# 3D Convolutional Neural Network for Falling Detection using Only Depth Information

Sara Luengo Sánchez, Sergio de López Diz, David Fuentes-Jiménez[a], Cristina Losada-Gutiérrez[b], Marta Marrón-Romera[c] and Ibrahim Sarker

*Department of Electronics, University of Alcalá, Polytechnics School, Campus Universitario S/N, Alcalá de Henares, Spain*

Keywords: Depth Information, Fall Detection, 3D-CNN, Top-view, Healthcare.

Abstract: Nowadays, one of the major challenges global society is facing is population aging, which involves an increment of the medical expenses. Since falls are the major cause of injuries for elderly people, the need of a low-cost falling detector has increased rapidly over the years. In this context, we propose a fall-detection system based on 3D Convolutional Neural Networks (3D-CNN). Due to the fact that the system only uses depth information obtained by a RGB-D sensor placed in a overhead position to avoid occlusions, it results in a less invasive and intrusive fall-detection method for users than systems based on wearables. In addition, depth information preserves the privacy of people since they cannot be identified from this information. The 3D-CNN obtains spatial and temporal features from depth data, which allows to classify users' actions and detect when a fall appears. Since there are no other available datasets for action recognition using only depth data from a top-view camera, the authors have recorded and labeled the GOTPD3, that has been made available to the scientific community. Thus, training and evaluation of the network has been carried out within the GOTPD3 dataset, and the achieved results validate the proposal.

## 1 INTRODUCTION

A fall is defined as "an unexpected event in which the participant comes to rest on the ground, floor, or lower level" (Ballinger and Payne, 2002). It is estimated that approximately, the third of people in the world over the age of 65 fall each year, and this proportion increases with the age (Organization, 2012). These falls can cause other health problems such as post-fall syndrome (Mathon et al., 2017) which may cause confusion, immobilization and depression, or even more serious injuries which may need surgery. At all events, the result is a loss of independence and autonomy for the person that falls. Moreover, when the fall occurs at home to a person who is alone, it may take a long time before any assistance arrives, which may aggravate the health condition of the person. To avoid this scenario, many researches have been working over the last decade to develop a low-cost fall-detection system (Mubashir et al., 2013; Pierleoni et al., 2015; Amin et al., 2016; Gia et al., 2018; Lapierre et al., 2018)

To address this topic, many research works have been carried out with different kinds of sensors. Most of the current work involves wearable inertial sensors such as accelerometers and gyroscopes (Pierleoni et al., 2015; Wu et al., 2015). However, these systems provide many false positives and fail when the person forgets to put the wearable on. In addition, if the person has circulatory problems, swelling limbs can become a problem when wearing these sensors.

To reduce the problem of false positives, several researches have developed systems with multiple sensors, mainly cameras and accelerometers (Ozcan and Velipasalar, 2016; Zerrouki et al., 2016). Although this kind of systems have proved to reduce the number of false positives, these are worse in other aspects since privacy is no longer preserved and wearables are still needed.

To avoid the problems associated to wearables, many works use cameras for falling detection (Baptista-Ríos et al., 2016; De Miguel et al., 2017). These systems provide high accuracy in suitable situations but their performance drops when lighting conditions change or occlusions occur. Moreover, privacy issues may appear considering that peo-

[a] https://orcid.org/0000-0001-6424-4782
[b] https://orcid.org/0000-0001-9545-327X
[c] https://orcid.org/0000-0001-7723-2262

Luengo Sánchez, S., de López Diz, S., Fuentes-Jiménez, D., Losada-Gutiérrez, C., Marrón-Romera, M. and Sarker, I.
3D Convolutional Neural Network for Falling Detection using Only Depth Information.
DOI: 10.5220/0009356205900597
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 590-597
ISBN: 978-989-758-402-2; ISSN: 2184-4321

ple can be recognized in the images. It has to be highlighted that privacy is specially important in environments in which a falling detector can be installed, such as hospitals (Banerjee et al., 2014) or elderly homes (Rougier et al., 2011).

In the last years, since RGB-D and depth cameras emerged in the sensors market, they have had a major role in solving privacy issues because people cannot be identified in depth maps. Furthermore, depth cameras include an infrared light source, so they do not require an external one, being more resilient to lighting changes in the captured scene. These properties have made depth cameras widely used in security systems (Chou et al., 2018; Luna et al., 2017).

