2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality

Thanh-Hai Tran*, Thi-Lan Le*, Dinh-Tan Pham*†, Van-Nam Hoang*,
Van-Minh Khong*, Quoc-Toan Tran*, Thai-Son Nguyen‡, Cuong Pham‡
*International Research Institute MICA, Hanoi University of Science and Technology
†Faculty of Information Technology, Hanoi University of Mining and Geology, Vietnam
‡Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Vietnam

***Abstract*—Over the last decade, a large number of methods have been proposed for human fall detection. Most existing methods were evaluated based on trimmed datasets. More importantly, these datasets lack variety of falls, subjects, views and modalities. This paper makes two contributions in the topic of automatic human fall detection. Firstly, to address the above issues, we introduce a large continuous multimodal multivew dataset of human fall, namely CMDFALL. Our CMDFALL dataset was built by capturing activities from 50 subjects, with seven overlapped Kinect sensors and two wearable accelerometers. Each subject performs 20 activities including 8 falls of different styles and 12 daily activities. All multi-modal multi-view data (RGB, depth, skeleton, acceleration) are time-synchronized and annotated for evaluating performance of recognition algorithms of human activities or human fall in indoor environment. Secondly, based on the multimodal property of the dataset, we investigate the role of each modality to get the best results in the context of human activity recognition. To this end, we adopt existing baseline techniques which have been shown to be very efficient for each data modality such as C3D convnet on RGB; DMM-KDES on depth; Res-TCN on skeleton and 2D convnet on acceleration data. We analyze to show which modalities and their combination give the best performance.**

I. Introduction

Falls happen frequently to patients and elderly people who stay at home alone. As a result, the demand for developing intelligent monitoring systems being able to detect falls has increased significantly in the health-care community. In the last few years, a large number of methods have been proposed. However there still remain some issues that strongly impact the performance of fall detection. Firstly, most of existing methods work on single modality of data observed by either ambient or wearable sensor such as camera or accelerometer. It still lacks of works which answer to the questions: Which modality gives the best performance ? Could the combination of them improve detection results ? Secondly, to answer to these questions, it needs to use a multi-modal dataset of fall which should be large in term of views, subjects and fall styles. This paper makes two main contributions in the topic of automatic fall detection. First, we introduce a large continuous multi-modal multi-view benchmark dataset for evaluating the automatic fall detection algorithms. Second, we investigate the role of modalities and their combination to obtain the best performance.

Related to human fall datasets, there are many fall datasets of single or multi-modality, captured by one or several sensors at different views [1]. However, these datasets are limited to a particular modality, fall styles, views, subjects and data continuity. To address these issues, we design and collect a new continuous and multi-modal multi-view fall dataset targeted at real-world applications. The new dataset is constructed using seven Kinects sensors and two accelerometers which are setup at different places and different views in a simulated home environment. Totally, this dataset contains 1000 samples of activities (400 fall and 600 daily activities) acquired from 50 subjects, each sample has four modalities (RGB, Depth, Skeleton, accelerometer). RGB, Depth and Skeleton have been collected from seven different view angles, acceleration data have been collected from two positions on subject body. All data are continuous, synchronized and annotated for research purpose. Falls are conducted in different orientations (backward, forward, left, right) and styles (fall when subject is lying on the bed, sitting on the chair or walking).

Our proposed dataset enables analysis on the role of modality and view. In this paper, we focus on investigating the role of modality on human fall detection. The analysis on the role of multiple views will be further work in the future. Single modality action recognition has been extensively explored [2], [3], [4]. It is obvious that each modality has different strength and limitations then analysis of different modality will help us to understand the complementary properties for a better performance of human activity recognition. In this work, we first select data from one view among seven available views for analyzing. For this view, we adopt the state-of-the-art techniques to analyze each modality. Specifically, 3D convolutional network is used on RGB data [5]; depth motion map (DMM) with improved kernel descriptor (KDES) is applied on depth data [6]; Res-TCN for skeleton data [7] and 2D convnet for acceleration data [8]. As multimodal data are time-synchronized, the detection results of all modalities are compared. Then different combinations of modalities are studied and the best is reported.

