An Image Representation of Skeletal Data for Action Recognition using Convolutional Neural Networks

Ioannis Vernikos
Department of Computer Science and Telecommunications
University of Thessaly
Lamia, Greece
ivernikos@uth.gr

Eirini Mathe*
Department of Informatics
Ionian University
Corfu, Greece
emathe@iit.demokritos.gr

Antonios Papadakis
Department of Informatics and Telecommunications
University of Athens
Athens, Greece
sdi1400141@di.uoa.gr

Evaggelos Spyrou
Institute of Informatics and Telecomunnications, NCSR – "Demokritos"
Athens, Greece
espyrou@iit.demokritos.gr

Phivos Mylonas
Department of Informatics
Ionian University
Corfu, Greece
fmylonas@ionio.gr

ABSTRACT

In this paper we present preliminary results of an approach for understanding human actions, based on a novel 2D image representation for 3D skeletal data. More specifically, motion information for human skeletal joints is transformed to a pseudo-colored image. A Convolutional Neural Network is then used for classification. Our approach is evaluated for actions that may be used in an ambient assisted living scenario.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; **Neural networks**;

KEYWORDS

human activity recognition, convolutional neural networks

ACM Reference Format:
Ioannis Vernikos, Eirini Mathe, Antonios Papadakis, Evaggelos Spyrou, and Phivos Mylonas. 2019. An Image Representation of Skeletal Data for Action Recognition using Convolutional Neural Networks. In *The 12th PErvasive Technologies Related to Assistive Environments Conference (PETRA '19), June 5–7, 2019, Rhodes, Greece.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3316782.3322740

1 INTRODUCTION

The problem of understanding human actions based on visual information has been increasingly attracting the interest of the research community in the fields of computer vision and pattern recognition. Main open challenges in this area include the representation, the analysis and the recognition of actions [1]. Main applications include surveillance, assisted living, human-machine interaction, affective computing etc.

Recent large-scale datasets such as PKU-MMD [6] comprised of large numbers of training video and depth sequences have enabled efficient training of deep learning approaches. Several approaches adopting pseudo-colored image representations of skeletal trajectories and working with CNNs have been proposed. Du et al. [2] proposed the use of images generated by corresponding chronologically arranged spatial coordinates to color components. Wang et al. [9] used saturation and brightness in a way that texture corresponded to motion magnitude of skeleton motion trajectories. Hou et al. [3] corresponded hue changes to the temporal variation of skeletal motion. Li et al. [5] encoded pair-wise joint distances in 2D planes, distances in the 3D space and hue was used to encode distance variations. Liu et al. [7] transformed skeletal sequences to ensure invariance to initial skeleton position and orientation and created images originating from 5D representation consisting of space coordinates, label and time. Finally, Ke et al. [4] opted for translation, rotation and scale invariant features by subsets of joints and proposed a representation based on concatenated cosine distances and normalized magnitudes from vector representations generated from pairwise relative positions between joints. In previous work [8] we proposed an image representation of skeletal data based on the DCT transform.

2 METHODOLOGY

We propose a novel visual representation of skeletal information corresponding to human actions. More specifically, we create pseudo-colored images capturing inter-joint distances during an action. We use as input the 3D trajectories of skeletal joints. From the x, y and z coordinates of each of the 25 available joints, we collect 75 signals for any given video sequence. We assume that each video segment contains exactly one action to be recognized. To address the problem of temporal variability between actions and between users,

*Also with Inst. of Informatics and Telecomunnications, NCSR – "Demokritos".

PETRA '19, June 5–7, 2019, Rhodes, Greece

ACM ISBN 978-1-4503-6232-0/19/06…$15.00
https://doi.org/10.1145/3316782.3322740

Figure 1: Representative pseudo-colored images. Top: *eat_meal_snack*, bottom: *falling*.

a linear interpolation step is imposed, setting the duration of all videos equal to $N = 150$ frames. From each video sequence, we calculate coordinate differences between consecutive frames. To create the pseudo-colored images, x, y, z coordinates correspond to R, G, B color channels, respectively. Let $x_i(n)$ denote the x-position of the i-th joint in the n-th frame. Let R denote the R channel of the color image. The value of $R(i, n)$ is calculated as: $R(i, n) = x_i(n+1) - x_i(n)$, where $i = 1, \ldots, N$. Similarly, B and G channels are constructed. As it is exhibited, the way these pseudo-colored images are formed, leads to preserving both the temporal and the spatial properties of the skeleton trajectories. For classification, we use a slightly modified deep CNN architecture, which has been proposed in our previous work [8]. An example is illustrated in Fig. 1.

