RGB-D Camera-based Activity Analysis

Chenyang Zhang and Yingli Tian
Department of Electrical Engineering
The City College of New York,
New York, USA 10031
E-mail: {czhang10, ytian}@ccny.cuny.edu Tel: +1-212-6508917

***Abstract*— In this paper, we propose a new activity analysis framework to facilitate the independence of elderly adults living in the community, reduce risks, and enhance the quality of life at home by using RGB-D cameras. Our contributions include two aspects: 1) recognizing 5 activities related to falling including standing, fall from standing, fall from sitting, sit on chair, and sit on floor. The main analysis is based on the depth information due to the advantages of handling illumination changes and identity protection. If the monitored person is out of the range of 3D camera, RGB-based video analysis module is employed to continue the activity monitoring. 2) Identifying the monitored person if there are multiple people in the camera view by combining both depth and RGB information. We have collected a dataset under different lighting conditions and ranges. Experimental results demonstrate the effectiveness of the proposal framework.**

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 [6]. 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 [7]. 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 [15]. In this paper, we focus on falling detection and attempt to recognize it from other similar activities related to falling such as sit on floor, etc.

We propose an activity analysis framework to recognize five activities related to falling event including standing, fall from standing, fall from sitting, sit on chair, and sit on floor by using RGB-D camera. Compared with traditional video surveillance cameras, RGB-D cameras have advantages of handling illumination changes and privacy protection. Our activity analysis depends on both depth information and appearance information. The kinematic features extracted from 3D information consist of two parts: 1) proposed structure similarity and 2) head-floor distance, which is defined as the vertical distance between the head and the floor plane. For user identification, from 2D appearance RGB information, we employ a background subtraction and tracking method and represent actions as histogram features. Classification on two different SVM schemes are performed and analysis. Experimental results demonstrate that our proposed framework is robust and efficient in falling event detection.

We further develop a patch-based histogram matching method by combining 3D information (depth) and appearance information (RGB) to identify different people. The effectiveness is evaluated on Cornell 3D Activity Dataset [13].

II. RELATED WORK

Helping people with special needs by human activity recognition is a hot topic in computer vision. Nait-Charif *et al.* developed a computer-vision based system to recognize abnormal activity in daily life [10] 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 [10], Wang *et al.* [14] proposed to use gestures by applying a deformable body parts model [4] 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 [8, 13, 18] especially after Microsoft released its SDK for Kinect cameras [9]. Li *et al.* [8] proposed a method by using 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 [13]. More recently, Yang and Tian [16] proposed to apply PCA and NBNN techniques to very discriminative skeleton features, EigenJoints, to represent game-interactive activities and their method outperforms the benchmark in [8]. Other

work also tried to recover more details such as head-pose from RGB-D videos [11]. Two-person interactions are studied by Yun *et al.* [17]. In this paper, our goal is to effectively recognize activities related to falling event by using both 3D depth and 2D appearance information.

III. Falling Event Detection and Recognition

A. Feature Extraction and Representation

1. Kinematic Feature Extraction

Microsoft Kinect SDK [9] provides 20 joints on human body tracked for each person in each depth frame. We select 8 joints on head and torso since intuitively other joints on limbs introduce more noise than useful information to distinguish whether a person has fallen or not. The selected 8 joints, as shown in Fig. 1(a) top row, keep a certain structure when a person is standing or sitting. The structure is not affected much when a person is performing normal activities. However, the structure is no longer reliable when a person has fallen (as shown in Fig. 1(a), L_1 and L_2 in bottom tow). We employ the statistics feature, structure similarity cost, which is calculated from the 3D coordinates of the 8 joints as the first feature. The other feature is the head-floor distance which measures the distance between user's head position and the floor plane.

TABLE I: FIVE ACTIVITIES RELATED TO FALLING EVENT RECOGNIZED IN THIS PAPER

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

2. Kinematic Feature Representation

Fig. 1(a) displays the initial (the first row) and final (the second row) 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 than the other three "non-falling" events. We define that the structure similarity cost $C(\xi)$ of a skeleton structure ξ to measure the degree of deformation as the summation of angles changed between the corresponding joints of the skeleton between the initial and final poses as following:

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

where $\theta(\xi_i, \xi_j)$ and $\theta(o_i, o_j)$ denote the angles between two joints i and j on skeletons 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 Euclidean distance between two joints i and j is denoted as *dist(i, j)*.

