International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:13, No:5, 2019

Deep Learning Based Fall Detection Using Simplified Human Posture

Kripesh Adhikari, Hamid Bouchachia, Hammadi Nait-Charif

***Abstract*—**Falls are one of the major causes of injury and death among elderly people aged 65 and above. A support system to identify such kind of abnormal activities have become extremely important with the increase in ageing population. Pose estimation is a challenging task and to add more to this, it is even more challenging when pose estimations are performed on challenging poses that may occur during fall. Location of the body provides a clue where the person is at the time of fall. This paper presents a vision-based tracking strategy where available joints are grouped into three different feature points depending upon the section they are located in the body. The three feature points derived from different joints combinations represents the upper region or head region, mid-region or torso and lower region or leg region. Tracking is always challenging when a motion is involved. Hence the idea is to locate the regions in the body in every frame and consider it as the tracking strategy. Grouping these joints can be beneficial to achieve a stable region for tracking. The location of the body parts provides a crucial information to distinguish normal activities from falls.

***Keywords*—**Fall detection, machine learning, deep learning, pose estimation, tracking.

I. Introduction

ELDERLY people are prone to fall hazard due to their physical capacity which weakens with the age. One in three adults who are aged 65 or above are likely to fall at least once every year [1], [2]. It is very distressful to know that this can increase with ageing. Fall can lead to a different level of injuries and can be even more serious if immediate help is not available. This can lead to even death ultimately in the absence of any support. Therefore, it is very important to look into a fall detection system that can support people by whistle-blowing immediately after a fall. The detection system also needs to be accurate and effective as it can reduce the duration of harm and prove to be cost-effective during treatment. According to [3] annual report, falls are the major cause of emergency admission in the hospital and the cost of treatment is about 4.4bn in the UK.

A person can be tracked using a computer vision technique during a fall. The pose and location information of the human body part can be used to analyse the occurrence of fall. But tracking using computer vision on real time are still an open area of research due to these challenges that are still huge hurdles like lighting variations, the motion of body parts, partial/ fully occlusion and pose deformations. The challenges are even more severe during falls. Following are the questions that demand answers from a fall detection system.

- Can a person be tracked during a fall?
- Does the tracking system handle light variations?
- Does the tracking system suffer from occlusion of any kind?
- Does the tracking system is able to identify and track human in the presence of another moving object like a pet or a toy?

A system that can defend itself from the above questionnaire is highly desirable. This paper will discuss further regarding our new approach and present its results in the later.

Kripesh Adhikari is with the Bournemouth University, Media School, Poole, Dorset, UK (e-mail: kadhikari@bouremouth.ac.uk).

Hamid Bouchachia is with the National Centre for Computer Animation, Bournemouth University, Dorset, UK (e-mail: abouchachia@bournemouth.ac.uk).

Hammadi Nait-Charif is with the National Centre for Computer Animation, Bournemouth University, Dorset, UK (e-mail: hncharif@bournemouth.ac.uk).

II. Related Work

Many researchers in the past have proposed a different strategy for the fall detection system. Tracking of body or body parts have been explored and applied in the past for fall detection. However, pose estimation strategy has never been applied in fall detection to our knowledge. Some of the traditional tracking methods include particle filter and Kalman filter. Reference [4] proposed overhead tracking strategy to recognise fall in inactivity regions such as floor using particle filter. Reference [5] proposed 2D human body tracking with structural Kalman filter which utilizes the relational information among sub-regions of a moving object. The model uses previous time frame sub-regions information to define sub-regions in the current frame. Similarly, [6] also uses the strategy to divide a human body into three sub-regions and detect fall by analysing the shape of the human silhouette achieved after background subtraction. The foreground blob is computed and then is divided into three portions with the ratio of 30:40:30 per cent. Most of the tracking strategy is performed with different preliminary conditions which can suffer when they are not met. Some more cases where human body parts were tracked to differentiate fall with other normal activities such as standing and sitting were proposed by [7]–[9]. Recently machine learning techniques are getting popular to differentiate fall from other normal activities. Reference [10] proposed vision-based fall detection approach where they combined several algorithms like background subtraction, Kalman filter and optical flow to achieve input features for a machine learning algorithm. This approach can suffer in speed as different algorithms are combined which brings complexity to the approach. Reference [11] proposed an adaptive deep learning approach to detect fall. They used deep learning to distinguish humans in the foreground from background and use adaptive learning by adjusting the

Fig. 1 Issues in predicting all the available joints

weight constraints when there is a change in the background. Similarly, [12] proposed convolutional neural networks based fall detection where they classified fall from non-fall activities using transfer learning technique. They used the pre-trained VGG-16 CNN model on Imagenet with the stack of optical flow images as the input to the CNN model. Another deep learning approach was exploited by [13] where they used 3D joint skeleton features achieved from Microsoft Kinect SDK to feed into an LSTM to identify fall. They also used the transfer learning technique to avoid the need for a huge dataset for training.

