RGB-D Camera-based Daily Living Activity Recognition

Chenyang Zhang, *Student Member, IEEE* and Yingli Tian, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a new activity analysis framework to facilitate the independence of older adults living in the community, reduce risks, and enhance the quality of life at home by recognizing activities of daily living (ADLs) by using RGB-D cameras. Comparing to the traditional RGB cameras, the depth information implicitly has advantages of handling illumination changes and protecting privacy. Our contributions include three aspects. First, to detect abnormal activities which are dangerous for elderly people, we recognize 5 activities related to fall including standing, fall from standing, fall from sitting, sit on chair, and sit on floor. Second, to recognize finer activities of daily living, we propose a discriminative representation of structure-motion features based on skeleton joints. Third, to continually track same person when there are multiple people appear in the same camera view, we further develop a binary classification based person identification method by combining appearance and depth information. The proposed framework is evaluated on a dataset we collected under different lighting conditions for fall detection and a benchmark dataset for daily living activity recognition. Experiment results demonstrate the effectiveness of the proposed framework and outperform the state-of-the-art method.**

***Index Terms*—RGB-D camera, Activity analysis, Fall detection, Privacy protection.**

I. Introduction

In 2008, about 39 million Americans were 65 years old or above. This number is likely to increase rapidly as the baby boomer generation ages. The older population increased elevenfold between 1900 and 1994, while the nonelderly increased only threefold, and the oldest old (persons of 85 or older) is the fastest growing segment of the older adult population [12]. Consequently, demands on programs of Medicare and Medicaid increase with the gradual retirement of baby boomers. The proportion requiring personal assistance with everyday activities increases with age, ranging from 9 percent for those who are 65 to 69 years old to 50 percent for those who are 85 or older. Furthermore, the likelihood of dementia or Alzheimer's disease increases with age over 65 [1]. In 2006, there were 26.6 million sufferers worldwide. These data indicate that the demand for caregivers will reach far beyond the number of individuals able to provide care.

One solution to this growing problem is to find ways to enable elders to live independently and safely in their own homes for as long as possible [14]. Recent technology developments in computer vision, digital cameras, and computers make it possible to assist the independent living of older adults by developing safety awareness technologies to analyze the elder's activities of daily living (ADLs) at home. Important activities that effect independence include ADLs (*e.g.*, taking medications, getting into and out of bed, eating, bathing, grooming/hygiene, dressing, socializing, doing laundry, cooking, cleaning). Among these activities, a few are rated as very difficult to monitor, including taking medication, falling and eating [24]. In our point of view, there are two aspects of ADL system application: one is to detect abnormal events from usual but similar actions such as falling down and sitting down, and the other is to log and record normal activities of subjects for further research and analysis [4, 13, 21].

In this paper, we focus on recognizing activities of daily living by developing a set of structure-motion based skeleton statistics features based on 3D information. We further utilize the combination of appearance and depth features to handle if observed subjects are apparently the same person. Our proposed research are designed for three tasks: 1) detecting falling event from other similar activities related to falling such as sit on floor, *etc.*, 2) classifying finer indoor activities, such as talking on the phone, *etc.*, and 3) user identification.

Task 1: Falling Event Detection and Recognition. As shown in Table 1, we recognize five activities related to falling event including *"standing"*, *"fall from standing"*, *"fall from sitting"*, *"sit on a chair"*, and *"sit on floor"* by using depth camera. Compared with traditional video surveillance cameras, depth cameras have implicit advantages of handling illumination changes and identity protection. We extract simple and discriminative kinematic features from 3D information which consist of two parts: 1) structure similarity and 2) head-floor distance, which is defined as the vertical distance between the head and the floor plane.

Task 2: ADL Activity Recognition. We recognize 13 finer activities including *"talking on the phone"*, *"writing on whiteboard"*, *"drinking water"*, *"rinsing mouth with water"*, *"brushing teeth"*, *"wearing contact lenses"*, *"talking on couch"*, *"relaxing on couch"*, *"cooking(chopping)"*, *"cooking(stirring)"*, *"open pill container"*, *"working on*

This work was supported in part by NSF grants IIS-0957016, Microsoft Research, and PSC-CUNY Award 64720-00 42.

Chenyang Zhang is with the City College, City University of New York, New York, NY 10031 USA (e-mail: czhang10@ccny.cuny.edu).

YingLi Tian is with the City College, City University of New York, New York, NY 10031 USA (phone: 212-650-7046; fax: 212-650-8249; e-mail: ytian@ccny.cuny.edu). Prior to joining the City College in September 2008, she was with IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA.

computer” and “*standing still / random*”. Compared to Task 1, the differences between actions for this task are more subtle and ambiguity. Therefore we propose to use low-level features based on tracked skeleton joints which combine both motion and structure features to learn a discriminative classifier for robust recognition.

Task 3: User Identification. We employ a background subtraction and tracking method and represent actions as histogram features based on 2D appearance RGB information. Classification on two different SVM schemes are performed and analysis. We further develop a patch-based histogram matching method by combining 3D information (depth) and appearance information (RGB) to identify different people.

