2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

An Efficient Part-Based Approach to Action Recognition from RGB-D Video with BoW-Pyramid Representation

Jih-Sheng Tsai, Yen-Pin Hsu, Chengyin Liu, and Li-Chen Fu, *Fellow, IEEE*

***Abstract*— In this paper, we propose an efficient part-based approach for action recognition. The main concept is to recognize human actions by less occluded parts without using a large set of part filters. Therefore, our approach is robust to occlusion and cost-effective. We extract spatio-temporal features from RGB-D videos, and assign a part-label to each feature. Then, for each part, a recognition score is computed for each action class by pyramid-structural bag of words (BoW-Pyramid) representation. The final result is determined by weighted sum of these scores and contextual information, which is based on the ratio of features between every pair of parts. Several contributions have been made in this work. First, the proposed part-based method is robust to occlusion and operates on-line. Second, our BoW-Pyramid representation can distinguish actions with reversed temporal orders. Third, recognition accuracy is increased by incorporating contextual information. The provided experimental results have verified effectiveness of our method and demonstrated high promise of surpassing performance of the state-of-the-art works.**

I. Introduction

Action recognition has become a popular field with a variety of applications, such as human-computer interaction, surveillance, and sports video analysis. Recognizing actions in realistic environments is especially of increasing interest in recent research [1–4]. A lot of existing works did much effort to deal with the problem of cluttered background, either static or dynamic. However, only fewer works tried to solve the problem of occlusion, which is an important issue to action recognition. In this paper, we focus on the problem of human action recognition under occlusion.

Although occlusion occurs commonly in real world, it is a difficult task to recognize human actions with some body parts being occluded. We consider three scenarios of occlusion while performing an action. The first one is that the subject is occluded by static objects, the second one is that the subject is occluded by moving persons, and the third one is that the camera lens is partially occluded. We evaluate the performance under occlusion using all of these three scenarios in our experiments.

Jih-Sheng Tsai is with the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC, (e-mail: r00922117@ csie.ntu.edu.tw).

Yen-Pin Hsu is with the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC, (e-mail: r01922124@ csie.ntu.edu.tw).

Chengyin Liu is with the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC, (e-mail: r01944040@ csie.ntu.edu.tw).

Li-Chen Fu is with the Department of Electrical Engineering and Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC (e-mail: lichen@ntu.edu.tw).

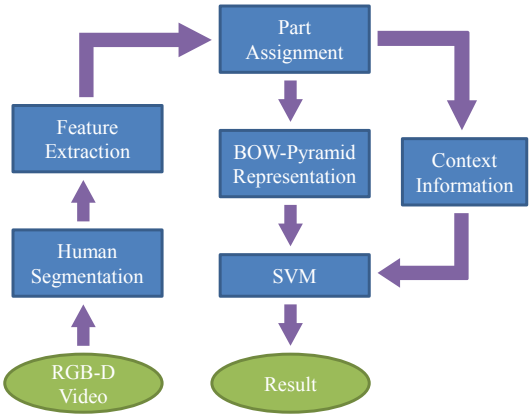


Fig. 1. System overview.

In our work, we define contextual information in the aspect of action itself. We observed that some actions are much associated with some particular body parts, whereas some other actions are associated with all body parts. For example, kicking is heavily associated with legs while running is associated with all body parts. With this observation, our contextual information is defined by the ratio of features between two body parts.

Recognizing actions from RGB videos suffers from some problems such as illumination variations and background noise. Owing to emerging of RGB-D cameras, we can get depth information besides intensity. Depth information is invariant to illumination and color of clothing. Moreover, it is helpful to occlusion detection and foreground-background segmentation. Since Kinect is a handy inexpensive RGB-D camera developed by Microsoft, we use it as our input sensor.

In this paper, we propose an efficient part-based approach that is robust to occlusion and can run on-line. The overview of our system is shown in Fig. 1. First, we segment human and extract spatio-temporal features from RGB video and depth video, respectively. Then, we assign a part-label to each feature. For each part, we use BoW-Pyramid to represent an action and train a Support Vector Machine (SVM) classifier for it. We also define contextual information according to the assignments of all features. Finally, each SVM classifier computes a recognition score, and the final result is determined by weighted sum of these scores and contextual information.