III. Our Approach

All the vision-based approach analyses images and videos to identify fall but tend to suffer heavily from the curse of extremely difficult poses that are demonstrated during a fall. Tracking can be accurate during normal activities where the pose is more stable but not during a fall. There are so much of variation in body pose, its speed and also localization. It becomes a big hurdle for a system to achieve better accuracy to fight against such series of frequently changing poses. Recently pose estimation have reached a new level with joint localization accuracy. References [14] and [15] are currently the state of the art in pose estimation arena setting a higher benchmark. They are able to achieve a very accurate level of human joint estimation. However, they also suffer a lot while predicting all the joints position mainly when poses are extremely challenging.

We tested the model from [14] on the challenging images that can be observed during all falls. Their model is capable of identifying 18 joints in the RGB images for multi-person. We are looking into a single person case and hence only single human cases are examined. The issues identified are shown in the figure 1. The above images are considered from [16] which the model has never seen before. As illustrated in Fig. 1, the accuracy of all the joints can easily deviate while dealing with difficult poses. Other than that, it is very difficult to predict those joints which are occluded by the body itself.

Therefore to solve this problem, this paper looks into a detection based tracking strategy. The idea is to find a stable position of a body part that could be detected all the time during a fall scenario. As shown in Fig. 2 below, 18 different joints that are available in COCO dataset are grouped into three subgroups to represent three different regions of the body: Upper region or Head region, mid-region or Torso region and lower region or leg region.

As shown in Fig. 2below, left eye, left ear, right eye, right ear and nose are grouped to represent upper-region or head. Similarly, left shoulder, left elbow, left wrist, right shoulder,

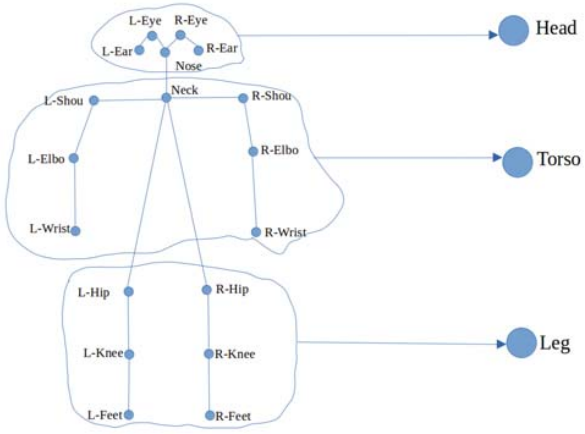

Fig. 2 Illustration of joints configuration in COCO dataset grouped into three regions

right elbow, right wrist and neck are grouped to represent mid-region or Torso. Lastly, left hip, left knee, left feet, right hip, right knee and right feet are grouped to represent lower region or leg. The grouping strategy is proposed with the aim to achieve a stable individual region of a human body. It is not important to identify all the joint positions or body facing side upward or downward to determine fall. To distinguish fall from other normal activities, all that is important is to identify the positions of primary parts of the human body. Simply the height of a head can provide a clue whether a person is standing, sitting or lying. Hence, instead of considering several individual joints, three points were created by grouping individual joints according to the section of the body they could represent.

IV. Training

Since the available joints now have been divided into three groups, these three points are now fed to CNN model. We use the same model and pre-trained weights as used by [14], but we apply a transfer learning technique to the model as their pre-trained weights have better learning capacity when trained on similar data. The top layer features are not specific to a particular object but are more general types of features such as edges and colours. Hence these features can be considered as transferable for different datasets and can be applicable to many other tasks [17]. We unfreeze only the bottom 12 layers to train the model. The higher layers have already learnt the basic features that are necessary for the pose estimation. However, the lower layers that were trained to predict 18 joints are now retrained to predict three sets of points only using the COCO dataset 2017 [18].

