A Novel CNN-LSTM Hybrid Architecture for the Recognition of Human Activities

Sofia Stylianou-Nikolaidou[1], Ioannis Vernikos[2,3], Eirini Mathe[3,4],
Evaggelos Spyrou[2,3](✉), and Phivos Mylonas[4]

[1] Department of Informatics and Telecommunications,
University of Athens, Athens, Greece
sdi1400195@di.uoa.gr

[2] Department of Computer Science and Telecommunications,
University of Thessaly, Lamia, Greece
{ivernikos,espyrou}@uth.gr

[3] Institute of Informatics and Telecommunications, National Center for Scientific Research – "Demokritos", Athens, Greece
emathe@iit.demokritos.gr

[4] Department of Informatics, Ionian University, Corfu, Greece
fmylonas@ionio.gr

Abstract. The problem of human activity recognition (HAR) has been increasingly attracting the efforts of the research community, having several applications. In this paper we propose a multi-modal approach addressing the task of video-based HAR. Our approach uses three modalities, i.e., raw RGB video data, depth sequences and 3D skeletal motion data. The latter are transformed into a 2D image representation into the spectral domain. In order to extract spatio-temporal features from the available data, we propose a novel hybrid deep neural network architecture that combines a Convolutional Neural Network (CNN) and a Long-Short Term Memory (LSTM) network. We focus on the tasks of recognition of activities of daily living (ADLs) and medical conditions and we evaluate our approach using two challenging datasets.

Keywords: Human activity recognition · Convolutional neural networks · Long short term memory networks · Multimodal analysis

1 Introduction

Human activity recognition (HAR) has attracted increasing research attention over the last years. Evidently, it consists one of the most prominent computer vision tasks, due to its many applications in e.g., video surveillance, assisted living, human-machine interaction, affective computing, etc. Recently, deep learning approaches, especially based on deep Convolutional Neural Network (CNN) architectures have been widely used for video-based human activity recognition, outperforming the majority of traditional machine learning approaches.

L. Iliadis et al. (Eds.): EANN 2021, INNS 3, pp. 121–132, 2021.
https://doi.org/10.1007/978-3-030-80568-5_10

Although a vast amount of research has been conducted on improving recognition performance, several principal challenges, such as the representation and the analysis of actions, still remain unresolved.

Additionally, with the advent of cost-effective sensors such as Microsoft Kinect, depth data have become available. This way, several challenging human activity datasets now provide multi-modal raw data, i.e., consisting of RGB video and depth information. The latter has also allowed for the extraction of a third modality, i.e., skeleton sequences that consist of 3D coordinates of human joints over time. Therefore, a large number of training videos are now an option for training deep neural network (NN) architectures. Note that the depth modality, unlike the conventional RGB, is invariant to illumination changes and also reliable for the estimation of body silhouettes. Nevertheless, RGB information contains colour and texture which are significant for discriminating several actions involving e.g., human-object interactions. Different modalities offer different perspectives of actions, thus, intuitively, a fusion of their complementary correlations should be meaningful. Moreover, the existence of skeletal information can be very helpful for accurately capturing the human body posture. However, in scenarios where the source of motion features is limited to sequence data, the challenge of CNN-based methods is to find efficient encoding techniques for representing skeleton sequences, while capturing spatio-temporal activity features.

In this paper we present a novel approach that utilizes multiple modalities for human activity recognition and incorporates RGB, depth and a visual representation of skeletal information. The latter has been proposed in our previous works [17,18] and is based on the Discrete Fourier Transform (DFT). More specifically, skeleton sequences are transformed to a sequence of 2D pseudocolored images for five subsets of skeletal joints corresponding to arms, legs and the trunk. All available modalities are then used for learning features using a hybrid network architecture that combines a CNN with a Long Short Term Memory network (LSTM). The extracted features are then fused and used for classification. Our proposed method is evaluated on subsets of two challenging 3D activity recognition datasets, namely a) the PKU-MMD dataset [13] on activities of daily living; and b) the NTU RGB+D [20] on medical conditions.

The rest of this paper is organized as follows: Sect. 2 presents recent research works closely related to the proposed approach. Section 3 presents the proposed visual representation of skeleton sequences and the hybrid CNN-LSTM architecture. The structure of the experiments and their results are presented in Sect. 4, while conclusions and plans for future work are included in Sect. 5.