II. Existing fall datasets and fall detection methods

A. Existing fall datasets

Fall can be considered as an human activity and some types of fall activities were included in other human activity datasets [16]. However, in those works, the authors did not

978-1-5386-3787-6/18/$31.00 ©2018 IEEE

TABLE I
COMPARISON BETWEEN CMDFALL DATASET WITH SOME OF THE OTHER PUBLICLY AVAILABLE DATASETS FOR FALL DETECTION

Dataset	#Falls	#ADL	#FallStyles	#Subjects	#Views	Modalities	Continuous	Year
UR [9]	22	24	na	1	8	RGB	No	2010
Le2i [10]	192	58	3	~ 8	1	RGB	No	2012
SDUFall [11]	300	1500	na	10	1	RGB + D + Skeleton	No	2014
OCCU [12]	30	80	2	5	2	RGB + D + Skeleton	No	2014
Cogent Lab [13]	448	1520	6	42	na	Acc. + Gyroscope	Yes	2015
EDF	160	100	8	10	2	RGB + D + Skeleton	No	2017
UF[14]	na	229-na	2	6	1	RGB + D + Skeleton	No	2017
SisFall[15]	1798	2707	15	38	na	Acc. + Gyroscope	Yes	2017
CMDFALL - Our dataset	**400**	600	**8**	**50**	7	RGB + D + Skeleton + **2 Acc.**	**Yes**	2017

pay attention on designing the fall activity in order to evaluate the performance of fall detection. This paper focuses on fall detection so we will present only fall focused datasets in this section. Fall can be captured by using either wearable sensors (e.g. accelerometer) or ambient sensors (e.g. camera) or combining both of them. Collecting a fall dataset from wearable sensors is view and location independent while that from ambient sensors such as cameras is view and location dependent. Table I summarizes some fall focused datasets that have been recently published in the literature. The limitations of those datasets are described as follows.

- **Limited modalities**: Most of datasets have single modality [9], [10] or captured by either wearable (accelerometer) or ambient (RGB, D, Skeleton) sensors [11], [12], [14]. There are no fall dataset capturing RGB, D, skeleton and accelerometer data at the same time.
- **Limited fall styles**: In most of datasets, subjects perform fall by standing at one position then falling down the floor at the center of scene. There are no fall from bed and only one dataset has falls from chair [10]. However, according to [1], falls from beds accounting for a high rate after only falls from tripping, slipping or stumbling.
- **Trimmed videos**: All fall videos have been trimmed that are very suitable for fall classification from different activities. However, those data not allow evaluating continuous fall detection.
- **Limited number of views**: The number of views observing the scene is usually limited from 1 to 2 views (OCCU [12], EDF[2]). In [9], the number of views are eight but only RGB data were captured. The view could be overlapped or non-overlapped then difficult for applying truly a multi-view approach.
- **Limited number of subjects**: Most of multi-modal datasets have been captured by a small number of subjects (ranging from 1 to 10). There are 300 fall and 1500 non-fall samples in [11]. However, that dataset has poor variation of fall styles by different subjects.

This motivated us to design and build a new dataset that addresses such issues to provide researchers a testbed to develop and evaluate their new algorithms.

B. Methods of fall detection

Fall detection using unimodal features: The first and the most widely used features are extracted from color images of RGB camera since RGB camera is inexpensive and easy to install. Color-based features can be computed from one sole camera or multiple cameras. Most color-based approaches based on shape features extracted from human region candidates [4]. Besides color images, RGB-D sensors provide depth and skeleton information which is independent of lighting condition. Therefore background subtraction becomes easier and more reliable. In [17], the authors proposed a method for fall detection from depth images. The skeleton-based works do not need to perform person detection because the skeleton is available whenever the person is detected. Then there are many works for fall detection based on skeleton using simple rules or machine learning techniques [18], [19]. Related fall detection using wearable sensors, a large number of methods have been proposed in [20].