The contributions of this paper are three-fold. First, the proposed part-based method is robust to occlusion and can run on-line. Second, with the BoW-Pyramid representation, we can distinguish actions with reversed temporal orders. Third, the contextual information is helpful to action recog-

nition by increasing recognition accuracy. To verify the mentioned contributions, three datasets are used in our experiments, namely KTH [5], RGBD-HuDaAct [6], and the dataset that we called OccData recorded by ourselves. The provided experimental results are satisfactory, and have demonstrated high promise of surpassing performance of the state-of-the-art works.

The rest of the paper is organized as follows. Section II introduces existing works related to our approach. Section III describes the proposed part-based BoW-Pyramid representation. Section IV presents the recognition scheme. Experimental results are shown in section V, and we conclude this paper in section VI.

II. Related Work

Several methods have been proposed to recognize human actions. Bobick and Davis [7] used motion history images (MHI) to represent actions. Gorelick *et al*. [8] represented actions as space-time volumes, and Lv and Nevatia [9] selected some key poses for recognition. However, so far these methods depend on extracted silhouettes, which are broken caused by occlusion. Hence, they cannot deal with occlusion robustly.

Part-based approaches are intuitively suitable for handling occlusion. Wang *et al*. [10] used histogram of oriented gradient (HOG) and Local Binary Pattern (LBP) as features, and trained part detectors to detect human with partial occlusion. Tran *et al*. [11] used a set of body-part detectors, and represented each part as a sparse motion descriptor image for action recognition. These approaches detect body parts by exhaustive sliding window search through the entire image, which is time-consuming. Instead of detecting body parts for each frame, we assign each feature to the nearest body part according to skeleton information provided by OpenNI [12]. Our method is cost-effective that can run on-line. Weinland *et al.* [13] used 3DHOG as features. They classified each embedded block descriptor individually and then combined the classification responses as the final result. But the training data with occlusion are made artificially, which are far from being realistic. In our work, we use non-occlusion data for training and use realistic occluded data for testing, and demonstrate promising results in our experiments.

Recently, methods based on local spatio-temporal features [14, 15] have shown promising results in RGB videos. Laptev *et al*. [2] followed the work by [15] to detect interest points and used histograms of oriented gradient (HOG) [16] and optic flow (HOF) as feature descriptor. Niebles *et al.* [17] followed the approach in [14] to detect interest points, and they used two models, *i.e*., probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), for recognition. After incorporating the additional depth information, spatio-temporal features can be extracted from not only RGB videos but also depth videos. Zhang and Parker [18] extended [14] to RGB-D videos and used LDA to recognize human actions, but features they used are not scale-invariant. In our work, we use features proposed by [2] and extend them to RGB-D videos.

Most methods mentioned above followed the bag of words (BoW) approach using spatio-temporal features [2, 17, 19]. However, BoW approach loses both spatial and temporal relations among features. To handle this problem, Laptev *et al*. [2] divided the whole space-time volume into spatio-temporal grids, and used SVM with a multi-channel χ^2 kernel to classify human actions. Ni *et al*. [6] used spatial pyramid match kernel (SPM) algorithm [20]. In our work, we use BoW-Pyramid to represent actions, which is inspired by temporal pyramids for object recognition developed by Pirsiavash and Ramanan [21]. BoW-Pyramid is a coarse-to-fine representation by concatenating all histograms produced from different time segments. Our action representation can distinguish actions with reversed temporal orders such as stand-up and sit-down. Experimental results demonstrate that our BoW-Pyramid representation is better than those in [2] or [6].

Some existing works defined context as scenes or objects associated with actions [22–24]. Such approaches recognize scenes or objects to assist in recognizing human actions in real world. However, generally speaking recognition of only some particular actions could benefit from the context. In other words, recognizing some actions may not get advantages from the context, and could even be jeopardized by it instead. There are two arguments for the above cases. First, the same action could occur at many places, and different subjects could take different objects while performing the same action. For example, running could occur on the street, at the gym, in the hallway, etc. Also, while a subject is running, he/she may carry a suitcase, a cellphone, or something else. Second, different actions could occur at the same place, and a subject may take the same object while performing different actions. For example, a subject could play basketball or volleyball at the gym. Also, a subject may take his/her cellphone while running or jumping. In our work, we define context by the content of actions so that it is general for most actions.