2 Related Work

In this section, we briefly review recent scientific literature on HAR using deep learning. Similar to this work, we focus on two categories, a) methods for depicting skeletal information by image-based representations; and b) models that utilize information from multiple modalities using CNNs and LSTMs.

2.1 Visual Skeletal Representations

Huynh-The et al. [5] proposed a technique named, "pose-transition feature to image," transforming skeletal information from video sequences to skeleton-based images. Two geometric features are extracted, i.e., joint-joint distance and joint-joint orientation. For each video frame, a row vector F is composed with four normalized values. The first two values correspond to the distance and orientation between two arbitrary joints, while the last two values correspond to the distance and orientation between two arbitrary joints of two consecutive frames, respectively. The RGB skeleton-based image is formed by stacking the feature vectors F of all skeleton frames in the video, and encoding the normalized values as color pixels. Similarly, Pham et al. [19] represented the skeletal information with an enhanced action map, named "enhanced Skeleton Posture-Motion Feature" (SPMF). Joint motion and posture are encoded using the aforementioned joint features, the Euclidean distance between two joints and the joint-joint orientation. Additionally, the Adaptive Histogram Equalization (AHE) which is a color enhancement method, is adopted for increasing contrast and highlighting the texture and edges of the motion maps.

In the work of Wang et al. [22], a representation called "joint trajectory maps" (JTM) was proposed, wherein skeleton data sequences are represented by three 2D images. The motion dynamics are captured as the image's texture and color. Specifically, motion direction is reflected as hue in the colored image, different body parts are represented by multiple color maps and last but not least, the motion magnitude of joints is reflected by the image's saturation and brightness. Another similar approach to representing skeletal information as color texture images is the method of "joint distance maps" (JDM), proposed by Li et al. [10]. Therein, pairwise distances of joints generate four JDMs. The first three maps correspond to distances in the three orthogonal planes (xy, yz and xz), while the fourth encodes distances calculated in the 3D space (xyz). The existence of the fourth JDM improved the robustness of the method when tested on multiple viewpoints. Again, the image's hue expresses the variations of joint distances.

Furthermore, Li et al. [11] presented a deep learning model that preserves spatial and temporal features, taking advantage of both LSTM and CNN architectures. Inspired by [19] and [22], spatial domain features and temporal domain features are extracted from skeleton sequence data. The former consist of relative position distances between joints and distances between joints and the lines connecting two joints. Temporal domain features are generated using the aforementioned methods to construct JDMs and JTMs. Spatial domain features are used as input to LSTM networks, while temporal domain features are used to train a CNN. To fuse resulting feature vectors a multiply score fusion is adopted.

Liu et al. [14] proposed a spatio-temporal representation for skeleton sequences that also incorporates the various durations of the different actions performed. The constructed images are based on a three-channel image patch, termed "Skepxel," which is composed by arranging the indices of the skeleton joints in a 2D grid and encoding their coordinate values along the third dimension. Each Skepxel might have a different arrangement of the joints but in order to keep the representation of the skeleton sequence compact, only a few, highly relevant arrangements are selected. A group of Skepxels are generated for a single skeleton frame and the final image is compactly constructed by concatenating the group of skepxels in a column-wise manner for an N-frame sequence.

An approach addressing the problem of view invariance was presented in the work of Liu et al. [15] where a transformation creating a 5D representation of joints has been proposed. They adopted a 5D space, consisting of the 3D coordinates of each joint and the additional two dimensions of time and joint label. Upon projection to a 2D image using the dimensions of time and joint label, the remaining three dimensions are used as R, G, B, channels to form pseudo-colored images. Similarly, Yang et al. [23] proposed a "tree structure skeleton image" (TSSI), based on the idea that spatially related joints in original skeletons have direct graph links between them. For their method, human skeleton graph structure is rearranged using a depth-first tree traversal order and therefore, the spatial correlations between joints are better preserved. Hou et al. [6] introduced an image-based representation called "joint skeleton spectra." The joint distribution maps are projected onto three Cartesian planes, reflecting the temporal variation of a skeleton sequence to hue values. Finally, Ke et al. [9] proposed a skeletal representation capable of extracting translation, rotation and scale invariant features. In their method, five subsets of joints are selected to represent the following body parts, arms, legs and trunk. For each body part the cosine distances and the normalized magnitudes are calculated, creating two feature arrays, which are then transformed into gray scale images.

