Early Fusion of Visual Representations of Skeletal Data for Human Activity Recognition

Ioannis Vernikos
Department of Computer Science and Telecommunications, University of Thessaly
Lamia, Greece
Institute of Informatics and Telecommunications, National Center for Scientific Research – "Demokritos"
Athens, Greece
ivernikos@uth.gr

Dimitrios Koutrintzes
Institute of Informatics and Telecommunications, National Center for Scientific Research – "Demokritos"
Athens, Greece
dkoutrintzes@iit.demokritos.gr

Eirini Mathe
Department of Informatics, Ionian University
Corfu, Greece
cmath17@ionio.gr

Evaggelos Spyrou
Department of Computer Science and Telecommunications, University of Thessaly
Lamia, Greece
Institute of Informatics and Telecommunications, National Center for Scientific Research – "Demokritos"
Athens, Greece
espyrou@uth.gr

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

ABSTRACT

In this work we present an approach for human activity recognition which is based on skeletal motion, i.e., the motion of skeletal joints in the 3D space. More specifically, we propose the use of 4 well-known image transformations (i.e., DFT, FFT, DCT, DST) on images that are created based on the skeletal motion. This way, we create "activity" images which are then used to train four deep convolutional neural networks. These networks are then used for feature extraction. The extracted features are fused, scaled and upon a dimensionality reduction step they are given as input to a support vector machine for classification. We evaluate our approach using two well-known, publicly available, challenging datasets and we demonstrate the superiority of the fusion approach.

CCS CONCEPTS

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

KEYWORDS

human activity recognition, early fusion, deep learning, convolutional neural networks

ACM Reference Format:
Ioannis Vernikos, Dimitrios Koutrintzes, Eirini Mathe, Evaggelos Spyrou, and Phivos Mylonas. 2022. Early Fusion of Visual Representations of Skeletal Data for Human Activity Recognition. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3549737.3549786

1 INTRODUCTION

Human activity recognition (HAR) is one of the most challenging problems in the area of computer vision and pattern recognition. Nowadays, several HAR-based applications exist, such as daily life monitoring, visual surveillance, assisted living, human-machine interaction, affective computing, augmented/virtual reality (AR/VR) etc. In this paper, we build upon our previous work [13] and propose the fusion of several 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 transformations is applied into each of the signal images, resulting to an "activity" image, which captures the spectral properties of signal images. For each image transformation category, we use a trained deep convolutional neural network (CNN) architecture for feature extraction. The extracted features are fused and then are used as input to a support vector machine

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9597-7/22/09…$15.00
https://doi.org/10.1145/3549737.3549786

(SVM), for classification. We evaluate the proposed approach using the challenging PKU-MMD [10] and NTU RGB+D [11] datasets and present results for single-view, cross-view and cross-subject cases.

The rest of this paper is organized as follows: section 2 presents related work, focusing on fusion of visual representations of human skeletal motion. Next, Section 3 presents the proposed feature extraction and fusion methodology. Experiments and results are presented in Section 4, while conclusions are drawn in Section 5, wherein plans for future work are also presented.

2 RELATED WORK

In recent years, several research works using image representations of skeletal data have been presented. Chen et al. [5] encoded spatial-temporal information into color texture images from skeleton sequences, referred to as Temporal Pyramid Skeleton Motion Maps (TPSMMs). The TPSMMs not only capture short temporal information but also embed the long dynamic information over the period of action. They evaluated their method on three distinct datasets. Experimental results showed that the proposed method can effectively utilize the spatio-temporal information of skeleton data. Silva et al. [14] mapped the temporal and spatial joints dynamics into a color image-based representation, wherein, the position of the joints in the final image is clustered into groups. In order to verify whether the sequence of the joints in the final image representation can influence the performance of the model, they conducted two experiments: in the former, they changed the order of the grouped joints in the sequence, while in the latter, the joints were randomly ordered. Tasnim et al. [16] proposed a spatio-temporal image formation (STIF) technique of 3D skeleton joints by capturing spatial information and temporal changes for action discrimination. To generate the spatio-temporal image, they mapped all the 20 joints in a frame with the same color, using the jet color map and then changed the colors as time was passing. Finally, they created the STIF, by connecting lines between joints in adjacent frames, subsequently. Huynh et al. [8] proposed a novel encoding technique, namely Pose-Transition Feature to Image (PoT2I), to transform skeleton information to image-based representation for deep convolutional neural networks (CNNs). This technique includes feature extraction, feature arrangement, and action image generation processes. The spatial joint correlations and temporal pose dynamics of action are exhaustively depicted by an encoded color image. Verma et al. [17] created skeleton intensity images, for 3 views (top, front and side) using a proposed algorithm from skeleton data. Caetano et al. [3] introduced a novel skeleton image representation, named SkeleMotion. The proposed approach encodes the temporal dynamics by explicitly computing the magnitude and orientation values of the skeleton joints. Different temporal scales were employed to compute motion values to aggregate more temporal dynamics to the representation making it able to capture long-range joint interactions involved in actions as well as filtering noisy motion values.

