Multi-camera and multi-person indoor activity recognition for continuous health monitoring using long short term memory

Meratwal, Mitali, Spicher, Nicolai, Deserno, Thomas

Mitali Meratwal, Nicolai Spicher, Thomas M. Deserno, "Multi-camera and multi-person indoor activity recognition for continuous health monitoring using long short term memory," Proc. SPIE 12037, Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, 120370A (4 April 2022); doi: 10.1117/12.2612642

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Multi-camera and multi-person indoor activity recognition for continuous health monitoring using long short term memory

Mitali Meratwal[a,b], Nicolai Spicher[b], and Thomas M. Deserno[b]

[a]Indian Institute of Technology, Bombay, India
[b]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

ABSTRACT

Human activities play a vital role in many health-related fields of research. For example, changes in activities of daily living can predict neurodegenerative diseases like Alzheimer's disease or the risk of suffering from a fall. Therefore, automatically recognizing activities in videos has become of scientific interest. In contrast to wearable devices, video cameras have the advantage of being truly unobtrusive without any physical contact, continuous as they do not require charging, and cannot be forgotten to wear. This work proposes a novel approach for multi-camera and multi-person human activity recognition (HAR) in videos. The aim is to classify each person in each frame of a video in one of the five classes: "Lying", "Sitting", "Standing", "Walking" or "Falling". We use a combination of YOLOv4 (person detection), DeepSort (tracking), a convolutional neural network (CNN, features extraction), and an attention-based multi-layer long short-term memory network (classification) to track actions of multiple subjects. Our entire dataset comprises of four publicly-available datasets (ETRI - Activity 3D, MSR Daily Activity 3D, HAR-UP Fall Dataset, High-Quality Fall Simulation Data) with a total file size of 300 gigabytes. In our experiments, we pick random subsets of 1% and 0.25% of data for training and testing respectively. We achieve a classification accuracy of 95%. In the presence of a single subject in the room, predictions from multiple cameras are combined using soft voting, which further improves the accuracy. In summary, HAR in videos is feasible using the proposed combination of machine learning techniques.

Keywords: Video processing, multiple cameras, video classification, activity recognition, fall detection, machine learning, long short-term memory, neural networks, deep learning

1. INTRODUCTION

Vision-based HAR is becoming increasingly popular due to its wide variety of applications, such as in human behaviour understanding, healthcare monitoring systems,[1] pattern inference from daily life activities, robotic intelligence for elder care[2] and social interaction, monitoring elderly in nursing homes, post-rehabilitation support and assessment etc. Advancements in deep learning and computer vision have enabled these tasks to be automated such that they can be carried out without the need for any human intervention while keeping them efficient, deployable, and easily accessible.

Some commonly used sensor-based modalities for HAR include sensors like radar, audio, accelerometer, or gyroscope, integrated into the living environment such as chairs or wearable devices for continuous health monitoring.[3] However, a significant downside of using these modalities for HAR is their obtrusiveness, which adds a layer of inconvenience in our daily lives and lacks visual intuitiveness. In contrast, (RGB, depth, thermal) camera-based HAR fosters unobtrusive, automatic, and continuous health monitoring, which becomes deployable in any private space such as a smart home.

Video-based HAR has witnessed many attempts using different techniques from optical flow,[4] hidden Markov models,[5] spatio-temporal correlation[6] to the ones based on joint skeleton, and more recently, deep learning.[7] Human pose or skeleton-based activity recognition primarily relies on extracting handcrafted features from body joints and using these feature vectors as input in support vector machines, random forests, or regression models. Some recent works have also explored graph-based Fourier transform[8] and directed and undirected

Send correspondence to Prof. Dr. Thomas M. Deserno, e-mail: thomas.deserno@plri.de, phone: +49 531 391-2130

Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications,
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 12037,
120370A · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2612642

graph neural networks.[9] However, the accuracy of skeleton-based methods for activity recognition depends on the quality of the pose estimator. To our knowledge, open-source pose estimators do not perform with sufficient accuracy in real-time for certain applications. For example, we found that multi-person pose estimators like OpenPifPaf,[10] AlphaPose,[11] and OpenPose[12] fail to localise and detect keypoints once the person has fallen or lied down. Another instance where keypoint detection algorithms provide poor performance is in the presence of partial occlusions, which occur from furniture and other objects in the surroundings. In our previous work,[13] handcrafted features were extracted from a five-point inverted pendulum model for each frame. The feature vector sequence then formed the input for the classification of the activity. Though handcrafted features simplify the representation of joints, we observed that the model overfitted easily and lacked generalisability on unseen data. Out of all the fall videos in High-Quality Fall Simulation Dataset,[14] only 43.5% were correctly classified as fall for the algorithm described in our previous work.[13]

Thus, inspired by the shortcomings of pose estimators, their inability to deal with occlusions, and in order to transcend from hand-crafted feature representation, we devise an alternative approach in this work. We recognise the action of multiple people while tracking unique subjects in all the frames in multiple cameras using YOLOv4,[15] DeepSort,[16] CNN based feature extractor, and attention-based multi-layer LSTM network. Our research questions are:

- RQ1: Can human activities (lying, sitting, standing, walking, falling) be detected accurately using the proposed methodology?
- RQ2: Can multiple camera views in combination with a simple voting mechanism be used to improve the detection accuracy?

2. MATERIAL AND METHODS

2.1 Person detection and tracking

The first stage in the pipeline (Fig. 1) detects all the subjects present in a frame at any instant and furthermore tracks each individual across all the frames processed so far. Object detection is performed on each incoming frame using the publicly available CNN-based object detector YOLOv4.[15] YOLOv4 produces state-of-the-art results on the MS COCO dataset and offers an efficient detector with real-time performance on a GPU, running faster than other object detectors with similar performance.

With the key objectives for our application being: tracking multiple objects, minimising identity switches, supporting tracking for longer periods of occlusions, DeepSort[16] is a popular and widely used framework. A primary advantage of DeepSort over other tracking algorithms is its independence from the object detector module. This allows easy integration with any state-of-the-art object detector. All bounding boxes with the label "person" are extracted from each frame using YOLOv4. Tracking is performed by linking these bounding boxes across all frames and assigning a unique ID to each person using DeepSort.

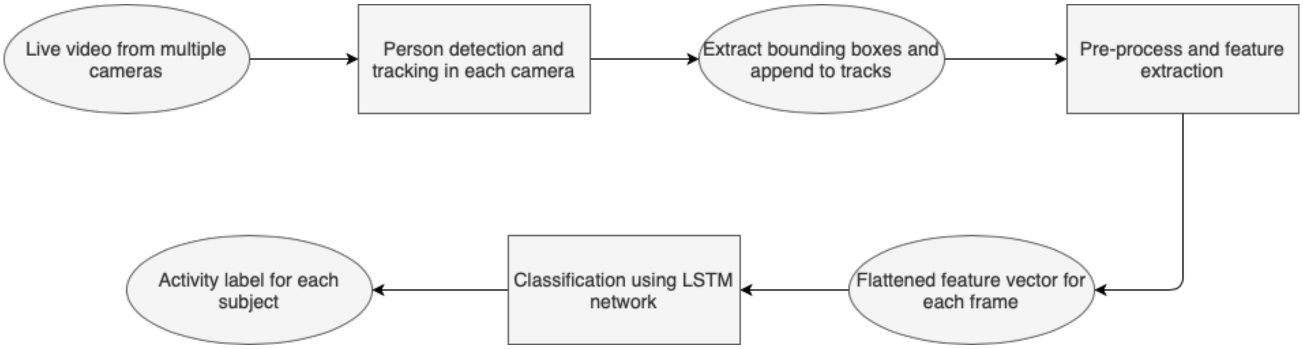

Live video from multiple cameras
Person detection and tracking in each camera
Extract bounding boxes and append to tracks
Pre-process and feature extraction
Flattened feature vector for each frame
Classification using LSTM network
Activity label for each subject