2.2 Multimodal Methods

In the work of Zhu et al. [24] both RGB and depth modalities are exploited for gesture recognition. Short term spatio-temporal features are learnt by a 3D CNN and then, long term spatio-temporal features are learnt based on the extracted features with the use of convolutional LSTM networks. Moreover, Haque et al. [3] follow an early fusion approach of RGB, Depth and thermal information, to capture complementary facial features related to pain. For feature extraction, a CNN-LSTM model is employed. Imran et al. [7] presented a multi-stream network for human action analysis, leveraging CNN and RNN networks, where features from RGB, depth and inertial data are incorporated. Sun et al. [21], fused feature elements of RGB and depth information which are then learnt through an enhanced two-stream LSTM network, called "Lattice-LSTM." A memory cell jointly trains both input gates and output gates, integrating motion patterns and temporal dependencies. In addition, Liu et al. [16] compensated with viewpoint variations, by utilizing RGB and depth information. Dense trajectories are extracted from RGB frames which are then processed by a non-linear knowledge transfer model for learning invariant features. Simultaneously, the depth stream is filtered by a CNN and a Fourier temporal pyramid is applied on the extracted features. The fusion of the invariant features is further used to train an L1-L2 classifier. Li et al. [12] suggested a method that uses three modalities from real-world data, specifically, depth information, microphone and RFIDs mobile sensors, for recognizing concurrent human activities. Each of the aforementioned modalities is processed by a CNN, followed by an LSTM for extracting spatial and temporal features respectively, which are later fused for the classification step. Finally, Hazirbas et al. [4] proposed an encoder-decoder network architecture for semantic labeling of indoor scenes. Although their approach is not addressing HAR, both RGB and depth information is taken into account and fused in the following way: two encoding branches are used for each modality and a fusion block consolidates the produced feature maps.

3 Methodology

The proposed method incorporates three different modalities, i.e., RGB, depth and skeletal information. The latter consists of the 3D motion of a set of 25 human skeleton joints. These modalities are processed independently through a deep hybrid architecture which is based on a 2D CNN and a LSTM network. To effectively leverage the capability of the hybrid network in mining discriminative features for the problem of recognition, skeletal data are encoded into five pseudo-coloured images, termed "activity images," each corresponding to a body part. In other words, a given activity performed by a human subject is represented by five sequences of activity images. Once again, these sequences are processed separately through the network and finally, the produced feature maps from all modalities are combined using a late fusion approach. In brief, the proposed approach consists of the following key components: a) input preprocessing and construction of the activity images; b) a 2D hybrid CNN-LSTM network architecture; and c) fusion of the produced feature maps.

3.1 Input Pre-processing

Firstly, input video sequences are resized. Although the resolution of RGB and depth videos is high (i.e., 1920×1080 and 512×424, respectively), following the common good practices we resized both modalities to 213×120 and 128×106, respectively. In order to create a more diverse training set, so as to reduce overfitting during the training procedure, we adopted a data augmentation strategy. More specifically, datasets were augmented by using random crops (wherein a random subset of a given image is created, while its original aspect ratio is preserved) and horizontal flips both for RGB and depth modalities. Finally, since the duration of different activities and of the performing speed of a given activity between different subjects vary, we downsampled all input video sequences into a fixed length of 15 frames, fulfilling the need for a fixed input size.

3.2 Skeletal Information

The skeletal input data consist of 3D spatial human joint coordinates, captured using Microsoft Kinect v2 cameras. In more detail, during the performance of a given action the Kinect sensors record the 3D position over time of 25 joint coordinates (x,y,z), for each detected human body in the scene. The human skeleton is modelled as a graph of skeletal joints and edges. Each joint corresponds to a body part such as head, shoulder, knee, etc. while edges connect these joints shaping the body structure. In the proposed method, we divide the human skeleton into five main body parts, namely, head-torso, right arm, left arm, right leg and left leg. For each of these joint groups we construct a sequence of activity images to be used as an input for our hybrid network. Skeleton joints and the selected body parts are illustrated in Fig. 1.