3 EXPERIMENTAL RESULTS

For the evaluation of the proposed approach, the large-scale PKU-MMD dataset [6] has been used, which contains instances recorded by 3 camera angles. Since our goal was to evaluate whether the proposed approach may be suitable for application in an ambient assistive living scenario, we selected 11 classes of PKU-MMD, which we believe are the most close to activities of daily living (ADLs) or events that should be monitored in such a use case. 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*. Note that we worked using only the available skeletal data, i.e., we discarded RGB, depth and infrared information.

We present results for the following cases: a) *Single-view*: both training and testing sets derived from the same camera; b) *Cross-view*: different viewpoints were used for training and testing; and c) *Cross-subject*: actors were split in training and testing groups. The goal of the second case was to evaluate the performance when abrupt viewpoint changes occur, while the goal of the third case was to evaluate the robustness into intra-class variations, i.e., when the algorithm is applied to *unseen* users. In all cases we measured classification accuracy. Results are presented in Table 1. It may be observed that compared to our previous work, it demonstrated superior results, both in single-view and in cross-view cases.

Table 1: 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	Proposed	[8]
Single-view	M	M	**0.88**	0.82
	L	L	**0.85**	**0.85**
	R	R	**0.84**	0.75
Cross-view	M	L	**0.71**	0.56
	M	R	**0.69**	0.64
	L	M	**0.72**	0.61
	L	R	**0.53**	0.40
	R	M	**0.70**	0.56
	R	L	**0.56**	035
	M,L	R	**0.76**	-
	M,R	L	**0.66**	-
	L,R	M	**0.81**	-
Cross-subject	M,L,R	M,L,R	**0.86**	-

4 CONCLUSIONS

In this paper we presented preliminary results of our approach for the recognition of human actions in video sequences, which was based on a novel representation of 3D skeletal trajectories using pseudo-colored images. A CNN was used for classification into actions that correspond to real-life ADLs or events that should be monitored into ambient assistive living use cases. Our initial results demonstrated the potential of our approach, since classification accuracy was satisfactory even in cross-view and cross-subject experiments. Among our plans for future are a) to further improve the representation, e.g., by incorporating features extracted from RGB and depth video data; b) modifications of the CNN architecture; c) extensive evaluation of the proposed approach on several publicly available datasets; and d) application into a real-like or even real-life ambient 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). This work has also been supported by the project VISOR "Virtual coachIng Services for OldeR adults" which is implemented under the "1st HFRI Call for Scholarships for PhD Candidates," funded by the General Secretariat for Research and Technology and the Hellenic Foundation for Research and Innovation (HFRI).

REFERENCES

[1] Berretti, S., Daoudi, M., Turaga, P., & Basu, A. (2018). Representation, analysis, and recognition of 3D humans: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(1s), 16.
[2] Du, Y., Fu, Y., & Wang, L. (2015). Skeleton based action recognition with convolutional neural network. IAPR Asian Conference on Pattern Recognition (ACPR).
[3] Hou, Y., Li, Z., Wang, P., & Li, W. (2018). Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 28(3), 807-811.
[4] Ke, Q., An, S., Bennamoun, M., Sohel, F., & Boussaid, F. (2017). Skeletonnet: Mining deep part features for 3-d action recognition. IEEE signal proc. letters, 24(6), 731-735.
[5] Li, C., Hou, Y., Wang, P., & Li, W. (2017). Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Proc. Letters, 24(5), 624-628.
[6] Liu, C., Hu, Y., Li, Y., Song, S., & Liu, J. (2017). PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. arXiv:1703.07475.
[7] Liu, M., Liu, H., & Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition, 68, 346-362.
[8] Mathe, E., Maniatis, A., Spyrou, E., & Mylonas, Ph. (2018). A Deep Learning Approach for Human Action Recognition using Skeletal Information. In Proc. of World Congress "Genetics, Geriatrics and Neurodegenerative Diseases Research" (GeNeDiS).
[9] Wang, P., Li, W., Li, C., & Hou, Y. (2018). Action recognition based on joint trajectory maps with convolutional neural networks. Knowledge-Based Systems, 158, 43-53.