Figure 1. Algorithm pipeline

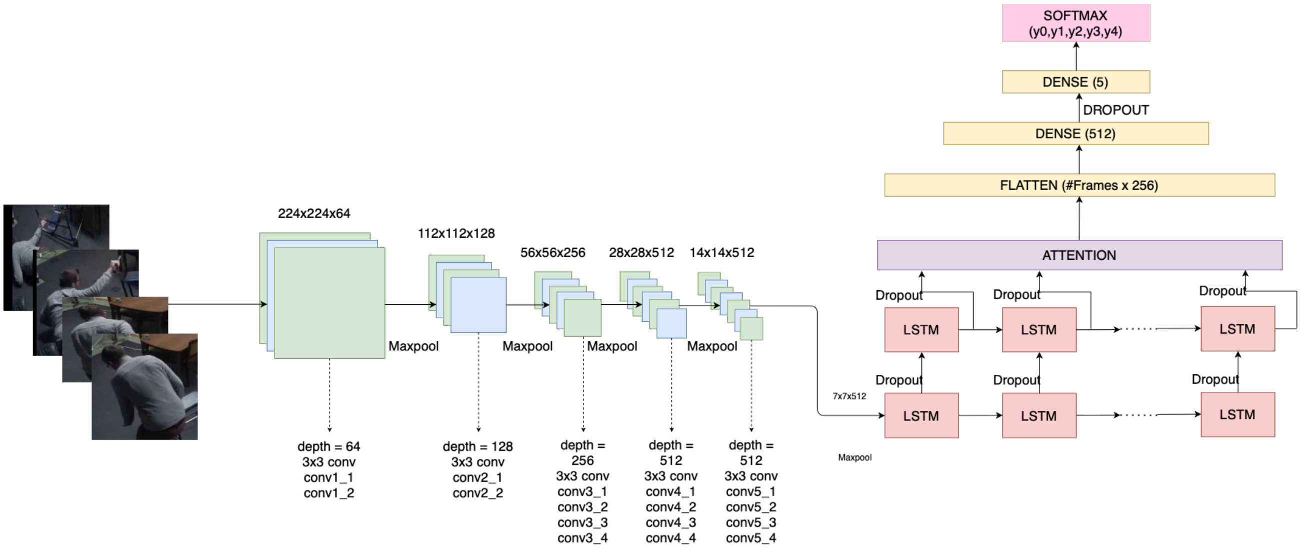


Figure 2. Model architecture

2.2 Pre-processing and feature extraction

The previous stage updates the bounding box position for each unique id in the current frame. A track is confirmed if at least in three consecutive frames a person with the corresponding track_id was detected. After iterating through all the tracks that have been confirmed and updated, an image patch from the current bounding box location is extracted for each track, thereby localising only people present in the image and eliminating the background. Each patch is resized to a standard dimension of 224 x 224 pixels, and the image is normalised using mean and standard deviation obtained from the ImageNet dataset. Subsequently, the patches are passed through the VGG19 model trained on the ImageNet dataset (Fig. 2). Features extracted are flattened to one long vector, thus reducing the dimensionality of an image patch. Low-level representation of an image any pre-trained model learns on being trained on a huge dataset, can be exploited to escape the computational cost of training the model again that produces similar feature representations from initial convolutional blocks.

2.3 Classification using LSTM

After sampling a fixed number of frames from the set of frames seen so far, we have a feature vector for each person in each frame from the previous step. The sampled frames for each individual are stacked and used for predicting the activity using an LSTM attention network (Fig. 2). The sequence of features extracted for each person is passed through a two-layer LSTM network where each cell is followed by a dropout. At every time step, the output of the first LSTM layer is the input to the next layer. In order to assign importance or "attention" to each frame in the sequence, the output sequence of the second LSTM layer is passed through an attention block followed by softmax activation. This yields a probability matrix which is multiplied with the output of the second layer. The operation effectively introduces sparsity by using a weighted sum. The output of the attention block is flattened, passed through two dense layers, and the final output of the model is a probability vector over all the classes. The class with maximum probability is predicted as the activity for the input frame.

