144 pixels. More specifically, the query action "put-objects" is correctly detected in cluttered background as shown in Figure 11(a). It should be emphasized that the proposed method can detect the query action even though the actor is merged with the other one. In Figure 11(b), a man calls someone by waving his hand while the other one is going past by him in the different direction. In such situation, the query action "call-people" is also detected correctly. One person pushes a button and then awaits the elevator in Figure 11(c). Although the local window is partially occluded by the other person, the query action is successfully detected. This example shows the robustness of our method to the partial

Table 1: False positive rate of each selected video.

	Put-objects	Call-people	Push-button
FPR	0.69%	0.22%	0.24%

occlusion in the complex scene. The accuracy of action recognition in surveillance systems is shown in Table 1. The false positive rate (FPR) is computed as follows:

$$\text{FPR} = \frac{\#\text{ of frames including misclassification in } W}{\text{Card}(W)}, \quad (10)$$

where W denotes a set of frames excluding the frames related to the query action in each surveillance video. Here the FPR is computed at the level where query actions are perfectly detected in each surveillance video. Based on the results of query action recognition, we confirm that the proposed method can be regarded as a useful indicator for smart surveillance system.

Furthermore, our proposed method can be applied for the event retrieval. Note that since the proposed method is originated for static camera applications as mentioned in Section 1, the large motion of camera is highly likely to yield unwanted detections. Thus, we demonstrate the performance of our method by using two query action videos captured with static camera, which are collected from broadcasting videos: turn-jump in ballet and pitching in baseball. Figure 12(a) shows the process of the query action recognition in the ballet sequence. The turn-jump action is correctly detected among various jump actions as shown in Figure 12(a). In Figure 12(b), the pitching action is also successfully detected in various baseball videos.

5. Conclusion

A novel method for human action recognition is proposed in this paper. Compared to previous methods, our proposed algorithm is performed very fast based on the simple ordinal measure of accumulated motion. To this end, AMI is firstly defined by using image differences. Then the rank matrix is generated based on the relative ordering of resized AMI values and distances from the rank matrix of query action video to the rank matrices of all local windows in the target video are computed. To determine the best match among the candidates close to the query action, we propose to use the energy histograms obtained by projecting AMI values in horizontal and vertical directions, respectively. Finally, experiments are performed on diverse videos to justify the efficiency and robustness of the proposed method. The classification results of our algorithm are comparable to state-of-the-art methods and further, the proposed method can be used for real-time applications. Our future work is to extend the algorithm to describe human actions in dynamic scenes.

Acknowledgments

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-(C1090-1011-0003)).

References

[1] A. Briassouli and I. Kompatsiaris, "Robust temporal activity templates using higher order statistics.," *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2756–2768, 2009.

[2] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 462–469, Beijing, China, October 2005.

[3] H. J. Seo and P. Milanfar, "Detection of human actions from a single example," in *Proceedings of the International Conference on Computer Vision (ICCV '09)*, October 2009.

[4] V. H. Chandrashekhar and K. S. Venkatesh, "Action energy images for reliable human action recognition," in *Proceedings of the Asian Symposium on Information Display (ASID '06)*, pp. 484–487, October 2006.

[5] M. Ahmad and S.-W. Lee, "Recognizing human actions based on silhouette energy image and global motion description," in *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG '08)*, pp. 1–6, Amsterdam, The Netherlands, September 2008.

[6] C. Kim, "Content-based image copy detection," *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 169–184, 2003.

[7] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 127–132, 2005.

[8] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[9] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 3, pp. 32–36, Cambridge, UK, August 2004.

[10] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 1, pp. 432–439, Nice, France, October 2003.

[11] N. Ikizler, R. G. Cinbis, and P. Duygulu, "Human action recognition with line and flow histograms," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, Tampa, Fla, USA, December 2008.

[12] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 166–173, Beijing, China, October 2005.

[13] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proceedings of International Conference on Computer Vision (ICCV '09)*, October 2009.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 886–893, San Diego, Calif, USA, June 2005.

[15] P. S. Dhillon, S. Nowozin, and C. H. Lampert, "Combining appearance and motion for human action classification in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 22–29, Miami, Fla, USA, June 2009.

[16] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1395–1402, Beijing, China, October 2005.

[17] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 984–989, San Diego, Calif, USA, June 2005.

[18] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[19] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 405–412, San Diego, Calif, USA, June 2005.

[20] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[21] H. Ning, T. X. Han, D. B. Walther, M. Liu, and T. S. Huang, "Hierarchical space-time model enabling efficient search for human actions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 6, pp. 808–820, 2009.

[22] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[23] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1905–1909, 2009.

[24] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, New York, NY, USA, 1987.