The model was trained for about a week as COCO datasets 2017 contains 118000 images as training dataset and 5000 as validation dataset. We trained our model using SGD with a batch size of 20, momentum of 0.9 and weight decay of 0.00005. We used an early stopping strategy to stop the training when the validation loss stops to improve by checking in 5 consecutive epochs. We used different data augmentation

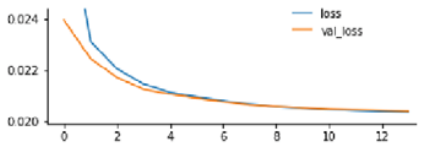

Fig. 3 Loss vs validation loss during training with three set of points using COCO dataset

technique to improvise the generalisation of learning. We used a rotation range of 30 degrees, a width shift range of 0.2, a height shift range of 0.2, a zoom range of 0.3 and the horizontal flip to augment the training data. This technique provides different changes to the original images so that while training the model gets to adopt all the possible changes that can be present in a test image. For training, we used an adaptive learning rate technique. The adaptive learning rate is a method to vary the learning rate during the training process. A pre-defined learning rate (LR) of 1e-5 was set which was slowly reduced by using step decay. A step decay drops the learning rate after certain epochs by a certain decay factor. We used a drop factor of 0.1 and drop after every 3 epochs during training. Adaptive learning rate can only be effective if the value of drop and epochs drop is sensible. That means if the drop is too high, it can hamper the learning rate will decrease faster during the process of training. Similarly, if it is too low, it can have no effect on learning by varying the learning rate. Furthermore, the epoch drop has to be set in such a way that the early stopping strategy is triggered otherwise again no effect will be experienced by the model during training.

In Fig. 4, the first row demonstrates the ground truth confidence heatmap for the three individual regions in three separate images as in [14]. The heatmap for each region is the average confidence map of the combination of several joint positions centred around that region. The second row demonstrates the ground-truth of Part Affinity Field (PAF) connections. Only 2 PAF connections are available in our case: the connection between the upper and middle region is represented by one PAF and the connection between the middle region and the lower region as the second PAF. PAF preserves both the connection and direction information of the body parts. PAF is very efficient in pose estimation to differentiate the sides like left and right body parts. However, in our case, we are trying to locate three centre sub-regions and hence the direction information from PAF for each connection is unidirectional. In the second row, therefore, we have 4 images demonstrating 2 direction for each PAF respectively.

V. Experiment Results and Discussion

We have tested the model performance against 4 different datasets: COCO dataset [18], extended Leeds Sports dataset [19], fall dataset [6] and our own dataset [16].

Fig. 5 below illustrates the output prediction heatmap on validation using the COCO Dataset for peak sub-regions in first row and their PAF connection with direction in the second row.

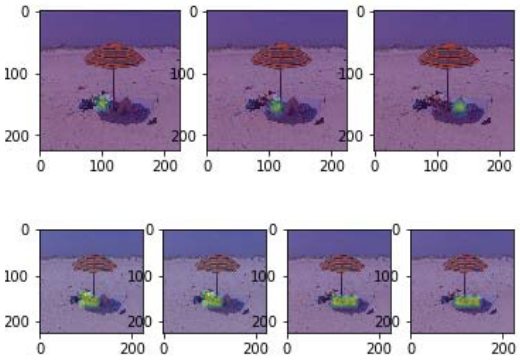

Fig. 4 Ground Truth on training using COCO dataset

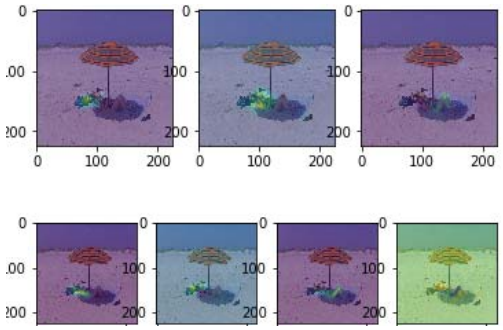

Fig. 5 Prediction on validation using COCO dataset

Finally, on testing for the same images with our solution on the issues discussed above in Fig. 1, we have successfully achieved a stable region of detection.

In the above Fig. 6, the three different regions detected are displayed in three colour points with red as the upper region, green as the middle region and blue as the lower region.

The rest of the sections looks into different scenario and conditions to analyse the performance of the model.

A. Complex Poses and Lighting Conditions

Fig. 7 below is the output from extended LSP datasets which contains challenging images from sports arena. The images in LSP dataset contains variations in poses, scale changes and illumination changes.

