Recognizing Human Actions Using 3D Skeletal Information and CNNs

Antonios Papadakis[1], Eirini Mathe[2,5], Ioannis Vernikos[3], Apostolos Maniatis[4], Evaggelos Spyrou[2,4(✉)], and Phivos Mylonas[5]

[1] Department of Informatics and Telecommunications, University of Athens, Athens, Greece
sdi1400141@di.uoa.gr
[2] Institute of Informatics and Telecommunications, National Center for Scientific Research - "Demokritos", Athens, Greece
{emathe,espyrou}@iit.demokritos.gr
[3] Department of Computer Science and Telecommunications, University of Thessaly, Lamia, Greece
ivernikos@uth.gr
[4] General Department, University of Thessaly, Lamia, Greece
amaniatis@teiste.gr
[5] Department of Informatics, Ionian University, Corfu, Greece
fmylonas@ionio.gr

Abstract. In this paper we present an approach for the recognition of human actions targeting at activities of daily living (ADLs). Skeletal information is used to create images capturing the motion of joints in the 3D space. These images are then transformed to the spectral domain using 4 well-known image transforms. A deep Convolutional Neural Network is trained on those images. Our approach is thoroughly evaluated using a well-known, publicly available challenging dataset and for a set of actions that resembles to common ADLs, covering both cross-view and cross-subject cases.

Keywords: Human action recognition · Convolutional neural networks · Skeletal data

1 Introduction

Understanding of human actions from video has attracted an increasing interest during the last few years. This research area lies in the broader field of human-centered activity recognition and combines ideas and techniques mainly from the fields of computer vision and pattern recognition. There exist several human action understanding tasks. Wang et al. [23] proposed a categorization into the following sub-problems: gesture, action, interaction and group activity recognition. The performance of a gesture requires a relatively small amount of time, while the performance of an action requires a significant amount of time and contrary to a gesture, it typically involves more than one body parts. Moreover,

J. Macintyre et al. (Eds.): EANN 2019, CCIS 1000, pp. 511–521, 2019.
https://doi.org/10.1007/978-3-030-20257-6_44

an interaction involves either a person and an object or two persons. Finally, a group activity may be defined as any combination of the aforementioned. Open challenges in the area of human action understanding include the representation, the analysis and the recognition of the actions [2], while a broad field of applications such as surveillance, assisted living, human-machine interaction, affective computing etc. have benefited.

Earlier recognition approaches such as the one of Schuldt et al. [19] were based on hand-crafted features, used to train traditional machine learning algorithms such as Support Vector Machines (SVMs). However, as it has been demonstrated, the accuracy of such approaches is drastically reduced, when the number of the actions significantly increases. Moreover, they may be unable to provide robustness to viewpoint changes, a case typically encountered in real-life scenarios. Recent advances in hardware and more particularly in graphics processing units (GPUs) have enabled fast training of deep neural network architectures [9]. Such approaches do not strictly require a feature extraction step; instead, features are "learnt" within the network. Also, their accuracy significantly increases with the increase of the available training examples.

In this paper, we build on previous works [17,18] and propose the use of visual representations of human actions, based on well-known 2D image transformations. More specifically, we use the Discrete Fourier Transform (DFT), the Fast Fourier Transform (FFT), the Discrete Cosine Transform (DCT) and the Discrete Sine Transform (DST). First, we create raw signal images which capture the 3D motion of human skeletal joints over space and time. Then, one of the aforementioned transforms is applied into each of the signal images, resulting to an "activity" image, which captures the spectral properties of signal images. For classification, we propose a deep CNN architecture. We evaluate the proposed approach using the challenging PKU-MMD dataset [15] and present results for cross-subject and cross-view cases. We demonstrate that the proposed approach may be used in real-like environments for the recognition of activities of daily living (ADLs) [12].

The rest of this paper is organized as follows: Sect. 2 presents related research works in the field of human action recognition, limited to those that are based on visual representations of 3D skeletal information and CNN. Next, Sect. 3 presents the proposed methodology, i.e., the 2D representation of a skeleton based on a set of 3D joints, the construction of image representations capturing skeletal motion in space and time. Also, the CNN architecture is presented therein. Experiments are presented in Sect. 4 and are discussed in Sect. 5, where plans for future work are also included.

