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

Recognizing Context-aware Activities of Daily Living using RGBD Sensor

Jie Fu, Chengyin Liu, Yen-Pin Hsu and Li-Chen Fu, *Fellow, IEEE*

***Abstract*—In this paper, we propose a Bayesian conditional probability with latent-structure model for context-aware activities of daily living (ADL) recognition. The proposed ADL recognition system takes RGBD sensor (Microsoft Kinect) as the input device. In ADL recognition, the object interacted with human is a sort of important context as well as human action. To better understand the activity, we model the interacted object and the human action together. As far as we known, many related works failed to take into account the relation between the context information and human action features, instead, most of them only consider the human action features, causing ambiguity in classifying the activities with similar human actions. In this paper, the context information and human action features are taken into consideration, concurrently, so that the performance of recognition can be greatly improved from previous works as has been demonstrated in our experimental results.**

I. INTRODUCTION

Recent years, as the RGBD sensor becomes more and more epidemic, such as Microsoft Kinect and ASUS Xtion, it has attracted a lot of attention on taking the advantage of depth information for action recognition and object detection. Now we can easily capture the skeleton feature and depth image from RGBD sensor via OpenNI [1] or Microsoft Kinect SDK [2]. How to exploit these information to build a highly discriminate activity recognition model is what we mostly concern about.

Automatically recognizing human activity is essential in many applications, such as human nursing robot, and daily life logger for health care. For instance, if the robot can understand the user working on computer, it will be able to remind the user to take a rest when the user focuses himself on the computer for too long. However, human activities are different from human actions, because: 1) Human actions usually take short period of time, and don't contain semantic meanings, whereas human activities often last for a long time, and contain semantic meanings, such as drinking water, brushing teeth, and cooking. 2) One activity may contain two, or three actions. For example, the activity of drinking water consist of raising of hand, approaching of hand to mouth, and lowering of hand, and 3) In addition, human activities usually involve special objects. As shown in Fig. 1, the user is working on a computer, and the special objects here include chair, table, and computer. The specific combination of different objects is defined as "context information" in our work. Here, we focus on recognizing activities of daily living (ADLs), and define the ADL recognition with specific context as context-aware ADL recognition.

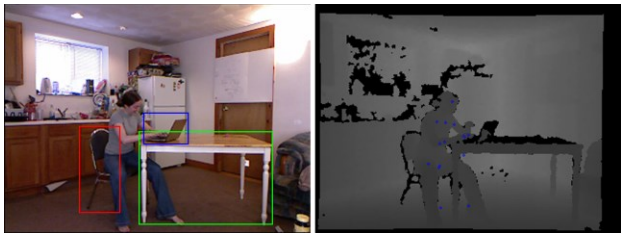

Fig. 1. The snapshot of working on computer of RGBD data from Kinect. We detect context information on RGB image, and extract skeleton features as well as Depth Histogram of Oriented Gradients (DHOG) features on depth image.

In this paper, we propose an approach combining the context information as well as human action features. With the context information, it is much more confident to classify different activities but with similar human action features. For example, "drinking water" and "making phone call" have similar human actions features, such as "raising of hand", "approaching of hand to head" and "lowering of hand". However, the context information of each is totally different from one the other; for instance, the cup is necessary when drinking, whereas a phone is necessary when a phone call is being made. Actually, in our daily living, many activities have similar human actions but with different specific context information. Alternatively, some activities have similar context information but with different human action features, like "rinsing of the mouth" and "drinking water". Both activities involve the same context information, but the user doesn't always need bend him-self down when drinking water. As far as we known, in ADL recognition, context information and human action features are both important, and they are not one less. Thus, our approach models them together for ADL recognition.

The rest of the paper is organized as follows. Section II presents the related work and some state-of-the-art recognition models. Section III describes the context-aware ADL recognition. Section IV discusses about the advanced context

Jie Fu is with the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC (e-mail: r00922140@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).

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).

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).

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

information extraction. The dataset and experimental results are presented in section V. Finally, conclusions are given in Section VI.

II. Related Work

From the ubiquitous computing community, there have been a fair amount of researches and works on activity recognition, such as Philipose, *et al*. [3]. In this field, many works address the problem from the view of "life-logging", such as Blum, *et al*. [4]. Most approaches ignore visual cues, but focus on alternate sensors such as RFID tags or accelerometers instead. This requires a fairly involved effort for instrumenting both the observer and the environment. One may argue that it is more practical to instrument the observer than instrument both the observer and environment, for example, Pirsiavash and Ramanan [5] suggest wearable camera may be a choice to place on observer, and observe the front scene.