B. Partial Occlusion

Fig. 8, consist of images from LSP dataset and videos from [6]. It can be noted in the images that the person is partially occluded in all the cases representing different

Fig. 6 Our model solution on the same image from Fig. 1

Fig. 7 Prediction on LSP dataset

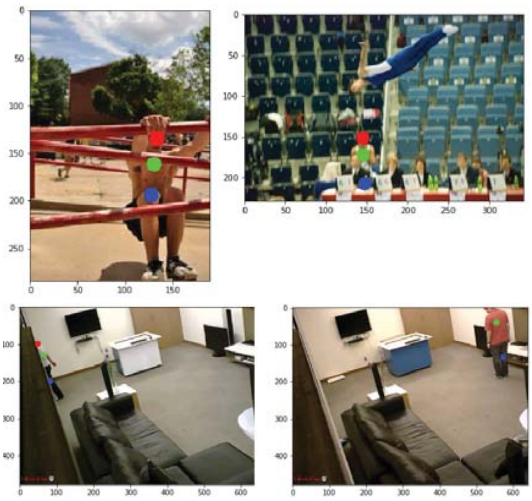

Fig. 8 Illustration of detection of three regions during a partial occlusion performed on LSP extended dataset and video 2 and 9 from [6]

possible occlusions. Our model was still able to predict the lower region of occlusion that occurred due to the presence of the chair in the scene.

C. Present of Other Moving Object

In Fig. 9, a person is playing with a toy that is moving along with person.Even in this case, the model was able to detect three sub-regions without any confusion. It can also be considered for the case where a pet may be present in the room and still the person can be detected successfully.

D. Sitting Scenario

Fig. 10 presents images of sitting scenario. Even in this case, the model was able to predict all the three regions during these scenes where the change of pose is significant compared to standing activities. It can be observed that it can also perform well during the bending poses

.

Fig. 9 Illustration of detection of three regions during the movement of a toy performed on fall dataset video 19 from [6]

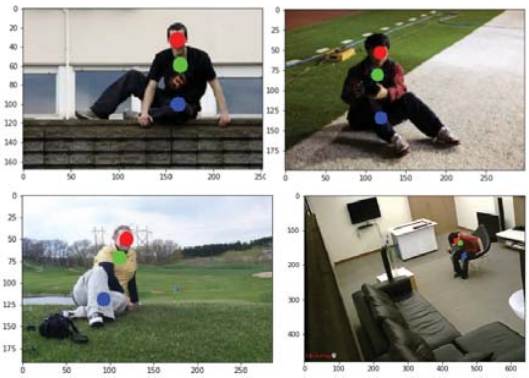

Fig. 10 Illustration of detection of three regions during a sitting pose performed on LSP extended dataset and fall dataset video 17 from [6]

Fig. 11 Illustration of detection of three regions during a lying pose performed on LSP extended dataset

E. Lying Scenario

Fig. 11 illustrates complex lying poses in different environments. Our model was able to perform well in all the cases. From this, we can already get some hint that the lying pose after a fall can be predicted in a more stable way. However, it is still important to test on the sequence of images that represents a fall to actually take up the challenge that it can generate due to the higher rate of change of these complex poses.

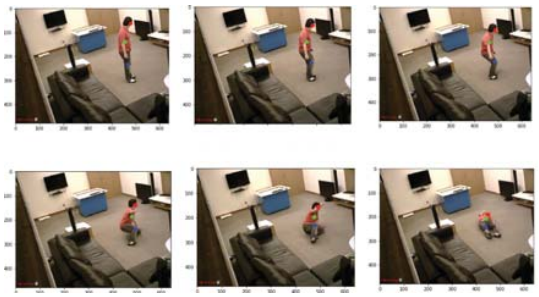

Fig. 12 Illustration of detection of three regions during a fall sequence performed on fall dataset video 5 from [6]

