Privacy Preserving Automatic Fall Detection for Elderly Using RGBD Cameras

Chenyang Zhang[1], Yingli Tian[1], and Elizabeth Capezuti[2]

[1]Media Lab, The City University of New York (CUNY), City College
New York, NY USA
{czhang10,ytian}@ccny.cuny.edu
[2]College of Nursing, New York University, New York, NY USA
ec65@nyu.edu

Abstract. In this paper, we propose a new privacy preserving automatic fall detection method to facilitate the independence of older adults living in the community, reduce risks, and enhance the quality of life at home activities of daily living (ADLs) by using RGBD cameras. Our method can recognize 5 activities including standing, fall from standing, fall from chair, sit on chair, and sit on floor. The main analysis is based on the 3D depth information due to the advantages of handling illumination changes and identity protection. If the monitored person is out of the range of a 3D camera, RGB video is employed to continue the activity monitoring. Furthermore, we design a hierarchy classification schema to robustly recognize 5 activities. Experimental results on our database collected under conditions with normal lighting, without lighting, out of depth range demonstrate the effectiveness of the proposal method.

Keywords: Privacy Preserving, Fall Detection, Video Monitoring, Elderly, Activities of Daily Living.

1 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 [1]. 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 [2]. 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 [3]. Recent technology developments in computer vision, digital cameras, radio frequency identification, and computers make it possible to assist the independent living of older adults by developing safety awareness technologies to analyze the activities of elders 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 [4]. In this paper, we focus on falling detection and recognize it from other similar activities such as sit on floor, *etc.*.

In this paper, we develop a privacy preserving activity analysis framework to recognize five activities related to falling event. Instead of using traditional video surveillance cameras, we utilize Kinect RGBD cameras, which are more easily accepted by older adults and their friends since it is designed for entertainment purposes. Analysis based on depth information has advantages of handling illumination changes and identity protection. If the monitored person is out of the range of 3D camera, RGB video is employed to continue the monitoring. From 3D depth information, kinematic model based features are extracted which consist of two parts: 1) structure similarity and 2) vertical height of the monitored people. In 2D RGB model, we integrate background subtraction and human tracking for activity monitoring and represent actions by quantized histograms of width-height ratios of the monitoring regions. To fulfill the need of privacy protection, only the foreground masks are used for visualization. In the classification phase, we design a hierarchy classification schema to robustly recognize the category of the activities. A comparison with traditional ”1-*vs.*-all” classifier structure is performed both theoretically and experimentally. Experimental results demonstrate that our proposed framework is robust and efficient for fall detection.

2 Related Work

Video-based human activity recognition is a hot research area in computer vision to help people with special needs. Nait-Charif *et al.* developed a computer-vision based system to recognize abnormal activity in daily life [5] in a supportive home environment. The system tracked activity of subjects and summarized frequent active regions to learn a model of normal activity. It detected falling events as abnormal activity, which is very important in patient monitoring systems. Unlike using location clues in [5], Wang *et al.* [6] proposed to use gestures by applying a deformable body parts model [7] to detect lying people in a single image, which is a strong cue of falling event.

Different from traditional RGB channel, recognizing activities using depth images has been demonstrated more straightforward and effective in recent years ([9], [10], and [11]), especially after Microsoft released the software development kit (SDK) for Kinect cameras [12]. RGBD images, in fact, are more similar to the visual perception mechanism of human beings since human has two eyes, which enable the depth information. Li *et al.* [10] proposed to use bag of 3D points to represent and recognize human actions which enables 3D silhouette matching. Hidden Markov Model (HMM) is employed with depth images to effectively recognize human activities [11]. In this paper, our goal is to effectively recognize activities related to falling from both 3D depth and 2D appearance information while preserving privacy of subjects.