2.4 Datasets

To increase variability in activities in terms of subjects, interaction with objects and surroundings, activities that are similar to fall but stem from natural actions, and improve the accuracy of the prediction model, we collected a dataset with a size of 300GB (Tbl. 1). We collect publicly available datasets on single-person daily activities recorded in a home-like environment. Across the dataset pool collected, all videos were relabelled as belonging to exactly one of the five classes: falling, lying, sitting, standing, or walking. However, training the model on 300GB is unfeasible, and therefore, we pick random subsets: 1% and 0.25% of all data for training and testing, respectively.

Table 1. Datasets used

Dataset	Resolution	FPS	#Cam	Size (GB)	#Subjects
ETRI-Activity 3D[2]	1920 x 1080	20	8	296	100
MSR DailyActivity3D[17]	640 x 480	30	1	2.2	10
HAR-UP Fall Dataset[18]	640 x 480	18	2	0.6	17
High Quality Fall Simulation Data[14]	800 x 480	30	5	0.06	-

2.5 Implementation details

Since the YOLOv4 object detector produces the same results on every run, instead of computing bounding boxes for each person every time the model is trained, the bounding box coordinates are saved and reused every time we want to retrain. We first execute the object detector module on the chosen random subset and store normalised bounding box coordinates for single individuals present in all frames before commencing training. This simplifies the training input to bounding box coordinates and target as the activity label for that video.

The length of the input sequence is also a hyper-parameter for the LSTM network. Hence, we train our model for different number of frames sampled uniformly from the set of all frames for any video. We compare the final results for different input sequence lengths (8, 16, 32) and select the model with the best validation accuracy.

2.6 Evaluation

We have trained three models with input sequence lengths as 8, 16, and 32 to inspect frame redundancy and obtain the optimal number of frames that produces good validation accuracy on keeping other model parameters constant. We employ early stopping and class weights to reduce overfitting and make the model generalizable. Weighted cross-entropy is used as the loss metric.

3. RESULTS

3.1 Training process (Accuracy & losses)

We achieve the best validation accuracy of 96.62% for 32 frames (Fig. 3), while we obtain 95.56% and 96.24% for 16 and 8 frames, respectively.

3.2 Activity recognition

Testing the models on 0.25% of the data, we find that the true positives for “Falling” are highest for 32 frames, lowest for 8 frames, while the model trained on 16 frames gives intermediate results (Fig. 4). Since our primary objective is high accuracy for fall detection while obtaining low false negatives, we use the model trained for 32 frames for subsequent experiments.

3.3 Impact of number of cameras on results

Next, we analyse if ensembling predictions for an activity recorded in multiple views improve the overall accuracy of the algorithm. Soft voting is used to ensemble all predictions of each instance. An instance is the activity of a single person captured by multiple cameras from different angles and positions. Testing on a subset of data, we found that activities initially incorrectly classified were assigned correct labels after voting predictions from more than one camera (Fig. 5).

