Proceedings of Machine Learning Research 116:39–51, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Privacy Preserving Human Fall Detection using Video Data

Umar Asif UMARASIF@AU1.IBM.COM
IBM Research Australia

Benjamin Mashford* BENJAMIN.MASHFORD@ANU.EDU.AU
IBM Research Australia

Stefan von Cavallar SVCAVALLAR@AU1.IBM.COM
IBM Research Australia

Shivanthan Yohanandan† SHIVANTHAN.YOHANANDAN@RMIT.EDU.AU
IBM Research Australia

Subhrajit Roy‡ SUBHRAJIT.ROY@AU1.IBM.COM
IBM Research Australia

Jianbin Tang JBTANG@AU1.IBM.COM
IBM Research Australia

Stefan Harrer SHARRER@AU1.IBM.COM
IBM Research Australia

Editors: Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

Falling can have fatal consequences for the elderly people especially if the fallen person is unable to call for help due to loss of consciousness or any other associated injury. Automatic fall detection systems can assist in overcoming this issue through prompt fall alarms which then allow the triggering of a third party response, and to minimize the fear of falling when living independently at home. Vision-based fall detection systems detect human regions in the scene and use information from these regions to train classifiers for fall recognition. However, the performance of these systems lack generalization to unseen environments due to factors such as errors in the human detection stage and the unavailability of large-scale fall datasets to learn robust features for fall recognition. In this paper, we present a deep learning based framework towards automatic fall detection from RGB images captured by a single camera. Our framework learns human skeleton and segmentation based fall representations purely from synthetic data generated in a virtual environment. This de-identifies personal information contained in the original images and preserves privacy which is highly desirable in health informatics. Experiments on challenging real-world fall datasets show that our framework performs successful transfer of fall recognition knowledge from synthetic to real-world data and achieves high sensitivity and specificity scores showcasing its generalization capability for highly accurate fall detection in unseen real-world environments.

* Now with Australian National University, Australia
† Now with RMIT University, Australia
‡ Now with Google

1. Introduction

Falling on the ground is considered to be one of the most critical dangers for the elderly people living alone at home which can cause serious physical injuries and restricts normal activities because of the fear of falling again Fleming and Brayne (2008). Automated fall detection systems can produce prompt alerts in hazard situations. Furthermore, these systems allow automatic collection and reporting of fall incidents which can be used to analyse the causes of the falls and thus improving the quality of life of the elderly people living alone at home. In this context, wearable devices based systems use sensors such as switches, accelerometers and gyroscopes (embedded in wrist bands, garments, or walking sticks). These devices capture high velocities (which occur during the fall), and provide alerts when abnormalities are detected in the sensor data. Although these devices are low-cost, they require frequent recharging and therefore pose problems for older people or persons with coginitive impairment. Vision-based fall detection systems provide a low cost solution to fall detection using videos or images. These systems do not cause sensory side effects on the human health and do not affect the normal routines of people Zhao et al. (2012). In a typical vision-based fall detection approach, features are extracted from the visual data and fed to a machine learning classifier for fall recognition. For instance, methods such as Mirmahboub et al. (2013); Huang et al. (2004) extracted human shape information from camera images and used different classification models to distinguish fall from other activities. The performance of these automatic systems depend on factors including: **i)** the quality of human-region detection in the scene, **ii)** the type of information extracted from the detected human-regions, and **iv)** the classifiers used to learn features for fall recognition. Furthermore, the data used to train the classifiers plays a critical role in learning robust features that can generalize to unseen environments. Due to the unavailability of large-scale public fall datasets, most of the existing fall detectors are trained and evaluated using simulated environments only or using restricted datasets (which cannot be shared publicly due to privacy concerns). Consequently, the existing systems lack generalization capabilities for fall detection in unseen real-world environments which is highly desirable for a commercial product in the health care industry.

In this paper, we explore ways to overcome the above mentioned challenges and improve human fall recognition generalization to unseen real-world environments while preserving privacy of the people. For this, we present a deep learning based framework for automatic fall detection using human pose and segmentation information from color images captured by a video camera. In summary, the main contributions of this paper are as follows:

1. We present HPES, a Human Pose Estimation and Segmentation module (Sec. 3.1), which combines multiple CNN structures to generate human proposals in the form of human bounding boxes, segmentation masks, and body joints estimates. We also present a novel method to refine the generated proposals based on joint confidence scores and convex hull based heuristic rules.