3 Method of Automatic Fall Detection

3.1 Kinematic Model Based Feature Extraction from Depth Channel

Selecting joints from major body parts. In Microsoft Kinect SDK [12], there are 20 body joints tracked for each person in each depth frame. Among them, we choose 8 joints on head and torso since other joints (on limbs) introduce more noise than useful information to distinguish whether a person is falling. As shown in Fig. 1 (a), the 8 joints keep a stable structure no matter a person is standing or sitting. However, the coordinates of these joints are no longer reliable when a person falls. Based on this observation, we compute the difference cost $C(\xi)$ of given joints ξ as the first part of our kinematic features.

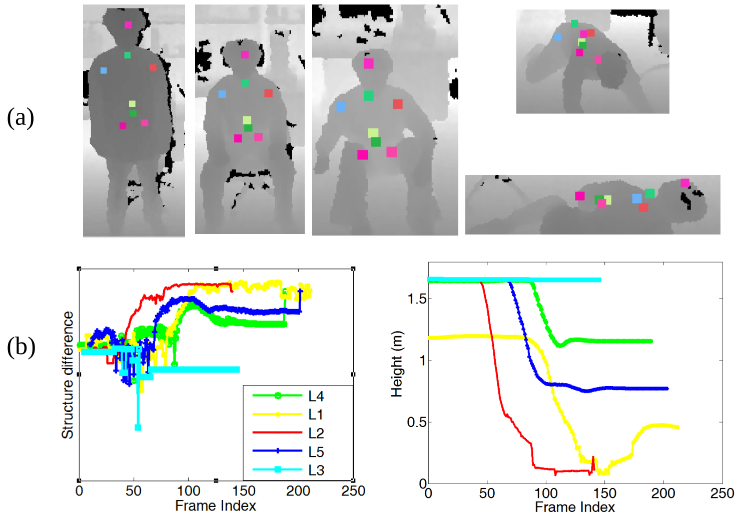

Fig. 1: (a) Structures of selected joints are stable for standing and sitting poses, but unstable for falling. (b) **Left:** logarithm of structure difference in of each event. 2. **Right:** Height sequences in each event. Five activities to be recognized are shown in Table 1.

Table 1: Five activities to be recognized in this work.

L_1	fall from chair	L_2	fall from standing	L_3	standing
L_4	sit on chair	L_5	sit on floor		

Computation of kinematic features: As shown in Fig. 1 (a), "falling" poses cause much larger deformation on the joint structures than other "non-fall" poses. We define the structure difference cost $C(\xi)$ as following:

$$C(\xi) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \|\theta(\xi_i, \xi_j) - \theta(\emptyset_i, \emptyset_j)\|, \tag{1}$$

$$\theta(i, j) = \arcsin(\frac{i_x - j_x}{dist(i, j)})/2\pi, \tag{2}$$

where $\theta(\xi_i, \xi_j)$ and $\theta(\emptyset_i, \emptyset_j)$ denote the angles between two joints i and j on two skeletons ξ and $\emptyset$, respectively, the geometry distance between two joints i and j is denoted as $dist(i, j)$.

The structure difference costs (in logarithm) of different activities in video sequences are displayed in Fig. 1 (b) (left graph). Red ("fall from standing") and yellow ("fall from chair") curves indicate significant costs as expected. We calculate the mean μ and variance σ of each activity as the first two feature elements.

Another feature for activity recognition is person height. As shown in the right graph of Fig. 1 (b). We take the highest value h and the minimum value l among person heights in all frame as the last two elements in our feature vector. Finally our kinematic feature vector is denoted as $[\mu, \sigma, h, l]^T$.

3.2 Appearance Model Based Feature Extraction

The depth range of a RGBD Kinect camera is less than 4 meters. When people are out of the range of depth sensor, we employ RGB video to continue human tracking and propose a histogram based feature representation based on background subtraction.