Fall detection using multimodal features: As each unimodal feature has its own advantages and disadvantages, some works try to combine/fuse more than one modality for fall detection. In [21], Mastorakis et al. exploited color and depth information to build 3D (height, width, depth) bounding box of the subject. Kwolek et al. [22] combined depth information from Kinect with accelerometer mounted on the human for fall detection. In our previous work [23], an efficient method that combines RGB and skeleton for fall detection has been proposed. To the best of our knowledge, there is no work that investigates the role of each modality and combines modalities to improve fall detection performance. Although several multimodal fall focused datasets have been constructed as described previously, it lacks methods that combines different modalities. In [24], the authors have also addressed similar questions to evaluate the importance of different modalities. However, that work has been done for general human activities, not for fall activity in particular.

III. THE PROPOSED DATASET: CMDFALL

A. Environmental and equipment setup

In this section we describe the main components of our acquisition system for collecting the multi-modal multi-view dataset of human fall. The system consists of seven Microsoft Kinect v1 and two WAX3 wireless accelerometers. Figure 1 shows the layout of our acquisition system. Six Kinects are installed at height of 1.8m surrounding a space of 3.6mx6.8m to simulate home environment. The 7^{th} Kinect is installed on the ceiling at height of 3m to observe the top-view of the scene. Two accelerometers are mounted on the subject's body, one on the left wrist and one on the left hip of the

[1]Fall Injuries among Older Adults in Oregon, 2008
[2]https://sites.google.com/site/kinectfalldetection/

TABLE II
LIST OF ACTIVITIES AND CATEGORIZATION

S_1	S_2: 6 groups	ID	S_3: 20 activities
Fall	Fall while walking	1	Front fall
		2	Back fall
		3	Left fall
		4	Right fall
	Fall while lying on the bed	5	Lie on bed then fall left
		6	Lie on bed then fall right
	Fall while sitting on the chair	7	Sit on chair then fall left
		8	Sit on chair then fall right
Non Fall	Horizontal movement of the whole body	9	Walk
		10	Run slowly
		11	Stagger
		12	Crawl
		13	Move chair
	Hands and legs movement	14	Move hand and leg
		15	Left hand pick up
		16	Right hand pick up
	Vertical movement of the whole body	17	Jump in place
		18	Sit on chair then stand up
		19	Sit on bed then stand up
		20	Lie on bed then sit up

subject. With this setup, every location in the space could be observed by all Kinects sensors. Microsoft Kinects collect three major modalities (RGB frames, Depth maps, 3D joints of Skeleton) at resolution of 640x480, 20fps. Skeleton consists of 3-dimensional locations of 20 major body joints for detected and tracked human bodies in the scene. Accelerations are collected at rate of 50 samples per second.

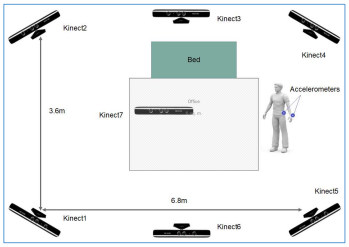

Fig. 1. Illustration of environment and material setup

B. List of activities

Twenty activities will be recorded among which eight are falls of different styles and twelve are daily and fall like activities. For the task of fall detection, we consider two main classes: fall and non-fall. In the group of falls, we have three types of fall, each has different styles and directions. In the group of non-fall activities, we categorize in term of movement of the whole body or only some body parts and direction of movement. Table II lists the twenty activities (S_3), six groups (S_2) and two main classes (S_1).

C. Subjects

To collect data, we invite 50 persons (30 males and 20 females in the range 21-40 years of age). The subjects wear their daily clothes without markers. Before a recording section, the subject is explained about activities to perform. The order of activities could be different from person to person but the transition between activities are smooth. That means the person can not make a fall then immediately run fast to sit on the chair. He/she will recover first, then stand up and walk slow to sit on the bed etc.