There is a rich history of activity recognition in the vision community. One common approach is to use space-time interest point to model the features in video, which is introduced by Laptev [6]. And many similar works like [7, 8]. However, this approach is only capable of classifying, rather than detecting activities. In addition, this approach often needs to consume a lot of time. Despite only 2D image features are extracted from the video, more and more approaches have attempted to capture skeleton features and depth image features as RGBD sensors have been widely used. Sung, *et al*. [9] propose to use a RGBD sensor (Microsoft Kinect) as the input sensor, and compute a set of features based on human pose, human motion, RGB image and depth information. Although this approach has achieved quite good performance, the result is still unsatisfactory when two activities involving similar human postures. As many human daily activities involve special objects, the interacted objects and other surrounding objects become a vital cue in human-object interaction ADL recognition [10, 11]. Pirsiavash and Ramanan [5] prove the essential of the objects near the observer, they only use the observed objects to model the temporal representation for the histogram of objects. However, due to use the wearable camera to observe the surrounding environment, it lacks the feature of human-self action.

Object detection has a long history in the field of computer vision. Dalal and Triggs [12] introduce the histogram of oriented gradients (HOG), which has been the most widely used approach on object detection. To deal with the occlusions and deformable objects in real environment, Felzenszwalb, *et al*. [13] propose a discriminative part-based model (DPM), which detects the whole object as well as the parts of the object. Within these years, the RGBD sensor is also used for object detection. Yu, *et al*. [14] propose a hierarchical representation with scarcity for RGBD object recognition approach. Lai, *et al*. [15] introduce the approach using 3D reconstruction.

To capture the temporal relations in the observation sequences of the input video, the recognition model should be well defined. Morency, *et al*. [16] propose a discriminative approach named Latent-Dynamic Conditional Random Field (LDCRF) model for gesture recognition. This approach incorporates hidden state variables which are able to model the sub-structure of a sequence and learn dynamics between class labels. Song, *et al*. [17] extend the LDCRF model for multi-view action detection.

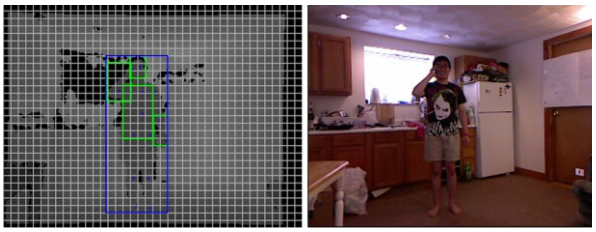

Fig. 2. We divide the whole depth image into blocks. And compute the histogram of oriented gradients for each block. The bounding boxes of body parts are computed according to joint positions, and add up the histograms of the blocks whitin the bounding box.

Our context-aware ADL recognition model is inspired by the excellent performance of DPM [13]. Although part-based detector has high time complexity, we can just detect the region near hand and foot rather than the whole image, since the region of interest (ROI) can be found with respect to human skeleton features. To emphasize the importance of the sub-activity model, LDCRF is employed to model our human action features. Inspired by Li, *et al*. [18] and Gupta, *et al*. [10], we model the human action and context information using the conditional probability. For the detail, our approach is introduced in Section III, IV, and is evaluated in Section V.

III. Context-Aware Daily Activity Recognition

A supervised learning approach is employed where we collect labeled data as ground-truth as training data. There are 12 functional activities, such as talking on the phone (TP), writing on whiteboard (WW), drinking water (DW), rinsing mouth with water (RM), brushing teeth (BT), wearing contact lenses (WL), talking on chair (TC), relaxing on chair (RC), cooking with chopping (CC), cooking with stirring (CS), opening pill container (OC), and working on computer (WC). Besides, another 2 activities are used for recognizing the neutral activity, including standing still and random actions. The input of our system are RGBD images from a Kinect sensor, from which we extract the features. Sung, *et al*. [9] introduce the features about the human actions, including body pose features, hand position, motion information as well as body shape features. However, combing these features results in a large dimension. We don't use motion feature and body pose features, because the difference of skeleton joints motion are not quite obvious in activities of daily living and different users may have different poses when performing the same activity. In addition, the foot position feature is also extracted, which similar to the hand