2. We present a human-pose based fall representation (Sec. 3.1.2) which is invariant to changes in human physical appearances (person identity), backgrounds, lighting conditions, and person spatial locations in the scene. This de-identifies personal information contained in the original images and preserves privacy which is highly

desirable in health informatics. Experiments show that the proposed fall representation enables a deep CNN to learn highly robust features which successfully generalize to unseen real-world environments for fall recognition.

3. We present FallNet (Sec. 3.2), an ensemble of multiple CNN structures which learn fall representations based on human pose and segmentation information. Given a multi-modal input data (in the form of human pose and segmentation masks), FallNet takes advantage of both modality-specific and the complimentary information among the two modalities and improves the quality of fall predictions compared to independent classifiers.

4. We present Synthetic Human Fall Dataset, a large-scale dataset (Sec. 3.4) containing synthetically generated human pose and segmentation data rendered using real-world human motion captures of fall and other events. Our dataset open up new possibilities for advancing human pose based fall detection using purely synthetic data.

5. We perform an ablation study of our framework in terms of different variations of our fall representation and present Human Fall Detection (HFD) models which learn highly robust mappings between the input representations and their corresponding fall or no-fall labels. Experiments show that our HFD models when trained with only synthetic data produced high fall recognition accuracy on an unseen real-world fall dataset.

2. Related Work

Existing vision-based fall detection approaches focus on detecting human regions in the scene through motion segmentation or background subtraction, and use the information from the detected regions to train classifiers for fall recognition. For instance, the methods of Miaou et al. (2006); Töreyin et al. (2005) used background subtraction to detect human bounding boxes and compared the boxes against different thresholds in consecutive frames of the MultiCam fall dataset Auvinet et al. (2010) to detect fall events. The methods of Mirmahboub et al. (2013) combined shape and context information from the human bounding boxes and used Support Vector Machine (SVM) for fall recognition. The method of Huang et al. (2004) used extreme learning machines with shape features and achieved fast computation. Most of the above mentioned methods strongly rely on the assumption that the change in the visual information between subsequent image frames is significant to achieve appropriate motion segmentation. This restricts their application in situations where the change in information between subsequent frames is not sufficient (e.g., if a person moves towards the camera). To overcome this limitation, 3D vision based methods used information from multiple cameras or depth sensors (such as Microsoft Kinect), and learned 3D features for fall recognition. For instance, the method of Hung et al. (2013) used visual data from multiple cameras and produced decisions through voting amongst different viewpoints. The methods of Gasparrini et al. (2014); Mastorakis and Makris (2014) used Kinect depth maps to extract 3D silhouettes and 3D bounding box based features for detecting falls. Although these multi-camera based systems produce more accurate fall detection results compared to the single-camera based methods, the performance of these

methods are largely affected by hardware limitations. For instance, multiple camera based methods require accurate synchronization between the individual cameras. On the other hand, depth camera based methods are affected by sensor inherent noise, narrow fields of view and limited depth sensing restrictions. Furthermore, many public places such as elderly care centres and health-care facilities restrict the use of depth-based camera systems due to health related concerns. Consequently, accurate fall detection from monocular images is considered to be highly relevant application domain in health care industry.

In this paper, we present a deep learning based framework which uses RGB images to detect a fall in the scene. Compared to the existing methods, our work differs in several ways. **First**, our framework integrates multiple CNN structures for human detection, pose estimation, and segmentation. It uses a refinement method to correct pose and segmentation errors and generates high quality human proposals especially for scenes with multiple people or scenes with partial occlusions, compared to the methods Miaou et al. (2006); Töreyin et al. (2005) which use background-foreground subtraction techniques for human region detection and produce low true positives in these challenging situations. **Second**, we using human-skeleton and segmentation based visual representations for deep feature learning. Our visual representations preserve human privacy and they are invariant to appearance variations and spatial translations of people in the scene. This enables our framework to successfully generalize to unseen real-world environments compared to the methods (e.g., Mirmahboub et al. (2013); Miaou et al. (2006)) which learn fall recognition using appearance data and suffer from poor generalization in the presence of large changes in appearance characteristics.