2 Related Work

As it has already been mentioned in Sect. 1, when working with deep learning approaches, a large-scale multi-class dataset may be the key to effectiveness and robustness. The first publicly available datasets such as the KTH [19], were limited to a small number of simple actions e.g., *walking*, *running*, *hand clapping*

etc.. Later, a next series of datasets such as the Hollywood dataset [11] targeted more realistic human actions e.g., *answer phone*, *get out of car*, *hand shake* etc., still being limited to a small number of classes. In less than a decade, more challenging datasets such as the UCF101 [21] and the HMDB [10] emerged, containing large numbers of more complex actions, including interactions with objects such as *playing cello*, *horse riding*, *swing baseball bat*, *fencing* etc. Recent large-scale datasets such as PKU-MMD [15] or the NTU [20], are comprised of large numbers of training video and depth sequences.

According to Wang et al. [23] human action recognition tasks may be divided into two major categories:

- **segmented recognition:** the given input video sequence contains *only* the action to be recognized. This means that any frame before/after the action, i.e., not depicting a part of the action, has been removed. In this case, Recurrent Neural Networks (RNNs) [5] or CNNs [13] are typically used.
- **continuous recognition:** the goal is to recognize actions within a given video; the video may or may not depict a single action. In that case, also known as "online" recognition, RNNs are typically used.

Note that when a CNN is used and the only available motion features are skeletal data, an intermediate visual representation of skeletal sequences is required. This representation should capture both spatial and temporal information regarding the motion of joints, i.e., in the 3D space over time. This information should be reflected to its color and/or texture properties. In this section our goal is to present research works that are based on visual representations of 3D skeletal data of human actions and training deep networks, i.e., an intermediate hand-crafted feature extraction step is not included in the process. Skeletal data typically consist of a set of skeletal joints moving in 3D space over time, i.e., for each joint 3 1D signals are generated per action. The extraction of joints from video requires depth information.

In the work of Du et al. [4], in order to preserve the spatial information the set of joints is split into five subsets corresponding to arms, legs and the trunk. Pseudo-colored images are generated by corresponding x, y and z spatial coordinates to R, G and B components, respectively. To preserve temporal information, spatial representations are chronologically arranged. Wang et al. [24] proposed a representation of "joint trajectory maps," wherein the motion direction is encoded by hue. Maps were constructed by appropriately setting saturation and brightness so as texture would correspond to motion magnitude; each was based on the projected trajectory of the skeleton to a Cartesian plane. Similarly, Hou et al. [6] transformed the extracted skeleton joints into a representation called "skeleton optical spectra," so that hue changes would reflect to the temporal variation of skeletal motion. Li et al. [14] proposed the representation of "joint distance maps," and opted for encoding the pair-wise joint distances in the 3 orthogonal 2D planes and also used a fourth one to encode distances in the 3D space, while hue was used to encode distance variations. Each map was separately classified and a late fusion scheme was adopted. In an effort for invariance to the initial position and orientation of the skeleton, Liu et al. [16] applied

transforms to skeletal sequences. The representation of each joint consisted of its 3D space coordinates, also adding time and joint label to create a 5D space. Upon projection to a 2D image using two of the aforementioned dimensions, the remaining three were used as R, G, B values to form pseudo-colored images. Finally, Ke et al. [8] did not extract 3D coordinates. Instead, they extracted translation, rotation and scale invariant features by subsets of joints as in [4]. From each, they extracted cosine distances and normalized magnitudes from vector representations generated from pairwise relative positions between joints. These representations were concatenated to form a 2D representation.

3 Human Action Recognition

3.1 Skeletal Information

The proposed approach requires as input 3D trajectories of skeletal joints (i.e., x, y and z coordinates for each) during an action. We work using data that have been captured with the Microsoft Kinect v2 sensor. More specifically, data consist of 25 human joints; up to 6 skeletons are simultaneously extracted in real time by using the Kinect SDK. A human skeleton corresponds to a graph; nodes correspond to body parts such as arms, legs, head, neck etc., while edges follow the body structure. Moreover, a parent-child relationship is implied. For example, the joint "HEAD" is parent of "NECK," while the "NECK" is the parent of "SPINE_SHOULDER," etc. Each joint consists a 3D signal capturing its 3D position over time. Equivalently, this signal may be seen as 3 1D signals; each corresponding to a coordinate. Therefore, 75 such 1D signals result from the set of 25 joints and for any given video sequence. Note that their duration may vary, since different actions may require different amounts of time. Also different persons or even the same one may perform the same action with similar, yet not equal duration.