Examples of the structure similarity costs (in logarithm) of different activities are displayed in Fig. 1(b) (left). Red ("fall from standing") and yellow ("fall from chair") curves obviously demonstrate significant costs as expected. We extract two statistics of the structure similarity cost (mean μ and the variance σ) to represent the action in a video sequence.

Another feature we use for activity recognition is head-floor distance. Given a floor plane constraint by $[A,B,C,D][x,y,z,1]^T=0$ and homogeneous representation of head 3D position $[\eta_x, \eta_y, \eta_z, 1]$, 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 of Fig. 1(b), head-floor distance is also a discriminative feature for fall related activity recognition. We employ the highest value h and the minimum value (lowest) l of head-floor distance at different skeleton poses as the last two elements in our kinematic feature vector. The kinematic feature vector from 3D depth information is denoted as $[\mu; \sigma; h; l]$.

Depth sensor is robust to handle illumination changes, however, when the user is out of the depth range, the skeleton structure from the depth information will not be available. In such situations, we will employ appearance information from RGB channels.

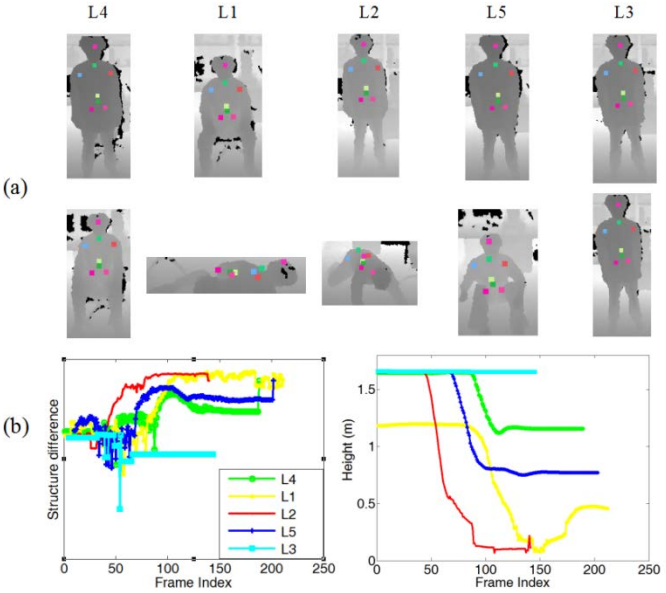

Fig. 1 Illustration of kinematic feature extraction: the structure similarity cost. (a) The initial pose (top row) and final pose (bottom row) of falling related activities to be recognized. (b) Two main elements we extracted from skeletons as features. Left: logarithm of structure similarity of each activity. Right: Height sequences in each activity. Five activities to be recognized are listed in Table 1.

3. Appearance Feature Extraction and Representation

In order to recognize falling events from appearance information, we perform a simple background subtraction method to detect moving people and then a tracking method to handle situations when people stay static for a long period (e.g., lying on the floor):

$$D_i = \|I_{i-\tau} - I_i\| \bigcap \|I_{i+\tau} - I_i\|, \qquad (3)$$

$$M_i = D_i \frac{1}{1 + e^{-(S-\lambda)}} + M_{i-1} \frac{1}{1 + e^{S-\lambda}}, \qquad (4)$$

where M_i is the foreground mask, merged by current frame difference D_i and the mask M_{i-1} of the last frame; S and λ indicate the current foreground region area and merge rate.

We represent a video sequence by a histogram of the ratio *width/height* of the bounding box of detected human.

In this paper, the action recognition is mainly accomplished by using kinematic features. Appearance indicator is employed to detect "falling" events when the user moves out of the range of the depth sensor.

B. *Activity Classification*

We employ a SVM classifier 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 (as shown in Fig. 2(a) and (b)). However, considering the semantic relationship between the five action classes, a structural SVM [5] is logically suitable for this problem, especially for our appearance indicator, whose discriminative power is high enough for the 1st layer classification (as shown as solid lines in Fig. 2(a) and (c)) ("falling" vs. "non-falling") yet limited in lower layers.