3. The Proposed Framework

Fig. 1 shows the overall architecture of our framework which has three main components. **i)** Human Pose Estimation and Segmentation (HPES) module, which uses multiple CNN structures to generate human proposals in the form of body joint estimates and segmentation masks in the scene. **ii)** Visual fall representation generation module, which encodes human pose information in the form of a skeleton representation and a corresponding segmentation mask. **iii)** FallNet, a CNN model which uses the skeleton and segmentation based visual representations and learns high-level feature embeddings for fall recognition. In the following, we describe in detail the individual components of the proposed framework.

3.1. The Proposed Human Pose Estimation and Segmentation (HPES) Module (Fig. 1-A)

Our HPES module is composed of two models termed Model 1 and Model 2 as shown in Fig. 1-A. These models generate multiple human proposals in the scene which are then refined to correct errors of the individual models.

3.1.1. Model 1 (Fig. 2-A) and Model 2 (Fig. 2-B)

Our Model 1 uses the CNN model of He et al. (2017) to generate human-specific region proposals as shown in Fig. 2-A. These proposals are then fed into a classification branch consisting of fully connected layers (f_{bbox}, f_{cls}) which estimate the bounding box coordinates

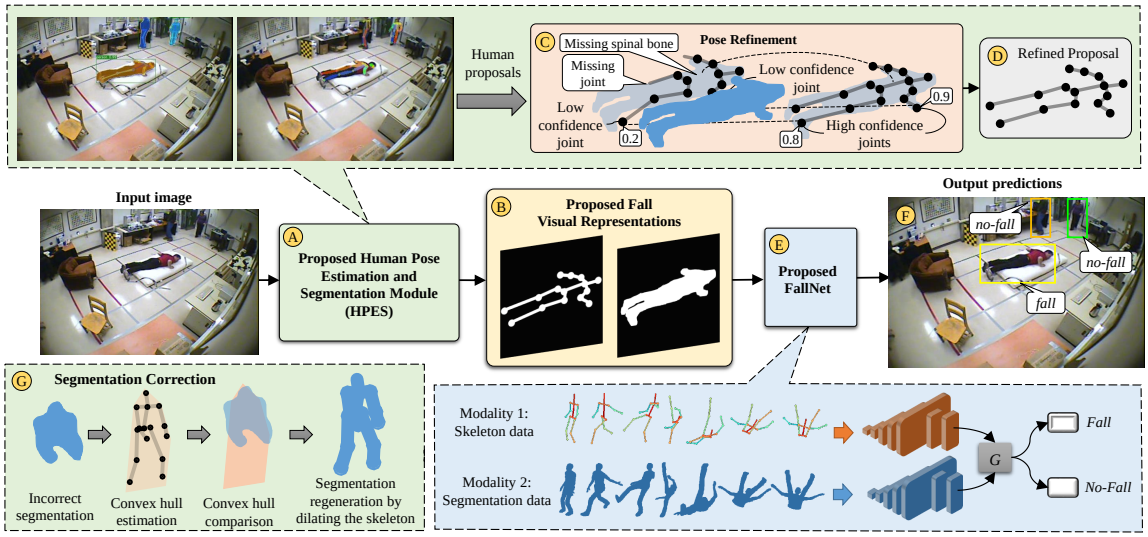

Figure 1: Overview of the proposed framework. Given an RGB image as input, our framework uses a Human Pose estimation and Segmentation (HPES) module (A) and generates human proposals with joint locations and segmentation masks. The proposals are refined through joints filtering (C) and segmentation correction (G). The refined proposals fed into the proposed FallNet (E), a CNN model with modality-specific layers, and a multi-modal embedding layer G. It learns high-level feature embeddings (from the input pose and segmentation data) to distinguish between fall and no-fall poses.

and confidence scores of the input proposals. The proposals are also fed into a segmentation branch consisting of fully convolutional layers (S_{mask}, S_{kps}) which predicts human-specific segmentation masks and joint locations. The output of Model 1 consists of human segmentation masks, joint locations, and their corresponding confidence scores. Our Model 2 uses the detector of Liu et al. (2016) to generate human-specific bounding boxes which are then fed into a stacked hourglass network Newell et al. (2016) based pose estimator which produces joint coordinates of the corresponding input proposals as shown in Fig. 2-B. It also uses a Spatial Transformation Network to select dominant proposals for pose estimation. The output of Model 2 consists of human-specific bounding boxes, joint estimates, and their corresponding confidence scores.