3.2 Convolutional Neural Networks

Deep learning is a recent trend in the broader field of machine learning. It is based on the idea to use multiple intermediate interconnected layers within a network, so as to non-linearly process its input. In a sense, these layers are used to "learn" how to extract features in multiple levels of abstraction. During the last few years they have started playing a dominant role in several applications in areas such as computer vision, audio analysis, speech recognition etc., having successfully replaced traditional machine learning approaches. The latter often exhibit a drop of performance when used in real-life applications. During the last few years, research in practical computer vision problems has shifted to deep architectures.

When dealing with computer vision problems, the most popular approach is to train CNNs. A CNN resembles to traditional feed-forward networks; training takes place by forward and backward propagation of input data and error,

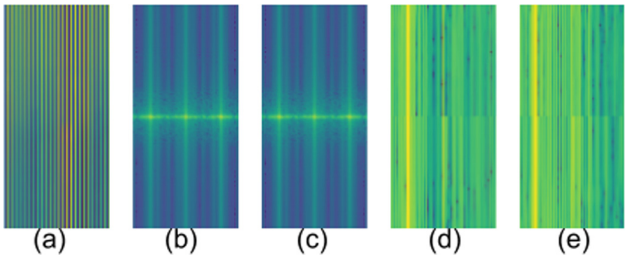

Fig. 1. (a) A signal image; activity image resulting upon (b) DFT; (c) FFT; (d) DCT; (e); DST. Action is *playing with phone/tablet*. DFT and FFT images have been processed with log transformation for visualization purposes. Figure best viewed in color. (Color figure online)

respectively. Their discrimination power lies to the fact that their *convolutional* layers are designed to learn a set of convolutional filters; during training, their parameters are learnt. Neurons are grouped into rectangular grids; each grid performs a convolution in a part of the input image. A *pooling* layer typically succeeds a single or a set of convolutional layers and sub-samples its input, to produce a single value from a small rectangular block. Finally, *dense* layers are those that are ultimately responsible for classification, based on the features that have been extracted by the convolutional layers and sub-sampled by the pooling ones.

3.3 Proposed Methodology

Our work has been partially inspired by the one of Jiang and Yin [7]. Similarly, we first create an image by concatenating the aforementioned 75 1D signals, which form the "signal" image. From each signal image we create an "activity" image, by applying one of the following transforms: (a) the 2D Discrete Fourier Transform (DFT); (b) the 2D Fast Fourier Transform (FFT); (c) the 2D Discrete Cosine Transform (DCT); and (d) the 2D Discrete Sine Transform (DST). From each transform we preserve only the magnitude, while we discard the phase. Note that DST and DCT are further processed by normalizing using the orthonorm. Obviously, in all cases the result is a 2D signal. Note that FFT is a fast implementation of DFT. Several implementations of FFT are approximations. Also even exact implementations are prone to floating point errors. Therefore, our goal was to assess whether the implementation of FFT we used showed a drop of performance compared to DFT. In Fig. 1 we illustrate an example signal image and the 4 corresponding activity images.

We herein remind that the goal of this work is limited to action classification. Therefore, it belongs to the category of segmented recognition (see Sect. 2), as it does not perform a temporal segmentation step. As we shall see in Sect. 4, we work using pre-segmented video sequences, aiming to only recognize the performed actions within each segment. We also assume that each segment contains exactly one action. To address the problem of temporal variability between

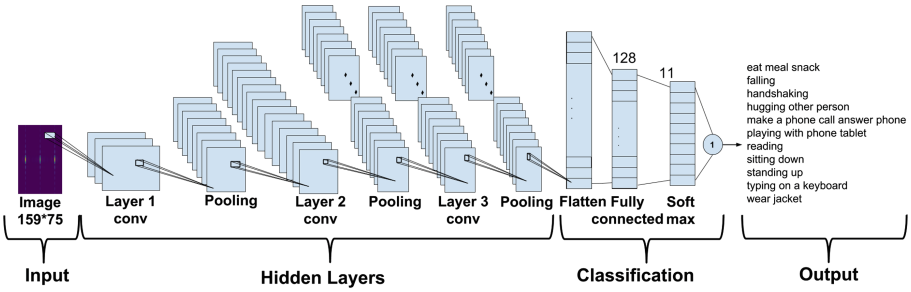

Fig. 2. The proposed CNN architecture.

actions and between users, an interpolation step is necessary. Upon experimentation, we set the duration of all videos equal to 159 frames, upon imposing a linear interpolation step. This way, the size of signal and activity images was fixed and equal to 159×75.