Moreover, several approaches dealing with the fusion of several representations have been proposed. Basly et al. [2] combined deep learning methods with traditional classifier hand-crafted features extractors. For feature extraction, they used a pre-trained CNN approach-based residual neural network (ResNet) model. The resulting feature vector was then fed as an input to an SVM classifier. Similarly, Koutrintzes et al. [9] used hand-crafted features and combined them with deep features. For classification, they also used an SVM. Karen et al. [15] trained two spatial and temporal CNNs. The softmax scores for each model were combined with a late fusion approach, i.e., by training a multi-class linear SVM. Ehatisham-Ul-Haq et al. [7] proposed a multimodal feature-level fusion approach for robust human action recognition. Their features include densely extracted histogram of oriented gradient (HOG) features from RGB/depth videos and statistical signal attributes from wearable sensors data. K-nearest neighbor and support vector machine classifiers were used for training and testing the proposed fusion model for HAR. Chaaraoui et al. [4] combined body poses estimation and 2D shape, in order to improve human action recognition. Using efficient feature extraction techniques, skeletal and silhouette-based features low-dimensional, real-time features were obtained. These two features were then combined by means of feature fusion. Finally, in previous work [18] we presented an approach for the recognition of human activity that combined handcrafted features from 3D skeletal data and contextual features learned by a trained deep CNN. To validate our idea, we trained a CNN using a dataset for action recognition and use the output of the last fully-connected layer as a contextual feature extractor. Then, an SVM was trained upon an early fusion step of both features.

3 PROPOSED METHODOLOGY

The proposed methodology is illustrated in Fig. 1. At the following, we present in detail all its steps, from sensor data to the final classification result.

3.1 Skeletal Information

The proposed approach requires as input 3D trajectories of skeletal joints during an activity. The data we are using have been captured using the Microsoft Kinect v2 sensor. More specifically, these data consist of 25 human joints (i.e., their x, y and z coordinates, over time). Considering each joint as an 1-D signal, 75 such signals result for any given video sequence. Each joint corresponds to a body part such as head, shoulder, knee, etc., while edges connect these joints shaping the body structure. For each of these joint we construct a "signal" image, by concatenating the aforementioned 75 signals. Note that the duration of these signals 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. To address the problem of temporal variability between actions and between users, 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 remains fixed and equal to 159×75.

3.2 Activity Image Construction

Based on our previous work [13] we create activity images upon applying onto the signal images the following well-known image transformations: a) the 2-D Discrete Fourier Transform (DFT); b) the 2-D Fast Fourier Transform (FFT); c) the 2-D Discrete Cosine Transform (DCT); and d) the 2-D Discrete Sine Transform (DST). We consider a segmented recognition problem, i.e., we assume