D. Data acquisition and annotation

Each person performed all 20 activities in about 7.5 minutes, yielding 375 minutes recording time in total. There are 1000 samples including 400 falls and 600 normal activities. Each sample has multi-modal data: RGB, Depth, Skeleton at 7 views and two acceleration values. Fig.2 shows a snapshot of synchronized multimodal data from 7 Kinects and two accelerometers. Totally, the size of dataset is around 350Giga bytes. All data are time synchronized. Starting and ending time of each action in all sequences are annotated for human activities classification and fall detection evaluation. This dataset is challenging due to the large number of subject styles and viewpoints. In addition, some activities are very confused that challenges the recognition. Specifically, many daily activities are similar for example fall-likes activities *hand left pick up and hand right pick up, sitting on the chair then stand up, sit on the bed and stand up*. The dataset is publicly available for research purpose at *http://www.mica.edu.vn/perso/Tran-Thi-Thanh-Hai/CMDFALL.html*.

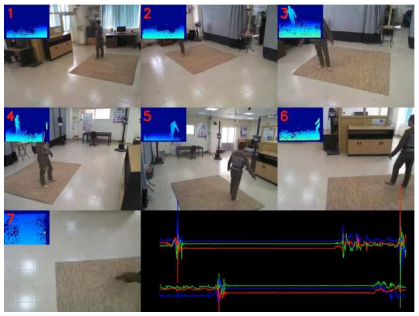

Fig. 2. Illustration of multi-modal data taken from 7 views and 2 accelerometers

E. Evaluation protocol

In a multi-modal multi-view system, one can evaluate a method using cross view or cross subject validation. In the context of modality evaluation, we evaluate human activity recognition and human fall detection in particular using cross subject validation on one view. We split the 50 subjects with ID from 1 to 50 into three sets. One set containing 25 subjects with odd IDs is used for training. One set containing 5 subjects randomly taken from 25 remaining subjects with even ID is used for validation. The remaining data containing 20 subjects is used for testing. According to the grouping of data (as shown in the Tab.II), three evaluations will be carried out: i) classification of 20 activities; ii) classification of 6 groups of activities; iii) classification of fall and non-fall.

IV. MODALITY ANALYSIS FOR ACTIVITY CLASSIFICATION AND FALL DETECTION

A. Baseline methods for activity classification and fall detection

One of objectives of this work is to evaluate the use of various modalities in the dataset. Therefore, we will take data from one view (the 3^{rd} view in this work) to investigate

the role of modality. In this section, we summarize methods that we use as baselines for evaluation of single modality. Continuous spotting and recognition from video stream is out of focus of this work.

1) C3D: 3D Convolutional Neural Network for RGB modality: In this work, we utilize the technique C3D (3D convolutional neural network) [5] due to its simplicity and efficiency compared to 2D ConvNets. The main idea of C3D is to use 3D convolutional operators to capture both temporal and spatial features of an activity. In our experiment, we use the same architecture of C3D network as described in [5]. Specifically, the network composes of 8 convolutions, 5 max-pooling and 2 fully connected layers. The number of filters for 5 convolutional layers from 1 to 5 are 64, 128, 256, 512, 512. All 3D convolution kernels are 3x3x3 with stride 1x1x1. We used the pre-trained network with I380K and fine tune on Sport1M dataset then finetune on our dataset with 4000 iterations. Each video of an activity will be divided into non-overlapped clips of 16 frames. These clips will pass through the C3D convnet and a feature vector will be extracted at FC6 layer. Finally, a video of an activity is represented by a feature vector which is average of all feature vectors extracted from clips of 16 frames. These feature vectors will be served for training and testing a multi-class SVM.