Person tracking by background subtraction: Our background subtraction includes two steps (as shown in Eq. 3 and 4): frame difference and tracking. Frame difference is to obtain changed area where a falling event most likely happens, which is given as follows:

$$D_i := \|I_{i-\tau} - I_i\| \bigcap \|I_{i+\tau} - I_i\|, \tag{3}$$

where D_i denotes the difference mask of the i^{th} frame. I_i denotes the intensity of current frame. $I_{i-\tau}$ and $I_{i+\tau}$ denote predecessor and successor with step τ of I_i. After morphology processing such as median filtering and connected component, a roughly foreground region is obtained as shown in Fig. 2 (a). Since static gestures may result in failure in foreground detection such as lying on the floor, we apply a simple merging strategy to merge current mask and former mask, which is formulated as following:

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

where S denotes the number of pixels in current foreground mask and λ denotes a parameter we set as a threshold to decide whether to update mask or keep former mask. M_i denotes foreground mask of the i^{th} frame and D_i is the same as in Eq. 3. A sample result of our merging strategy is shown in Fig. 2 (b).

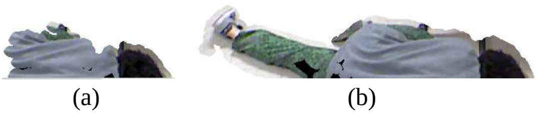

Fig. 2: (a) Fragmentary mask due to static legs. (b) Integral mask after merging.

Histogram represented features: We observe that the ratio between the width and height of a foreground bounding box can effectively indicator a falling activity. We represent an activity by a histogram of the width-height ratios during a video sequence.

3.3 Hierarchy classification

Let $\mathbf{L} \equiv \{L_1, L_2, ..., L_k\}$ be the labels of a set of activity categories. If each pair of categories requires one classifier to distinguish as in "1-*vs.*-1" manner, there will be totally $\binom{k}{2}$ classifiers. However, if prior knowledge is available or a clustering process is

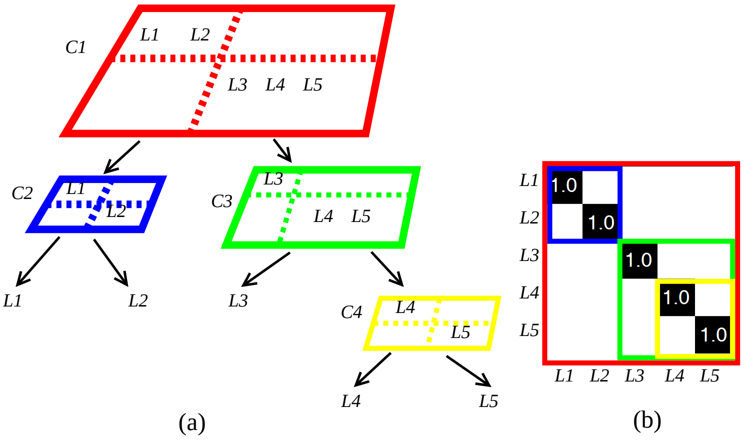

Fig. 3: (a) Structure of our hierarchy classifier set and scope of each classifier. C_1, C_2, C_3 and C_4 denote four classifiers we trained in our model. (b) The mapping relationship from our hierarchy structure to a confusion matrix to clarify the definition of "scope". Labels' meanings are shown in Table 1.

taken beforehand, a hierarchy binary classifier set can only consist of k-1 classifiers instead of $\binom{k}{2}$ classifiers. Compared with "1-*vs.*-all" manner, which requires k classifiers, hierarchy SVM classifier is more efficient when the number of categories is large. Our experimental result shows that a well-defined hierarchy classifier structure (as shown in Fig. 3) can match "1-*vs.*-all" when features are distinguishable enough for all five labels (kinematic feature) while outperforming it when features are distinguishable between "falling" and "non-falling" labels but not so distinguishable among finer labels (appearance feature).

4 Experimental Results

4.1 Dataset

We collect a dataset containing five types of activities performed by five different subjects under three different conditions: 1) subject is ***within*** the range of 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 of the range of depth sensor (> 4 meters distance between the subject and the camera) and ***with*** normal illumination. There are total of 200 video sequences, including 100 videos for condition 1, 50 videos for condition 2, and 50 videos for condition 3. Each video contains one activity. In our experiments, we select 50 videos which include all subjects and all types of activities for training. The remaining 150 sequences are used for testing. Some of the data are shown in Fig. 6.