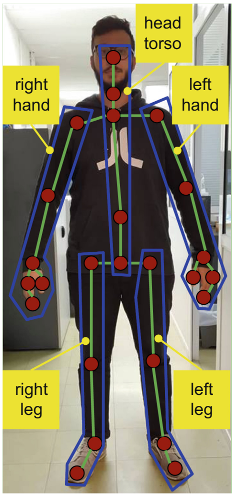

Fig. 1. The 25 skeletal joints divided into five main body parts

3.3 Activity Images

The activity images we use to capture spatiotemporal properties of skeletons have been partially inspired by the work of Jiang et al. [8], who presented a similar concept for the representation of signal sequences obtained by accelerometers and gyroscopes. Moreover, Papadakis et al. in [17,18] composed an activity image representing a single activity, wherein joint coordinates are considered as three separate 1D signals; this way a given video sequence consists of 75 1D signals. All signals for a given activity sample are concatenated forming a signal image which is transformed to the activity image by a spectral transformation.

In this work we propose a quite similar implementation of activity images. We remind that they are constructed so as to be visual representations of 3D human body motion over time. In other words, they capture both spatial and temporal dependencies that are reflected by the color and texture of the images. Joint coordinates are likewise considered as 1D signals and a spectral transformation is utilized to form the final activity image. However, a key difference in the presented approach is that each action is depicted by five sequences of activity images corresponding to five body parts, i.e., arms, legs and head-torso. To this goal, joint coordinates are appropriately grouped based on the body part to which they belong. This way, head-torso is associated with the joints head, neck, spine-shoulder, spine-mid and spine-base. Arms are associated with spine-shoulder left/right-shoulder, elbow, wrist, thumb, hand and hand-tip. Lastly, legs are related with hip, knee, ankle and foot.

Each video frame is translated as a row vector, composed of the 3D joint coordinates (x,y,z). These row vectors are then concatenated, composing a signal image. Image width is equal to the number of joints participating, while image height is equal to the number of frames. The height parameter is affected by the user-defined sequence length: to create a sequence of N activity images for a single action, it is necessary that frame rows are split into N equal sets. Larger sequence length induces less frames for each image and thus, a smaller height. It is also worth mentioning that joint coordinates are arranged in chronological order, i.e., the first row corresponds to the first video frame etc. As a result, a sequence of signal images corresponds to representations of consecutive segments of an activity. Furthermore, since performed activities suffer from temporal variations, an interpolation step is required and a set of signal images of fixed height and sequence length is created for each body part. Finally, activity images are constructed by imposing the 2D Discrete Fourier Transform (DFT) on the interpolated signal images, discarding their phase.

3.4 Network Architecture

As it has already been mentioned, the proposed approach is based on a 2D hybrid CNN-LSTM network. The motivation for our approach is that a) CNNs have been widely used for learning *spatial* features; and b) LSTM networks have been successfully used for *sequential modeling*. Our approach combines both network architectures for capturing *spatiotemporal* correlations. Particularly, the different typed data (RGB, Depth, body-part based Activity Images) are filtered by a 2D-CNN for learning short-term spatiotemporal features and are then fed to an LSTM for extracting long-term spatiotemporal dependencies. The output feature maps are concatenated and a final dense layer with a softmax activation is applied for the classification.

The implemented architecture of the CNN is presented in Fig. 2 and includes the following layers: a) a convolutional layer with 32 kernels of size 7×7 with stride 2, with batch normalization and ReLU activation; b) convolutional layer with 32 kernels of size 3×3, again with batch normalization and ReLU activation; c) pooling layer, downsampling the image with 2×2 max-pooling with stride 2; d) convolutional layer with 32 kernels of size 3×3 with batch normalization and ReLU activation; e) pooling layer, downsampling the image with 2×2 max-pooling with stride 2; f) flatten layer. For the aforementioned layers, time distributed layers are used to extract features from the entire sequence of images, creating the appropriate input for g) an LSTM layer with 256 units. Finally, the output features that are extracted from the LSTM layer are concatenated for each type of data (i.e., RGB, depth, activity Images) and used as input to h) a dense layer with softmax activation that produces the final classification results. The proposed approach is summarized in Fig. 3.

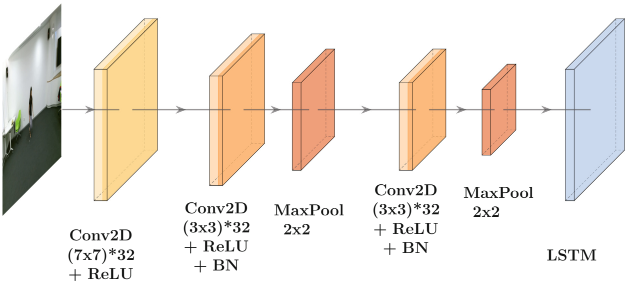

Fig. 2. The proposed hybrid CNN-LSTM architecture.