2) DMM-KDES: Depth Motion Map and Kernel Descriptor for Depth modality: In our previous works [6], [23], we have shown that motion map with improved kernel descriptor is a good combination for action representation. In this paper, we would like to investigate the role of depth map on a more challenging dataset. The main idea of the proposed methods is following. Firstly, the video sequence of an activity is represented by a motion map as follow. Given a sequence of N depth maps $D_1, D_2, ..., D_N$, the depth motion map is defined by:

$$DMM = \sum_{i=1}^{N-1} (|D_{i+1} - D_i| > \epsilon) \quad (1)$$

where ϵ is a threshold to make binary the difference between two consecutive maps D_{i+1} and D_i. The binary map of motion energy indicates motion regions or where movement happens in each temporal interval. So the DMM represents sum of motions through entire video sequence. Then gradient based kernel descriptor [6] is computed on the DMM. After that, we utilize the bag of word technique to quantize kernel descriptors into 1000 codewords and perform the classification by a simple MPL neural network. The network contains two hidden layers with 100 neurons for each layers. We use neuron with ReLU activation function, train the network over 200 iterations with log-loss function using stochastic gradient descent.

3) Res-TCN: a convnet for skeleton data: To evaluate the activity recognition using skeleton data, we employ Temporal Convolutional Neural Networks with Residual units (Res-TCN) presented in [7]. This network provides spatial-temporal representation of sequential skeleton frame and achieves state-of-the-art results on NTU-RGBD - the largest dataset on 3D human activity. The input of network is sequence of skeleton frame where in each frame, all the x, y, z coordinates of each joint will be concatenated into 150-D vector (2 subjects x 25 joints x3). The network is a stack of convolutional layers with residual unit. Each layer consists of temporal convolutions, a non-linear activation function, and max pooling across time. We train Res-TCN networks from scratch with our skeleton dataset over 200 epochs, batch size = 32, learning rate = 0.01, using SGD optimizer with momentum = 0.99.

4) 2D convolutional neural network for acceleration modality: Inherited from our previous work [8], the acceleration signals are segmented into 2 second-sliding windows with 50% overlapped between two consecutive windows. Each sliding window roughly contains 100 samples as our WAX sensors operate at the sampling frequency of 50 Hz. With two accelerometers worn on right-hand and hip, each accelerometer has 3 X,Y,Z signals and therefore we have 6 signals streaming in total. In order to analyze accelerometer data, the CNN model used in this experiment is a stack of convolutional, max-pooling, fully-connected and softmax layers. The input of the CNN model are 6 channels for X-axis, Y-axis, and Z-axis acceleration signals from two accelerometers. Each channel accepts 1-D arrays consisting of frames of 100 samples as inputs. We use 60 filters for convolutional layer 1 and 128 filters for the next convolutional layer. The dimension of the fully-connected layer is set to 1000. At the training phase, mini-batches of 128 frames are used and the negative log likelihood is minimized using stochastic gradient descent optimizer provided in TensorFlow. The fully-connected layer directly connects to the softmax layer to produce the output probabilities of the activity classes.

B. Late fusion of modalities

To understand the complementary properties of modalities, we evaluate the performance of all possible combinations of modalities using a simple late fusion technique. Specifically, for each activity, each single modality based classifier gives N scores $\{P_1^i, P_1^i, ..., P_N^i\}$ where N is the number of classes in each split (N = 20 for S_1, N = 6 for S_2, N = 2 for S_3), $i \in [1, 4]$ is the single i^{th} modality classifier. In this work, we have four classifiers which are RGB based; Depth based; Skeleton based and Acceleration based. We take the biggest score among all scores provided by every classifier to produce the final scores for a given video. Then a label will be assigned to a test sample if it has the maximal score from the final scores. From this result, the classifier which gives the best classification is also indicated (eq.2).

$$\{Classifier, ClassLabel\} = \underset{i \in [1,4], j \in [1,N]}{argmax} \{P_j^i\} \quad (2)$$

According to this equation, only the best modality will contribute to the classification. In reality, one modality could be more important than others. In this work, we also employed weighted scores fusion technique by searching for the best weight for each modality from validation test. For example, suppose that we would like to combine M classifiers. We

search M non negative weights $\lambda_1, \lambda_2, ..., \lambda_{M-1}, \lambda_M = 1 - \sum_{i=1}^{M-1} \lambda_i$ so that the performance of the combined classifier is the best on the validation data. Specifically:

$$\{ClassLabel\} = \underset{i \in [1,4], j \in [1,N]}{argmax} \{\sum \lambda_i * P_j^i\} \quad (3)$$

The experimental results will show performance of combination of modalities on the test set following these two fusions techniques.