Parameter setting: In our experiments, parameters in appearance model include background subtraction difference threshold φ, frame step τ, the pixel number threshold λ, maximum acceptable value of width/height ratio m and bin size b in the histogram representation. These parameters remain same throughout all our experiments: φ = 5, τ

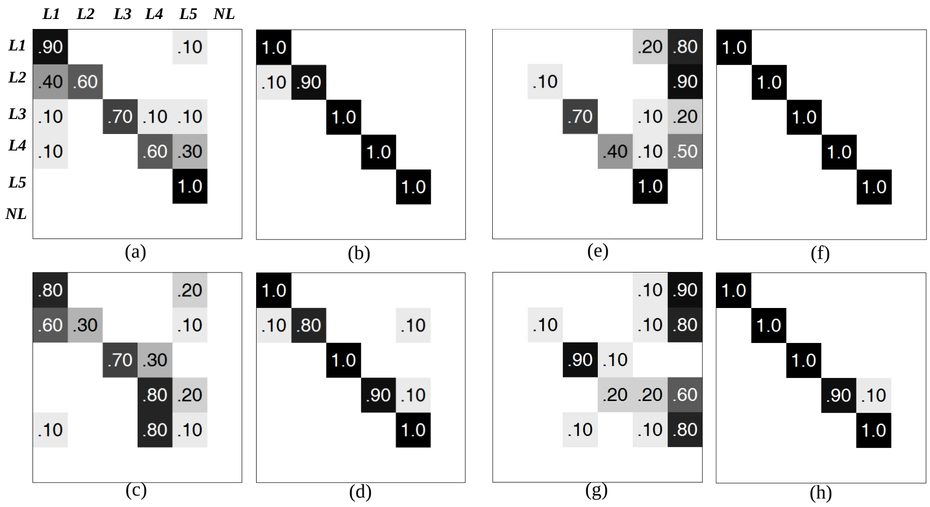

Fig. 4: Performances of proposed methods under different situations. (a)-(d) confusion matrices of activity recognition using hierarchy SVM classifier set. (e)-(h) confusion matrices of activity recognition using "1-*vs.*-all" SVM classifier. (a) and (e) Appearance model in normal case. (b) and (f) Kinematic model in normal case. (c) and (g) Appearance model with sufficient illumination but out of depth range. (d) and (h) Kinematic model with insufficient illumination and within depth range. Meaning of each label is shown in Table 1. ***NL*** means "no label".

= 5 and $\lambda = 0.05$ if a person is in the depth-range and 0.005 for out-range cases; whereas the maximum accepted values and bin widths {m, b} in histogram representation for different classifier layers are {4, 0.5}, {2, 0.5}, {2, 0.1} and {2, 0.5} for C_1, C_2, C_3 and C_4, respectively. For kinematic model, there is no manually tuned parameter.

4.2 Performance Analysis

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 videos with normal condition which is used to train both hierarchy and "1-*vs.*-all" classifiers. 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), appearance features are effective to distinguish activities in a coarse scale between "falling" {fall from chair, fall from standing} and "non-falling" {standing, sit on chair, sit on floor}, and achieve an accuracy rate of 94%. However for activity classification in a finer scale, the appearance model achieves an average accuracy rate at 76% while the kinematic model achieves a much higher accuracy rate of 98% as expected.

As shown in Fig. 4 (c), the accuracy of appearance model based coarse action classes is 92% (C_1), which is comparable to that in Fig . 4 (a). Apparently, recognition accuracy decreases for activity classification in a finer scale, as expected. For kinematic model, as shown in Fig. 4 (b) and (d), we observe that the accuracy of each classifier is high, which demonstrates that our proposed kinematic features are strong for each classifier.