II. Related work

Helping people with special needs by human activity recognition is a hot research area in computer vision. Recognizing Activity in Daily life (ADL) is a potential field where computer vision can really help elderly people to improve the quality of their lives [18]. Nait-Charif *et al.* developed a computer-vision based system to recognize abnormal ADL [17] in a supportive home environment. The system tracked human activity and summarized frequent active regions to learn a model of normal activity. It detected falling as an abnormal activity, which is very important in patient monitoring systems. Unlike using location cues in [17], Wang *et al.* [23] proposed to use gestures by applying a deformable body parts model [11] to detect lying people in a single image. To detect certain parts of human body, Buehler *et al.* [2] proposed to fit an upper-body model for sign language recognition. Different from traditional RGB channel, recognizing activities using depth images is a new trend in recent research [15, 21, 25, 26] especially after Microsoft released its SDK for Kinect cameras in year 2010 [16]. Recently, Li *et al.* [15] proposed to use bag of 3D points to represent and recognize human actions based on 3D silhouette matching. Hidden Markov Model (HMM) is employed with depth images to effectively recognize human activities in [21].

In this paper, we propose a new activity analysis framework and develop a set of structure-motion based skeleton statistics features based on 3D information to recognize activities of daily by using RGB-D cameras.

III. Falling Event Detection and Recognition

A. Feature Extraction and Representation

Kinematic Feature Extraction: The Microsoft Kinect SDK [16] provides 20 joints on human body tracked for each person in each depth frame. In order to detect and recognize activities related to falling event, we select 8 joints on head and torso since joints on limbs introduce more noise than useful information to distinguish whether a person falls or not. The chosen 8 joints, as shown in Figure 1(a), keep a relative stable structure model when a person is standing or sitting, in other words, the structure model is not affected much when a person is performing normal activities. However, the structure model is no longer reliable when a person falls. In this paper, we extract kinematic features including the structure similarity and head-floor distance.

Table I: Five activities related to falling event

L_1	Fall from sitting	L_2	Fall from standing
L_3	Standing	L_4	Sit on chair
L_5	Sit of floor		

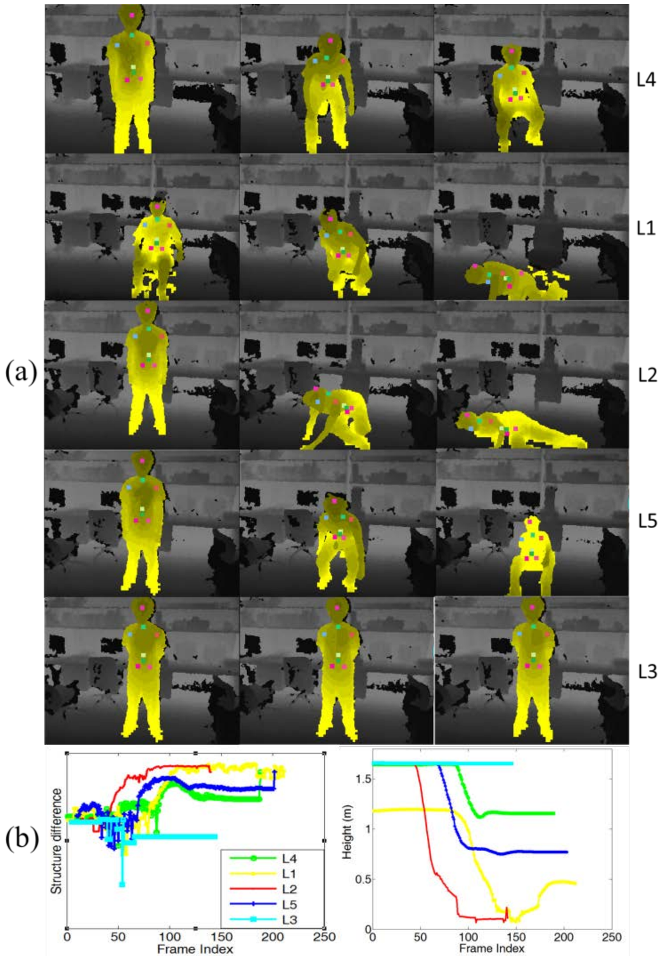


Figure 1 Illustration of kinematic feature extraction: the structure difference cost. (a) From top to bottom, each label’s initial (left), intermediate (middle) and final pose (bottom). Human segmentation and skeleton joints are also displayed. (b) Two main elements we extracted from skeletons as features. Left: logarithm of structure similarity. Right: head-floor distance.