TABLE I. OBJECTS AND ACTIVITIES OF DAILY LIVING

Objects	Activities of Daily Living											
	TP	*WW*	*DW*	*RM*	*BT*	*WL*	*TC*	*RC*	*CC*	*CS*	*OC*	*WC*
Table	0	0	0	0	0	0	0	0	**1**	**1**	0	**1**
Chair	0	0	0	0	0	0	**1**	**1**	0	0	0	**1**
Computer	0	0	0	0	0	0	0	0	0	0	0	**1**
Cup	0	0	**1**	**1**	**1**	0	0	0	0	0	0	0
Teeth brush	0	0	0	0	**1**	0	0	0	0	0	0	0
Phone	**1**	0	0	0	0	0	0	0	0	0	0	0
White board	0	**1**	0	0	0	0	0	0	0	0	0	0
Container	0	0	0	0	0	0	0	0	0	0	**1**	0
Contact lenses	0	0	0	0	0	**1**	0	0	0	0	0	0
Bucket	0	0	0	**1**	0	0	0	0	0	0	0	0

Talking on the phone (TP), Writing on whiteboard (WW), Drinking water (DW), Rinsing mouth with water (RM), Brushing teeth (BT), Wearing contact lenses (WL), Talking on chair (TC), Relaxing on chair (RC), Cooking with chopping (CC), Cooking with stirring (CS), Opening pill container (OC), and Working on computer (WC)

1: The activity involves the object

0: The activity doesn't involve the object

position feature. Here, HOG features are extracted for depth image. The context information is extracted from object detection using DPM [13], which is the state-of-the-art approach to object detection problem. A conditional random field-like model is trained to capture the relations among activity, action and context, including activity-action relation, action-context relation, and activity-context relation.

A. Human Action Features

Hand position. Hand is the most common body part performs action in activities of daily living. And as an end-effector that interacts with object directly, hand plays an important role. We compute the relative positions of left and right hands with respect to head position and torso position, respectively.

Foot position. Body pose can be inferred from relative foot position with respect to torso position. For example, it can be inferred from the relative positions of feet whether the user is standing or sitting. Thus, we compute the relative positions of both feet, respectively.

HOG feature. Although the relative positions of hand and foot describe human's pose, we still need some features to describe the shapes of body parts. Here, the depth image is regarded as gray scale image from which HOG features are extracted. HOG feature descriptors, which gives 32 features in a block that count how often certain gradient orientations are seen in specified bounding boxes of an image, is widely used in human detection. To boost the feature extraction, firstly, the whole depth image is divided into blocks with 8 pixels high and 8 pixels wide, and compute the HOG feature for each block, as shown in Fig. 2. Secondly, the regions of body parts are located in terms of joint positions, including head, left arm, right arm and torso, and add up the features of every block in the body part region. Thirdly, the histogram which represents the HOG feature is normalized between 0 and 1.

B. Context Information

The context information is extracted from the RGB image. It is the combination of specific objects which are detected using DPM [13]. Activities often involve specific context information, as shown in Table I. In normal object detection problem, it is a hard task to detect the location and the category of the object, which also consumes a lot of time. But in our work, the location is limited to the neighborhood of the region of body part. That will reduce much search space for a sliding window method in object detection.

C. Model Formulation

The object categories and the human actions are estimated, simultaneously. Then the activity can be inferred base on the object categories and human actions. We model the relations as shown in Fig. 3.

To simplify our model, we assume the activity A and combination of objects $\mathbf{O} = \{O_1, O_2, ..., O_m\}$ are conditional independent under the observation of object $\mathbf{x}_o$, and the

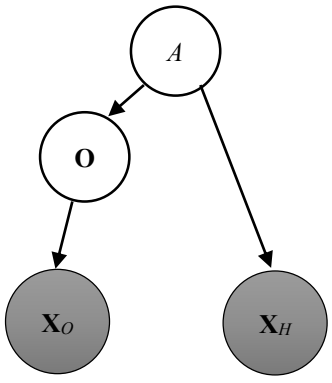

Fig. 3. Graphical Model for activity recognition. The gray nodes represent observations, **O** node represents detected object categories and A node represents activity.

observation of human action $\mathbf{x}_H$. The joint probability is decomposed using conditional independence relations:

$$P\left(A, \mathbf{O} \mid \mathbf{x}_H, \mathbf{x}_O\right) = P(A \mid \mathbf{x}_H) P(\mathbf{O} \mid \mathbf{x}_O) \quad (1)$$

where $P(A \mid \mathbf{x}_H)$ computes the candidates of the activity recognition according to the human actions, and $P(\mathbf{O} \mid \mathbf{x}_O)$ computes the result of object detection for the surrounding objects.