There are several papers that deal with the detection of actions from depth information (Megavannan et al., 2012; Liang and Zheng, 2015), both from skeletal joints and depth maps. However, very few of them do so with the camera in zenithal position (Lin et al., 2015; Tang-Wei Hsu et al., 2016), as in the proposal described in this work. In (Lin et al., 2015; Tang-Wei Hsu et al., 2016), the authors propose a novel feature for activity recognition from top-view depth image sequences, based on representative body points, that is then validated using their own dataset. However, as far as the authors of are concerned, the dataset used in (Lin et al., 2015; Tang-Wei Hsu et al., 2016) is not available for comparing the obtained results.

Recently, researches have started applying novel deep learning techniques to improve the results obtained with classic approaches. In this context, 2D-CNN have been used in a frame-by-frame basis to perform image classification. These networks do not have the capacity to obtain motion features since their input is a still image. Some authors propose the use of descriptors that include data from several frames, such as Depth Motion Maps (DDM) described in (Yang et al., 2012), or multi-view dynamic images (Wang et al., 2017; Wang et al., 2018), that summarise video-temporal features.

On the other hand, to take advantage of temporal features, traditional Deep Neural Networks (DNNs) have been replaced by Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTM) and 3D Convolutional Neural Networks (3D-CNN). In particular, 3D-CNN are able to analyze a video or set of images as a single input and apply filters to get temporal information as well as spatial characteristics from it. These features improve the system performance in video applications (Ji et al., 2012).

DNNs are classifiers that extract low level features in the first layers and complex characteristics in the next ones. Then, in the last layer, the network makes a prediction about each class occurrence probability. To obtain adequate features, the network is trained so filters learn each class characteristics features.

In this paper, we propose a 3D-CNN based fall-detection system which only uses depth information acquired by a Time of Flight (ToF) sensor placed in an overhead position of the scene to analyze. The top-view configuration reduces occlusions, and with the use of depth information preserves people's privacy.

Furthermore, the proposal detects not only when a person falls down, but also other of his daily actions such as 'Walk', 'Run' or 'Stand up'. The spatial and temporal feature extraction is carried out by *"3D Convolutional" ("Conv 3D")* and *"Max Pooling"* layers, while the fall (and other actions) detection is done through a classification stage based on two *"Fully Connected"* (dense) layers. The entire architecture of the proposed neural network is shown in figure 1, and it is explained in detail in section 2.

Due to the lack of available datasets for action recognition from a top-view depth image sequences, it has been necessary to record and label a new dataset, called GOTPD3 (Macias-Guarasa et al., 2018) which has been made available to the scientific community. This new GOTPD3 dataset has been used for training, validating and testing the proposal. This dataset provides several sequences for 5 actions, that are performed by 7 different people: 'Walk', 'Walk fast', 'Run', 'Fall down' and 'Stand up under the sensor'. It is worth highlighting that the provided dataset includes walking at different speeds (normal and fast), as well as running because of the strong relationship between walking speed and the people's health, tiredness and sadness (Sundelin et al., 2015)

The rest of the paper is organized as follows. Section 2 explains the network architecture and its internal operations. After that, in section 3 the training phase is described, including a brief description of the GOTPD3 dataset. Then, the results obtained in the test stage are presented in section 4. Finally, section 5 introduces the main conclusions and future work.

## 2 NETWORK ARCHITECTURE

As mentioned in the introduction, the proposed system for falling detection is based on a 3D Convolutional Neural Network (3D-CNN), which allows the extraction of temporal and spatial features, and has has proven its effectiveness for action recognition in RGB and RGB-D image sequences (dipakkr, 2018).

The general architecture of the used DNN can be seen in Figure 1, whereas a summary including the different layers of the network, as well as their output sizes and fundamental parameters is shown in Table 1.