4 Experiments

4.1 Datasets

For network training and experimental evaluation, our method was tested on a) activities that resemble to "activities of daily living" (ADLs) and are part of the PKU-MMD dataset; and b) on "medical conditions" that are part of the NTU RGB+D dataset. PKU-MMD [13] is a large-scale dataset for continuous multi-modality 3D human action understanding, captured via the Kinect v2 sensor. It contains action instances providing color and depth images, infrared sequences and human skeleton joints. For the evaluation of our model we selected 11 classes that are considered to be mostly related to ADLs: *eat meal snack*, *falling*, *hand-shaking*, *hugging other person*, *make a phone call answer phone*, *playing with phone tablet*, *reading*, *sitting down*, *standing up*, *typing on a keyboard* and *wearing a jacket*. Moreover, NTU RGB+D [20] is a large scale benchmark dataset for 3D Human Activity Analysis. RGB, depth, infrared and skeleton videos for each performed action have been also recorded using the Kinect v2 sensor. We selected the medical-condition-related category consisting of 12 classes, namely: *sneeze/cough*, *staggering*, *falling down*, *headache*, *chest pain*, *back pain*, *neck pain*, *nausea/vomiting*, *fan self*, *yawn*, *stretch oneself* and *blow nose*.

4.2 Implementation and Network Training Details

Experiments were performed on a personal workstation with an Intel^{TM}i7 5820K 12 core processor on 3.30 GHz and 16 GB RAM, using NVIDIA^{TM}Geforce GTX 2060 GPU with 8 GB RAM and Ubuntu 18.04 (64 bit). The deep architecture has been implemented in Python, using Keras 2.2.4 [2] with the Tensorflow 1.12 [1] backend. All data pre-processing and processing steps have been implemented in Python 3.6 using NumPy, SciPy and OpenCV. For training, we set the batch size to 16. We used the Adam optimizer and set the learning rate to 0.0001.

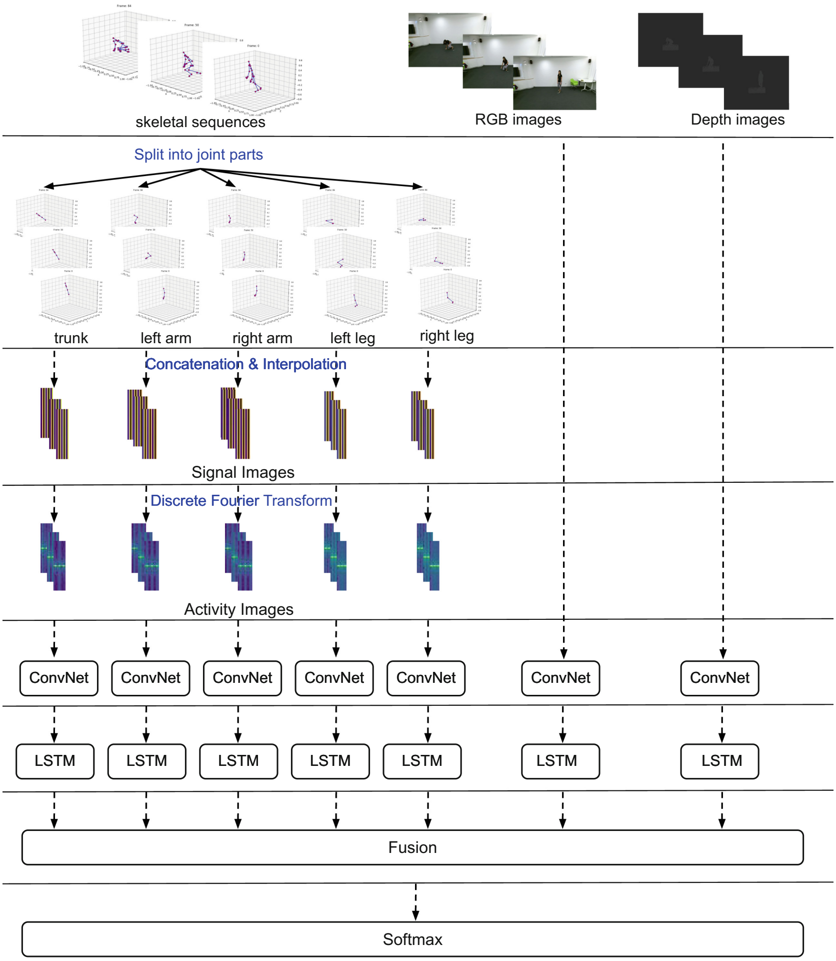

Fig. 3. An overview of the proposed approach.