Kinematic Feature Representation: Figure 1(a) displays the initial (the leftmost column), intermediate (middle column) and final (the rightmost column) poses of the five activities to be recognized. Obviously, the two “falling” events (L_1: Fall from sitting and L_2: Fall from standing) have much larger deformation on the skeleton structure model than the other three “non-falling” events. We define that the structure similarity as the difference cost $C(\xi)$ of a skeleton structure ξ to measure the degree of deformation as the summation of angle changes between the corresponding joints of the skeleton between the initial and final poses as following:

$$C(\xi) = \sum_{i=1}^{n}\sum_{j=i+1}^{n}\left\|\theta\left(\xi_i,\xi_j\right) - \theta\left(o_i,o_j\right)\right\|, \qquad (1)$$

where $\theta\left(\xi_i,\xi_j\right)$ and $\theta\left(o_i,o_j\right)$ denote the angles between two joints i and j on skeletons of ξ and o, respectively, which is given as:

$$\theta(i,j) = \frac{\arcsin\left(\frac{i_x - j_x}{dist(i,j)}\right)}{2\pi}, \qquad (2)$$

where the geometry distance between two joints i and j is denoted as *dist(i, j).*

Examples of the structure similarity (in logarithm) for videos with different activities are displayed in Figure 1(b) (left graph). Red ("fall from standing") and yellow ("fall from chair") curves obviously demonstrate significant costs as expected. We extract two statistical features of the structure similarity (the mean μ and the varianceσ) to represent the action in a video sequence.

Another feature we use for activity recognition is head-floor distance, which will change dramatically if a person sits or falls. Given a floor plane [A B C D][x y z 1]^T=0 and homogeneous representation of head 3D position $[\eta_x,\eta_y,\eta_z,1]$, the head-floor distance can be estimated as $[\eta_x,\eta_y,\eta_z,1]$[A,B,C,D]^T, where the parameters of floor plane can be fitted using RANSAC algorithm. As shown in the right graph of Figure 1(b), the head-floor distance is a discriminative feature for fall related activity recognition. We employ the highest value h and the minimum value l of the head-floor distance at different skeleton poses as the last two feature elements in our kinematic feature vector. The kinematic feature vector from 3D depth information is denoted as $[\mu\ \sigma\ h\ l]^T$.

B. Activity Classification

We employ a set of SVM classifiers to recognize different actions by using a "1-*vs.*-all" structure. ``1-*vs.*-all'' is applied to kinematic features since the inter-class difference can be well represented by our modeling.

IV. Activities of Daily Living Recognition

A. Feature Computation for Finer Activity Recognition

In addition to the abnormal activities such as falling down, there are many activities of daily living such as drinking water, talking on the phone, *etc.* Different from falling activities as discussed above, these activities are more subtle which cannot be recognized by only using simple features such as the head-floor distance and skeleton structure similarity. We define these actions as "finer" activities of daily living (ADL). For example, comparing "drinking water" and "talking on the phone", the poses of both actions are very similar, i.e., holding an object beside mouth or ear. Thus to recognize activities with such subtle differences, we need to architect more distinguishable features as well as representation.

In the work of [21], the authors propose to model the actions with sub-actions and resolve the recognition problem by using a two-layer Maximum Entropy Markov Model (MEMM). In our work, however, we resolve this problem using a popular and powerful Bag of Features (BoF) model with simpler representation and less features and outperforms the results reported in [21].

Skeleton-based action recognition becomes more feasible when recent stereo camera technology evolution [16, 19]. To represent an activity by using a series of skeletons, we extract two types of features: one type of features for joint-wise motion to describe the movement of body parts during adjacent video frames and another type of features for multi-joint structure to describe the activity pose in a single video frame. The combination of the motion features and the structure features is sufficient to describe an activity represented by skeleton points. Considering each action contains a series of certain poses, the two types of features we propose in this paper are able to model both temporal information (joint-wise motion) and spatial information (multi-joint structure).

Assume that there are J stable joints tracked on a skeleton and each joint be represented by its Cartesian coordinate $[x,y,z]^T$, we model the motion features between two adjacent frames i and j for joint $k \in J$ as:

$$motion(i,j,k) = \begin{bmatrix} x_{i,k} \\ y_{i,k} \\ z_{i,k} \end{bmatrix} - \begin{bmatrix} x_{j,k} \\ y_{j,k} \\ z_{j,k} \end{bmatrix}, \qquad (3)$$

Thus for every joint the motion is described with 3 elements, for all J joints there are $3 \times J$ elements to represent motion information.

For multi-joint structure of a skeleton, we model the structure for every two joints $k,l \in J$ in frame *I* as:

$$structure(i,k,l) = \begin{bmatrix} x_{i,k} \\ y_{i,k} \\ z_{i,k} \end{bmatrix} - \begin{bmatrix} x_{i,l} \\ y_{i,l} \\ z_{i,l} \end{bmatrix}, \qquad (4)$$

Thus there are $3 \times \binom{2}{J}$ elements to represent the spatial information of the whole structure in a certain frame.

In our work, we employ 15 joints. Thus the total number of features after concatenating motion features and structure features of any two consecutive frames is: $3 \times 15 + 3 \times 15 \times 14 \div 2 = 360$, which is much smaller than 715 used in [21].

