# Short-term Recognition of Human Activities using Convolutional Neural Networks

Michalis Papakostas
*Department of Computer Science*
*University of Texas*
*Arlington, USA*
*michalis.papakostas@mavs.uta.edu*

Theodoros Giannakopoulos
*Institute of Informatics & Telecommunications,*
*National Center for Scientific Research Demokritos*
*Athens, Greece*
*tyiannak@gmail.com*

Fillia Makedon
*Department of Computer Science*
*University of Texas*
*Arlington, USA*
*makedon@uta.edu*

Vangelis Karkaletsis
*Institute of Informatics & Telecommunications,*
*National Center for Scientific Research Demokritos*
*Athens, Greece*
*vangelis@iit.demokritos.gr*

***Abstract*—This paper proposes a deep learning classification method for frame-wise recognition of human activities, using raw color (RGB) information. In particular, we present a Convolutional Neural Network (CNN) classification approach for recognising three basic motion activity classes, that cover the vast majority of human activities in the context of a home monitoring environment, namely: sitting, walking and standing up. A real-world fully annotated dataset has been compiled, in the context of an assisted living home environment. Through extensive experimentation we have highlighted the benefits of deep learning architectures against traditional shallow classifiers functioning on hand-crafted features, on the task of activity recognition. Our approach proves the robustness and the quality of CNN classifiers that lies on learning highly invariant features. Our ultimate goal is to tackle the challenging task of activity recognition in environments that are characterized with high levels of inherent noise.**

***Keywords*-activity recognition; ADLs; deep learning; health monitoring**

## I. Introduction

During the last years, human activity recognition has become one of the most challenging tasks in the computer vision community. The abstract nature of human generated data, along with the additional obstacles added by the physical environment (e.g., illumination, occlusion and point of view) makes the task of activity classification computational demanding and hard to generalize. Several approaches have been proposed that tried to exploit hand-crafted features to describe activities. In [1, 2] the authors proposed cuboid structures for describing spatiotemporal interest points (STIPs) extracted from such features. Those approaches are based on color image sequences (RGB color space) and the cuboid descriptors were defined over a deterministic time window. In a similar manner, in [3] the authors tried to acquire and describe STIPs extracted from depth image sequences. In another STIP-based approach, the authors in [4] combined the two information channels (depth and color) and proposed two different modality-fusion methods, offering also the first multimodal publicly available dataset on activities of daily living (ADLs). Other methods [5, 6] focused on extracting and modeling normal vectors, either from regions of interest or from whole frames. Even though the aforementioned references reported promising and highly accurate results, they significantly suffer from physical problems such as, occlusion, illumination, frame rate and point of view. Moreover, the usage of hand-crafted features and histogram-based shallow classifiers makes those techniques hard to be generalized in multiple domains.

On the the other hand, works that focus explicitly on ADLs usually tackle the problem by making unrealistic assumptions regarding the nature of the data. In [7] the authors provide valuable results and conclusions regarding the affect of the objects in the ADL recognition task. However their method is based on data acquired by a chest-mounted camera, an assumption that is both invasive and impractical, especially for the case of a home monitoring and unobtussive application. Similarly in [8, 9] the subjects are always in front of the camera sensor, acting in most cases from the same point of view, thus providing very clean and noise-free data. In [10] the authors propose a very robust and well defined framework, which is however based on hand-crafted features and fuzzy activity estimation, hence making it vulnerable when the point of view changes or in the presence of occlusion.

Our approach tries to exploit the highly invariant characteristics of deep architectures, and especially convolutional neeural networks (CNNs), to provide a more generalizable activity recognition framework that can be easily tailored for recognising ADLs. Towards that direction [11, 12], used CNNs on Depth and Binary Motion Images respectively in order to predict the performed activity. However, the datasets that both works used, even though they are very popular among previous works on activity recognition, they are quite naive and unrealistic, thus making it very hard to generalize on a more obtrusive scenario. Our work was mainly inspired by approaches such as [13, 14] or more recently [15, 16], where the authors exploited deep

architectures under more realistic experimental conditions. Such frameworks are characterized by high invariance and are very promising candidates for the demanding task of recognising ADLs.