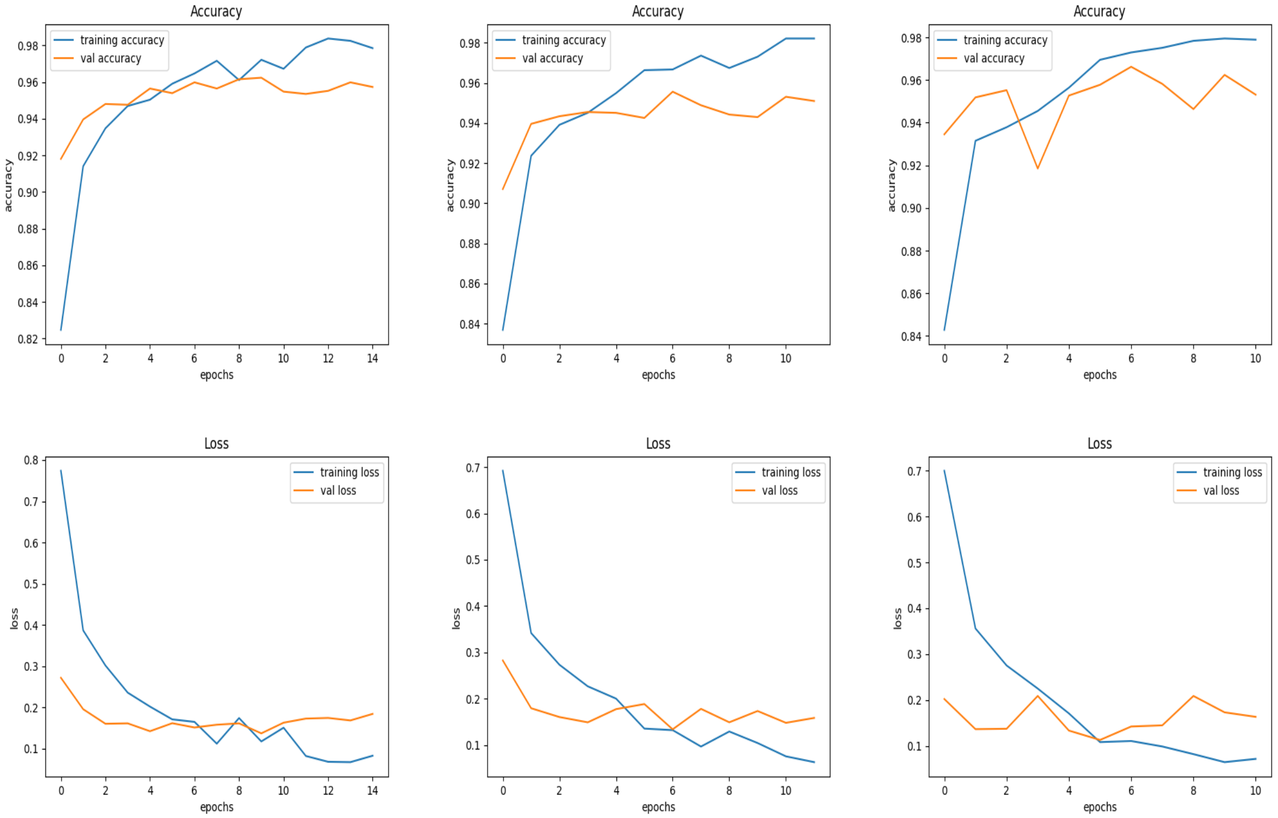


Figure 3. Accuracy (top row) and losses (bottom row) for 8 (left), 16 (centre), 32 frames (right)

