ICIC Express Letters
Part B: Applications
Volume **12**, Number **3**, March **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 215–221

HUMAN ACTIVITY RECOGNITION USING INTER-JOINT FEATURE FUSION WITH SVD

Byung Woo Yoon[1], Erdal Genc[2], Ömer Faruk Ince[3]
and Mustafa Eren Yildirim[1,4]

[1]Department of Electronics Engineering
Kyungsung University
309, Suyeong-ro, Nam-gu, Busan 48434, Korea
{ meyeren3; bwyoon }@ks.ac.kr

[2]Hasso Plattner Institute
University of Potsdam
Am Neuen Palais 10, House 9, Potsdam 14469, Germany
erdal.genc@hpi.de

[3]Center for Intelligent and Interactive Robotics
Korea Institute of Science and Technology
5, Hwarang-ro 14-gil, Seongbuk-gu, Seoul 02792, Korea
023967@kist.re.kr

[4]Department of Electrical and Electronics Engineering
Bahcesehir University
Yıldız, Çırağan Cd., 34349, Beşiktaş, Istanbul, Turkey
mustafaeren.yildirim@eng.bau.edu.tr

Received August 2020; accepted October 2020

Abstract. *In this paper, a multi-feature descriptor for human activity recognition (HAR) was presented. The joints of the human skeleton were extracted from RGB images by using OpenPose to develop a robust multi-feature descriptor. Three features which are joint-joint angle, joint-joint horizontal distance, and joint-joint vertical distance were calculated. For the ease of computational cost, the singular value decomposition (SVD) was performed. In order to obtain singular values representing one full cycle of activity without information loss, the matrix sizes were equalized by zero paddings and row shifting. The singular values obtained from SVD form the final descriptor. The authors evaluated the performance of the proposed method on the well-known KTH and Weizmann datasets. The experimental results showed that the proposed descriptor gives out state-of-the-art results in human action recognition.*

Keywords: Human activity recognition, Inter joint relation, KTH dataset, Weizzmann dataset, Singular value decomposition

1. **Introduction.** Human activity recognition (HAR) is one of the most popular research topics in computer vision for the past twenty years and has been used in many different fields such as robotics [1-3], healthcare systems [4], and industrial applications [5, 6]. To identify human activities with the utmost precision HAR performs an accurate diagnosis of activity patterns collected from various sensors. Sensors used in HAR can be digital cameras, wearable sensors, and gyro sensors [7-12].

It is possible to separate sensors for HAR into two categories as vision-based and non-vision-based sensors. In a non-vision sensor used HAR systems, the relevant features are first calculated based on the data coming from the sensor. Then, the chosen classifier recognizes the activity concerning the features obtained. On the other hand, vision sensors used HAR systems follow three steps, namely, pre-processing, feature extraction

DOI: 10.24507/icicelb.12.03.215

and selection, and classification, to recognize the activity. Both approaches have their advantages and disadvantages. Compared to vision-based sensors, wearable sensors are more beneficial in terms of obtaining signals, for the following reasons: (a) unlike vision-based sensors, wearable sensors would not be affected by environmental limitations such as fixed scenes and lack of illumination; (b) wearable sensors are better at delivering the signal while protecting the privacy of subjects. However, wearable sensors are not cost-effective and cannot be used in daily life conditions. As a conclusion, we decide to use the vision-based sensor for this study due to the advantages in daily life usage.

Lately, the vision-based HAR problem has been studied extensively and the current approaches draw attention with their high accuracy rates. Authors in [13] present an HAR system that relies on weighted segmentation and feature selection approach. They extract the red channel of the frame and apply some filtering to dealing with background variations. Also, they introduce a weighted mechanism that extricates a person by assigning weights for the foreground and the background. They use the rank correlation-based method to select the most relevant features. Another approach to recognize human activities is using the bag-of-visual-words model. In [14], authors utilize SURF descriptors and bag of visual words for human activity recognition. Different from similar studies, they use grayscale images for feature extraction. They also conduct four different machine learning classifiers for comparison. In addition to all these, various studies use skeletal information in the time domain to recognize the activity. Authors in [15] propose a biometric system that extracts the human skeleton and body joints using an RGB-Depth sensor. They store joint angles in a queue and reduce the dimension of the feature vector by their proposed thresholding method. They also perform various machine learning classifiers to compare the performance of their algorithm.

In this paper, authors propose a novel solution to the vision-based human activity recognition problem. We intend to develop an accurate biometric system which uses a multi-feature descriptor constructed from the skeletal data. The block diagram of the proposed method is shown in Figure 1. We use the OpenPose which is a deep-learning-based algorithm for the extraction of the human skeleton. We calculate the joint-joint