The major contribution of the proposed work compared to the aforementioned publications is that we focus specifically on the application of home-based activity monitoring using convolutional neural networks. We aim to identify core ADL activities such as standing, sitting and walking that carry valuable information, when targeting ADLs of higher abstraction such as watching TV, eating etc., which is our far reaching goal. Moreover instead of evaluating on pre-segmented videos that represent a single target class, we adopted a frame-wise activity identification approach where different activities are swapped consistently within the same recording. This approach acts as a foundation for our next steps on temporal modeling of activities when evaluating on continuous video streams.

The motivation of the current work stems from the application domain of medical monitoring. Specifically, we base our evaluation setup on monitoring the activities of daily living (ADLs) of elder people in their home environment, using a robot that functions both as a companion and as a moving monitoring device. This work is part of the RADIO EU research program www.radio-project.eu/ that aims at utilizing a robotic platform, equipped with multimodal sensors (microphones, cameras and laser range finders), to unobtrusively monitor seniors. In order to have a guideline about what type of information is used by medical doctors to assess medical condition of interest, we use the *interRAI Long-Term Care Facilities Assessment System (interRAI LTCF)*. interRAI has been analysed previously in order to identify assessment items (mood and ADL) that are useful to medical personnel [17].

A very important type of information always assessed by medical experts monitoring senior patients consists of measures related to the ability of the subject to *move independently*. Questions like "how much time is it required for the person to move self to standing position (from sitting)?" are vital for the health assessment procedure. In that context, our goal here is to provide a fast and accurate method for basic human activity recognition that takes into account the resource and implementation restrictions, defined in such a scenario. For instance, realtime extraction of estimated classes and limited processing power and bandwidth are instructing towards a simple and straightforward solution. In this work, we have focused on three basic but important activity classes that cover the vast majority of everyday activities in the context of a home environment. In particular, our goal is to recognize the following activity classes: *sitting, standing up and walking*. It is important to emphasize that the estimation of each class is performed *per frame*, based *only on color information*. Our preference on focusing explicitly on the RGB data, springs as a requirement of the RADIO project, which demands user identification and tracking. In addition, the proposed framework is able to exploit low cost equipment (ie. regular webcam) instead of a more expensive depth sensor.

## II. Methodology

For recognising the activities, we decided to utilize a CNN classifier that describes activities in a frame-wise manner, based on RGB information. As recent literature has shown, deep hierarchical visual feature extractors can significantly outperform shallow classifiers trained on hand-crafted features and are more robust and generalizable when countering problems that include significant levels of inherent noise. The architecture of our deep CNN was initially proposed in [16]. The model is mainly based on the Caffenet [18] reference model, which is similar to the original AlexNet [19]) and the network proposed in [20]. For our experiments we used the *BVLC Caffe* deep-learning framework.

The network architecture consists of two convolution layers with stride of 2 and kernel sizes equal to 7 and 5 respectively, followed by max pooling layers. As a next step, a convolution layer with three filters of kernel size equal to 3 is applied, followed again by a max pooling layer. The next two layers of the network are fully connected layers with dropout, followed by a fully connected layer and a softmax classifier, that shapes the final probability distribution. All max pooling layers have kernel size equal to 3 and stride equal to 2. For all the layers we used the ReLu as our activation function. The output of the network is a distribution on our three target classes, while the output vector of the semifinal fully connected layer has size equal to 4096. We have adopted a 1000-iterations fine-tuning procedure, with an initial learning rate of 0.001, which decreases after 700 iterations by a factor of 10.

Since training a new CNN from scratch would require big loads of data and high computational demands, we used transfer learning to fine-tune the parameters of a pre-trained architecture. The original CNN was trained on the 1.2M images of the ILSVRC-2012 [21] classification training subset of the ImageNet [22] dataset. Following this approach, we manage to decrease the required training time and to avoid overfitting our classifier by ensuring a good weight initialisation, given the relatively small amount of available data. Finally, the data are preprocessed by augmenting the frame dimensionality to 240x320. The input to the network corresponds to the 227x227 center crops and their mirror images.