V. Experimental results

A. Evaluation on RGB modality

The table III shows the experimental results obtained by each baseline method on the single modality. In overall, C3D on RGB data achieved the best recognition performance on all splits of data (20 activities, 6 groups or 2 classes). The F1-score on 20 activities is only 68.35%. This is a reasonable result due to many activities having similar movement and appearance. Most confusion appears at activities *left hand pick up* and *right hand pick up*. Besides, the method can not distinguish fall activities from different sides during walking (*left fall, right fall, back fall, front fall*). This is interesting to notice that the C3D convnet represents motion and appearance of activity, but it is difficult to capture difference of movement orientation. This conclusion is confirmed again when the F1-score has increased significantly to 95.98% in case of classification of 6 groups. This time, the C3D can not discriminate some activities with hands and legs movement with activities performed by horizontal movement of the whole body. In case of fall and non-fall classification, C3D produces the highest F1-score up to 96.82%. It shows that C3D could distinguish very well fall and non-fall activities, even there are many fall styles or fall-like activities.

B. Evaluation on Depth modality

Using depth modality, the F-1 score is about 47.03% (20 activities), 75.94% (6 groups) and 87.07% (fall detection). Compared to C3D on the RGB modality, the technique DMM-KDES-MLP has lower performance although it can obtain the good performance on other datasets such as MSRAction3D, MSRGesture3D [6]. The reasons are multiple. Firstly, depth captured by Kinect sensor v1 is quite noise. Secondly, some of activities were not performed at space center but surrounding, leading to the noise and missing of depth data. Finally, as the movement of activities is quite complex and confused, the motion map can not capture small variation of the motion. Motion map is only suitable for the observation of movement in one direction with one style of fall.

C. Evaluation on Skeleton modality

Similar to depth, performance of activity classification based on skeleton is much lower than that of RGB and still lower than depth. The F1-score obtained for 20 activities is 39.38%, for 6 groups is 58.43% and 2 main groups is 76.06%. It is worth to note that ResTCN outperforms state of the art methods on NTU RGB+D dataset. Our dataset is more challenging than NTU-RGBD dataset since it contains a number of activities with "non-standing" postures where skeleton is not always well estimated such as lying, bending, sitting. Moreover, the duration of activity varies largely from one subject to the other. Another reason is that the network was trained from scratch which is still a small dataset for training deep neural networks. Data augmentation should be done for a better performance.

TABLE III
Comparison of F1-score using single modality

Modality	20 Activities	6 groups	Fall and Non-Fall
RGB	**0.6835**	**0.9598**	**0.9682**
Depth	0.4703	0.7594	0.8707
Acc.	0.3897	0.6403	0.8916
Skeleton	0.3938	0.5843	0.7606

D. Evaluation on Acceleration modality

The situation for acceleration data is slightly worse. Overall, F1 scores for 20 activities are approximately 38.97%, whose is reasonable as the set of activities addressed in this study composes of extremely fine grained and complex activities, in which the motions of several activities are highly similar. For example, lie on bed and sit up, sit on bed and stand up etc. Therefore, based only on movement without utilizing the appearance data is highly confused. This makes our dataset highly challenging for fall detection using a single modality such as accelerometer only. Consequently, F1-score for fall detection is 89.16%. However, acceleration based classification is better than skeleton based in case of 6 groups. It is better than both depth and skeleton based methods in case of fall and non-fall detection.