III. Part-Based BoW-Pyramid Representation

In this work, we extract spatio-temporal features, which combine both intensity and depth information from RGB-D videos, after preprocessing. Then, we classify the extracted features into different parts and organize them into BoW-Pyramid to represent an action for each part.

A. Preprocessing

The depth information provided by Kinect is 14-bit values. To form depth images, we convert the raw data to 8-bit values so that their range is consistent with intensity images. We also calibrate both color camera and depth camera of Kinect to coordinate the corresponding intensity image and depth image. Before extracting spatio-temporal features, we segment human from depth image and intensity image implemented by OpenNI [12] for each frame. Fig. 2 shows one example of segmented result. Due to the segmentation, noise in the background can be removed.

Fig. 2. Segmented result of kicking.

B. Spatio-Temporal Feature

For the spatial-temporal features, we apply the approach proposed by [15] that extended Harris operator to space-time space. We detect interest points separately from sequence of intensity images and sequence of depth images, where human has been segmented. Since shapes and motions are two important cues for action recognition, HOG and HOF are concatenated as feature descriptor. Features are extracted at multiple levels of spatio-temporal scales. Fig. 3(a) shows an example of extracted features from RGB video and from depth video. The red dots represent detected interest points, and the sizes of the corresponding circles represent scales of interest points. Note that there is some noise on the boundary of silhouette due to imperfect segmentation, so we perform a denoising process before assigning part-labels to these features. To this end, we check the depth of each pixel within a window centered at an interest point. An interest point is considered as noise if the ratio of non-human pixels within the window is greater than a threshold. We discard these noisy features and keep the remainder as valid features. Fig. 3(b) shows the denoising result of Fig. 3(a).

C. Part Assignment

We assign a part-label to each valid feature using Nearest Neighbor Classifier (NNC) according to skeleton information, which consists of 15 joint positions, provided by OpenNI [12]. Although such skeleton information provides body configuration of a human, it is quite unstable due to tracking failure, change of viewpoints, and occlusion. Weng and Fu [25] pointed out that some of postures might be lost in tracking, especially for those with self-occlusion. In fact, the more occluded parts, the more severe failures that might occur. Some examples of unstable skeleton are shown in Fig. 4.

Therefore, we use skeleton information to detect which joints are occluded instead of recognizing actions. A joint is regarded as occluded if most of its neighbors are non-human pixels. This technique is the same as denoising process of features described in the last subsection. We predefine the part to which each joint belongs, denoted by l_j for joint j. Then, we classify features into parts using non-occluded joints. Given the set of non-occluded joints $\boldsymbol{U}$ and a feature f, we find the closest non-occluded joint j^* of f using NNC

$$j^* = \arg\min_{j \in \boldsymbol{U}} d(j, f) \quad (1)$$

where d is the Euclidean distance function. We then assign l_{j^*} to f as its part-label.

Since the configuration of human body is articulated, we treat some adjacent joints as the same part. Moreover, in order

(a) (b)

Fig. 3. (a) Features extracted from RGB video and depth video. (b) The denoising result of (a).

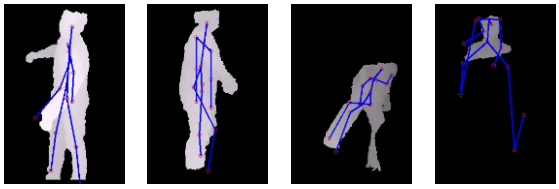
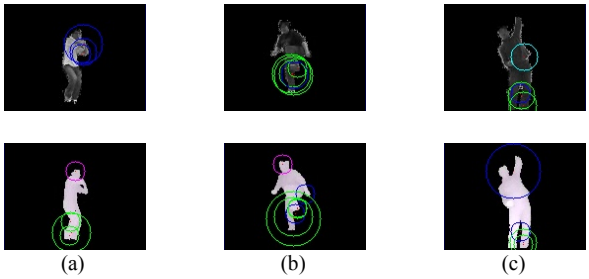


Fig. 4. Some examples of unstable skeleton. The rightmost two are under occlusion.