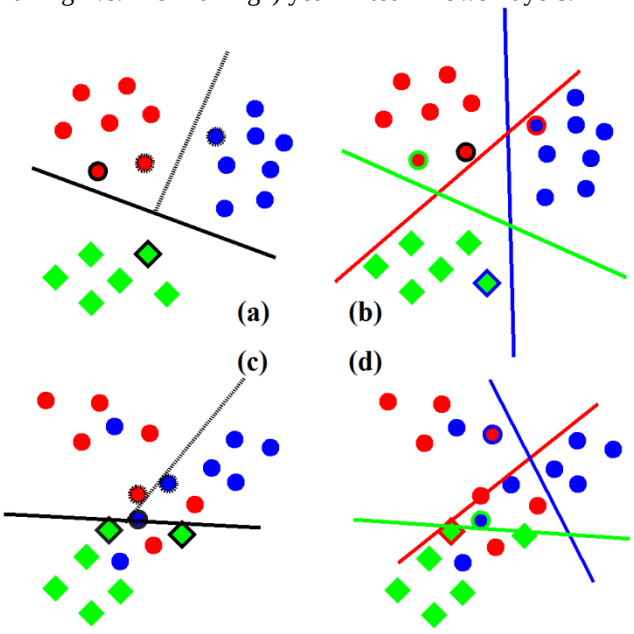

Fig. 2 (a) and (b) are of well distinguishable feature points (as kinematic features) while (c) and (d) are more clutter. Solid and dashed lines in (a) and (c) are two different layered classification boundaries, respectively and lines of different colors in (b) and (d) are ``1-*vs.*-all'' classifier boundaries. In (a) and (c), the 1st layer classification can be done based on shapes and second layer classification can be done based on colors. Since lack of such semantic information, performance of (d) is apparently worse than (c).

IV. Identify Multiple Users

If more than one user appears in the view of a camera or cross different cameras, both RGB channels and D (depth) channel will be combined to perform user identification.

Although some embedded user identification functions are available in both Microsoft SDK for Kinect [9] and PrimeSense OpenNI [12] to track a user. However, this tracking 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. The combination includes two meanings: 1) we extract 4 patches in color image according to certain skeleton joints, which are from depth channel. The 4 patches are as shown in Fig. 3, one along shoulders, one on torso and two on two upper legs. 2) A weighted strategy is applied on each pixel inside patches based on their depth value, as described in the Section below.

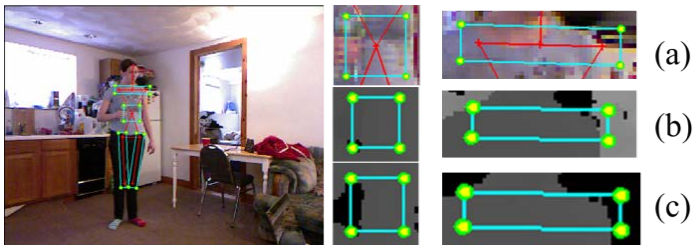

Fig. 3 Left: A sample image from Cornell Dataset with 4 patches we designed for identifying multiple users based on color information. (a) Two patches on upper body. (b) Corresponding depth channel. (c) Mask of weighting.

A. *Color-histogram-based user appearance representation*

Human detection and calibration (detected certain joints on skeleton, such as head, shoulder, torso *etc.*) for RGB-D images are provided by built-in functions in Microsoft SDK [9] and PrimeSense OpenNI libraries [12]. To identify user, as shown in Fig. 3, we extract four patches from RGB video based on skeleton joints from the depth channel: one on the shoulder, one on the torso, and two on the upper legs.

In our method, we assign the pixels with different weights according to their distance to the local joints on the Z (depth) direction. Local joints are defined as the joints inside the current patch, for example: in the patch along the shoulders (as shown in Fig. 3), the local joints are ones on the two shoulders. We denote the weight as w_i:

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

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

Sometimes the tracked joints of skeleton may locate on the background instead on the body due to fast motion or false alignment. We select the measure point m with following rule:

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

where m_k is the median depth of all 8 joints. By doing so, we can get rid of the affection on weight of false-aligned joints.

To handle illumination changes, we transform the RGB to HSV color space and only use H and S channels. We quantize