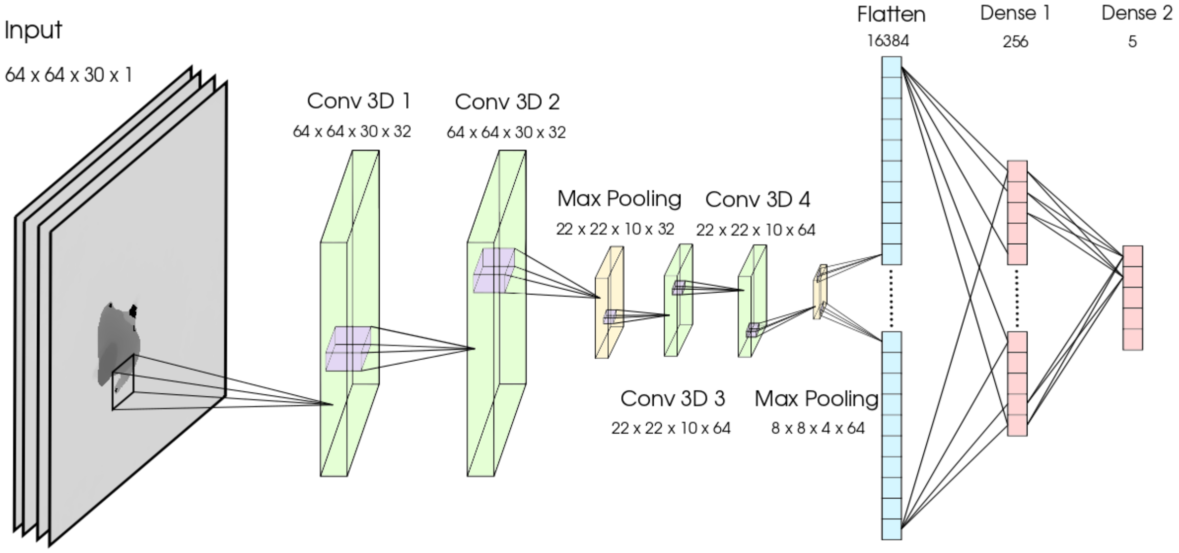

Figure 1: 3D Convolutional Neural Network (3D-CNN) architecture.

Table 1: Proposed network architecture and tensor sizes at each layer.

| 3D Convolutional Neural Network layers | | |
|---|---|---|
| **Layer** | **Output size** | **Parameters** |
| Input | $64 \times 64 \times 30 \times 1$ | - |
| Conv3D 1 | $64 \times 64 \times 30 \times 32$ | kernel=(3, 3, 3)<br>strides=(1, 1, 1) |
| Activation | ReLU | |
| Conv3D 2 | $64 \times 64 \times 30 \times 32$ | kernel=(3, 3, 3)<br>strides=(1, 1, 1) |
| Activation | ReLU | |
| MaxPooling | $22 \times 22 \times 10 \times 32$ | size=(3, 3, 3) |
| Dropout | 0.15 | |
| Conv3D 3 | $22 \times 22 \times 10 \times 64$ | kernel=(3, 3, 3)<br>strides=(1, 1, 1) |
| Activation | ReLU | |
| Conv3D 4 | $22 \times 22 \times 10 \times 64$ | kernel=(3, 3, 3)<br>strides=(1, 1, 1) |
| Activation | ReLU | |
| MaxPooling | $8 \times 8 \times 4 \times 64$ | size=(3, 3, 3) |
| Dropout | 0.15 | |
| Flatten | 16384 | - |
| Dense 1 | 256 | - |
| Activation | ReLU | |
| Dropout | 0.35 | |
| Dense 2 | 5 | - |
| Activation | SoftMax | |

The network input has been tuned to be a set of 30 depth images (corresponding to one second of video) with a size of 64x64 pixels. The length of the video-segments (one second) has been set experimentally to get a compromise between the processing time, which increases as the length of the video-segment grows, and the precision of the action recognition procedure. Since images from the dataset have a size of 512x424 pixels, a preprocessing stage is required in which the original depth input images are cropped to a square image and then resized to the required dimensions.

The system has been designed to operate through a long sequence where multiple actions may appear (as in a real scenario). To achieve this goal, the proposal has to be able to recognize an action with just a part of its execution. In addition, since the network input is a group of 30 depth images, the input sequence has to be divided into 30-frames segments that are analyzed by the 3D-CNN. The designed algorithm, thus, employs a sliding window to select the frames to be inserted at each time to the network. The window has a size of 30 frames with a stride of 15, creating an overlap of 15 frames between consecutive input sets. Each 30-frame segment is then processed to obtain a final vector including its classification probability of belonging to each of the five possible actions.

In the proposed network (in Figure 1), the first two layers (*"Conv3D 1"* and *"Conv3D 2"*) are convolutional layers with 32 filters each, and a kernel of dimensions (3,3,3). These two layers extract spatio-temporal features from the input depth sequence. To avoid reducing the dimensionality here, padding is used in both of them. Next, there is a pooling layer (*"Max Pooling 1"*) for reducing the dimensionality. The output tensor generated is then introduced into another pair of convolutional layers *"Conv3D 3"* and *"Conv3D 4"* with 64 filters each and a kernel of dimensions (3,3,3), in order to extract features with a higher level of abstraction. Again, padding is used at this stage to avoid dimensionality reduction. These high-level features are then introduced in the *"Max Pooling 2"* layer to reduce their dimensionality, followed by the layer *"Flatten"* and two dense layers, *"Dense 1"* and *"Dense 2"*, with 256 and 5 neurons respectively.