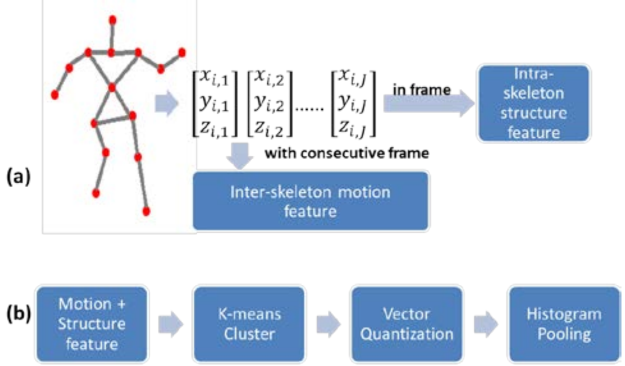
B. Bag of Features based Motion-structure Feature Representation

As described in the above section, we model the current pose of human and its temporal variations by fusion of structure features and motion features. A certain activity can be viewed as a certain composition of such combination of poses and variation trends, *i.e.*, sub-actions.

As illustrated in Figure 2, we employ the traditional Bag of Features (BoF) model to represent an activity with a fixed-size histogram and then apply a set of support vector machine (SVM) based classifiers to resolve the classification in a supervised learning manner.

We first perform K-means algorithms on training samples to learn a set of (K) vector centers. Since our vector dimension is fixed (360), the computation cost mainly determined by different K values, which is also named *codebook size*. Then we represent each video (a set of features of dimension 360) with a K-dimensional histogram, which is obtained by pooling all features the video contains into the K centers in a nearest-neighbor manner; this step is also named *histogram pooling* as illustrated in Figure 2 (b). Finally, a feature vector of size *K* is used to represent a video sequence and train a linear SVM classifier for finer activity recognition.

The framework of our proposed method is illustrated in Figure 2. The 15 joints and related positions are shown as red dots in Figure 2 (a). We use “average pooling” instead of “max pooling” in the histogram pooling step. Our method of feature extraction and representation is discriminative to classify different finer activities with subtle differences.



(a)
$\begin{bmatrix}x_{i,1}\\y_{i,1}\\z_{i,1}\end{bmatrix}\begin{bmatrix}x_{i,2}\\y_{i,2}\\z_{i,2}\end{bmatrix}\ldots\ldots\begin{bmatrix}x_{i,J}\\y_{i,J}\\z_{i,J}\end{bmatrix}$
in frame
Intra-skeleton structure feature
with consecutive frame
Inter-skeleton motion feature
(b)
Motion + Structure feature
K-means Cluster
Vector Quantization
Histogram Pooling

Figure 2 The proposed framework of feature computation and BoF representation of motion-structure features. (a) 15 joints are extracted on skeleton in a frame and represented as their Cartesian coordinates. Spatial structure is described by the translation vectors of each pair of joints. Temporal motion is described by joint-wise translation of skeletons in two consecutive frames. (b) BoF framework is applied to represent the final feature of the whole video.

V. Identify Multiple People

In order to handle multiple people appear in the view of a camera or cross different cameras, we develop a method of people identification by employing both RGB channels and D (depth) channel.

Although some embedded user identification functions are provided in both Microsoft SDK for Kinect [16] and PrimeSense OpenNI [14] to track a user, this tracking method can only answer questions like “How many users are there?” “Is the tracked user lost?” or “Is there a new user?” *etc.* When a person is out of the camera view and then re-enter the view, it is unable to tell whether this person is a new user or not.

In our approach, we combine 3D information (depth channel) and appearance information (HS channels in HSV color model) to accomplish user identification. First, we extract 4 patches in color image according to certain skeleton joints, which are available from depth channel, as shown in Figure 3, one along shoulders, one on torso, and two on two upper legs. Then we apply a weighted strategy on each pixel inside patches based on their depth value, as described in next section.

A. *User Identification based on Color Histogram*

Human detection and skeleton joins (for RGBD images are provided by built-in functions in the Microsoft SDK [16] and the PrimeSense OpenNI libraries [14]. To identify people, we extract four patches (see Figure 3) from RGB video based on skeleton joints from the depth channel: one on the shoulder, one on the torso, and two on the lower body.

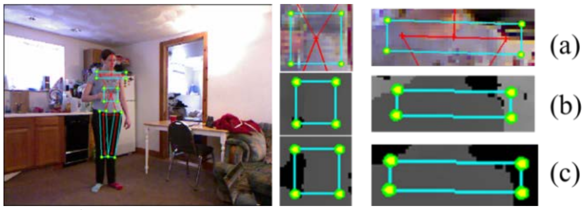


Figure 3 Left: 4 patches are extracted in color image according to certain skeleton joints. (a) The two patches on the upper body. (b) Corresponding depth channel. (c) Mask of weighting.

In our method, we assign the pixels of these patches with different weights according to their distance to the local joints on the Z (depth) coordinate. A local joint is defined as the joints in current patch, for example, in the patch along the shoulders (as shown in Figure 3), the local joints are two shoulders. We denote the weight as $\mathrm{w_i}$:

$$\mathrm{w_i} = \mathrm{e}^{-(\mathrm{z_i}-\mathrm{m})^2/\sigma^2}, \qquad (5)$$

where z_i is the depth value of i^{th} pixel in the patch and m is the measure point.