3.1.2. The Proposed Fall Visual Representations (Fig. 1-B)

Here, we build visual representations from the body joints estimates and human segmentation masks. We considered two visual representations as shown in Fig.1-B. A binary skeleton, where we construct a binary image by drawing the joint estimates and bones connecting the joints for a target pose, and a segmentation mask representing the silhouette information of the target pose.

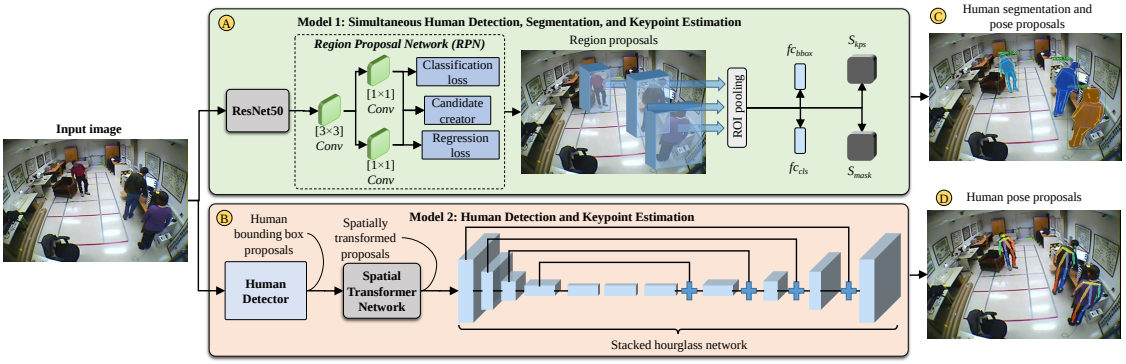

Figure 2: Overview of the proposed Human Pose Estimation and Segmentation (HPES) module. It is composed of two main models. Model 1 starts with a Region Proposal Network to generate human-specific proposals, which are fed into fully connected layers fc_{bbox} and fc_{cls} to produce human-centered bounding boxes and their confidence scores. The proposals are also fed into a segmentation branch consisting of fully convolutional layers S_{mask} and S_{kps}, which estimate human segmentation masks and joint coordinates. Model 2 uses an independent human detector to generate human proposals, which are spatially corrected through a Spatial Transformation Network (STN), and finally fed into a Stacked Hourglass Network to produce joint estimates.

3.1.3. The Proposed Pose Refinement (Fig. 1-C and Fig. 1-G)

Human detectors inevitably make errors in joint estimates and pixel-wise segmentation predictions, which in turn produce incorrect skeletons and segmentation information. To overcome these challenges, we present a pose refinement method which eliminates pose redundancies and reject pose proposals with low-confidence joints and the proposals with joints less than the minimum joints threshold ($\delta_{kps} = 7$). Fig. 1-C illustrates our pose refinement method. Let us denote poses P_i^h and P_i^{h+1} generated by models h and $h+1$. The pose P_i has m joints denoted by $\{[k_i^1, c_i^1], ..., [k_i^m, c_i^m]\}$, where k_i^j and c_i^j denote the jth location and confidence score of the joints respectively. We compare the poses P_i^h and P_i^{h+1} by joint types, and reject a joint in proposal P_i^h if its score c_i is less than the score of its corresponding joint type in P_i^{h+1}. We iteratively process all the joints in P_i^h and reject the proposal P_i^h if the number of filtered joints in P_i^h is less than those in P_i^{h+1}. Similar to joint estimation errors, pixel-wise segmentations produced by Model 1 (Fig. 2-A) are prone to errors due to occlusions and variations in appearance characteristics. This can lead to incorrect pose encodings for fall recognition. To corrects errors in segmentation predictions, we compare the segmentation mask of a proposal with the convex hull estimation of its corresponding joints data and reject the segmentation if the intersection area between the segmentation and the convex hull is less than a threshold. Subsequently, we generate a new segmentation using the joints data (through image dilation and binary thresholding) as shown in Fig. 1-G. The refined joints and segmentation proposals are then used to build the proposed visual fall representations as shown in Fig. 1-B.