We consider the following terms.

- $P(\mathbf{O} \mid \mathbf{x}_O)$: The context term is modeled by detecting the probability of object categories given the observation $\mathbf{x}_O$. Thanks to the skeleton model from depth image, joint positions of the user can be obtained. Thus we only consider the nearby regions around the user using a mask to crop the whole image. *K* object models are used to detect the objects in every frame, and score the pixel at particular position (*r, c*) as well as scale *s*. So the score of the k[th] object model in frame *i* at the particular location and scale can be written as (2). The maximum score of the k[th] object model at the particular (*r, c, s*) in frame i is regarded as the score of the k[th] object model in frame i. The score of the k[th] object in the whole observation can be represents as (3), which is the average score of the object in all frames. Then, the score of the k[th] object is converted to conditional probability using (4). We make a normalized probability vector **p** as (5), which includes the detection result for each object. At last, the probability of context is computed by Bhattacharyya distance [19] as shown in (6), where $\mathbf{N}_A$ represents the normal vector of the context information according to Table I for activity *A*.

$$\mathrm{g}(k,i) = \max_{r,c,s} \mathrm{score}_i^k(r,c,s) \quad (2)$$

$$\mathrm{G}(k) = \frac{1}{|T|} \sum_{i \in T} \mathrm{g}(k,i) \quad (3)$$

$$P(O_k \mid \mathbf{x}_O) = 0.5 + 0.5 \frac{\mathrm{G(k)}}{\max_{i \in K} \mathrm{G}(i)} \quad (4)$$

$$\mathbf{p} = \{P(O_1 \mid \mathbf{x}_O), P(O_2 \mid \mathbf{x}_O), ..., P(O_K \mid \mathbf{x}_O)\} \quad (5)$$

$$P(\mathbf{O} \mid \mathbf{x}_O) \propto -\ln\left(\sum_{i \in K} \sqrt{\mathbf{p}^{(i)} \cdot \mathbf{N}_A^{(i)}}\right) \quad (6)$$

- $P(A \mid \mathbf{x}_H)$: We model this human action term using latent-structure model as shown in Fig. 4. Here the bottom depth images represent the observations $\mathbf{x}_H$. The system learns a mapping between a sequence of observations $\mathbf{x}_H = \{x_1, x_2, ..., x_n\}$ and a sequence of labels $\mathbf{y} = \{\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_n\}$. Each y_j is a class label for the j[th] frame of the depth image sequence and is a member of a set y of possible class labels. Each frame observation x_j is represented by human action feature vector $\phi(x_j) \in R^d$. For each sequence, we also assume a vector of "sub-structure" variables $\mathbf{h} = \{\mathrm{h}_1, \mathrm{h}_2, ..., \mathrm{h}_m\}$. This sub-structure can be regarded as sub-activity in our problem, and an entire activity is consist of several sub-activities. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model. With the definition above, we model the term as (7), which is first proposed in [16] named Latent-Dynamic Conditional Random Field (LDCRF).

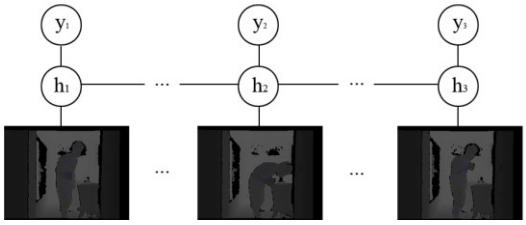

Fig. 4. We model the human action feature using latent-struct model. The entire activity can be divided into sevral sub-activities.

$$P(A \mid \mathbf{x}_H) = P(\mathbf{y} \mid \mathbf{x}_H, \theta) = \sum_{\mathbf{h}} P(\mathbf{y} \mid \mathbf{h}, \mathbf{x}_H, \theta) P(\mathbf{h} \mid \mathbf{x}_H, \theta) \quad (7)$$

where θ is the set of parameters of the model.

IV. Advanced Context Information

In the previous section, we manually assign the involved objects to the related activities. It will be too naïve when the more and more objects are involved as the activity become more and more complex. In this section, we introduce an advanced method to present the relations between objects and activities.

Latent semantic indexing (LSI) [20] is a method widely used in the field of Information Retrieval. LSI is based on the principle that words occur in the same contexts of a document tend to have similar meanings.