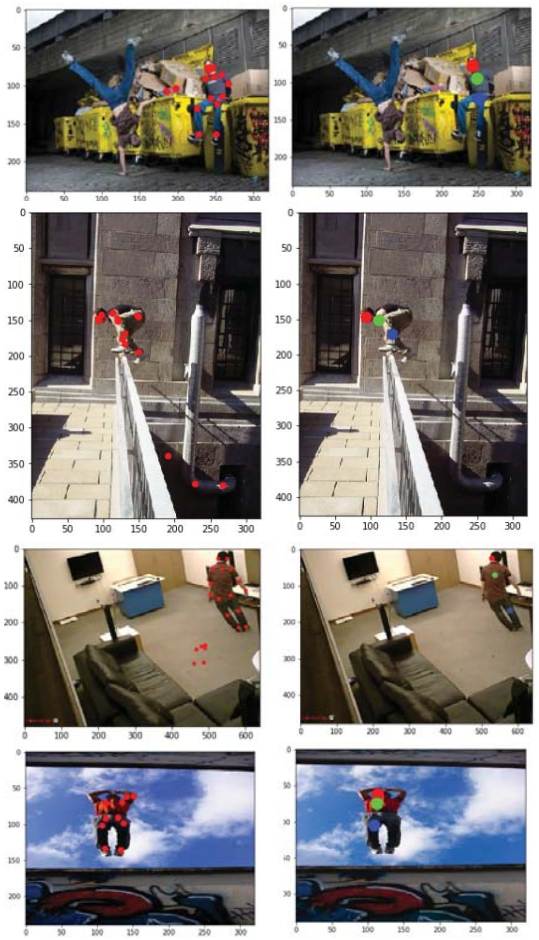

Fig. 13 Comparing prediction of 18 joints using the model from [14]in column (a) with our approach of predicting three regions in column coloumn (b) on LSP extended dataset and video1 from [6]

F. Falling Scenario

Finally, a falling scenario consisting of another 6 consecutive frames from the video 5 of fall dataset [6] was chosen to analyse the performance of the model. Here too our model was able to detect the sub-regions all the way until the person is on the floor.

In all the four sets, we can observed a much stable

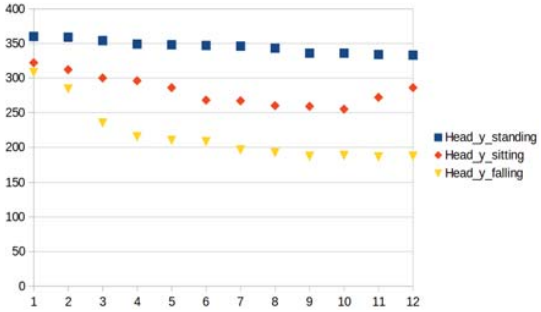

Fig. 14 Illustration of head position during a standing,sitting and falling scenario performed on videos from [6]

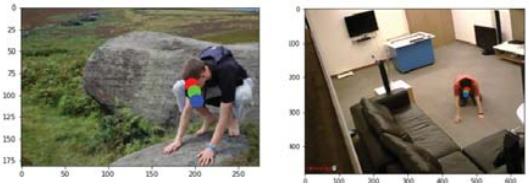

Fig. 15 . Failure case 1

prediction in our case. All these performance are analysed on the never seen dataset and there is also no ground truth available for joints in this dataset.From these observations we can summarize that our model is able to produce a competitive result against challenges like light variations, pose changes, scale changes, multiple views, movement of another object with the subject of interest and partial occlusion.

VI. Detection of Head Position

Let us also analyse the position of the head for the three scenarios: standing, sitting and falling to distinguish their characteristics.

Fig. 14 illustrates the height of the head during standing, sitting and falling scenario considered on 12 consecutive frames from different videos of [6]. In the x-axis, 1-12 represents the successive frame number and in the y-axis, 0-400 is the height in pixels. As expected,it can be clearly noted that the plot of the head height during the falling scenario is very low compared to the other two scenarios.

VII. Failure Cases

Although the model has performed very well to detect the three regions during different scenarios, it has also failed in some cases.

In the first case shown in Fig. 15, there are two images:one from LSP dataset and other from video 11 of dataset [**?**]. These image represents the problem of deformation of the body that can easily bring complexity in pose estimation. In this case, the body regions were not completed distinguished and the regions seems to overlap.

In case 2 shown in Fig. 16, in the first image from LSP dataset, the person is facing upside down. That means the direction of the head facing towards the bottom side of the image. Similarly, even in the other three images, the person is

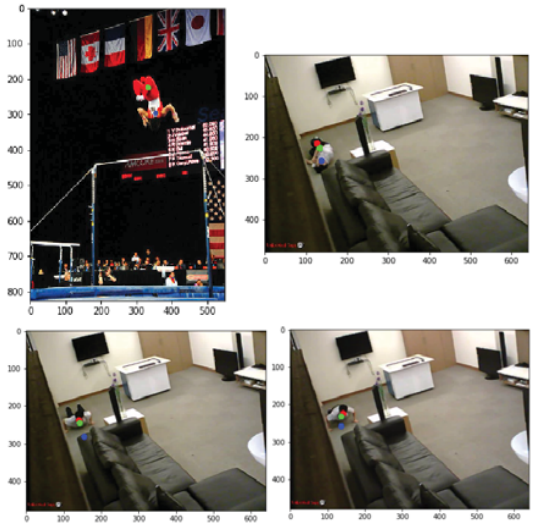

Fig. 16 . Failure case 2

lying again with the direction of the head towards the bottom of the image.