3.2. The Proposed FallNet (Fig. 1-E)

FallNet consists of two sub-modules: a modality-specific module $F_\phi, \phi \in \{\mathcal{Q}, \mathcal{S}\}$, and an embedding module G as illustrated in Fig. 1-E. The terms $\mathcal{Q}$ and $\mathcal{S}$ denote the skeleton and segmentation based visual representations, respectively. The modality-specific module F_ϕ has a CNN structure similar to ResNet18 He et al. (2016). It is composed of bottle-neck convolutions (with 3×3 and 1×1 sized filters) with Batch Normalization (BN) and Rectified Linear Units (ReLUs). It produces $512 \times 7 \times 7-$dimensions feature maps which are squeezed to $512-$dimensional vectors through a global average pooling operation. There is one F_ϕ module for each visual representation. Output feature vectors from the modality-specific modules are then fed into the embedding module G which combines the multi-modal feature inputs through summation and uses a Softmax operation on the combined feature vectors and produces probabilistic distributions with respect to the target classes (fall and no-fall). Let ρ_i denote the outputs of G for the i^{th} image. We define the loss over $\boldsymbol{K}$ images as:

$$Loss = \sum_{i \in \boldsymbol{K}} \mathcal{L}_{cls}(\rho_i, \rho_i^*), \tag{1}$$

where ρ_i^* represent the ground-truths. The term $\mathcal{L}_{cls}$ is a Cross Entropy Loss defined as:

$$\mathcal{L}_{cls}(\mathrm{x}, C) = -\sum_{C=1}^{N_C} \mathcal{Y}_{\mathrm{x},C} \log(p_{\mathrm{x},C}), \tag{2}$$

where $\mathcal{Y}$ is a binary indicator if class label C is the correct classification for observation x, and p is the predicted probability of observation x of class C.

3.3. Training and Implementation

We initialized the weights of the networks of the HPES module with the weights pre-trained on MS COCO Keypoints dataset Lin et al. (2014), which contains 64K images including 260K person instances and 150K instances with key-point annotations. For FallNet, we initialized the weights of the convolutional layers with the weights pre-trained on ImageNet and initialized the weights of the embedding layers with zero-mean Gaussian distributions (standard deviations were set to 0.01 and biases were set to 0). We trained the convolutional and the embedding layers in an end-to-end manner for 150 epochs. The starting learning rate was set to 0.01 and divided by 10 at 50% and 75% of the total number of epochs. The parameter decay was set to 0.0005 on the weights and biases. Our implementation is based on the framework of Torch library Paszke et al. (2017). Training was performed using ADAM optimizer and an Nvidia Tesla K80 GPU.

3.4. The proposed Synthetic Human Fall Dataset

We present a synthetic human fall dataset which contains around 767K samples with body pose and their corresponding segmentation ground truths, categorized into fall and no-fall classes. Specifically, there are around 480K poses representing fall and 287K poses for no-fall. For data generation, we first used the MakeHuman tool to generate 3D humanoid models with associated pose skeletons. Next, we used Blender to create a rich library of scene

Table 1: Ablation study of the proposed framework in terms of different visual representations for fall recognition on the MultiCam fall dataset Auvinet et al. (2010) and the Le2i fall dataset Charfi et al. (2013). Our Human Fall Detection Models were trained only on the synthetic data and evaluated on the test datasets.

Human Fall Detection models	MultiCam fall dataset			Le2i fall database		
	F1Score	Precision	Recall	F1Score	Precision	Recall
Skeleton	0.8677	0.8685	0.8671	0.8517	0.8668	0.8529
Segmentation	0.8071	0.8448	0.7964	0.8529	0.8535	0.8529
Multi-modal (MM)	**0.8708**	**0.8703**	**0.8715**	**0.9244**	**0.9245**	**0.9244**
MM (without pose refinement)	0.8384	0.8380	0.8441	0.9034	0.9031	0.9032

templates containing humanoid meshes with the required cameras and lights configurations. Finally, we simulated the humanoid models using the MoCap data from CMC (2003) (which provides motion capture pose data of human individuals performing everyday life activities in an indoor environment). For each MoCap motion sequence, we generated outputs in the form of **i)** segmentation masks, **ii)** body joint locations, and **iii)** a label representing the pose as "fall" or "no-fall" for each pose in the motion sequence. Fig. 3-left shows some sample frames from the proposed Synthetic Human Fall dataset.