Here, the objects are regarded as indexing terms, and the activities are regarded as context of the document. LSI is employed to extract the latent relations of objects by establishing associations between those objects that occur in similar activities. For example, when the user is brushing teeth, the necessary context information is the combination of toothbrush, and cup, then we assume toothbrush and cup have similar meaning. However, the cup is also involved in the activity of rinsing mouth with water. Thus we concern whether the toothbrush relates to rinsing mouth with water latently. Actually, the user often brushes teeth and rinsing mouth with water in the bathroom, and the toothbrush can be

observed in both situations.

The content of Table I is transfer to matrix **A**, and decomposed using Singular Value Decomposition (SVD) as shown below,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{8}$$

where **U** is a unitary matrix, **D** is a diagonal matrix with sorted eigenvalues, and $\mathbf{V}^T$ is a conjugate matrix. The rank of **D** is *r*. Rank-reduce is applied to set the eigenvalues to 0 whose value is under a threshold, and thus the rank of **D** becomes *k,* written as $\mathbf{D}_k$. Then matrix **A** is converted to matrix $\mathbf{A}_k$ as shown below.

$$\mathbf{A}_k = \mathbf{U}\mathbf{D}_k\mathbf{V}^T \tag{9}$$

We use $\mathbf{A}_k$ as the new context information matrix.

V. EXPERIMENTS

A. Dataset

Sung, *et al*. [9] provide Cornell Activity Datasets (CAD-60) recorded by Microsoft Kinect with RGB images, depth images and skeleton data. The dataset includes 14 activities performed by 4 different subjects. Among the activities, there are 12 functional activities, and 2 activities for recognizing the neutral activity. These activities are common in our daily living, like brushing teeth, drinking water and talking on the phone. Shown in Table II, the length of a collection is between 400 and 2000 frames. The data was collected in different parts of regular household with no occlusion of arms and body from the view of sensor. We split each activity into 150 to 200 frames, and use leave-one-out cross validation to train and test our model.

It is difficult to train a reliable object detection model even though the state-of-the-art DPM is employed. Because the objects interacted with human during the activity are highly occluded, and have arbitrary orientations. So, we collect the cropped object image among part of CAD-60 dataset, which are used for training. Besides, we obtain more images from Google image search as our positive samples, and use some negative data from INRIA and CALTECH dataset.

B. Experimental results

We compare our Bayesian conditional probability model with Hierarchical Maximum Entropy Markov Model (HMEMM) [9]. Different settings are evaluated to show the outstanding performance of involving context information and latent-structure model. Table III shows the overall average performance of our approach with different settings and HMEMM. Since the original work of HMEMM aims to classify un-segmented data, so we only compare with it over un-segmented data. A sliding window method is applied with

TABLE II. STATISTIC OF DATA LENGTH

Activity name	mean of length (frames)	std. dev. of length (frames)
Talking on the phone	1500	200
Writing on whiteboard	1500	100
Brushing teeth	1350	300
Cooking (stirring)	1200	200
Working on computer	1300	50
Rinsing mouth with water	1500	100
Wearing contact lenses	400	50
Relaxing on a chair	1300	200
Opening a pill container	400	200
Drinking water	1500	100
Cooking (chopping)	1600	200
Talking on a chair	1400	300
Still	1000	100
Random	1900	200

TABLE III. OVERALL AVERAGE OF DIFFERENT APPROACH

Approach	Pre-segmented		Non-segmented	
	Prec.	*Rec.*	*Prec.*	*Rec.*
Context+LDCRF	95.4	95.3	90.1	89.5
LSI Context+LDCRF	95.3	94.1	88.4	89.1
Context+CRF	92.5	91.3	80.4	80.3
LDCRF	87.4	86.7	83.1	80.4
HMEMM	N/A	N/A	84.7	83.2

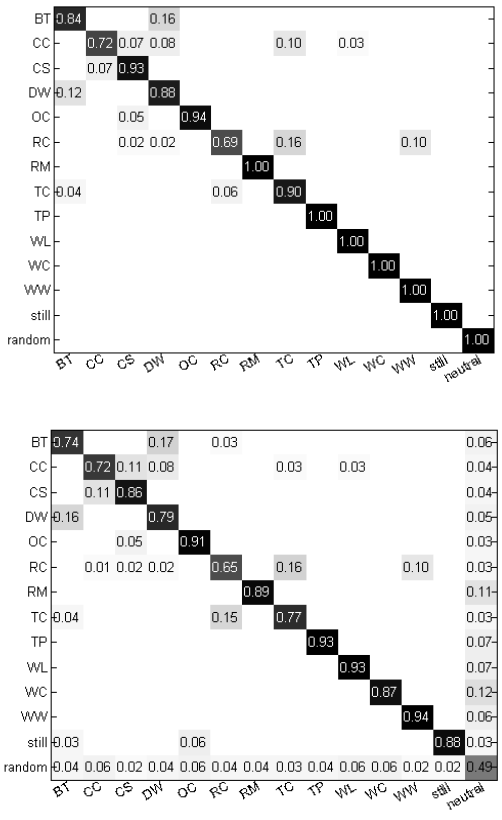

Fig. 5. Confusion matrix of Context+LDCRF approach for pre-segmented data and non-segmented data.