E. Evaluation on fusion of modalities

Tab.IV shows the results obtained using multiple modalities data. It is interesting to see that using multiple modalities could help to increase the performance compared to the use of single modality. For example, for the weak modalities such as Acceleration, Skeleton, Depth, the combination of two or three modalities improved performance compared to using single modality. However, for the late fusion method based only on max score of all classifiers, the combination of two, three or four modalities does not improve recognition performance comparing to the use of single RGB data.

In case of weighted scores and average scores based fusion, a set of weights which gives the best performance on the validation set is used to determine the performance on the test set. We observe that on the column *weighted score*, classification performance using multiple modalities has been improved significantly comparing to max score fusion or average score fusion. In case of 20 activities, F1-score increases from 65.54% to 73.53%, from 91.47% to 97.13% for 6 groups classification and from 95.27% to 98.29% for fall and non-fall classification. This shows that even with the combination of modality, the role of modality is different and could be carefully taken into account.

TABLE IV
COMPARISON OF F1-SCORE USING MULTIPLE MODALITIES

Multiple Modalities	20 activities			6 groups			Fall and Non-Fall		
	Max score	Average Score	Weighted Score	Max score	Average Score	Weighted Score	Max score	Average Score	Weighted Score
RGB+Depth	0.6639	0.6754	0.6815	0.9562	0.9672	0.9635	0.9659	0.9776	0.9789
RGB+Skeleton	0.6466	0.6386	0.7096	0.9138	0.9229	0.9608	0.9407	0.9436	**0.9829**
RGB+Acc	0.6554	0.6503	0.6977	0.9574	0.9645	0.9674	0.9714	0.9776	0.9776
Depth + Skeleton	0.4578	0.4848	0.5348	0.7236	0.7356	0.7943	0.8584	0.8707	0.9125
Depth + Acc	0.5214	0.5338	0.5935	0.8187	0.8564	0.8462	0.9304	0.9362	0.9104
Skeleton + Acc	0.5017	0.4932	0.4875	0.6833	0.6945	0.7188	0.8766	0.8666	0.9017
RGB+Depth+Acc	0.6495	0.6875	0.6897	0.9572	0.9684	0.9711	0.9691	0.971	0.9776
RGB+Skeleton + Acc	0.6554	0.6848	**0.7353**	0.9147	0.9188	**0.9713**	0.9538	0.9606	0.9763
RGB+Depth+Skeleton	0.6319	0.6693	0.7096	0.9135	0.9335	0.9581	0.9384	0.9603	**0.9829**
Depth+Skeleton+Acc	0.5253	0.5726	0.6234	0.7628	0.8275	0.86	0.9111	0.9404	0.9433
RGB+Depth+Skeleton+Acc	0.6452	0.7035	0.7273	0.9146	0.9589	0.962	0.9527	0.9697	**0.9829**

VI. CONCLUSION

We have presented a unique multi-modal multi-view dataset of human fall which is, to the best of our knowledge, biggest fall focused dataset in term of number of human subjects, fall styles, viewpoints and modalities. We have investigated the role of each modality in the task of human activity classification from one camera view. The experiments show that each modality has strength for best recognizing some of activities. In overall, C3D on RGB achieved the best performance in all splits. As depth is quite noisy, skeleton is unreliable due to the complex and various human poses, leading to lower performance when using depth or skeleton. Acceleration is good for quickly detecting a fall candidate for further verification by another modality. Itself, the acceleration can not distinguish a fall from many other fall-like activities. The late fusion experiment confirmed that combination of modalities could improve performance of classification.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-17-1-4056.

REFERENCES

[1] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. A survey on vision-based fall detection. In *Int. Conf. on Pervasive Technologies Related to Assistive Environments*, page 46. ACM, 2015.

[2] Weiguo Feng, Rui Liu, and Ming Zhu. Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera. *Signal, Image and Video Processing*, 8:1129–1138, 2014.

[3] Yixiao Yun and Irene Yu-Hua Gu. Human fall detection via shape analysis on riemannian manifolds with applications to elderly care. In *Int. Conf. on Image Processing (ICIP)*, pages 3280–3284, 2015.