The output of *"Dense 2"* layer is a vector (*S(a)* in equation 1), including 5 classification probability real values ($a_1$ to $a_5$, one for each action). This vector is then introduced into the final *"Softmax"* layer that normalizes it through the probability distribution in equation 2, in order to obtain the final 5 probabilities ($S_1$ to $S_5$) in equation 1.

$$S(a) = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_5 \end{bmatrix} \rightarrow \begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_5 \end{bmatrix} \qquad (1)$$

$$S_j = \frac{e^{a_j}}{\sum_{k=1}^{9} e^{a_k}} \quad j \in 1..5 \qquad (2)$$

The four "Conv 3D" and "Dense 1" layers use ReLU (*"Rectified Linear Unit"*) activation functions at their outputs, due to its high efficiency. Moreover, this kind of activation provides the non-linearity necessary in the classifier for action detection. It is also used *"Dropout"* to ignore nodes in the network randomly at each training stage, to prevent over-fitting in the learning process.

## 3 TRAINING STAGE

In this work, the chosen dataset for training and evaluating the proposal is GOTPD3 (Macias-Guarasa et al., 2018), recorded and labeled by the authors. It includes 34 video sequences, with a resolution of 512x424 pixels, which represent 7 different people performing 5 different daily activities. At each frame, pixels represent the distance from a 3D point in the scene to the sensor in millimeters, coded in 16 bits unsigned integer.

The actions in the dataset correspond to behaviours that can be commonly seen in the smart spaces scenarios of interest: 'Walk at normal speed', 'Walk fast', 'Run', 'Fall down' and 'Stand up for a while under the sensor'. Thus, the proposed system has been trained to recognize other actions of the dataset in addition to the falling one.

Each sequence has a single ground truth label associated although more than one action may appear along it. This occurs with actions 'Fall down' and 'Stand up', where there is a part of the sequence with a person walking before and after performing the main action. Since the network analyzes only a part of the sequence at a time, and not the whole one, some of its images do not belong to the main action.

For training, each input sequence should include only one action. Hence, it has been necessary to trim the training sequences. In consequence, the frames in which the person is walking in sequences labeled as 'Standing up' and 'Falling down' are removed.

Information about the number of frames in final trimmed sequences is shown in Table 2. Its column "Original" presents the total number of frames of the sequences corresponding to each action in the dataset before any process is carried out, whereas the column "Trimmed" shows the number of frames after the trimming process.

Since the dataset size is small, data augmentation has been needed to perform a correct training. In order to do that, images in each sequence have been mirrored vertically, horizontally and vertically plus horizontally at the same time. After this data augmentation process, the dataset contains 136 video-sequences where people may appear from different angles and go through different paths, which, in addition, improves the generalization of the network and increases its invariability against rotations. The result in the number of frames can be seen in the column "Augmented" of Table 2.

Table 2: Number of frames for each action in the original dataset (column 2), after the trimming process (column 3) and after the data augmentation (column 4).

| Action | Original | Trimmed | Augmented |
|---|---|---|---|
| Walk | 2820 | 1324 | 5296 |
| Run | 1377 | 462 | 1848 |
| Walk fast | 1737 | 690 | 2760 |
| Fall down | 1932 | 711 | 2844 |
| Stand up | 1883 | 546 | 2184 |
| Total | 9749 | 3733 | 14932 |

In Table 2, it can be observed that the number of frames in the training dataset varies significantly from one action to another. Action 'Walk' includes double of frames (∼5300) than 'Walk fast' or 'Fall down' (∼2800), which has a significantly larger number of frames than 'Run' and 'Stand up' (∼2000). This imbalance in the classes size makes the network predicting some classes more often than others, leading to wrong classification results. To avoid that, a weighting of the classes is done (Blamire, 1996).

As mentioned in previous section, images are preprocessed before being inserted in the neural network. As in other works, such as (Lin et al., 2015), first, each image is cropped out to a square shape of 355 × 355 pixels, then, it is resized to the network input size ($64 \times 64$ pixels). Figure 2 shows some examples of different input images (on the left) and their corresponding cropped ones (on the right column), including several people performing diverse actions. In particular, the first and second images correspond to a person walking, whereas the third and fourth ones presents a person falling down and standing up respectively.