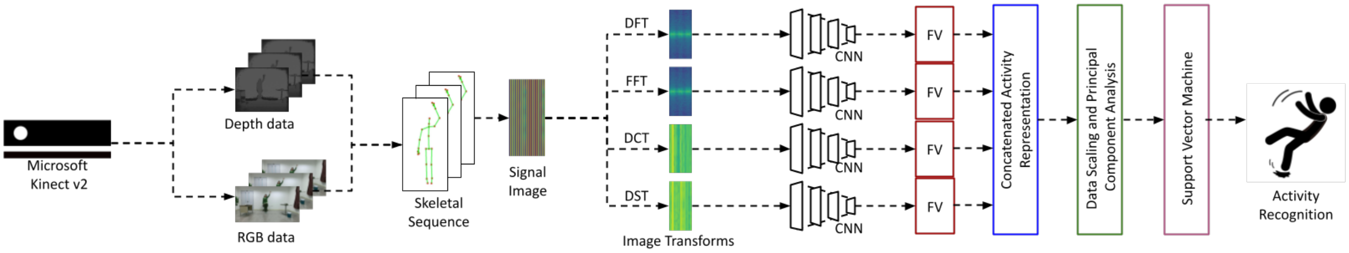

Figure 1: A visual overview of the proposed approach.

that each segment contains exactly one action to be recognized. Then, we train a CNN for each of the 4 image transformations. After the training of the network, we extract the features from the images using the above models and we fuse them. These fused features are then scaled and upon principal component analysis (PCA) their dimension is reduced. This reduced vector is then used for classification with an SVM classifier.

3.3 Network Architecture

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 subsampling. The second convolutional layer filters the 78 × 36 resulting image with 64 kernels of size 3 × 3. A second pooling layer uses "max-pooling" to perform 2 × 2 subsampling. A third convolutional layer filters the 38 × 17 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 size 18 × 17 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. Note that this layer is omitted when the network is used as feature extractor.

4 EXPERIMENTAL RESULTS

4.1 Datasets

For the experimental evaluation of the proposed approach we used two publicly available, large scale, challenging motion activity datasets. More specifically, NTU RGB+D [11] 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. They collected data from 106 distinct subjects and they managed to record more than 114 thousand video samples and 8M frames for three camera angles. This dataset contains 120 different action classes including daily, mutual, and health-related activities. PKU-MMD [10] is a large-scale benchmark focusing on human action understanding and containing approx. 20K action instances from 51 categories, spanning into 5.4M video frames. 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.

4.2 Experimental Setup and Network Training

The 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 RTX 2060 GPU with 8 GB RAM and Ubuntu 20.04 (64 bit). The deep CNN architecture has been implemented in Python, using Keras [6] with the Tensorflow [1] backend. We split the data for training, validation and testing as it is proposed from the datasets' authors [10, 11]. For the training of the network, we used batch size 8 for 150 epochs. For the SVM configuration we used the RBF kernel, with $\gamma = 0.001$ and $C = 100$. To evaluate our method, in case of the PKU-MMD dataset we used the augmented set of samples that we have created in the context of our previous work [12], wherein we augmented the data with four angles, i.e., ±45°, ±90°. In case of the NTU-RGB+D dataset, due to a plethora of camera positions that had been used, we omitted the augmentation step, as it was experimentally proved that it caused a drop of performance.

4.3 Results

For the evaluation of the proposed fusion approach we performed three types of experiments: First, we performed experiments per camera position (single view) in this case both training and testing sets derived from the same viewpoint. Secondly, we performed cross-view experiments, where different viewpoints were used for training and for testing. And finally, in the third experiment, we performed cross-subject experiments, where subjects were split into training and testing groups. In Tables 1 and 2, we present the results for the PKU-MMD and the NTU-RGB+D dataset, respectively. As it may be observed, in all cases the fusion approach leads to a significant increase of accuracy, thus we may assume that the four image transformations may capture complementary features of human motion.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented a fusion approach to the problem of human activity recognition. Our approach was based on image transformations that have been applied on signal image. Convolutional neural networks have been used as feature extractors, while a support vector machine has been used for classification of the fused features. We experimentally demonstrated that the proposed fusion approach outperforms previous work based on a single image transformation. Thus, all transformations capture complementary features.