The architecture of the proposed CNN is presented in detail in Fig. 2. The first convolutional layer filters the 159×75 input activity image with 32 kernels of size 3×3. The first pooling layer uses "max-pooling" to perform 2×2 sub-sampling. The second convolutional layer filters the 76×34 resulting image with 64 kernels of size 3×3. A second pooling layer uses "max-pooling" to perform 2×2 sub-sampling. A third convolutional layer filters the 36×15 resulting image with 128 kernels of size 3×3. A third pooling layer uses "max-pooling" to perform 2×2 sub-sampling. Then, a flatten layer transforms the output image of the last pooling to a vector, which is then used as input to a dense layer using dropout. Finally, a second dense layer produces the output of the network. To avoid overfitting, the most popular approach, which is also adopted in this work is the use of the dropout regularization technique [22]: at each training stage several nodes are "dropped out" of the network. This way overfitting is reduced or even prevented, since complex co-adaptations on training data are prevented.

3.4 Implementation Details

For the implementation of the CNN we have used Keras [3] running on top of Tensorflow [1]. All data pre-processing and processing steps have been implemented in Python 3.6 using NumPy[1], SciPy[2] and OpenCV.[3]

4 Experiments

4.1 Dataset

For the experimental evaluation of our approach we used the PKU-MMD dataset [15]. As it has already been mentioned, PKU-MMD is a large-scale benchmark

[1] http://www.numpy.org/.
[2] https://www.scipy.org/.
[3] https://opencv.org/.

focusing on human action understanding and containing approx. 20K action instances from 51 categories, spanning into 5.4M video frames. Note that 66 human subjects have participated in the data collection process, while each action has been recorded by 3 camera angles, using the Microsoft Kinect v2 camera. For each action example, raw RGB video sequences, depth sequences, infrared radiation sequences and extracted 3D positions of skeletons are provided.

Our experiments are divided into two parts. In the first part our goal was to assess whether the proposed approach may be used for ambient assistive living scenarios and more specifically for the recognition of ADLs. Therefore, we selected 11 out of the 51 classes of PKU-MMD, which we believe are the most close to ADLs or events in such a scenario. The selected classes are: *eat meal snack*, *falling*, *handshaking*, *hugging other person*, *make a phone call answer phone*, *playing with phone tablet*, *reading*, *sitting down*, *standing up*, *typing on a keyboard* and *wear jacket*. In Fig. 3 we illustrate sample signal and activity