(a) (b) (c)

Fig. 5. Results of part assignments for (a) baseball striking, (b) kicking, (c) tennis serving. Top row are intensity images and bottom row are depth images. Different colors of circles represent different parts. In this paper, we use four parts: *head*, *torso*, *arm*, and *leg*, which are represented by magenta, cyan, blue, and green, respectively.

to take within-class variation into account such as some persons are right-handed while some are left-handed, we also treat the symmetric body parts as the same part. By doing so, we can get the correct assignments of features despite of incorrect configuration of skeleton in most cases. We use four parts in our work, and Fig. 5 shows some results of part assignments. Note that there are still few false assignments due to incorrect configuration of skeleton, but it does not affect our final result severely.

D. BoW-Pyramid Representation

We organize features that belong to the same part into BoW-Pyramid representation from RGB video and depth video separately. Then we fuse RGB BoW-Pyramid and depth BoW-Pyramid by concatenation, and the resulting feature vector is our representation of an action.

The traditional bag-of-words approach for action recognition calculates the distribution of visual words over the entire video. The visual words are made by running clustering algorithm over all features so that each feature can be assigned to a visual word. An action is represented by a histogram where each bin is the occurrence of a visual word. However, different actions with reversed temporal orders cannot be distinguished by such representation since their distributions of visual words over the entire video are very similar. To encode temporal information, we calculate the distributions

of visual words using different time segments from a video based on the pyramid-like structure as Fig. 6 shows. There are L+1 levels in the pyramid and the whole video is divided into 2^l segments at level l where $L \geq l \geq 0$. A histogram is computed from each segment, and BoW-Pyramid is represented by concatenating normalized histograms from segments of all levels. Note that level 0 produces the same histogram as the one produced by traditional bag-of-words model. Therefore, BoW-Pyramid can be treated as the generalization of bag-of-words model. With BoW-Pyramid representation, we can distinguish different actions with reversed temporal orders at finer levels although they are indistinguishable at level 0. There is no need to divide spatial grids in our work since we have already divided features into different parts.

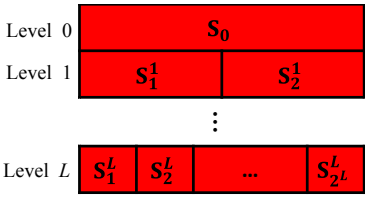


Fig. 6. Pyramid-like structure of a video. The whole video is denoted by S_0, and the k^{th} segment at level l is denoted by S_k^l where $2^l \geq k \geq 1$ and $L \geq l \geq 1$.

IV. Recognition Scheme

The main concept of our recognition scheme is to recognize actions by parts. We also take contextual information into consideration, which can be treated as a prior distribution to each action class. A recognition score is computed from each part of view and the final result is combined in a weighted sum scheme as shown in Fig. 7. In this section, we describe contextual information followed by the weighted sum scheme.

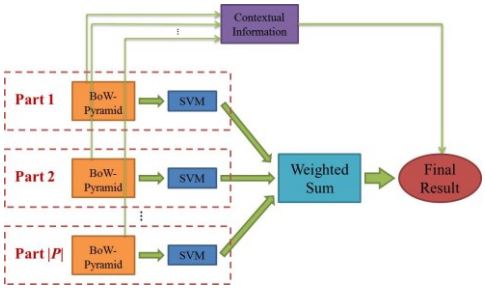


Fig. 7. Recognition scheme.

A. Contextual Information

We get contextual information based on the ratio between two parts from all training videos. The ratio between part p and part q in video v is defined as

$$r_{pq,v} = \frac{|p_v|}{|q_v|+\varepsilon} \tag{2}$$

where $|\cdot_v|$ denotes the number of features belong to a certain part in video v, and ε is a small real number to prevent from dividing by zero. We divide all videos into one of three sets according to the value of this ratio and a positive threshold δ

$$\begin{aligned} \boldsymbol{V}_1 &= \{v | r_{pq,v} > \delta\}, \\ \boldsymbol{V}_2 &= \left\{v \middle| \delta \geq r_{pq,v} \geq \frac{1}{\delta}\right\}, \\ \boldsymbol{V}_3 &= \{v | r_{pq,v} < \frac{1}{\delta}\}. \end{aligned} \tag{3}$$