[4] Yixiao Yun and Irene Yu-Hua Gu. Human fall detection in videos via boosting and fusing statistical features of appearance, shape and motion dynamics on riemannian manifolds with applications to assisted living. *Computer Vision and Image Understanding*, 148:111–122, 2016.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[6] Thanh-Hai Tran and Van-Toi Nguyen. How good is kernel descriptor on depth motion map for action recognition. In *Int. Conf. on Computer Vision Systems*, pages 137–146. Springer, 2015.

[7] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. *CoRR*, abs/1704.04516, 2017.

[8] Cuong Pham and Tu Minh Phuong. Real-time fall detection and activity recognition using low-cost wearable sensors. In *Int. Conf. on Computational Science and Its Applications*, pages 673–682. Springer, 2013.

[9] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall dataset. *DIRO-Université de Montréal, Tech. Rep*, 1350, 2010.

[10] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Definition and performance evaluation of a robust svm based fall detection solution. In *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, pages 218–224. IEEE, 2012.

[11] Xin Ma, Haibo Wang, Bingxia Xue, Mingang Zhou, Bing Ji, and Yibin Li. Depth-based human fall detection via shape features and improved extreme learning machine. *Journal of biomedical and health informatics*, 18(6):1915–1922, 2014.

[12] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. Evaluating depth-based computer vision methods for fall detection under occlusions. In *Int. Symp. on Visual Computing*, pages 196–207. Springer, 2014.

[13] Korbinian Frank, Maria Josefa Vera Nadales, Patrick Robertson, and Tom Pfeifer. Bayesian recognition of motion related activities with inertial sensors. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct*, pages 445–446. ACM, 2010.

[14] Rami Alazrai, Mohammad Momani, and Mohammad I Daoud. Fall detection for elderly from partially observed depth-map video sequences based on view-invariant human activity representation. *Applied Sciences*, 7(4):316, 2017.

[15] Angela Sucerquia, José David López, and Jesús Francisco Vargas-Bonilla. Sisfall: a fall and movement dataset. *Sensors*, 17(1):198, 2017.

[16] Jing Zhang, Wanqing Li, Philip O. Ogunbona, Pichao Wang, and Chang Tang. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60(Supplement C):86 – 105, 2016.

[17] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji, and Y. Li. Depth-based human fall detection via shape features and improved extreme learning machine. *IEEE Journal of Biomedical and Health Informatics*, 18(6):1915–1922, Nov 2014.

[18] Thanh-Hai Tran, Thi-Lan Le, and Jeremy Morel. An analysis on human fall detection using skeleton from microsoft kinect. In *Int. Conf. on Communications and Electronics (ICCE)*, pages 484–489, 2014.

[19] Martha Magali Flores-Barranco, Mario-Alberto Ibarra-Mazano, and Irene Cheng. Accidental fall detection based on skeleton joint correlation and activity boundary. In *Int. Sym. on Visual Computing*, pages 489–498. Springer, 2015.

[20] Yueng Santiago Delahoz and Miguel Angel Labrador. Survey on fall detection and fall prevention using wearable and external sensors. *Sensors*, 14(10):19806–19842, 2014.

[21] Georgios Mastorakis and Dimitrios Makris. Fall detection system using kinect's infrared sensor. *Journal of Real-Time Image Processing*, 9(4):635–646, 2012.

[22] Bogdan Kwolek and Michal Kepski. Improving fall detection by the use of depth sensor and accelerometer. *Neurocomputing*, 168:637–645, 2015.

[23] Thanh-Hai Tran, Thi-Lan Le, Van-Nam Hoang, and Hai Vu. Continuous detection of human fall using multimodal features from kinect sensors in scalable environment. *Computer Methods and Programs in Biomedicine*, 2017.

[24] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *Workshop on Applications of Computer Vision*, pages 53–60. IEEE, 2013.