4. Experiments

To evaluate the generalization capability of our framework for fall detection in unseen real-world environments, we trained framework using the proposed synthetic dataset and tested framework on the public MultiCam fall dataset of Auvinet et al. (2010). The MultiCam dataset consists of 24 different scenarios where each scenario is comprised of a video sequence of people performing a number of activities. Each scenario is recorded using 8 different cameras installed at different locations in an indoor environment. For all scenarios, the video data is annotated for 9 different activities (such as walking, falling, lying on the ground, crouching, moving up/down, sitting, lying on a sofa, and moving horizontally). The dataset is challenging for single camera based fall detection because, different camera viewpoints produce occlusions and significant variations in the spatial locations, scale, and orientations of the fall events. We also used the the Le2i fall dataset Charfi et al. (2013) for our evaluations. The Le2i dataset contains 221 videos of different actors performing fall actions and various other normal activities in different environmets. The dataset is challenging due to variable lighting conditions and occlusions. To quantify the recognition performance of our framework, we computed the weighted F1 score, precision and recall scores. We used these measures as they are not biased by imbalanced class distributions which make them suitable for the test datasets where the number of fall samples are considerably small compared to the number of non-fall samples.

4.1. Results

Here, we evaluated the generalization capability of our framework for fall detection in unseen real-world environments. For this, we trained our models using only the proposed