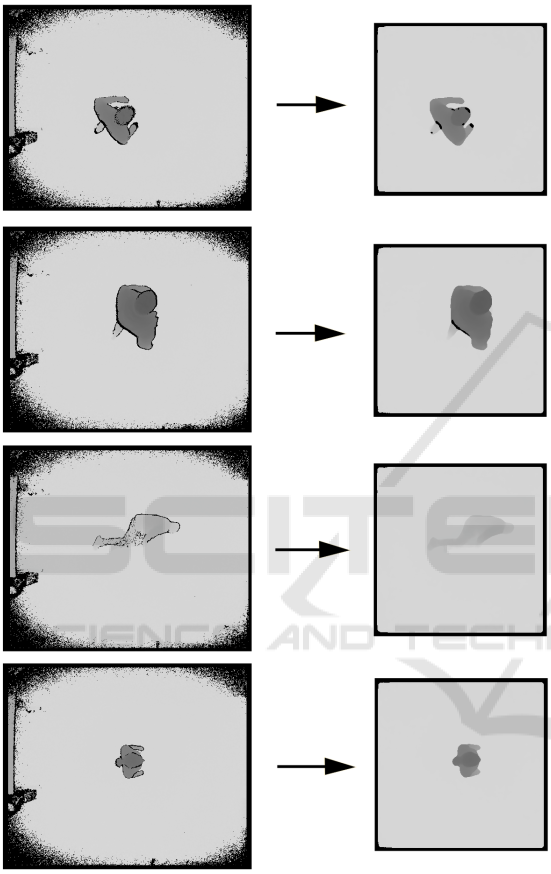

Figure 2: Before and after the preprocessing.

The dataset is divided into three separated parts for training, validating and testing the proposal. First, 75% of the available sequences is used for training (50%) and validating (25%) the neural network, whereas the other 25% is employed for testing and obtaining the experimental results. Each part contains sequences from the original dataset and from the augmented one, which increases the system robustness, by improving its generalization capability. The final number of frames of each division are shown in Table 3, that also includes a number (code) identifying each action in the dataset.

Table 3: Number of frames per action for each stage of the proposal.

| Action | Code | Train | Val. | Test |
|---|---|---|---|---|
| Walk | 000 | 2648 | 1324 | 1324 |
| Run | 001 | 913 | 473 | 462 |
| Walk fast | 002 | 1380 | 690 | 690 |
| Falling down | 003 | 1350 | 783 | 711 |
| Standing up | 004 | 1092 | 546 | 546 |
| Total | - | 7383 | 3816 | 3733 |

To evaluate the performance of the network, it has been employed the Categorical Cross-Entropy (CCE) as the loss function to compare the predictions made by the system with the ground-truth, using equation 3. The value $P_i$ is the network output after the *"SoftMax"* layes, that is, the proposal predicted probability for each class, whereas $t_i$ is the *one hot* vector value that represents the corresponding correct class in the ground-truth.

$$CCE = -\sum_{i=0}^{4} t_i log\left(P_i\right) \qquad (3)$$

To optimize the loss function, it has been chosen the algorithm *Adam* (Kingma and Ba, 2014), because it provides an adaptive method to adjust the learning rate based on its gradient first and second moments. This method starts with a learning rate provided by the user and extracts information of the loss function to calculate the next learning rates. Thus, the training speed is boosted allowing a fastest training phase.

The batches for the network are produced by a generator which selects sets of 30 frames and adapts them to the network input. The sets are selected randomly among the sequences related to each stage. The batch size is set experimentally to 25, through an exhaustive analyses, taking into account the size of the dataset. This value may be enough for Adam algorithm to get accurate information for the adaptive learning rate calculus. To make sure Adam is training correctly, an external decay factor is applied to the learning rate.

Figure 3 and 4 show the accuracy and loss function of the proposed model. By comparing training (in blue color) and validation (in orange) results in each chart it can be concluded that the system does not overfit and achieves a high level of generalization during the training stage. Moreover, it can be observed that the system stabilizes at 60 epoch with optimal performance.

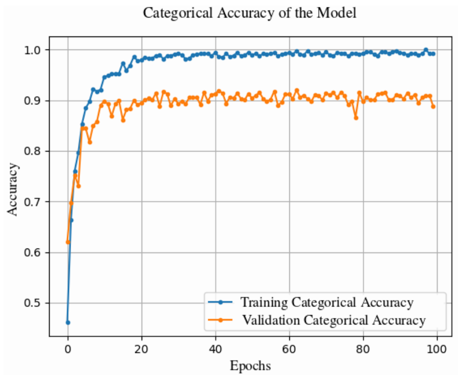

Figure 3: Categorical accuracy curves obtained in the training stage.