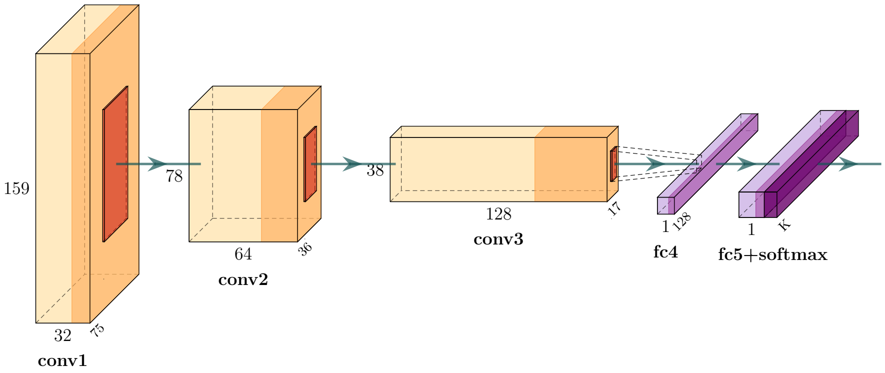

Figure 2: The CNN architecture that has been used in this work; "conv" denotes a convolutional layer, "fc" denotes a fully-connected layer.

	Train	Test	DFT	FFT	DCT	DST	F
CV	**LR**	**M**	0.75	0.76	0.85	0.84	**0.92**
	LM	**R**	0.70	0.69	0.70	0.79	**0.86**
	RM	**L**	0.68	0.69	0.78	0.62	**0.86**
	M	**L**	0.64	0.63	0.68	0.74	**0.85**
	M	**R**	0.63	0.62	0.76	0.64	**0.85**
	R	**L**	0.58	0.58	0.66	0.46	**0.78**
	R	**M**	0.67	0.65	0.78	0.66	**0.87**
	L	**R**	0.58	0.59	0.40	0.64	**0.74**
	L	**M**	0.66	0.66	0.67	0.73	**0.86**
CS	**LRM**	**LRM**	0.70	0.69	0.79	0.79	**0.85**
SV	**L**	**L**	0.62	0.60	0.75	0.72	**0.83**
	R	**R**	0.62	0.61	0.75	0.72	**0.82**
	M	**M**	0.65	0.66	0.79	0.75	**0.85**

Table 1: Experimental results for the PKU-MMD dataset. Numbers denote accuracy, L, R and M denote left, right and middle camera position. Best result per case is indicated by bold. CV, CS and SV correspond to cross-view, cross-subject and single-view, respectively. F denotes the fusion of DFT, FFT, DCT, DST.

A possible application of this work in AR environments so as to measure user experience and assess user engagement. For example, within a museum environment, the detection of a visitor making a phone call while interacting with an AR application could be an indicator of low engagement. In contrast, when the visitor is reading in front of an AR screen, this could be an indicator of high engagement. Among our plans for future work are to investigate other deep architectures and fusion techniques and test our method with novel representations capturing motion properties of several modalities. Finally, we would like to perform evaluation using several other datasets and also perform real-life experiments within the AR environment of the Mon Repo project[1].

	Train	Test	DFT	FFT	DCT	DST	F
CV	**LR**	**M**	0.47	0.48	0.44	0.45	**0.62**
	LM	**R**	0.44	0.45	0.42	0.42	**0.56**
	RM	**L**	0.53	0.54	0.52	0.51	**0.70**
	M	**L**	0.38	0.38	0.30	0.30	**0.43**
	M	**R**	0.47	0.47	0.38	0.39	**0.59**
	R	**L**	0.43	0.45	0.31	0.34	**0.53**
	R	**M**	0.37	0.38	0.28	0.30	**0.45**
	L	**R**	0.43	0.43	0.34	0.36	**0.53**
	L	**M**	0.44	0.45	0.37	0.40	**0.57**
CS	**LRM**	**LRM**	0.50	0.51	0.52	0.52	**0.68**
SV	**L**	**L**	0.48	0.49	0.48	0.50	**0.67**
	R	**R**	0.43	0.44	0.40	0.42	**0.60**
	M	**M**	0.45	0.44	0.48	0.45	**0.65**

Table 2: Experimental results for the NTU-RGB+D dataset. Numbers denote accuracy, L, R and M denote left, right and middle camera position. Best result per case is indicated by bold. CV, CS and SV correspond to cross-view, cross-subject and single-view, respectively. F denotes the fusion of DFT, FFT, DCT, DST.