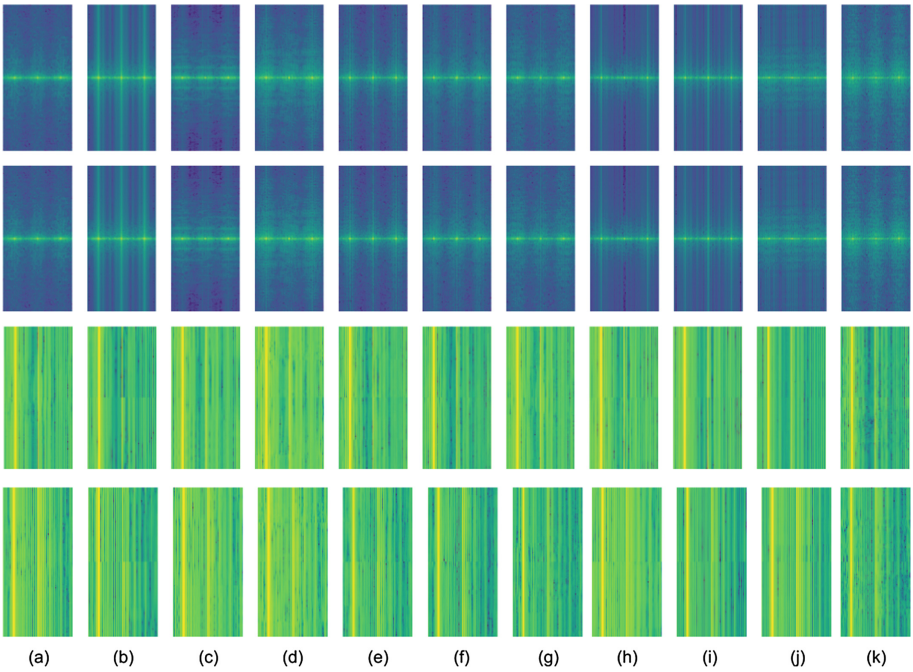

Fig. 3. Examples of activity images from 11 classes and for the 4 transforms used. 1st row: DFT; 2nd row: FFT; 3rd row: DCT; 4th row: DST. (a) eat meal/snack; (b) falling; (c) handshaking; (d) hugging other person; (e) make a phone call/answer phone; (f) playing with phone/tablet; (g) reading; (h) sitting down; (i) standing up; (j) typing on a keyboard; (k) wear jacket. DFT and FFT images have been processed with log transformation for visualization purposes. Figure best viewed in color. (Color figure online)

images from these 11 classes and for all types of transforms. In the second part, we performed experiments with the whole dataset, i.e., with all 51 classes. In both parts, we worked only based on the skeletal data, discarding RGB, depth and infrared information.

4.2 Results

The evaluation protocol we followed is as follows: we first performed experiments per camera position; in this case both training and testing sets derived from the same viewpoint. Then, we performed cross-view experiments, where different viewpoints were used for training and for testing. The goal of these experiments was to test the robustness of the proposed approach in terms of transformation (e.g., a translation and a rotation), which could correspond to abrupt viewpoint changes which typically occur in real-life situations. Finally, we performed cross-subject experiments, where subjects were split in training and testing groups, i.e., any actor "participated" only into one of the groups. The goal of this subject was to test the robustness of our approach into intra-class variations. In real-life situations this is expected to happen when a system is trained e.g., within a laboratory environment and is deployed into a real ambient-assistive living environment. Note that in all cases we measured classification accuracy. Detailed results are depicted in Tables 1 and 2 for 11 and 51 classes, respectively. As it may be observed, in the first case, DST showed best accuracy for the majority of single- and cross-view experiments, followed by DCT. In cross-view experiments, while in the cross-subject case, DFT showed best accuracy, once again followed

Table 1. Experimental results denoting accuracy of the proposed approach in the 11 selected classes of the PKU-MMD dataset. M, L and R denote the middle, left and right camera angles, respectively.

Experiment	Train	Test	DFT	FFT	DCT	DST
Single-view	M	M	**0.89**	0.84	0.83	0.86
	L	L	0.76	0.82	0.78	**0.84**
	R	R	0.84	0.85	**0.89**	0.87
Cross-view	M	L	0.62	0.61	0.63	**0.64**
	M	R	0.58	0.61	**0.65**	0.63
	L	M	0.65	0.66	0.64	**0.72**
	L	R	0.41	0.38	0.32	**0.43**
	R	M	0.56	0.59	**0.64**	0.63
	R	L	0.32	0.37	0.33	**0.39**
	M, L	R	0.60	0.60	**0.62**	**0.62**
	M, R	L	**0.60**	0.57	0.59	**0.60**
	L, R	M	0.77	0.77	0.81	**0.82**
Cross-subject	M, L, R	M, L, R	**0.85**	0.79	0.83	0.81

Table 2. Experimental results denoting accuracy of the proposed approach in the 51 classes of the PKU-MMD dataset. M, L and R denote the middle, left and right camera angles, respectively.