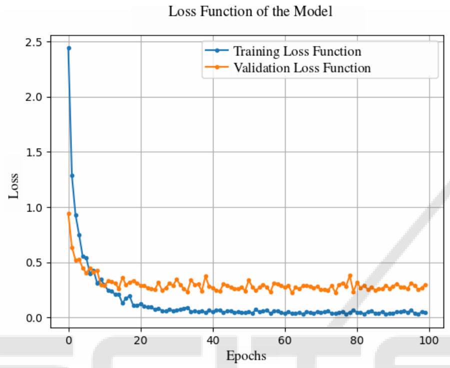

Figure 4: Loss function curves obtained in the training stage.

# 4 EXPERIMENTAL RESULTS

As it has been said in the introduction, as far as the author are concerned, there are no other public dataset for action recognition from top-view depth data available for comparison. Due to that, in order to evaluate the fall-detection system, the GOTPD3 (Macias-Guarasa et al., 2018) dataset has been used, which has been labeled and recorded by the authors and made available to the scientific community.

The have been used the 25% of the GOTPD3 videos reserved for test. Due to the fact that the system also provides an action detection feature that allows the recognition of multiple activities, the results for the complete functionality are shown first. Further on, fall results will be analyzed in depth. Finally, it is addressed the results for the other actions.

The performance of the system has been measured through its accuracy in the test data. The accuracy has been computed comparing the predicted label of each set of 30 the frames with the label assigned to the whole sequence. It is worth highlighting that, at this point there may appear problems if the input sequences include actions different from the one labeled, for example if a falling down sequence include some frames in which the person is walking. This could be solved assigning to each video the label that obtained for most of the segments, instead of providing an individual label for each of these segments. However, it would prevent the proposal for working in a real scenario in which multiple actions may exist in a given period of time.

As it has been explained before, each video-sequence is processed using a sliding window, obtaining a result (label) for each window. This sliding window has a size of 30 frames and a stride of 15, hence some frames are processed in two different windows, obtaining two possible results. If the results in the two analysis differ, the probabilities for each action are obtained by computing the arithmetic mean over the two results. Then, the system assigns the label of the action with the highest calculated probability.

Regarding the computational cost of the proposal, the average time for processing a 30-frames segment of video (that corresponds to 1 second) is 7 milliseconds on a conventional Linux desktop PC, with a Processor Intel®Core(TM) i7-6700K CPU @ 4.00 GHz with 64 GB of RAM, and an NVIDIA GTX-1080 TI GPU. Furthermore, the maximum time it takes to process a 30-frames segment is 18 milliseconds, thus, it is able to work in real time.

The results of the test are shown in Figure 5. In this confusion matrix, the true actions are presented in the rows whereas the predicted actions are shown on the columns. In this image, each action has been represented by their code from Table 3.

It can be observed that all classes have an accuracy over 75%. In particular, 'Fall-down' (003) are detected 84% of the time of the sequences. It should be noted down that the system barely confuses 'Fall down' with 'Walk' (000), and when it happens it is always due to the fact that the person walks before and after falling down. It is important to highlight that, although there can be some wrong detections at the beginning and ending of the falling sequences, the system always correctly detects the 'Fall down' in the middle frames of every falling sequence. In addition, it should be pointed out that there are close to none false positives, making the falling detections extremely reliable. The same situation happens with the action 'Walk+Stop' (004), where the person walks before and after stopping below the camera.

The worst accuracy corresponds to the 'Walk fast' action (002). Here, the confusion appears with 'Walk' (000) and 'Run' (001), which can be expected since the difference lies in the walking speed of the person. Thus, if someone walks faster or slower than the average the system might confuse this action with

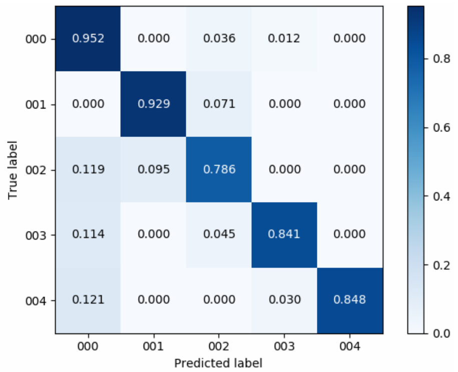

Figure 5: Confusion matrix of the 3D-CNN depth based action recognition classification proposal.