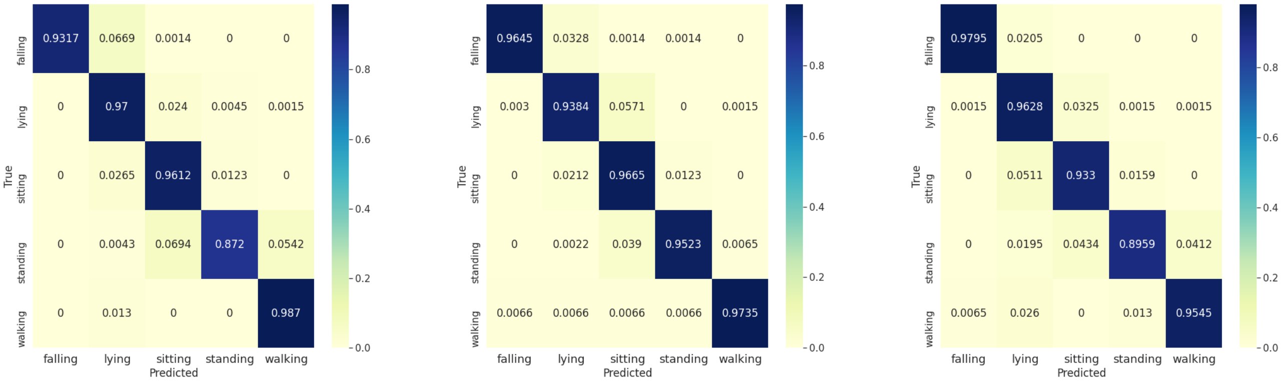


Figure 4. Confusion matrices for 8 (left), 16 (centre), 32 frames (right)

3.4 Comparing Results

Fig .6 presents the output for a frame from a video in the ETRI-Activity 3D dataset with multiple subjects. The proposed method works even if the subjects are interacting with each other (Fig. 6 (a)) as the model is trained and tested on image patches extracted for each person. Next, we shows results on some videos from ETRI-Activity 3D and High Quality Fall SImulation dataset for which the algorithm suggested in[13] produces

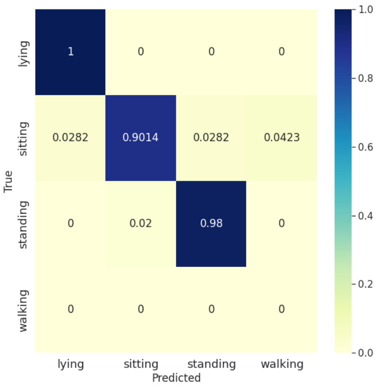


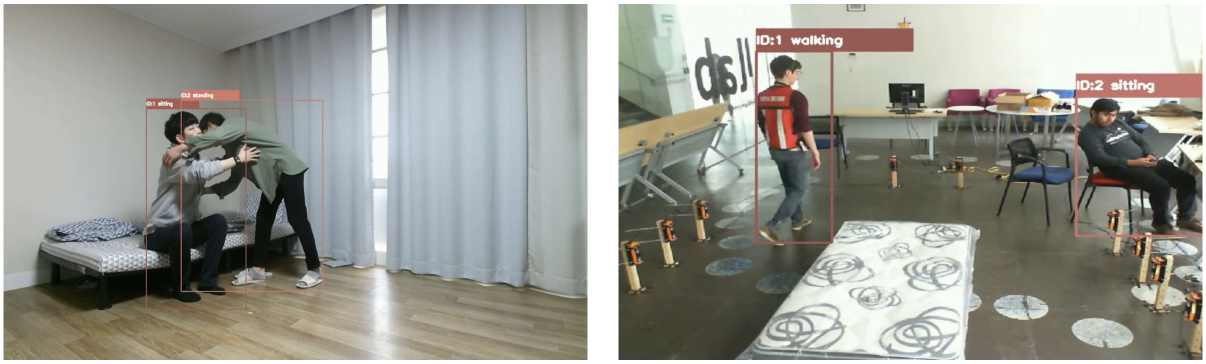
Figure 5. Results with multiple cameras on a subset of the data (no walking activity, 32 frames only)

false positive and false negative results (Fig. 7) . The scheme suggested in this paper is able to reduce false positives and false negatives by making correct predictions.

4. CONCLUSIONS

We answer our research questions as follows: Activity recognition from videos is feasible using the proposed algorithm (RQ1). Furthermore, multiple cameras and a voting mechanism increase accuracy (RQ2). We proposed a new methodology for multi-person, multi-camera action recognition by using and extending established technologies in machine learning. With the training and testing, we have obtained commendable results on realistic and simulated fall datasets like High-Quality Fall Simulation Data and UP Fall Detection Dataset. We offer a solution that is fast and at the same time provides good accuracy. The model can be trained on any dataset and extended to any number of classes, irrespective of the differences in the FPS of the recorded videos across different datasets.