Comparing columns 1 (Fig. 4 (a, c)) and 3 (Fig. 4 (e, g)), the merit of using a hierarchy SVM construction instead of using a “1-*vs.*-all” SVM construction is manifested. Due to the unbalancedness of classifier structure, it tends to classify a test data into negative group when the input features are not strong enough. And when feature is strong enough, kinematic ((b) and (d)) and “1-*vs.*-all” ((f) and (g)) structures reach 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) Hierarchy based 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 phase, kinematic approach is much faster than appearance approach. In test phase, kinematic approach takes a little longer (0.0194s) than appearance approach (0.0074s) to answer a query video with 120-220 frames.

Privacy Protection. One of the applications of the proposed models is to monitor a nursing room or a home of elder people. The benefits of the proposed models are twofold. Firstly, it can handle special cases such as when light is turned off (insufficient illumination) or people walk far from the camera (out of depth range) but is still in the view of RGB cameras. Secondly, our models are privacy preserving by only displaying 3D depth information or foreground mask as illustrated in Fig. 5 (a).

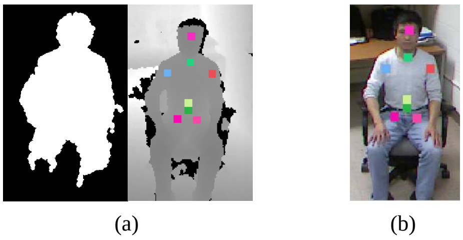

Fig. 5: Privacy preserving without reveal person identification. (a) Foreground mask (left) and depth image (right) used in our proposed models for displaying activities, (b) RGB image which reveals personal privacy.

5 Conclusion

In this paper, we have proposed an effective activity recognition framework based on RGBD cameras. Experiment results have demonstrated that our feature extraction and representation of human activity are robust and efficient. Our future work will focus on integrating RGB and depth information and recognizing more activities, including taking medicines, group activities, and human interactions.

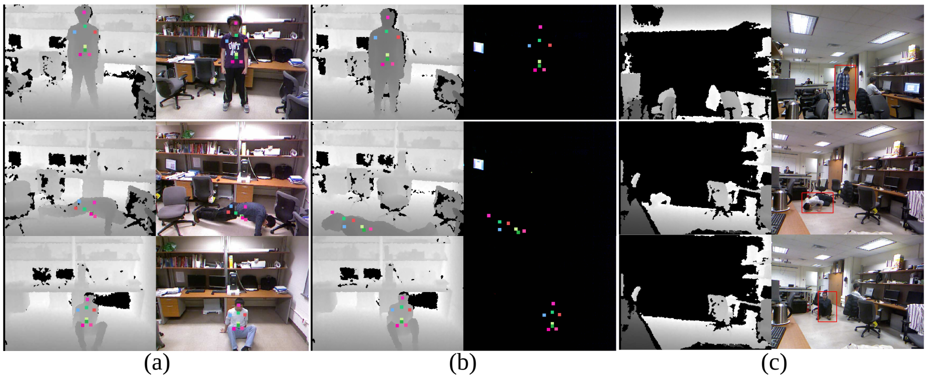

Fig. 6: Examples of actions and extracted features in our dataset. In each image pair, depth image is shown on the left and RGB image is shown on the right. (a) Sufficient illumination within the range of depth sensor. (b) Insufficient illumination within depth range. (c) Sufficient illumination but out of the range of depth sensor.

References

1. Hobbs, F.B.:The elderly population. In: U.S. Bureau of the Census. `http://www.census.gov/population/www/pop-profile/elderpop.html`
2. 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)
3. 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)
4. 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)
5. Nait-Charif, H. and McKenna, S.J.: Activity summarisation and fall detection in a supportive home environment. In Proc: Pattern Recognition (ICPR), International Conference on, vol. 4, pp. 323–326, IEEE (2004)
6. 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)
7. 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)
8. 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)
9. 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)
10. 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)
11. 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)
12. Microsoft Research: Windows Kinect SDK Beta from Microsoft Research, Redmond WA.