'Run' (001) or 'Walk' (000) respectively. The accuracy would improve significantly by fusing the actions: 'Walk', and 'Walk fast' or 'Walk fast' and 'Run' in the same class, but since this work is focused on falling detection, specially for old people, and it has been proved that the walking speed is strongly related with the general health of people (Sundelin et al., 2015), we have considered that it is important differentiating between the three classes.

The results obtained during the test stage allows validating the proposal, showing a high accuracy despite the complicated task. In particular, the fact that the system always detect a fall in falling sequences and that there are almost no false positives, make this system a great option for falling detection.

In addition, since the only information used by the network comes from depth maps, this proposal protects people privacy, not intruding in their daily life. Also, other actions which may be useful to recognize are detected by the network with a high accuracy, complementing the main function of the system.

Overall, the proposal achieves great results taking into account the difficulty introduced by the use of overhead position as observation point of view, and the use of only depth information to complete the privacy protection of the surveillance system.

## 5 CONCLUSIONS

This paper propose a fall detection system based on a 3D Convolutional Neural Network (3D-CNN) which employs convolutions over spatial and temporal dimensions to predict the action in the scene from only depth information acquired by a RGB-D sensor placed in an overhead configuration.

The use of depth maps and the top-view position of the sensor make it possible to preserve the privacy of the people, since their identity cannot be recognized. This fact is crucial due to the application scope, where some privacy requirements may be demanded (private houses, nursing homes, hospitals, etc.)

The evaluation of the system has been carried out using the GOTPD3 dataset, provided by the GEINTRA research group, that has been available to other researches. The results allow us to validate the proposal with a 84% accuracy in falling detection (being most of the errors due to frames in which the person appears walking in sequences labeled as 'Fall down'). Moreover, considering the results for the whole sequences (not for each 30-frame window), in the 100% of the falling sequences the falling down action is correctly detected in more than one 30-frame window. It is important to note that these results are achieved with close to none false positives. The system also predicts other actions which may be interesting to recognize, such as walking or running, with an accuracy over 90% for those two actions and around 80% for the others. In addition, the proposal can work in real scenarios, since it allows obtaining results for each 30-frames video segment, hence processing sequences that includes different actions.

The fall detection is an open research line where people are still working on with multiple fronts such as the reduction of false positives through different techniques, the increase of generalization during the training phase (so it can be applied in different environments with a higher accuracy) and the implementation and test in real environments.

Besides, another line of future work is the time labeling of the dataset, in order to have information about the starting and ending times for each action in the available sequences, which will allow improving the experimental evaluation.

## ACKNOWLEDGEMENTS

This work has been supported by the Spanish Ministry of Economy and Competitiveness under project HEIMDAL-UAH (TIN2016-75982-C2-1-R), by the University of Alcalá under project ACERCA (CCG2018/EXP-029) and under the Initiation to Research Grant.

## REFERENCES

Amin, M. G., Zhang, Y. D., Ahmad, F., and Ho, K. D. (2016). Radar signal processing for elderly fall detection: The future for in-home monitoring. *IEEE Signal Processing Magazine*, 33(2):71–80.

Ballinger, C. and Payne, S. (2002). The construction of the risk of falling among and by older people. *Ageing & Society*, 22(3):305–324.

Banerjee, T., Enayati, M., Keller, J. M., Skubic, M., Popescu, M., and Rantz, M. (2014). Monitoring patients in hospital beds using unobtrusive depth sensors. In *2014 36th Annual International Conf. of the IEEE Engineering in Medicine and Biology Society*, pages 5904–5907. IEEE.

Baptista-Ríos, M., Martínez-García, C., Losada-Gutiérrez, C., and Marrón-Romera, M. (2016). Human activity monitoring for falling detection. a realistic framework. In *2016 International Conf. on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–7.

Blamire, P. A. (1996). The influence of relative sample size in training artificial neural networks. *International Journal of Remote Sensing*, 17(1):223–230.

Chou, E., Tan, M., Zou, C., Guo, M., Haque, A., Milstein, A., and Fei-Fei, L. (2018). Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *arXiv preprint arXiv:1811.09950*.

De Miguel, K., Brunete, A., Hernando, M., and Gambao, E. (2017). Home camera-based fall detection system for the elderly. *Sensors*, 17(12).

dipakkr (2018). 3d-cnn for action recognition. https://github.com/dipakkr/3d-cnn-action-recognition.