4.3 Experimental Results and Analysis

Initially, we assessed the contribution of different body parts to the accuracy of classification. It is obvious that the majority of the aforementioned activities mainly consist of upper body motion. A few also involve significant leg motion. Our experiments indicated that all parts were needed to maximize accuracy. When legs were omitted, a small, yet significant drop of performance occurred, especially in classes such as "sitting down" and "standing up." We also performed several experiments regarding image sizes and sequence lengths. The adjustment of these parameters depends on the duration of each activity performed. Our goal was to conclude with parameters that are suitable for representing the average duration of the activities. Upon this experimental evaluation, we ended up with the following setup: RGB sequences consisting of 15 frames with dimension 213 × 120, depth sequences of 15 frames with dimension 128 × 106, image sequences of 7 activity images from a) head-torso with dimension 15 × 53; b) arms with dimension 18 × 53; and c) legs with dimension 12 × 53. Within the evaluation using both datasets, we adopted the same evaluation protocol, which includes single-view, cross-view and cross-subject evaluation criteria. This approach has been followed in order to investigate our approach's competency in dealing with view-independent action recognition and intra-class variations among different subjects. In both datasets each scene is captured by three different angles, thus in a single-view setting train and test sets are derived from the same camera, while in a cross-view evaluation one/two viewpoints are used for training and the remaining for testing. In cross-subject experiments, subjects are split into training and testing groups.

Table 1 summarizes the results achieved for the classification of the 11 ADLs from the PKU-MMD dataset, the classification of the 12 medical conditions from the NTU RGB+D dataset as well as the comparison between the proposed approach and that of our previous work [17] in terms of accuracy scores. It is evident that the accuracy of the latter has been outperformed in every evaluation setup. The highest accuracy scores were achieved in the following setups: a) LR-M which is indicative of a cross-view evaluation, b) the cross-subject evaluation and c) M-M which is indicative of a single view evaluation. Specifically, our approach achieved an accuracy of 0.95 in the LR-M setup that is 19% higher than the compared approach. In the cross-subject setup it achieved an accuracy of 0.94 that is 9.6% higher while in the M-M setup it achieved an accuracy of 0.94 that is 5.3% higher. The experimental results on the different subsets have shown

Table 1. Experimental results of the proposed approach. P, R, F_1 and Acc. denote Precision, Recall, F_1 score and Accuracy, respectively.

Experiment	Viewpoint		PKU-MMD					NTU			
	Train	Test	P	R	F_1	Acc.	Acc. of [17]	P	R	F_1	Acc.
Cross view	**LR**	**M**	0.95	0.95	0.95	0.95	0.77	0.75	0.75	0.75	0.75
	LM	**R**	0.89	0.88	0.88	0.88	0.60	0.72	0.71	0.71	0.71
	RM	**L**	0.87	0.86	0.86	0.86	0.60	0.64	0.62	0.62	0.62
	M	**L**	0.86	0.85	0.84	0.85	0.62	0.63	0.63	0.63	0.63
	M	**R**	0.90	0.90	0.90	0.90	0.58	0.67	0.68	0.68	0.68
	R	**L**	0.65	0.65	0.63	0.65	0.32	0.57	0.57	0.56	0.57
	R	**M**	0.86	0.86	0.86	0.86	0.56	0.70	0.69	0.69	0.70
	L	**R**	0.73	0.72	0.72	0.72	0.41	0.64	0.64	0.64	0.64
	L	**M**	0.87	0.86	0.86	0.86	0.65	0.68	0.68	0.68	0.68
Cross subject	**LRM**	**LRM**	0.94	0.94	0.94	0.94	0.85	0.63	0.64	0.65	0.64
Single view	**L**	**L**	0.90	0.90	0.90	0.90	0.76	0.68	0.67	0.67	0.67
	M	**M**	0.94	0.94	0.94	0.94	0.89	0.69	0.68	0.68	0.68
	R	**R**	0.92	0.91	0.91	0.91	0.84	0.69	0.68	0.68	0.68

the ability of this approach to effectively classify human activities. However we should indicate that in the cross-view setup a performance gap occurs when the viewpoint used for training differs significantly from the one used for testing (e.g. R-L: 0.65, L-R: 0.72).