## III. Dataset

In order to train and evaluate our system, a dataset that consists of real-world recording sessions of RGBD data has been created. For data acquisition, an XTion sensor [1] has been used as part of the robotic sensoring infrastructure of the Radio platform. In total, 12 humans have participated in different 8 scenarios. Each recording session has been repeated on 1 to 3 different days (random number of repeats for each user). This has been done in order to ensure a certain diversity of lighting conditions.

[1] https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/

272 recordings have been recorded and annotation in total. In each scenario the recording starts with the user sitting on a chair, after a while he/she stands up and starts walking to the exit of the room.

The 8 (2x2x2) different scenarios per user are acquired as follow:

1) 2 walking directions on predefined pathways (landmarks have been used to make the annotation process easier)
2) 2 different lighting conditions (natural and artificial) and
3) with and without obstacles between the visual sensor and the user

During the annotation process of each video, the following information has been manually provided by the human annotators: the timestamps of the onset and offset of the standing-from-chair activity (b) the timestamp of the moment that the 4-meter distance has been covered.

In the context of this research work, we are interested in demonstrating the ability of the adopted classification method to discriminate between three basic human motion classes that are closely related to the aforementioned mobility measures. These classes are: sitting, standing up and walking. For training and evaluating our frame-based classifiers, the aforementioned videos and respective manual annotations result in a set of directories of JPEG images organized and indexed per classes and recordings. Figure 1 shows some examples of images from the compiled dataset, demonstrating the different conditions, in terms of lighting (natural and artificial), walking routes and obstacles. The dataset is also openly available, along with the annotation files at Note that this is only part of the whole activity recognition dataset, which is more than 10x times larger. For access on the full dataset please contact the authors.

## IV. Experiments

We have adopted the following types of experimentation based on the dataset described at Section III:

1) Evaluation of the frame-wise classification method using a cross-validation procedure that splits training and testing data based on individual recordings. In other words, according to that approach, all frames of each recording are either used for training or testing. Three random subsampling cross-validation repetitions have been conducted, each repetition corresponding to a different random permutation of the videos in the dataset.
2) Evaluation using a cross-validation procedure that splits training and testing data based on subject IDs. According to that setup, the frames of all videos that belong to the same person are either used for training or testing. This has been conducted in order to evaluate the method in terms of subject-independence. Again, three random subsampling cross-validation repetitions have been conducted, each repetition corresponding to a different random permutation of the subjects (humans).
3) Evaluation against different noise ratios and comparison to a Support Vector Machine classifier using hand-crafted features. Towards this end, we have added Gaussian noise of several SNR ratios to the images before testing. In addition, an SVM classifier using typical visual features (HOGs, LBPs, color histograms) has been adopted for comparison reasons. The SVM classifier has been fine-tuned in terms of the C parameter, while a linear kernel has been adopted.

Table I: Experimental setup 1: Average Initial Confusion Matrix (normalized to sum up to 100%)

| | Sitting | Standing | Walking |
|---|---|---|---|
| Sitting | 18.84 | 0.58 | 0 |
| Standing | 2.89 | 11.2 | 0.98 |
| Walking | 0.2 | 1.17 | 64.54 |

Table II: Experimental setup 1: Row-Normalized Confusion Matrix. Diagonal elements represent the respective recall rates. Note that these recall rates are not exactly equal to the recall rates presented in III: this is due to the fact that these recall - precision rates are averaged *per video recording*, not per frame.

| | Sitting | Standing | Walking |
|---|---|---|---|
| Sitting | 97.01 | 2.99 | 0 |
| Standing | 19.18 | 74.32 | 6.50 |
| Walking | 0.3 | 1.78 | 97.92 |

Table III: Experimental setup 1: Per class and average Recall, Precision and F1 and overall Accuracy, computed over all frames of the testing dataset.

| | Sitting | Standing | Walking | Average |
|---|---|---|---|---|
| Precision | 0.86 | 0.86 | 0.99 | 0.91 |
| Recall | 0.97 | 0.76 | 0.98 | 0.90 |
| F1 | 0.91 | 0.81 | 0.99 | 0.90 |
| Average Accuracy | | 0.95 | | |