Gia, T. N., Sarker, V. K., Tcarenko, I., Rahmani, A. M., Westerlund, T., Liljeberg, P., and Tenhunen, H. (2018). Energy efficient wearable sensor node for iot-based fall detection systems. *Microprocessors and Microsystems*, 56:34–46.

Ji, S., Xu, W., Yang, M., and Yu, K. (2012). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Lapierre, N., Neubauer, N., Miguel-Cruz, A., Rincon, A. R., Liu, L., and Rousseau, J. (2018). The state of knowledge on technologies and their use for fall detection: A scoping review. *International journal of medical informatics*, 111:58–71.

Liang, B. and Zheng, L. (2015). A survey on human action recognition using depth sensors. In *2015 International Conf. on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8.

Lin, S., Liu, A., Hsu, T., and Fu, L. (2015). Representative body points on top-view depth sequences for daily activity recognition. In *2015 IEEE International Conf. on Systems, Man, and Cybernetics*, pages 2968–2973.

Luna, C. A., Losada-Gutierrez, C., Fuentes-Jimenez, D., Fernandez-Rincon, A., Mazo, M., and Macias-Guarasa, J. (2017). Robust people detection using depth information from an overhead time-of-flight camera. *Expert Systems with Applications*, 71:240–256.

Macias-Guarasa, J., Losada-Gutierrez, C., and Fuentes-Jimenez, D. (2018). GEINTRA Overhead ToF People Detection 3 (GOTPD3) database: Human activity detection. Available online: http:// www.geintra-uah.org/datasets/gotpd3 (Last accessed: 09-Oct-2019.

Mathon, C., Beaucamp, F., Roca, F., Chassagne, P., Thevenon, A., and Puisieux, F. (2017). Post-fall syndrome: Profile and outcomes. *Annals of Physical and Rehabilitation Medicine*, 60:e50–e51.

Megavannan, V., Agarwal, B., and Babu, R. V. (2012). Human action recognition using depth maps. In *2012 International Conf. on Signal Processing and Communications (SPCOM)*, pages 1–5.

Mubashir, M., Shao, L., and Seed, L. (2013). A survey on fall detection: Principles and approaches. *Neurocomputing*, 100:144–152.

Organization, W. H. (2012). Good health adds life to years. global brief for world health day 2012. 2012.

Ozcan, K. and Velipasalar, S. (2016). Wearable camera- and accelerometer-based fall detection on portable devices. *IEEE Embedded Systems Letters*, 8(1):6–9.

Pierleoni, P., Belli, A., Palma, L., Pellegrini, M., Pernini, L., and Valenti, S. (2015). A high reliability wearable device for elderly fall detection. *IEEE Sensors Journal*, 15(8):4544–4553.

Rougier, C., Auvinet, E., Rousseau, J., Mignotte, M., and Meunier, J. (2011). Fall detection from depth map video sequences. In *International Conf. on smart homes and health telematics*, pages 121–128. Springer.

Sundelin, T., Karshikoff, B., Axelsson, E., Höglund, C. O., Lekander, M., and Axelsson, J. (2015). Sick man walking: Perception of health status from body motion. *Brain, Behavior, and Immunity*, 48:53 – 56.

Tang-Wei Hsu, Yu-Huan Yang, Tso-Hsin Yeh, An-Sheng Liu, Li-Chen Fu, and Yi-Chong Zeng (2016). Privacy free indoor action detection system using top-view depth camera based on key-poses. In *2016 IEEE International Conf. on Systems, Man, and Cybernetics (SMC)*, pages 004058–004063.

Wang, P., Li, W., Gao, Z., Tang, C., and Ogunbona, P. O. (2018). Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Transactions on Multimedia*, 20(5):1051–1061.

Wang, P., Wang, S., Gao, Z., Hou, Y., and Li, W. (2017). Structured images for rgb-d action recognition. In *Proceedings of the IEEE International Conf. on Computer Vision*, pages 1005–1014.

Wu, F., Zhao, H., Zhao, Y., and Zhong, H. (2015). Development of a wearable-sensor-based fall detection system. *Int. J. Telemedicine Appl.*, 2015:2:2–2:2.

Yang, X., Zhang, C., and Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international Conf. on Multimedia*, pages 1057–1060. ACM.

Zerrouki, N., Harrou, F., Sun, Y., and Houacine, A. (2016). Accelerometer and camera-based strategy for improved human fall detection. *Journal of Medical Systems*, 40(12):284.