Then for each set, the prior probability of action class c in $\boldsymbol{V}_i$, i =1,2,3 is defined as

$$p_{c,\boldsymbol{V}_i} = \frac{|\boldsymbol{V}_i^c|}{|\boldsymbol{V}_i|} \tag{4}$$

where $|\boldsymbol{V}_i^c|$ denotes the number of videos that belong to class c in $\boldsymbol{V}_i$. There are such prior probabilities for every two parts, but since most human actions are mainly different from limbs, and also the extracted features mainly focus on limbs as Fig. 5 shows, only two parts, namely *hand* and *leg*, are used to get contextual information in our implementation.

B. Weighted Sum Scheme

A SVM classifier, which is implemented by LIBSVM [26] in our work, is trained for each part using training data. To recognize actions, we use the probability yielded by each SVM as the recognition score given a video. The final result is the action class with the maximal weighted-sum value

$$c^* = \arg\max_c p_{c,\boldsymbol{V}_j} \cdot \left(\sum_{k \in \boldsymbol{P}} w_k \, s_{k,c}\right) \tag{5}$$

where $p_{c,\boldsymbol{V}_j}$ is the prior probability according to contextual information that the given video belongs to $\boldsymbol{V}_j$, $\boldsymbol{P}$ is the set of all parts, w_k is the weight for part k, and $s_{k,c}$ is the score of class c produced by SVM of part k. The weight w_k is determined by the fraction of non-occluded joints that belong to part k, defined as

$$w_k = \frac{N_k}{\sum_{t \in \boldsymbol{P}} N_t} \tag{6}$$

where N_t denotes the number of non-occluded joints belong to part t. The more non-occluded joints a certain part has, the greater its weight is. Thus, the final result is dominated by these parts with less occlusion while parts which are severely occluded have only little contribution to the final result.

V. Experimental Results

Our experiment is carried out under a computer with Intel Core i5 CPU and 4GB RAM. Three datasets are used in experiments, i.e., KTH [5], RGBD-HuDaAct [6], and OccData recorded by ourselves. OccData contains eight types of actions, which are baseball striking, boxing, jumping, kicking, running, basketball shooting, swimming, and tennis serving.

There are 875 RGB-D video clips of actions without occlusion performed by 12 subjects, and 159 RGB-D video clips of occluded actions performed by 3 subjects out of the 12. All actions are performed in cluttered background with significant intra-class variation and inter-class variation. The viewpoints and the distances between camera and each subject are slightly different. There are three cases of occlusion in OccData, that is, the subject is occluded by static objects, the subject is occluded by moving persons, and the camera lens is partially occluded. Fig. 8 shows some sample frames of OccData dataset, where the subject is segmented shown in the depth videos. We set L=1 in the following experiments.

To evaluate the performance of our BoW-Pyramid representation, we use KTH and RGBD-HuDaAct for testing. Because skeleton information is not provided by these two datasets, our features are not classified into parts, and contextual information is not used. We use the original data without any preprocess. To take fairness into account, we use the same evaluation scheme and performance measure as the compared work. The comparisons are shown in Table 1 and Table 2. Note that our BoW-Pyramid representation is better than the spatial-temporal grid proposed by [2] with the same feature used, and the accuracy is much improved compared to [6] since there are many actions with reversed temporal orders, i.e., enter the room versus exit the room, go to bed versus get up, take off the jacket versus put on the jacket, and stand up versus sit down. Therefore, it can highlight the advantages of BoW-Pyramid representation using RGBD-HuDaAct. Such improvement can be observed in Fig. 9. The main confusion of our approach is between drink water and eat meal since the motions between these two actions are very similar. Also note from Table 2 that the result of our approach using RGB features is sufficient good enough, and the additional depth features help only a little. One reasonable conjecture is due to the background noise in depth videos that makes the BoW-Pyramid representation less informative. We believe that such improvement would increase if human is segmented from depth videos.