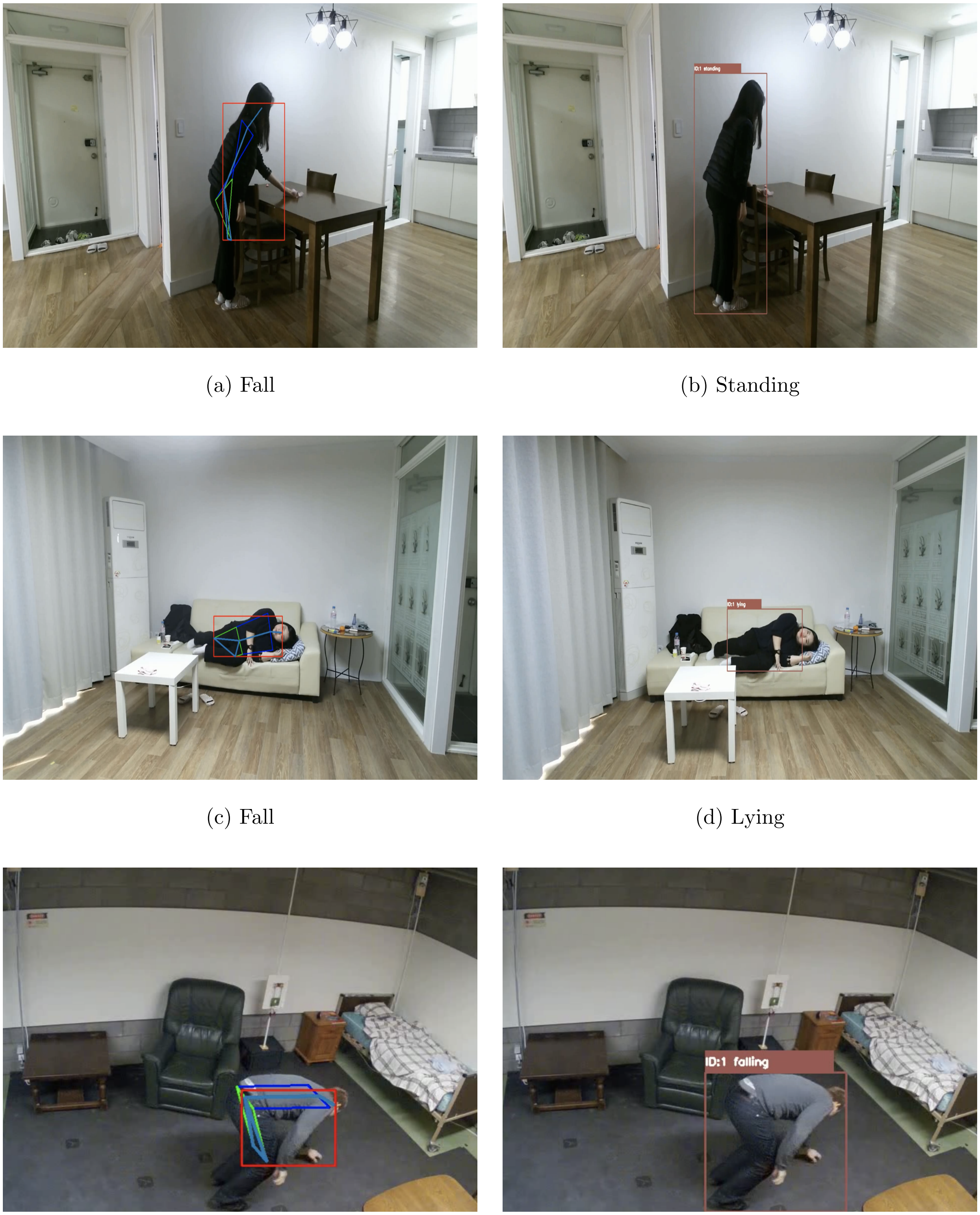
In future work, we wish to examine the performance of the presented algorithm on videos recorded in our smart home laboratory. Eventually, the proposed algorithm could be used to detect falls and automatically send a message to all relevant systems in the rescue chain, i.e. the rescue service and hospital, by using the recently proposed International Standard Accident Number (ISAN).[19,20] Furthermore, since the accuracy of the LSTM network depends on the features representing the input sequence, we will experiment with different feature extractors to optimize memory and processing requirements as well as the system's precision.