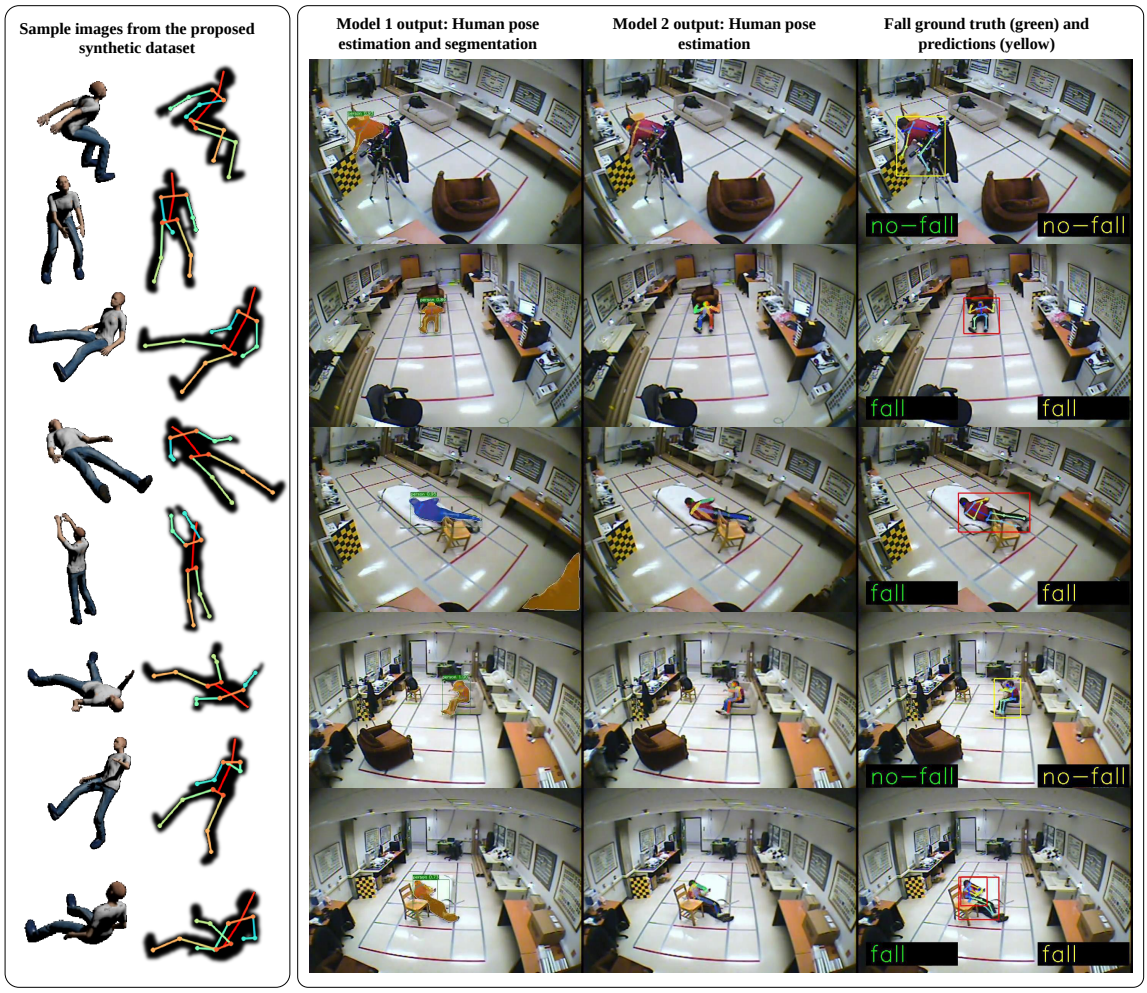

Figure 3: Left: Sample frames from our Synthetic Human Fall dataset with a variety of poses and viewpoints. Right: Qualitative results of our fall detection on the MultiCam fall dataset.

synthetic data and evaluated the models on the test datasets. Table 1 shows the results of these experiments, where we present different variants of our framework termed "Human Fall Detection (HFD) models", trained using different visual representations. The results show that our framework achieved high precision, recall, and F1 scores which demonstrate the generalization capability of our framework in successfully transferring fall recognition knowledge learnt purely from synthetic data to unseen real-world environments. Fig. 3-right shows qualitative results on sample images from the MultiCam fall dataset Auvinet et al. (2010). From the results, we observe that our fall detector is robust to partial occlusions, and variations in the spatial locations, scale, and orientations of falls in the scene. We attribute this successful generalization of our fall detector from synthetic-to-real data to our three key contributions as described in the following sub-sections.

Table 2: Ablation study of our framework in terms of model generalization when trained on one dataset and tested on a different dataset for different input data modalities.

Training data	Modality	Testing data					
		MultiCam fall dataset			Le2i fall database		
		F1Score	Precision	Recall	F1Score	Precision	Recall
MultiCam	RGB	**0.9860**	**0.9860**	**0.9861**	0.7351	0.7604	0.7405
	Multi-modal	0.9627	0.9627	0.9628	**0.8449**	**0.8512**	**0.8456**
Synthetic	RGB	0.8631	0.8671	0.8699	0.6421	0.7874	0.6775
	Multi-modal	**0.8708**	**0.8703**	**0.8715**	**0.9244**	**0.9245**	**0.9244**

4.1.1. Improved Human Proposal Generation

Our framework utilizes the domain knowledge of human pose estimation learnt from the large scale MS COCO database and produces multiple human proposals which are refined by our pose refinement method. Consequently, our HPES module produces highly accurate and reliable human proposals compared to motion-segmentation based proposals Wang et al. (2016) which suffer errors if there is no significant change between subsequent frames of the video sequence. Table 1 shows the improvements in fall recognition precision and recall scores obtained using the proposed pose refinement method.

4.1.2. Multi-modal CNN Architecture

Our FallNet combines modality-specific convolutions and multi-modal embedding layers and takes advantage of both the modality-specific and complimentary information in the body pose and segmentation based visual representations for discriminating fall and no-fall poses. This enables our model to learn features which are more robust compared to the methods such as Hung et al. (2013) which only utilize modality-specific feature learning for fall recognition. Table 1 shows the improvements in fall recognition precision and recall scores obtained using the proposed multi-modal architecture compared to independent data modalities.