100 frames long window size when testing un-segmented data. After involving the context information, the precision and recall are both improved about 5%. Comparing with the LDCRF setting and CRF setting, the improvement is remarkable. However, not all objects can contribute to improve the performance of the ADL recognition, for example, contact lens and teeth brush are too small to be detected using DPM, and these kinds of activities are mainly recognized by human's actions. In our approach, the human action dominates the result of activity recognition, the context information plays a role of supporter to refine the result of ADL recognition.

We evaluate the approach with LSI, however, the performance is not higher than that without LSI, even lower in several cases. Because the activities in the dataset don't involve very similar combination of objects, and in the most case, the activity only involve one object according to Table I. LSI needs more similar combination of objects associated with the different activities.

In activities of daily living, many of them are quite similar on the view of human actions. Such as talking on the phone and drinking water. In our approach, the context information is taken into consideration. Fig. 5 shows the confusion matrix of our recognition result for pre-segmented data and non-segmented data. From the results of confusion matrix, the similar human action but with different context information, such as brushing teeth and talking on the phone, can be distinguished well in our experiment. As the activities involving specific objects, like working on computer and writing on white board, they can be recognized at a high accuracy. At the same time, the activities that have similar context information but with different human actions can be also classified, due to our latent structure human action model.

VI. Conclusion

In this paper, we considered the problem of ADL recognition in home environment. The RGBD sensor (Microsoft Kinect) is used as the input sensor, which is inexpensive to build applications like human nursing robot, and daily living logger for health care. We presented a Bayesian conditional probability model which combining DPM and LDCRF. Our approach involves context information in activity recognition, which is different from the traditional activity recognition method only concerns human actions. During inference, our algorithm exploited the nature of human-object interaction in activities to classify the category of the activity. In our approach, the human action dominates the result of activity recognition, the context information plays a role of supporter to refine the result of ADL recognition. We tested our algorithm extensively on twelve different activities performed by four different people with different objects around the people. It achieves good recognition performance in both inter-user and intra-user.

References

[1] *OpenNI*. Available: http://www.openni.org/

[2] *Microsoft Kinect SDK*. Available: http://www.microsoft.com/en-us/kinectforwindows/

[3] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, *et al.*, "Inferring activities from interactions with objects," *IEEE Pervasive Computing,* vol. 3, pp. 50-57, 2004.

[4] M. Blum, A. Pentland, and G. Troster, "Insense: Interest-based life logging," *IEEE MultiMedia,* vol. 13, pp. 40-48, 2006.

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

[6] I. Laptev, "On space-time interest points," *International Journal of Computer Vision,* vol. 64, pp. 107-123, 2005.

[7] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1234-1241.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

[9] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 842-849.

[10] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, pp. 1775-1789, 2009.

[11] B. Yao and L. Fei-Fei, "Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1691-1703, 2012.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886-893.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 1627-1645, 2010.

[14] K.-T. Yu, S.-H. Tseng, and L.-C. Fu, "Learning hierarchical representation with sparsity for RGB-D object recognition," in *IEEE Conference on Intelligent Robots and Systems*, 2012, pp. 3011-3016.

[15] K. Lai, L. Bo, X. Ren, and D. Fox, "Detection-based object labeling in 3d scenes," in *IEEE Conference on Robotics and AutomationI*, 2012, pp. 1330-1337.

[16] L. P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[17] Y. Song, L. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2120-2127.

[18] G. Li, C. Zhu, J. Du, Q. Cheng, W. Sheng, and H. Chen, "Robot semantic mapping through wearable sensor-based human activity recognition," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 5228-5233.

[19] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā: The Indian Journal of Statistics (1933-1960),* vol. 7, pp. 401-406, 1946.

[20] S. T. Dumais, G. Furnas, T. Landauer, S. Deerwester, and S. Deerwester, "Latent semantic indexing," in *Proceedings of the Text Retrieval Conference*, 1995.