Sometimes the tracked joints of skeleton may locate on the background instead on body due to fast motion. Thus, we select the measure point m with the following rule:

$$\mathrm{m} = \begin{cases} \frac{\mathrm{z_i}+\mathrm{z_j}}{2} & \text{both i and j are located on body} \\ \mathrm{m_k} & \text{otherwise} \end{cases}, \quad (6)$$

where m_k is the median depth of all joints.

To better handle illumination changes, we transform the RGB to HSV color space and only use H and S channels for person identification. We quantize each channel in each patch into 20 bins, each pixel votes one bin with its weight w_i as calculated above.

B. *Identifying People by SVM-based Classifier*

For each patch, we generate a histogram in *H* and *S* channels as the feature representation respectively. We concatenate the histograms of four patches and two channels together and use the bin-wise difference as the input of a binary SVM classifier to identify if same person appears at different time under one camera view or under different camera views.

VI. Experimental Results

A. *Falling Detection and Recognition*

1) Dataset

In order to evaluate the proposed method for falling event detection and recognition, we collected a dataset by using RGBD camera. The dataset contains five activities related to falling event performed by five different subjects under two different conditions: one is with sufficient illumination, and the other is under insufficient lighting. In total there are 150 video sequences including 100 videos for condition 1, 50 videos for condition 2.

In our experiments, we select 50 videos which covering all 5 subjects and 5 types of activities in condition with sufficient lighting for training. The rest 100 video sequences (50 for each condition) are used for testing.

2) *Performance Analysis of Activity Recognition*

The results of falling event detection and recognition are displayed in Figure 4. We observe that our proposed approach achieves high accuracy to recognize activities related to falling event even for environments with insufficient illuminations. Some example frames of RGB and depth images and the tracked skeleton joints under different conditions as mentioned in Section VI-A-1 are demonstrated in Figure 5.

In test phase, the recognition speed of our fall detection runs based approach is about 37 frames per second for real time applications.

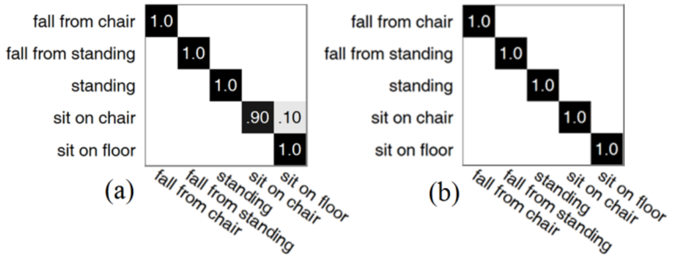


Figure 4 Performance of the proposed falling event detection and recognition. (a) Classification results of videos with "insufficient illumination". (b) Classification Results of videos with "sufficient illumination". We observe that illumination has almost no influence in the performance of our kinematic model.

B. *Finer ADL Activity Recognition*

1) *Dataset and Experiment setups*

We evaluate the proposed algorithm for finer ADL activity recognition on the Cornell 3D activity dataset [21] (some of the subjects and skeletons can be seen in Figure 9). This dataset contains 4 subjects at different poses from different view-angles under different lighting conditions. The subjects are requested to perform 13 different activities such as typing on a computer, writing on a white board, and drinking water *etc*.

To comparison the results in [21], we follow the same experiment setting, *i.e.* grouping the 12 classes of activities and the two special classes (still and random activities) into 5 sub sets by the location of the activity performed: bathroom, bedroom, kitchen, living room, and office. The two tests, "Have Seen" and "New Person" are used to compare the performance.

2) *Performance and Comparison*

We train a linear kernel SVM classifier [5] to map extracted features into corresponding labels of activities

We first investigate the effects of the value *K* for the K-means algorithm by randomly selecting 40% of all data as training set and the remaining 60% as test set. The recognition accuracies (vertical axis) over *K* values from 2^4 to 2^{10} (horizontal axis) are shown in Figure 6. Since the final performance is partially affected by the result of K-means, which may differ in each experiment. To reach a comparable stable performance, we run one experiment with the same parameter setting 10 times and report the averaged performance in Figure 6.

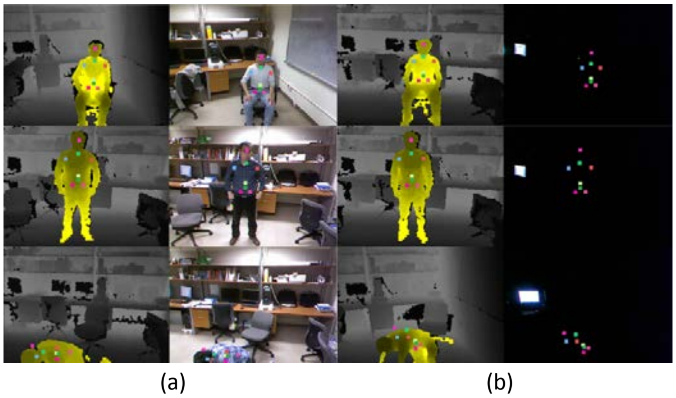


Figure 5 Example frames with tracked skeleton joints of our dataset for fall detection and recognition under different conditions. (a) With sufficient illumination (left column: depth images; right column: RGB images). (b) With insufficient illumination (left column: depth images; right column: RGB images). The depth camera is robust to different illumination changes, which can significantly benefit ADL recognition.