Experiment	Train	Test	DFT	FFT	DCT	DST
Single-view	M	M	0.46	0.49	**0.66**	0.65
	L	L	0.46	0.52	0.60	**0.65**
	R	R	0.55	0.51	**0.69**	0.67
Cross-view	M	L	0.33	0.34	0.33	**0.36**
	M	R	0.32	0.31	**0.36**	0.32
	L	M	0.33	0.30	0.35	**0.39**
	L	R	**0.22**	**0.22**	0.13	0.14
	R	M	0.32	0.34	**0.36**	0.34
	R	L	**0.20**	0.19	0.13	0.14
	M, L	R	0.33	**0.34**	0.33	**0.34**
	M, R	L	0.32	0.33	0.32	**0.34**
	L, R	M	0.44	0.44	**0.55**	0.52
Cross-subject	M, L, R	M, L, R	0.50	0.52	**0.64**	0.63

by DCT. In the second case, DCT and DST showed best accuracy in the majority of cases, apart from the extreme cross-view cases where left angle was used for training and right for testing or vice versa, where DFT showed best accuracy, followed by FFT.

5 Conclusions and Future Work

In this paper we presented a novel approach which aims to recognize human actions in videos. Our approach is based on a novel representation of skeletal 3D motion which uses spectral image transformations and also on a novel CNN architecture. More specifically, a CNN was trained on images which resulted upon (a) concatenation of raw 1D signals corresponding to 3D motion of skeletal joints' coefficients and (b) application of a transform to the created image.

We evaluated the proposed approach using a state-of-the-art and challenging dataset, which consisted of sequences corresponding to 51 human actions. These sequences had been captured with 3 Kinect v2 cameras, under different camera angles and the skeletal joints of the human actors involved had been extracted. We performed experiments involving either a single camera (single-view) or more than one (cross-view). We also performed cross-subject experiments to evaluate the robustness of the approach. We mainly focused on a subset of 11 actions which in our opinion are the most close to real-life ADLs. However, we also experimented with the whole dataset. Our initial results indicate that the proposed approach may be successfully applied to human action recognition in

real-like conditions, yet a drop of performance is expected when significant changes of viewpoint occur.

Among our plans for future are the following: (a) investigation on methods for creating the signal image, possibly with the use of other types of sensor measurements such as wearable accelerometers, gyroscopes etc.; (b) investigation on image processing methods for transforming the signal image to the activity image; (c) use of simplified activity images by considering symmetries, e.g., in DFT and FFT (d) exploitation of other types of visual modalities in the process, such as RGB and depth data; (e) evaluation of the proposed approach on several other public datasets; and (f) application into a real-like or even real-life assistive living environment.

Acknowledgment. We acknowledge support of this work by the project SYNTELESIS "Innovative Technologies and Applications based on the Internet of Things (IoT) and the Cloud Computing" (MIS 5002521) which is implemented under the "Action for the Strategic Development on the Research and Technological Sector", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale maching learning. In: Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI) (2016)
2. Berretti, S., Daoudi, M., Turaga, P., Basu, A.: Representation, analysis, and recognition of 3D humans: a survey. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **14**(1S), 16 (2018)
3. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
4. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 579–583. IEEE (2015)
5. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE (2013)
6. Hou, Y., Li, Z., Wang, P., Li, W.: Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Trans. Circuits Syst. Video Technol. **28**(3), 807–811 (2018)
7. Jiang, W., Yin, Z.: Human activity recognition using wearable sensors by deep convolutional neural networks. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1307–1310 (2015)
8. 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)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

10. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision, pp. 2556–2563. IEEE (2011)
11. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
12. Lawton, M.P., Brody, E.M.: Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontol. **9**(3 Part 1), 179–186 (1969)
13. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. 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)
15. 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)
16. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognit. **68**, 346–362 (2017)
17. Mathe, E., Mitsou, A., Spyrou, E., Mylonas, Ph.: Arm gesture recognition using a convolutional neural network. In: Proceedings of International Workshop Semantic and Social Media Adaptation and Personalization (SMAP) (2018)
18. Mathe, E., Maniatis, A., Spyrou, E., Mylonas, Ph.: A deep learning approach for human action recognition using skeletal information. In: Proceedings of World Congress "Genetics, Geriatrics and Neurodegenerative Diseases Research" (GeNeDiS) (2018)
19. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 03, pp. 32–36. IEEE Computer Society (2004)
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 the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019 (2016)
21. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
22. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
23. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: RGB-D-based human motion recognition with deep learning: a survey. Comput. Vis. Image Underst. **171**, 118–139 (2018)
24. Wang, P., Li, W., Li, C., Hou, Y.: Action recognition based on joint trajectory maps with convolutional neural networks. Knowl.-Based Syst. **158**, 43–53 (2018)