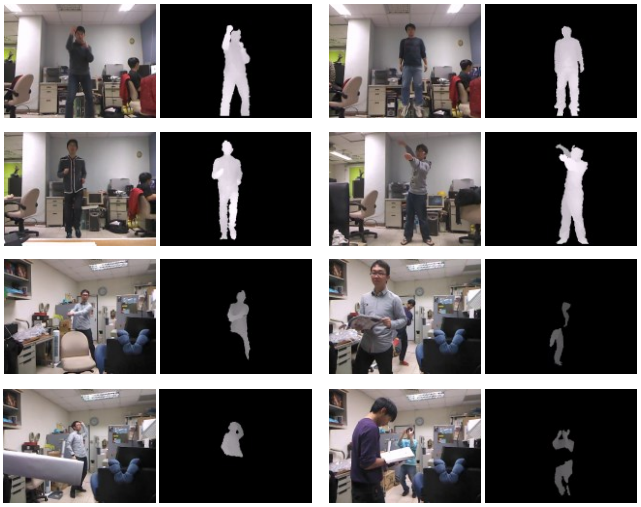


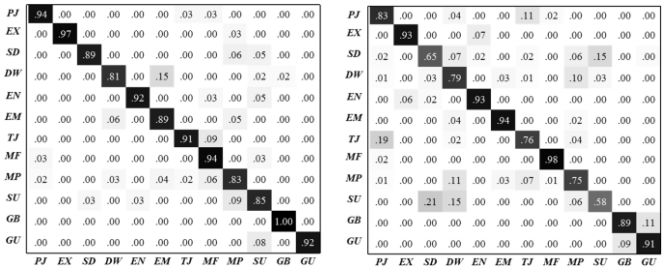
Fig. 8. Sample frames of OccData dataset.

Work	Average accuracy
Ours	**92.3%**
Laptev *et al.* [2]	91.8%

Table 1. Comparison result of KTH.

Work	Average accuracy
Ours (RGB-D)	**91.7%**
Ours (RGB)	90.1%
Ni *et al.* [6]	82.8%
Zhao *et al.* [27]	89.1%

Table 2. Comparison result of RGBD-HuDaAct.



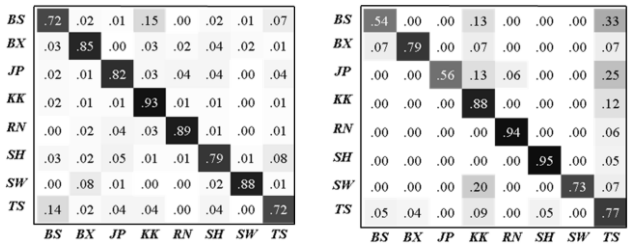
(a) Ours (b) Ni *et al.* [6]

Fig. 9. Confusion matrix of our approach and Ni *et al.* [6]. Class of background activity is not used and thus not shown in (b). For better view, we use two characters to represent each action category, i.e., ***PJ***: put on the jacket, ***EX***: exit the room, ***SD***: sit down, ***DW***: drink water, ***EN***: enter the room, ***EM***: eat meal, ***TJ***: take off the jacket, ***MF***: mop the floor, ***MP***: make a phone, ***SU***: stand up, ***GB***: go to bed, and ***GU***: get up.

To evaluate the robustness of our part-based method to occlusion and the benefit of contextual information, we use OccData dataset for testing. Our method is on-line that the frame rate is above 16 using video clips under resolution 144×108. The video clips of training data are mirrored to handle variation between left-handed and right-handed. We test non-occluded data in leave-one-subject-out scheme, and use all non-occluded video clips as training data for testing occluded data. The result is shown in Table 3. As it can be seen, the recognition accuracy is promising even under such low resolution. Two observations can be made from Table 3. First, the accuracy can be improved with contextual information for both occluded test and non-occluded test thus verify the benefit of our contextual information. Second, the accuracy doesn't drop much when there is occlusion and thus verify the robustness to occlusion of our approach. Fig. 10 shows the confusion matrices using contextual information. For non-occluded test, baseball striking is confused with kicking because some subjects raise or stick out one of legs during performing the striking motion, and the confusion between tennis serving and basketball shooting is due to the similar motions. For the occluded test, the accuracies of most action classes drop while some increase compared to the non-occluded test. This might be due to whether correct prior probability is got from contextual information according to the computed ratio from (2).