Table IV: Experimental Setup 2: Average Initial Confusion Matrix (normalized to sum up to 100%)

| | Sitting | Standing | Walking |
|---|---|---|---|
| Sitting | 21.2 | 0.79 | 0 |
| Standing | 2.37 | 11.75 | 0.72 |
| Walking | 0.02 | 0.65 | 62.5 |

Table V: Experimental setup 2: Row-Normalized Confusion Matrix. Diagonal elements represent the respective recall rates. Note that these recall rates are not exactly equal to the recall rates presented in VI: this is due to the fact that these recall - precision rates are averaged *per video recording*, not per frame.

| | Sitting | Standing | Walking |
|---|---|---|---|
| Sitting | 96.41 | 3.59 | 0 |
| Standing | 15.97 | 79.18 | 4.85 |
| Walking | 0.03 | 1.03 | 98.94 |

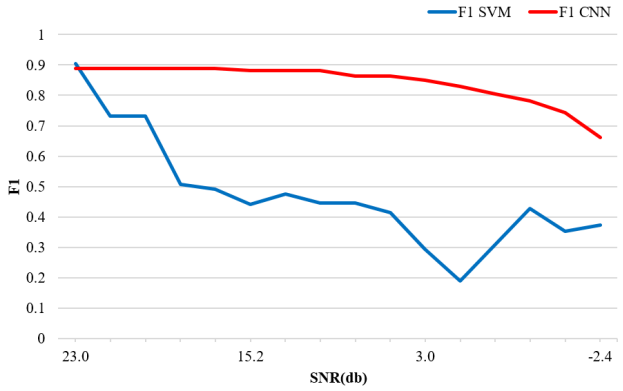
Figure 1: Examples of the compiled and annotated dataset's frames. Differnt rows correspond to separate users (humans). The first column corresponds to the "sitting" class, the second to the "sit to stand" class, while the third and fourth columns show walking examples. It can be seen that different walking directions, lighting conditions and obstacles have been used to increase diversity in conditions

Table VI: Experimental setup 2: Per class and average Recall, Precision and F1 and overall Accuracy

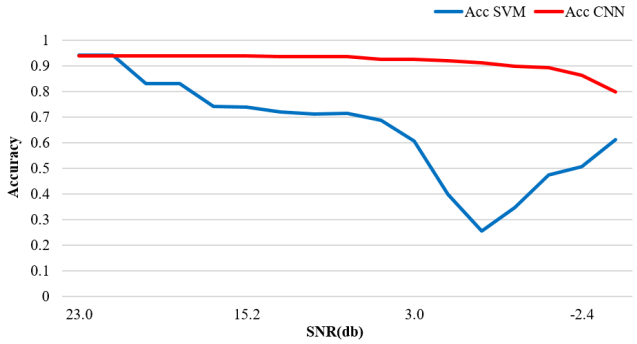
| | Sitting | Standing | Walking | Average |
|---|---|---|---|---|
| Precision | 0.9 | 0.89 | 0.99 | 0.93 |
| Recall | 0.96 | 0.79 | 0.99 | 0.92 |
| F1 | 0.93 | 0.84 | 0.99 | 0.92 |
| Average Accuracy | | 0.95 | | |

Tables I, II and III present the initial confusion matrix, the row-normalized confusion matrix and the performance measures (Recall, Precision, F1, Accuracy) respectively, for the first experimental setup. Similarly, tables IV, V and VI present the confusion matrix, the row-normalized confusion matrix and the performance measures, for the second experimental setup. These results prove that the CNN classifier is robust and independent to subject-specific characteristics, since the performance in the two first experimental setups is similar.