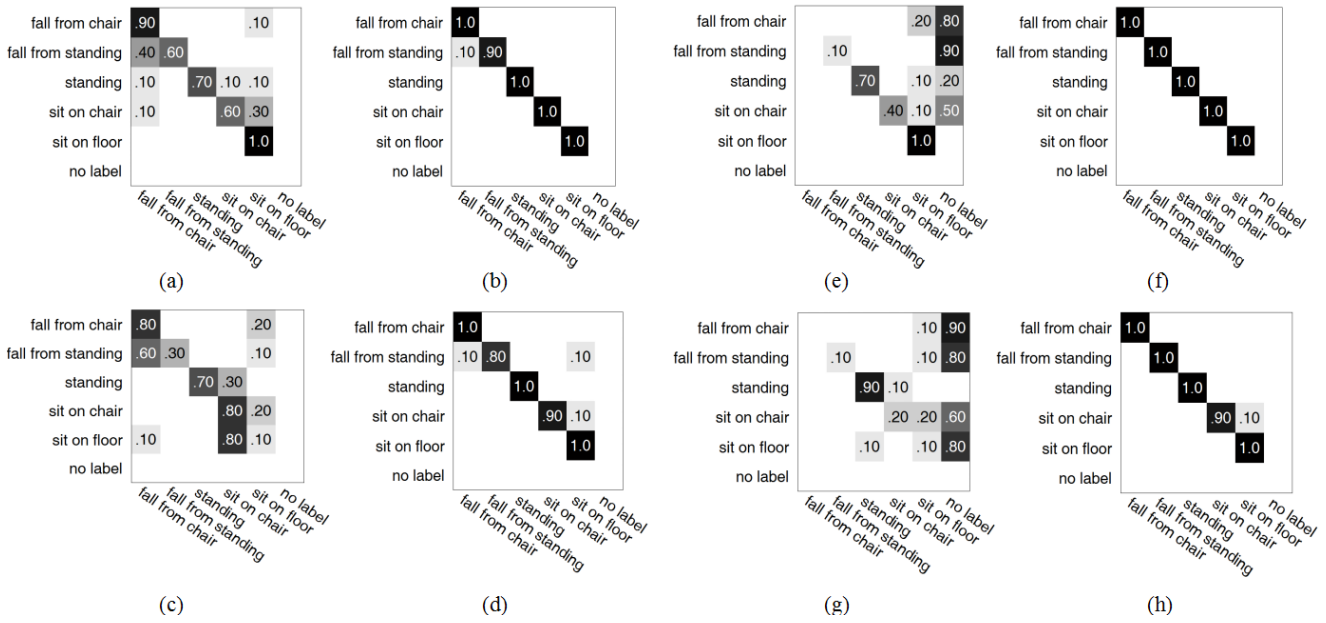

Fig. 4 Performances of the proposed method on different situations. (a)-(d) are using structure SVM classifier set. (e)-(h) are using "1-vs.-all" SVM classifier. (a)(e) Appearance model in normal case. (b)(f) Kinematic model in normal case. (c)(g) Appearance mode with sufficient illumination but out of depth sensor's range. (d)(h) Kinematic model with insufficient illumination and within depth sensor's range.

each channel in each patch into 20 bins, each pixel votes one bin with its weight w_i as calculated above. Then a normalization step is conducted.

B. Identifying user using SVM 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 the same person appears at different time under one camera view or under different camera views.

V. EXPERIMENTAL RESULTS

A. Falling Detection and Recognition

1. Experimental Setup and Dataset

We collected a dataset containing five actions performed by five different subjects under three different conditions: 1) subject is within the range of the 3D depth sensor (<4 meters distance between the subject and the camera) and with normal illumination; 2) Subject is within the range of depth sensor but without enough illumination; and 3) subject is out the range of the 3D depth sensor (>4 meters distance between the subject and the camera) and with normal illumination. In total there are 200 video sequences including 100 videos for condition 1, 50 videos for condition 2, and 50 videos for condition 3. Each video consists of one activity. Some examples of our dataset are shown in Fig. 5.

In our experiments, we select 50 videos which covering all 5 subjects and 5 actions for training. The remaining 150 sequences are used for testing. The parameters setting in appearance model are background subtraction difference threshold $\varphi = 5$, frame step $\tau = 5$, the pixel number threshold $\lambda = 0.01$, maximum acceptable value of width/height ratio m and bin size b in the histogram representation are {4, 0.5} {2, 0.5} {2, 0.1} and {2, 0.5} for each classifier in structure classifier set. For the kinematic model, there is no manually tuned parameter.