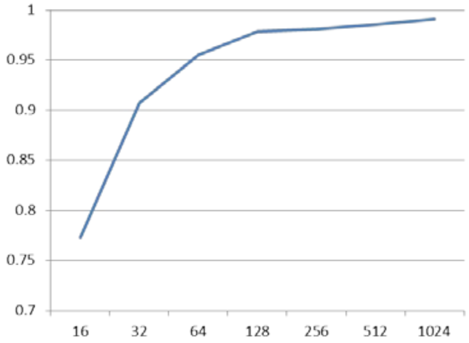


Figure 6 Performance changes of finer ADL recognition over different codebook sizes. The vertical axis is the accuracy rate while the horizontal axis indicates the codebook size. The average accuracy reaches 97.9% when codebook size is 128, which is high enough for our application.

From Figure 6, we observe that *K = 128* is apparently a proper value to be used as the default parameter since the performance increases are limited when *K* is larger than 128 but the computation cost will significantly increase, which is mainly manifested in the *K*-means phase.

The detailed recognition results of our method for finer ADL recognition are shown in Figure 7. In "Have Seen" test (the upper figure), our method correctly labels most classes of activities and achieves an average accuracy rate of 98.1%. For the activity of *"cooking (stirring),"* some of the videos are mislabeled as *"cooking (chopping)"*, because these two types of activities are very similar and very difficult to distinguish only from the skeleton features we have extracted. This observation is more obvious in "New Person" test (the lower figure of Figure 7), where the test person has not been appeared in the training phase. We observe that the performance decreases for the activities with similar actions. For example, as the activities of *"talking on the phone"*, *"drinking water,"* and *"rinsing mouth with water"* contain the common movement of *"raise something to or near the head"*. We will address the problem in future by extract more detailed features of hands. The average accuracy of our method for "New Person" test is 81.79%.

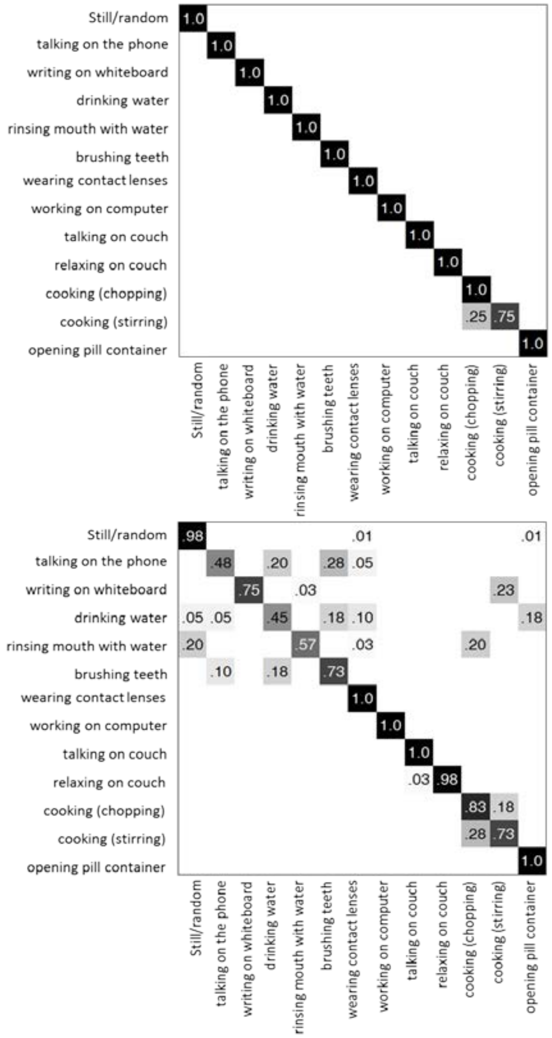


Figure 7 Detailed recognition accuracy results of our method with codebook size as 128. Upper: in "Have Seen" test, we correctly classify most activities except of two very similar action classes: "cooking (chopping)" and "cooking (stirring)". Lower: in "New Person" set, the results are not as good as in "Have Seen" especially in some very similar activities such as "talking on the phone" and "drinking water".

We further compare our method with the benchmark of this dataset [21]. Comparison in terms of precision and recall which are reported in [21] is shown in Figure 8. As shown in Figure 8, our method performs much better results than the one in [21] in almost all classes of activities for both "Have Seen" and "New Person" tests. Moreover, our method is more efficient since we only use the skeleton information while in paper [21], they used a combination of both skeleton and RGB HOG features with a much higher dimension than our features.

We also observe that both our method and [21] obtain better results in "Have Seen" test than in "New Person" test because new subjects are appeared for "New Person" test. In our application of home-assistant systems, only a certain group of people will be monitored.