4.1.3. Invariance to Physical Appearances and Background

Here we further evaluate the generalization capability of our fall detector by training the model on one dataset and testing the model on a different dataset. Table 2 shows the results of these experiments. From Table 2, we see that an RGB-based detector (using color information of human proposals for fall recognition), when trained on the MultiCam dataset Auvinet et al. (2010) produced good fall recognition accuracy on the MultiCam test set but did not generalize well on the Le2i dataset Charfi et al. (2013). This is due to the differences in the physical appearance of human actors and different backgrounds of Le2i dataset Charfi et al. (2013) compared to the MultiCam dataset Auvinet et al. (2010). On the other hand, our multi-modal skeleton-segmentation based representation produced competitive fall recognition performance when trained and tested on the MultiCam fall dataset Auvinet et al. (2010). Also, our model produced improvements of at least 11% in the

average weighted f1 scores on the Le2i dataset demonstrating its generalization capability across multiple real-world datasets. Similarly, an RGB-based model trained on our synthetic dataset produced inferior generalization across the MultiCam and Le2i datasets compared to our skeleton-segmentation based model. These improvements are attributed to our human pose-based fall representation which is invariant to appearance characteristics, thus making our framework robust to different human actors and unknown background in real-world scenes. On the contrary RGB-based fall detector fails to generalize to scenes which have large variations in the appearance characteristics of people and backgrounds.

5. Conclusion and Future Work

In this paper we present a deep learning framework towards automatic human fall detection from images captured by a single camera. Our framework produces human proposals with body joint locations and segmentation information. These proposals are refined and transformed into multi-modal visual representations for input to FallNet, a CNN model which uses modality-specific and multi-modal layers and learns highly discriminative feature embeddings for fall recognition. We also present a human fall dataset which consists of human pose and segmentation data synthetically generated under different camera viewpoints. Experiments on challenging public fall datasets show that our framework trained using only synthetically generated pose data successfully generalizes to unseen environments and achieves high precision and recall scores for fall recognition. Trained on pure synthetic data, our framework is highly robust to variations in appearance characteristics, scale changes, and different camera viewpoints. This opens up new possibilities for advancing privacy preserving human fall detection which is highly desirable in health informatics. In future, we plan to expand our framework for the recognition of other activities to enhance its potential for general human activity recognition. We also plan to reduce the computational burden of our fall detector through parameter-pruning and memory efficient CNN structures for low-powered GPU devices.

References

Cmu graphics lab: Carnegie-mellon motion capture (mocap) database. http://mocap.cs.cmu.edu, 2003.

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

Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and adaboost-based classification. *Journal of Electronic Imaging*, 22(4):041106, 2013.

Jane Fleming and Carol Brayne. Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. *Bmj*, 337:a2227, 2008.

Samuele Gasparrini, Enea Cippitelli, Susanna Spinsante, and Ennio Gambi. A depth-based fall detection system using a kinect sensor. *Sensors*, 14(2):2756–2775, 2014.

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: a new learning scheme of feedforward neural networks. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pages 985–990. IEEE, 2004.

Dao Huu Hung, Hideo Saito, and Gee-Sern Hsu. Detecting fall incidents of the elderly based on human-ground contact areas. In *2013 2nd IAPR Asian Conference on Pattern Recognition*, pages 516–521. IEEE, 2013.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

Georgios Mastorakis and Dimitrios Makris. Fall detection system using kinect's infrared sensor. *Journal of Real-Time Image Processing*, 9(4):635–646, 2014.

S-G Miaou, Pei-Hsu Sung, and Chia-Yuan Huang. A customized human fall detection system using omni-camera images and personal information. In *Distributed Diagnosis and Home Healthcare, 2006. 1st Trans disciplinary Conference on*, pages 39–42. IEEE, 2006.

Behzad Mirmahboub, Shadrokh Samavi, Nader Karimi, and Shahram Shirani. Automatic monocular system for human fall detection based on variations in silhouette area. *IEEE Transactions on Biomedical Engineering*, 60(2):427–436, 2013.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

B Uğur Töreyin, Yiğithan Dedeoğlu, and A Enis Çetin. Hmm based falling person detection using both audio and video. In *International Workshop on Human-Computer Interaction*, pages 211–220. Springer, 2005.

Shengke Wang, Long Chen, Zixi Zhou, Xin Sun, and Junyu Dong. Human fall detection in surveillance video based on pcanet. *Multimedia tools and applications*, 75(19):11603–11613, 2016.

Guoru Zhao, Zhanyong Mei, Ding Liang, Kamen Ivanov, Yanwei Guo, Yongfeng Wang, and Lei Wang. Exploration and implementation of a pre-impact fall recognition method based on an inertial body sensor network. *Sensors*, 12(11):15338–15355, 2012.