(a)

(b)

Figure 6. Prediction on multi-person videos using approach described in this paper.



(a) Fall

(b) Standing

(c) Fall

(d) Lying

(e) No Fall

(f) Falling

Figure 7. Comparing results: (a), (c), (e) are predictions based on approach described in.[13] (b), (d), (f) are predictions of proposed method.

REFERENCES

[1] Wang J, Spicher N, Warnecke JM, Haghi M, Schwartze J, Deserno TM. Unobtrusive health monitoring in private spaces: The smart home. Sensors. 2021;21(3).

[2] Jang J, Kim D, Park C, Jang M, Lee J, Kim J. ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020. p. 10990–7.

[3] Wang J, Warnecke JM, Haghi M, Deserno TM. Unobtrusive health monitoring in private spaces: The smart vehicle. Sensors. 2020;20(9).

[4] Sun S, Kuang Z, Sheng L, Ouyang W, Zhang W. Optical flow guided feature: A fast and robust motion representation for video action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018. p. 1390–9.

[5] Brand M, Oliver N, Pentland A. Coupled hidden markov models for complex action recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 1997. p. 994–9.

[6] Nebisoy A, Malekzadeh S. Video action recognition using spatio-temporal optical flow video frames. arXiv preprint arXiv:210305101. 2021;.

[7] You J, Korhonen J. Attention boosted deep networks for video classification. In: 2020 IEEE International Conference on Image Processing (ICIP); 2020. p. 1761–5.

[8] Kao JY, Ortega A, Tian D, Mansour H, Vetro A. Graph based skeleton modeling for human activity analysis. In: 2019 IEEE International Conference on Image Processing (ICIP); 2019. p. 2025–9.

[9] Shi L, Zhang Y, Cheng J, Lu H. Skeleton-based action recognition with directed graph neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 7904–13.

[10] Kreiss S, Bertoni L, Alahi A. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. arXiv preprint arXiv:210302440. 2021;.

[11] Fang HS, Xie S, Tai YW, Lu C. Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2334–43.

[12] Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence. 2019;43(1):172–86.

[13] Taufeeque M, Koita S, Spicher N, Deserno TM. Multi-camera, multi-person, and real-time fall detection using long short term memory. In: Deserno TM, Park BJ, editors. Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications. vol. 11601. International Society for Optics and Photonics. SPIE; 2021. p. 35–42.

[14] Baldewijns G, Debard G, Mertes G, Vanrumste B, Croonenborghs T. Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms. Healthcare technology letters. 2016;31:6–11.

[15] Bochkovskiy A, Wang CY, Liao HYM. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:200410934. 2020;.

[16] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. 2017 IEEE International Conference on Image Processing (ICIP). 2017;p. 3645–9.

[17] Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2012. p. 1290–1297.

[18] Martínez-Villaseñor L, Ponce H, Brieva J, Moya-Albor E, Núñez-Martínez J, Peñafort-Asturiano C. Up-fall detection dataset: A multimodal approach. Sensors. 2019;19(9). Available from: https://www.mdpi.com/1424-8220/19/9/1988.

[19] Spicher N, Barakat R, Wang J, Haghi M, Jagieniak J, Öktem GS, et al. Proposing an international standard accident number for interconnecting information and communication technology systems of the rescue chain. Methods of Information in Medicine. 2021;60(S01):e20–e31.

[20] Haghi M, Barakat R, Spicher N, Heinrich C, Jageniak J, Öktem GS, et al. Automatic information exchange in the early rescue chain using the International Standard Accident Number (ISAN). Healthcare. 2021;9(8):996.