In summary, since our method combines both temporal motion and spatial structure information of skeleton joints and applies powerful Bag of Features (BoF) model, our classification results (both cross-subject and non-cross-subject) outperform benchmark performance [21] with features of lower dimension.

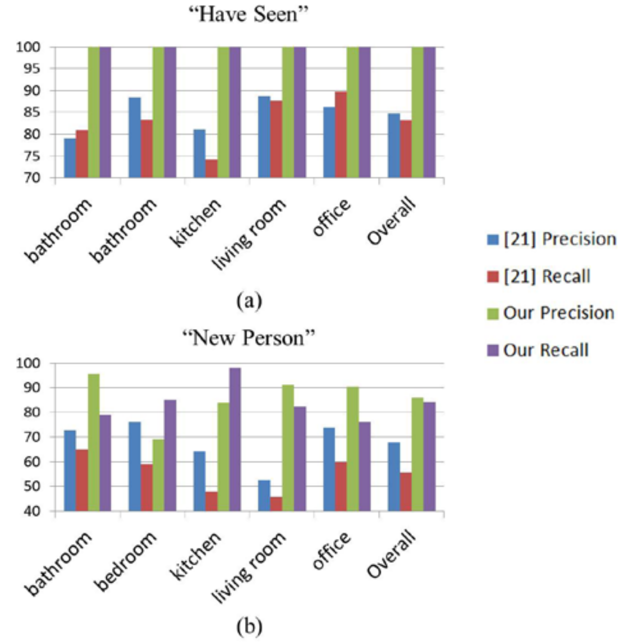


Figure 8 Comparison our method with paper [21]. (a) Comparison of experiment in "Have Seen" test, where all subjects are involved in both training and testing phases. (b) Comparison in "New Person" test, where test subject is not seen in training phase.

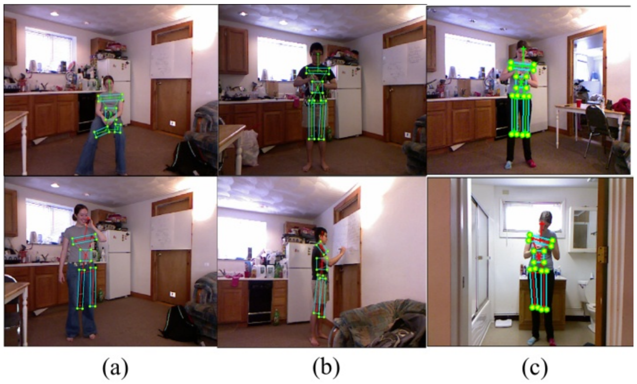


Figure 9 Examples of skeletons and extracted patches in our method of people identification. (a) Pose variation. (b) Viewpoint variation. (c) Illumination variation due to different locations.

C. *User Identification*

In this task, we also employ the dataset of [21] and some of the frames in the dataset are shown in Figure 9. The training set contains 2000 images with 1000 positive samples (*i.e.*, two images are selected from the same person) and 1000 negative samples (*i.e.*, two images are selected from different persons). In experiments, the color histogram of each patch for each channel (H and S channel) is quantized into 20 bins. The color histograms are then input to a SVM classifier for user identification.

Since we only employ the Hue and Saturation channels in HSV color space to form the color histogram representation of human body appearance, this representation is robust to illumination change. In addition, to eliminate the effects of the background pixels inside those patches, we apply the depth-adaptive weighting strategy on each pixel.

Our person identification approach achieves an accuracy rate of 99.6%. Our model by combining RGB channels and Depth channel can effectively handle people identification problem.

VII. Conclusion

In this paper, we have proposed a framework for recognizing activities of daily living to facilitate the independence of older adults living in the community, reduce risks, and enhance the quality of life at home by using RGB-D cameras. Experiments demonstrate that our framework is effective and robust to recognize activities related to falling event and finer activities of daily living. Our RGBD camera-based framework can handle lighting changes and pose variations, as well as provide a good solution for privacy protection.

References

[1] Brookmeyer, R., Gray, S. and Kawas, C.: Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. In: American journal of public health, vol. 88, pp. 1337, Am Public Health Assoc (1998)

[2] Buehler, P., Everingham, M., Huttenlocher, D.P. and Zisserman, A.: Upper Body Detection and Tracking in Extended Signing Sequences. In: International Journal of Computer Vision (IJCV), pp. 1–18, Springer (2011).

[3] Catz, A., Itzkovich, M., Agranov, E., Ring, H. and Tamir, A.: SCIM-spinal cord independence measure: a new disability scale for patients with spinal cord lesions. Spinal Cord, 35(12):850–856, 1997.

[4] Capezuti, E., Wagner, L.M., Brush, B.L., Boltz, M., Renz, S., & Secic, M. (2008). Bed and toilet heights as potential environmental risk factors. Clinical Nursing Research, 17 (1), 50-66.

[5] Chang, C.C. and Lin, C.J.: LibSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[6] Courtney, K.L, Privacy and senior willingness to adopt smart home information technology in residential care facilities. Methods in Informatics Medicine, 2008; 47(1):76-81.