2. Performance Analysis of Activity Recognition

To evaluate the performance of both kinematic model and appearance model under different conditions, we conduct 8 combinations of conditions and classifier structures (2 models times 2 classifier structures times two situations, normal and special). The training set contains 50 video sequences with normal condition. We use this training dataset to train both structure and "1-vs.-all" classifier sets. Performances of two classifier structures as well as models of kinematic and appearance are also compared using corresponding test datasets. The activity recognition accuracies of the proposed methods are displayed in Fig. 4. As shown in Fig. 4(a) and (b), since the features we used in appearance model is not as discriminative as in kinematic model, the appearance model achieves an average accuracy at 76% while the kinematic model achieves a much higher accuracy at 98%. Therefore, appearance features are mainly proposed to distinguish coarse classes between "falling" (i.e., fall from chair, fall from standing) and related "non-falling" events (i.e., standing, sit on chair, sit on floor). For this coarse classification to distinguish "falling" from "non-falling", as shown in Fig. 4(c), the accuracy of appearance model based coarse action classes is 92% (C1), which is comparable with that of kinematic model as in Fig. 4(a), 94%. Apparently, recognition accuracy

Fig. 5 Examples of image pairs (left: depth image; right: RGB image) for different actions and extracted skeleton features employed in our dataset under three different conditions: (a) subject is within the range of the 3D depth sensor (<4 meters distance between the subject and the camera) and with normal illumination; b) Subject is within the range of depth sensor but without enough illumination; and c) subject is out the range of the 3D depth sensor (>4 meters distance between the subject and the camera) and with normal illumination.

decreases as the class layer goes finer just as our expectation. For kinematic model, as shown in Fig. 4(b) and (d), we observe that the accuracy of each classifier is relative high, which demonstrates that our proposed kinematic features are strong for each classifier. Comparing columns 1 ((a) and (c)) and 3 ((e) and (g)), the effect we mentioned in Section III.B and Fig. 2 is manifested. Kinematic (Fig. 4(b) and (d)) and "1-vs.-all" (Fig. 4(f) and (g)) structures achieve almost the same performance. The experiments demonstrate that: 1) the proposed kinematic model is robust in each activity class according to Fig. 4(b), (d), (f), and (h). 2) Structure classifier is more robust than "1-vs.-all" classifier when using appearance model according to comparison between Fig. 4(a, c) and (e, g). In feature extraction and training phase, kinematic approach is much faster than appliance approach since its dimension is quite small (only 4). In the test phase, kinematic approach takes an average *19.4ms* and the appearance approach takes an average *7.4ms* to classify a video sequence. The length of each video is between 120-220 frames.

B. People Identification

1. Dataset

We evaluate the proposed people identification algorithm using the Cornell 3D activity dataset [13]. This dataset contains 4 subjects, different poses, and different lighting conditions, performing different activities such as typing on a computer, writing on a white board and drinking water *etc*. The training set contains 2000 samples with 1000 positive samples (*i.e.*, two images are from the same person) and 1000 negative samples (*i.e.*, two images are from different persons). In experiments, we set the patch size as 4 and quantize each channel (H and S channel) into 20 bins. In each condition, parameters of RBF kernel of SVM classifier are optimized by grid search and cross validation during training phase as advised in libSVM manual [3]. In the test phase, we calculate the accuracy as well as recall and precision. Some examples of this dataset are shown in Fig. 6.

2. Performance Analysis

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

VI. CONCLUSIONS

In this paper, we have developed a framework for fall detection using RGB-D camera by combining both 3D information (depth) and appearance information (RGB). We have recognized five categories of falling related events. Our framework can identify different users. Experiments demonstrated that our framework is effective and robust to lighting changes and pose changes. Our future work will focus on recognizing more activities, including group activities and people interactions.

(a) (b) (c)

Fig. 6 Examples of the Cornell 3D activity dataset [13]. (a) Pose variation. (b) Viewpoint variation. (c) Illumination variation. Rectangles on the body illustrate the locations and scales of patches where we calculate the color histogram.

ACKNOWLEDGMENT

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

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] 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

[4] 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)

[5] Finley, T. and Joachims, T.: Training Structural SVMs when Exact Inference is Intractable, ICML (2008)

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

[7] Lee, H., Kim, Y. T., Jung, J. 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)

[8] 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)

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

[10] NaitCharif, 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)

[11] Padeleris, P., Zabulis, X., and Argyros, A. A.: Head pose estimation on depth data based on Particle Swarm Optimization. IEEE Workshop on CVPR for Human Activity Understanding from 3D Data, (2012)

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

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

[14] 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)

[15] 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)

[16] Yang, X. and Tian, Y.: EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. IEEE Workshop on CVPR for Human Activity Understanding from 3D Data, (2012)

[17] Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. IEEE Workshop on CVPR for Human Activity Understanding from 3D Data, (2012)

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