ACKNOWLEDGMENTS

This research was co-financed by the European Union and Greek national funds through the Competitiveness, Entrepreneurship and Innovation Operational Programme, under the Call «Special Actions "Aquaculture" - "Industrial materials" - "Open innovation in culture"»; project title: "Strengthening User Experience & Cultural Innovation through Experiential Knowledge Enhancement with Enhanced Reality Technologies — MON REPO"; project code: $T6YB\Pi$ - 00303; MIS code: 5066856

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. {TensorFlow}: A System for {Large-Scale} Machine Learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 265–283.

[2] Hend Basly, Wael Ouarda, Fatma Ezahra Sayadi, Bouraoui Ouni, and Adel M Alimi. 2020. CNN-SVM learning approach based human activity recognition. In

[1] https://monrepo.online/

International Conference on Image and Signal Processing. Springer, 271–281.

[3] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. 2019. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–8.

[4] Alexandros Chaaraoui, Jose Padilla-Lopez, and Francisco Flórez-Revuelta. 2013. Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices. In *Proceedings of the IEEE international conference on computer vision workshops*. 91–97.

[5] Yanfang Chen, Liwei Wang, Chuankun Li, Yonghong Hou, and Wanqing Li. 2020. ConvNets-based action recognition from skeleton motion maps. *Multimedia Tools and Applications* 79, 3 (2020), 1707–1725.

[6] Francois Chollet. 2021. *Deep learning with Python*. Simon and Schuster.

[7] Muhammad Ehatisham-Ul-Haq, Ali Javed, Muhammad Awais Azam, Hafiz MA Malik, Aun Irtaza, Ik Hyun Lee, and Muhammad Tariq Mahmood. 2019. Robust human activity recognition using multimodal feature-level fusion. *IEEE Access* 7 (2019), 60736–60751.

[8] Thien Huynh-The, Cam-Hao Hua, Trung-Thanh Ngo, and Dong-Seong Kim. 2020. Image representation of pose-transition feature for 3D skeleton-based action recognition. *Information Sciences* 513 (2020), 112–126.

[9] Dimitrios Koutrintzes., Eirini Mathe., and Evaggelos Spyrou. 2022. Boosting the Performance of Deep Approaches through Fusion with Handcrafted Features. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - ICPRAM,*. INSTICC, SciTePress, 370–377. https://doi.org/10.5220/0010982700003122

[10] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. 2017. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*. 1–8.

[11] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2684–2701.

[12] Antonios Papadakis, Eirini Mathe, Evaggelos Spyrou, and Phivos Mylonas. 2019. A geometric approach for cross-view human action recognition using deep learning. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 258–263.

[13] Antonios Papadakis, Eirini Mathe, Ioannis Vernikos, Apostolos Maniatis, Evaggelos Spyrou, and Phivos Mylonas. 2019. Recognizing human actions using 3d skeletal information and CNNs. In *International Conference on Engineering Applications of Neural Networks*. Springer, 511–521.

[14] Vinícius Silva, Filomena Soares, Celina P Leão, João Sena Esteves, and Gianni Vercelli. 2021. Skeleton driven action recognition using an image-based spatial-temporal representation and convolution neural network. *Sensors* 21, 13 (2021), 4342.

[15] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* 27 (2014).

[16] Nusrat Tasnim, Mohammad Khairul Islam, and Joong-Hwan Baek. 2021. Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints. *Applied Sciences* 11, 6 (2021), 2675.

[17] Pratishtha Verma, Animesh Sah, and Rajeev Srivastava. 2020. Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition. *Multimedia Systems* 26, 6 (2020), 671–685.

[18] Ioannis Vernikos, Eirini Mathe, Evaggelos Spyrou, Alexandros Mitsou, Theodore Giannakopoulos, and Phivos Mylonas. 2019. Fusing handcrafted and contextual features for human activity recognition. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. IEEE, 1–6.