[7] Courtney, K.L. Matthews, J.T., Beach, S.R., Downs, J., de Bruin, W.B., Mecca, L.P., & Schulz, R. Preference and concerns for quality of life technology among older adults and persons with disabilities: National survey results. Technology and Disability, 2010, 5-15.

[8] Courtney, K.L., Demiris, G., Rantz, M., & Skubic, M., Needing smart home technologies: the perspectives of older adults in continuing care retirement communities. Informatics in Primary Care, 2008, 16(30: 195-201.

[9] Demiris, G., Oliver, D.P., Giger, J., Skubic, M., & Rantz, M. Older adults' privacy considerations for vision based recognition methods of eldercare applications, *Technology and Health Care* 17 (2009), 41–48.

[10] Demiris, G., Hensel, B.K., Skubic, M., & Rantz, M. Senior residents' perceived need of and preferences for "smart home" sensor technologies. International Journal of Technology Assessment Health Care, 2008, 24 (1), 120-124.

[11] Felzenszwalb, P., McAllester, D. and Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In Proc: Computer Vision and Pattern Recognition (CVPR). IEEE Conference on, pp. 1–8, IEEE (2008)

[12] Hobbs, F.B.: The elderly population. In: U.S. Bureau of the Census.*http://www.census.gov/population/www/pop-profile/elderpop.html*

[13] Kopp, B., Kunkel, A., Flor, H., Platz, T., Rose, U., Mauritz, K., Gresser, K., McCulloch, K. and Taub, E.: The Arm Motor Ability Test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living. Arch. of physical medicine and rehab. 78(6), 1997.

[14] Lee, H.Y., Kim, T.J., Jung, W., Park, K.H., Kim, D.J., Bang, B. and Bien, Z.Z.: A 24-hour health monitoring system in a smart house. In: Gerontechnology, vol. 7, pp. 22–35 (2008)

[15] Li, W., Zhang, Z. and Liu, Z.: Action recognition based on a bag of 3D points. In: Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on, pp. 9–14, IEEE (2010)

[16] Microsoft Research, Microsoft® Kinect™ for Windows® Software Development Kit (SDK) Beta from Microsoft Research, Remond, WA USA (2011)

[17] Nait-Charif, H. and McKenna, S.J.: Activity summarization and fall detection in a supportive home environment. In Proc: Pattern Recognition (ICPR), International Conference on, vol. 4, pp. 323–326, IEEE (2004)

[18] Pirsiavash, H. and Ramanan, D.: Detecting activities of daily living in first-person camera views, CVPR(2012)

[19] PrimeSense Ltd, OpenNI, *www. openni.org*

[20] Quinn, K. Methodological considerations in surveys of older adults: Technology matters. International Journal of Emerging Technologies and Society, 2010, 8(2), 114-133.

[21] Sung, J., Ponce, C., Selman, B. and Saxena, A.: Human activity detection from RGBD images. In: AAAI workshop on Pattern, Activity and Intent Recognition (PAIRW) (2011)

[22] Tomita, M.R., Russ, L.S., Sridhar, R., Naughton, B.J. Chapter 8: Smart home with healthcare technologies for community-dwelling older adults (pp 139-158). In Smart Home Systems (Ed: M. A. Al-Qutayr), 2010.

[23] Wang, S., Zabir, S. and Leibe, B.: Lying Pose Recognition for Elderly Fall Detection. In: Proceedings of Robotics: Science and Systems, Los Angeles, CA, USA (2011)

[24] Wilson, D.H., Consolvo, S., Fishkin, K.P. and Philipose, M.: Current practices for in-home monitoring of elders' activities of daily living: A study of case managers. Citeseer (2005)

[25] Yang, X. and Tian, Y.: EigenJoints-based Action Recognition Using Naive-Bayes-Nearest-Neighbor, International workshop on Human Activity Understanding From 3D Data (2012)

[26] Zhang, H. and Parker, L.E.: 4-dimensional local spatio-temporal features for human activityrecognition. In: Intelligent Robots and Systems (IROS), IEEE/RSJ International Conferenceon, pp. 2044–2049, IEEE (2011).

Chenyang Zhang (S'12) received his B.s. degree in the Pilot School of Software Engineering of Tianjin University, Tianjin, China, in 2011. Since 2011, he is a Ph.D. student in Department of Electrical Engineering at the City College of New York, the City University of New York.

His research focuses on action recognition and activity analysis. His research interests include machine learning, pattern recognition and artificial intelligence.

YingLi Tian (M'99–SM'01) received her BS and MS from TianJin University, China in 1987 and 1990 and her PhD from the Chinese University of Hong Kong, Hong Kong, in 1996. After holding a faculty position at National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, she joined Carnegie Mellon University in 1998, where she was a postdoctoral fellow of the Robotics Institute. Then she worked as a research staff member in IBM T. J. Watson Research Center from 2001 to 2008. She is currently an associate professor in Department of Electrical Engineering at the City College of New York. Her current research focuses on a wide range of computer vision problems from motion detection and analysis, to human identification, facial expression analysis, and video surveillance. She is a senior member of IEEE.