	With context	Without context
Non-occluded test	82.5%	79.2%
Occluded test	77%	74.6%

Table 3. Recognition accuracy of OccData.



(a) (b)

Fig. 10. Confusion matrix of (a) non-occluded test and (b) occluded test with contextual information. For better view, we use two characters to represent each action category, i.e., ***BS***: baseball striking, ***BX***: boxing, ***JP***: jumping, ***KK***: kicking, ***RN***: running, ***SH***: basketball shooting, ***SW***: swimming, ***TS***: tennis serving.

In this paper, we propose an efficient part-based approach to recognize actions from RGB-D videos. The proposed part-based recognition scheme is robust to occlusion and can run on-line. Our system recognizes an action mainly by spatio-temporal features lying on less occluded parts. Instead of using a set of part filters to detect each part for each frame, we directly classify features into parts so that a lot of computational cost can be reduced. We use BoW-Pyramid to represent an action so that actions with reversed temporal orders could be distinguished. Moreover, we define the contextual information based on the content of actions so that it is applicable to most actions and helpful to action recognition. Experimental results demonstrate that BoW-Pyramid is comparable to the state-of-the-art approaches, and also show the robustness to occlusion as well as the benefit of contextual information.

REFERENCES

[1] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," *IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 726–733, Oct. 13-16, 2003.

[2] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, June 23-28, 2008.

[3] J. Liu, J, Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1996–2003, June 20-25, 2009.

[4] M. S. Ryoo, C.-C. Chen, J. K. Aggarwal, and A. Roy-Chowdhury, "An overview of contest on semantic description of human activities (SDHA) 2010," *Recognizing Patterns in Signals, Speech, Images and Videos*, *LNCS* vol. 6388, pp. 270–285, 2010.

[5] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," *Int. Conf. Computer on Vision and Pattern Recognition*, vol. 3, pp. 32–36, Aug. 23-26, 2004.

[6] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: a color-depth video database for human daily activity recognition," *IEEE Int. Conf. on Computer Vision Wksp.*, pp. 1147–1153, Nov. 6-13, 2011.

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

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

[9] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and Viterbi path searching," *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, June 17-22, 2007.

[10] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," *IEEE 12th Int. Conf. on Computer Vision*, pp. 32–39, Sept. 29-Oct. 2, 2009.

[11] K.N. Tran, I.A. Kakadiaris, and S.K. Shah, "Part-based motion descriptor image for human action recognition," *Pattern Recognition*, vol. 45, pp. 2562–2572, 2012.

[12] http://www.openni.org/

[13] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," *Euro. Conf. on Computer Vision, LNCS* vol. 6313, pp. 635–648, 2010.

[14] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *2nd Joint IEEE Int Wksp. on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, Oct. 15-16, 2005.

[15] I. Laptev, "On space-time interest points," *Int. J. Computer Vision*, vol. 64, pp. 107–123, Sept. 2005.

[16] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 886–893, June 25, 2005.

[17] J. C. Niebles, H. Wang, L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Computer Vision*, vol. 79, pp. 299–318, Sept. 2008.

[18] H. Zhang and L.E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 2044–2049, Sept. 25-30, 2011.

[19] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," *Proc. of 15th Int. Conf. on ACM Multimedia*, pp. 357–360, 2007.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, 2006.

[21] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," *IEEE. Conf. on Computer Vision and Pattern Recognition*, pp. 2847–2854, June 16-21, 2012.

[22] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2929–2936, June 20-25, 2009.

[23] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," *IEEE 12th Int. Conf. on Computer Vision*, pp. 1933–1940, Sept. 29-Oct. 2, 2009.

[24] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," *Euro. Conf. on Computer Vision, LNCS* vol. 6311, pp. 494–507, 2010.

[25] E.-J. Weng and L.-C. Fu, "On-line human action recognition by combining joint tracking and key pose recognition," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 4112–4117, Oct. 7-12, 2012.

[26] C.-C. Chang and C.-J, Lin, "LIBSVM: a library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 27, Apr. 2011

[27] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing RGB and depth map features for human activity recognition," *Signal & Information Processing Association Annual Summit and Conference*, pp. 1–4, Dec. 3-6, 2012.