5 Conclusion

In this paper an effective method for human action recognition has been proposed. Our goal was to incorporate multiple modalities and exploit complementary features from pre-segmented videos. For skeletal information we used an image-based spectral representation aiming to capture spatio-temporal features. Considering that CNNs work efficiently with highly dimensional data, such as images, we utilized a CNN architecture to extract spatial features. In addition to capture the temporal dependencies among the extracted features an LSTM was employed. We assessed the performance of our approach in actions related to daily living activities and to medical conditions. The experimental results on the different subsets have shown the competence of this approach for learning and classifying human activities. Our plans for future work include investigation on methods for creating the signal image, possibly with the use of other types of sensor measurements such as wearable accelerometers, gyroscopes etc. and evaluation of the proposed approach on several other public datasets, and for other types of activities. Finally, we would like to perform an evaluation into a real-like or even real-life assistive living environment.

Acknowledgments. This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No. 273 (Funding Decision:ΓΓΕΤ122785/I2/19-07-2018).

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: Proceedings of USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (2016)
2. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
3. Haque, M.A., et al.: Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities. In: Proceedings of IEEE International Conference on Automatic Face & Gesture Recognition (2018)
4. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Proceedings of ACCV (2016)
5. Huynh-The, T., Hua, C.H., Ngo, T.T., Kim, D.S.: Image representation of pose-transition feature for 3D skeleton-based action recognition. Inf. Sci. **513**, 112–126 (2020)
6. Hou, Y., Li, Z., Wang, P., Li, W.: Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Trans. CSVT **28**(3), 807–811 (2016)

7. Imran, J., Raman, B.: Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition. J. Ambient Intell. Hum. Comput. **11**(1), 189–208 (2020)
8. Jiang, W., Yin, Z.: Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of ACM International Conference on Multimedia (2015)
9. Ke, Q., An, S., Bennamoun, M., Sohel, F., Boussaid, F.: Skeletonnet: mining deep part features for 3-D action recognition. IEEE Signal Process. Lett. **24**(6), 731–735 (2017)
10. Li, C., Hou, Y., Wang, P., Li, W.: Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Process. Lett. **24**(5), 624–628 (2017)
11. Li, C., Wang, P., Wang, S., Hou, Y., Li, W.: Skeleton-based action recognition using LSTM and CNN. In: Proceedings of IEEE ICME Workshops (2017)
12. Li, X., et al.: Concurrent activity recognition with multimodal CNN-LSTM structure. arXiv preprint arXiv:1702.01638 (2017)
13. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475 (2017)
14. Liu, J., Akhtar, N., Mian, A.: Skepxels: spatio-temporal image representation of human skeleton joints for action recognition. In: CVPR Workshops (2019)
15. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn. **68**, 346–362 (2017)
16. Liu, J., Akhtar, N., Mian, A.: Viewpoint invariant RGB-D human action recognition. In: Proceedings of International Conference on DICTA (2017)
17. Papadakis, A., Mathe, E., Vernikos, I., Maniatis, A., Spyrou, E., Mylonas, P.: Recognizing human actions using 3D skeletal information and CNNs. In: Proceedings of EANN (2019)
18. Papadakis, A., Mathe, E., Spyrou, E., Mylonas, P.: A geometric approach for cross-view human action recognition using deep learning. In: Proceedings of ISPA (2019)
19. Pham, H.H., Salmane, H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.: Spatio-temporal image representation of 3D skeletal movements for view-invariant action recognition with deep convolutional neural networks. Sensors **19**(8), 1932 (2019)
20. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of CVPR (2016)
21. Sun, L., Jia, K., Chen, K., Yeung, D.Y., Shi, B.E., Savarese, S.: Lattice long short-term memory for human action recognition. In: Proceedings of ICCV (2017)
22. Wang, P., Li, Z., Hou, Y., Li, W.: Action recognition based on joint trajectory maps using convolutional neural networks. In: Proceedings of ACM-MM (Oct 2016)
23. Yang, Z., Li, Y., Yang, J., Luo, J.: Action recognition with spatio-temporal visual attention on skeleton image sequences. IEEE Trans. CSVT **29**(8), 2405–2415 (2018)
24. Zhu, G., Zhang, L., Shen, P., Song, J.: Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access **5**, 4517–4524 (2017)