During a lying pose, y-information is almost same or negligible among the sub-regions. That means only the x-information contributes to learning during a lying pose. It is possible that the training dataset has images where the direction of the human head is mainly towards the top side of the images. Therefore it can be confusing for the model when the head is towards the downside of the image. This logic can also be justified by another information of the PAF. Usually, PAF is considered for preserving the direction information in pose estimation [14] where the model learns to identify connections with directions (left and right) among the human body joints. In our case, the sub-regions are the average value represented at the centre of their group of joints. Hence our model understands only the connection between these three regions in one direction and that could be the major reason for this kind of failure. Fine tuning the model with more images having such cases where the human head is present at the lower end of the image and also using vertical flip during data augmentation can be beneficial.

VIII. Conclusion

This paper presents a stable region based detection strategy that can be used as an alternative to the tracking of human body regions during a fall. These three regions can provide significant clues to locate the human body part to distinguish fall from other activities. Experimental results indicate our proposed technique can be used as an alternative approach to detect and track people during a fall in real life scenarios with better accuracy and stability. These regions can be considered as features that indicate the position of the body to distinguish fall from other normal activities. Therefore, we will explore the possibility of using these region features for further classification of fall and non-fall activities using other classification techniques like SVM or LSTM.

References

[1] M. Kangas, I. Vikman, J. Wiklander, P. Lindgren, L. Nyberg, and T. Jämsä, "Sensitivity and specificity of fall detection in people aged 40 years and over," *Gait & posture*, vol. 29, no. 4, pp. 571–574, 2009.
[2] W. C. H. A. a Fall, "Important facts about falls," 2016.
[3] U. Age, "Later life in the united kingdom," *Age UK Factsheet*, 2018.
[4] H. Nait-Charif and S. J. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4. IEEE, 2004, pp. 323–326.
[5] D.-S. Jang, S.-W. Jang, and H.-I. Choi, "2d human body tracking with structural kalman filter," *Pattern Recognition*, vol. 35, no. 10, pp. 2041–2049, 2002.
[6] J.-L. Chua, Y. C. Chang, and W. K. Lim, "A simple vision-based fall detection technique for indoor video surveillance," *Signal, Image and Video Processing*, vol. 9, no. 3, pp. 623–633, 2015.
[7] Z.-P. Bian, J. Hou, L.-P. Chau, and N. Magnenat-Thalmann, "Fall detection based on body part tracking using a depth camera," *IEEE journal of biomedical and health informatics*, vol. 19, no. 2, pp. 430–439, 2015.
[8] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Monocular 3d head tracking to detect falls of elderly people," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2006, pp. 6384–6387.
[9] M. Yu, S. M. Naqvi, and J. Chambers, "Fall detection in the elderly by head tracking," in *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*. IEEE, 2009, pp. 357–360.
[10] K. de Miguel, A. Brunete, M. Hernando, and E. Gambao, "Home camera-based fall detection system for the elderly," *Sensors*, vol. 17, no. 12, p. 2864, 2017.
[11] A. Doulamis and N. Doulamis, "Adaptive deep learning for a vision-based fall detection," in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. ACM, 2018, pp. 558–565.
[12] A. Núñez-Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless Communications and Mobile Computing*, vol. 2017, 2017.
[13] A. Shojaei-Hashemi, P. Nasiopoulos, J. J. Little, and M. T. Pourazad, "Video-based human fall detection in smart homes using deep learning," in *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*. IEEE, 2018, pp. 1–5.
[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.
[15] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," *arXiv preprint arXiv:1802.00434*, 2018.
[16] K. Adhikari, "Fall detection dataset," 2017, last accessed 10 December 2018. [Online]. Available: http://www.falldataset.com/
[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.
[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[19] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proceedings of the British Machine Vision Conference*, 2010, doi:10.5244/C.24.12.