Finally, Figures 2a and 2b present respectively the mean F1 and accuracy measures for the two methods (CNN




(a) F1 of both SVM and CNN methods, for different Signal to Noise Ratios




(b) Accuracy of both SVM and CNN methods, for different Signal to Noise Ratios

Figure 2: CNN and SVM comparison against different noise levels. Both CNN and SVM models are trained based on the first experimental scenario. Similar behaviours were observed in the models trained on the second experimental scenario. The robustness and consistency of the CNN classifier is obvious against traditional SVM-based methods trained on hand-crafted features.

and SVM) and for different levels of signal-to-noise ratio (SNR, in dB). The CNN models illustrated in those figures are the ones trained one the first experimental scenario. We can easily infer similar behavior on the models trained on the second experimental scenario. The robustness of the CNN approach is obvious: both F1 and accuracy fall dramatically for the SVM model as the SNR ratio is reduced. On the other hand, CNN is much more robust to noise: even for 0dB SNR, the F1 measure is kept above $80\%$, while the respective measure for the SVM case is below $50\%$. In particular, the SVM classifier with hand-crafted features seems totally unstable in terms of both overall accuracy and F1 measure, for all levels of noise bellow 20dB SNR. Finally, we have also experimented using a temporal median filtering to smooth the outputs of the frame-wise classifier using various kernel sizes. However, this did not lead to any worth noticing improvements.

## V. Conclusions

We have presented a Convolutional Neural Network classification approach for frame-wise recognition of three basic activity classes, that cover the vast majority of human activities in the context of a health monitoring environment. A real-world dataset has been compiled and annotated in a smart home environment and experimentation has highlighted the benefits of deep learning architectures against traditional shallow classifiers functioning on hand-crafted features. Our ongoing research efforts on the subject focus on extending the deep learning approach to more detailed activity taxonomies that will manage to describe a greater variety of ADL-related classes. In particular, we are currently implementing an important extension of the dataset described in the current paper, focusing on more frame-wise activity classes as well as high-level activity classes (e.g. cooking, wathcing TV, preparing food, drinking, etc). Additionally, we are working on modelling temporal models of the visual information, in order to both boost the performance of the classifiers, but also to extract correlations between long-term ADLs and short-term human activity classes. Finally, we conduct research towards adopting other modalities in the activity recognition cues. In particular, the we are building a CNN that also takes into consideration depth information from the visual sensor. In addition, audio-based decisions will be extracted from the audio channel in order to be fused in a late fusion scheme that extracts high level events based on several modalities.

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 643892 and is also based upon work supported by NSF under award numbers CNS 1338118, 1035913 Please see http://www.radio-project.eu for more details.

## References

[1] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[2] Konstantinos Avgerinakis, Alexia Briassouli, and Ioannis Kompatsiaris. Activity detection and recognition of daily living events. In *Health Monitoring and Personalized Feedback using Multimedia Data*, pages 139–160. Springer, 2015.

[3] Lu Xia and JK Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841, 2013.

[4] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[5] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723, 2013.

[6] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–811, 2014.

[7] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012.

[8] Chenyang Zhang and Yingli Tian. Rgb-d camera-based daily living activity recognition. *Journal of Computer Vision and Image Processing*, 2(4):12, 2012.

[9] Jie Fu, Chengyin Liu, Yen-Pin Hsu, and Li-Chen Fu. Recognizing context-aware activities of daily living using rgbd sensor. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2222–2227. IEEE, 2013.

[10] Tanvi Banerjee, James M Keller, Mihail Popescu, and Marjorie Skubic. Recognizing complex instrumental activities of daily living using scene information and fuzzy logic. *Computer Vision and Image Understanding*, 140:68–82, 2015.

[11] Pichao Wang, Wanqing Li, Zhimin Gao, Jing Zhang, Chang Tang, and Philip Ogunbona. Deep convolutional neural networks for action recognition using depth map sequences. *arXiv preprint arXiv:1501.04686*, 2015.

[12] Tushar Dobhal, Vivswan Shitole, Gabriel Thomas, and Girisha Navada. Human activity recognition using binary motion image and deep learning. *Procedia Computer Science*, 58:178–185, 2015.

[13] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[14] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[15] Kishore Konda, Pramod Chandrashekhariah, Roland Memisevic, and Jochen Triesch. Real-time activity recognition via deep learning of motion features. In *Proceedings*, page 427. Presses universitaires de Louvain, 2015.

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[17] RADIO Project. D2.2: Early detection methods and relevant system requirements. Available at http://radio-project.eu/deliverables, 2015.

[18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.