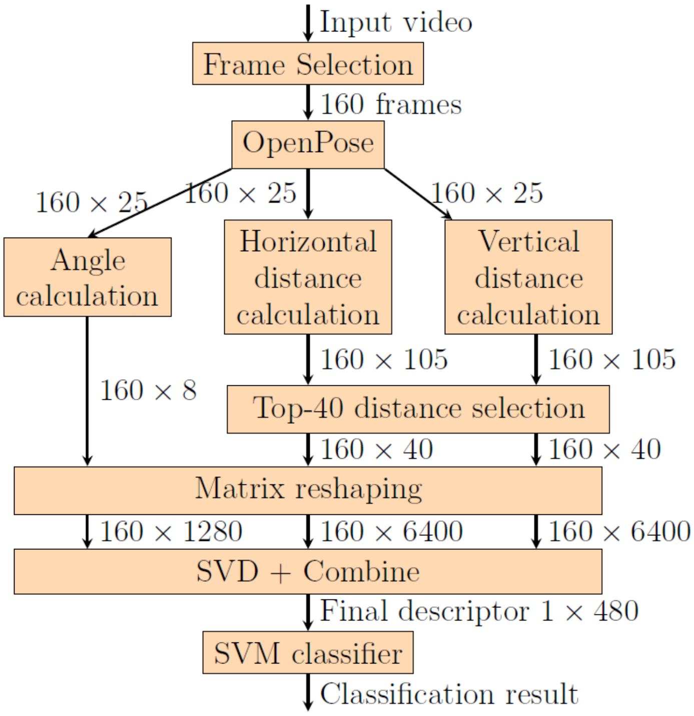

Figure 1. The block diagram of the proposed method

angle, joint-joint horizontal distance, and joint-joint vertical distance. Singular value decomposition (SVD) [16] is used to represent the high-dimensional feature matrices by their unique singular values. We reshape the matrices to the same size by row shifting and zero paddings, to obtain the same amount of singular values without loss of essential information. Lastly, a support vector machine (SVM) is used for recognition of the activity. According to the results of tests conducted on KTH and Weizmann datasets, our algorithm performed state-of-the-art performance.

The paper is organized as follows. The research method with all of its substeps is explained in Section 2. The experiments and benchmarking results on public datasets are given in Section 3. Lastly, the paper is concluded in Section 4.

2. Research Method.

2.1. **Feature extraction.** [15] has shown the eight joint-joint angles which are informative about human activity representation. The locations of these angles are shown with dashed circles in Figure 2. In this paper, we use the same angles for the final descriptor. The calculation of the angles is made as below:

$$a_{j_L,j_R} = \cos^{-1}\left(\frac{||mj_R||^2 + ||mj_L||^2 - ||j_L - j_R||^2}{2 \cdot ||mj_R|| \cdot ||mj_L||}\right) \tag{1}$$

where $||mj_R||$ and $||mj_L||$ represent the distances of reference joint m to the right joint j_R and left joint j_L, respectively. $||j_L - j_R||$ is the distance between joints j_R and j_L. During the analysis, we observe that a period of activity takes 160 frames for both KTH and Weizmann datasets. Thus, in one video, the size of the matrix obtained from the calculation of the joint-joint angle stage is 160 × 8.

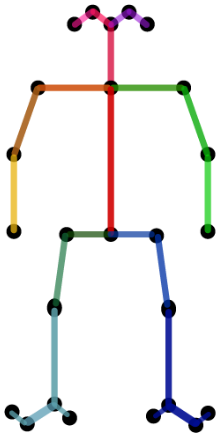

FIGURE 2. Human skeleton map using OpenPose framework

Next, we explain only the extraction of joint-joint horizontal distance feature since the method for the vertical distance feature is the same as the horizontal one. The captured body joints by OpenPose are shown in Figure 2. However, vertical or horizontal distances between some of the joints do not carry distinctive information related to the activities. For example, distances from the left eye to the left ear or from the right ear to the nose do not change during any activity. This is because the activities we focus on do not contain any changes in the face. Therefore, we do not consider the joints left eye, left ear, right ear, and right eye in this study. Only the nose is used as a reference point for the head. A similar relationship is observed on the feet of a human. The distance between the big toe and small toe does not differ among the activities. Furthermore, the distance between the ankle and heel is almost constant. Thus, we keep the joints right-ankle and left-ankle as foot references but remove the joints $\{19, 20, 21, 22, 23, 24\}$. The 2-combination of the remaining 15 joints gives 105 distance features. The calculated distances are normalized among the dataset to be scale-invariant.

Rather than using 105 distance features, we selected the most distinctive features. The same frames used in the angle calculation are also used in this stage. Feature-wise standard deviation for each distance feature is calculated. This gives the change in distance between joints during one period of activity. This procedure is repeated for every video in the dataset. In the next step, we calculate the activity-based average of all standard deviation values. After the average standard deviation of each feature for each class is obtained, min-max normalization is conducted and the inter-class variance is calculated. While a large variance refers to a more distinctive feature, a small one refers to a non-distinctive feature. After sorting in descending order, the top-40 features are selected for the descriptor. The value of 40 is determined by observation of variance values. It is observed that there is a dramatic drop in variance values after 40. Thus, the size of the joint-joint horizontal distance matrix of one video is 160 × 40.

The above procedure is also carried for the calculation of joint-joint vertical distance, and it gives an output matrix with a size of 160 × 40. Table 1 shows the selected top-40 joint pairs for horizontal and vertical distances. As seen in Table 1, there is a strong correlation between the vertical and horizontal pairs.

2.2. **SVD on shifted feature matrix.** We use SVD to reduce the dimensions of the three feature sets and merge the resulting vectors to create the final descriptor. If we recall from previous sections, we extract three matrices with sizes 160 × 8, 160 × 40, 160 × 40 from each video, in which frames are the rows and features are columns. If the SVD is applied to these matrices it will give out non-negative singular values with amounts of 8, 40, and 40, respectively. However, rows in the matrices are sequentially located feature vectors of consecutive frames. Thus, information from each row has to be used to represent one full cycle of the action. On the other hand, the application of SVD on these matrices would lead to data loss. Thus, we increase the matrix sizes to obtain 160 singular values from each matrix. For a matrix $\mathbf{A}$ with size of $160 \times K$, the n^{th} row is $\mathbf{a}_n = \left[a_n^1, \dots, a_n^K\right]$ where $n \in [1, 160]$. In (2), each row entry is shifted by $(n-1)$ to the right. While each value in the diagonal is a feature vector of that frame, all the values outside the diagonal are set to zero. The new size of the reshaped matrix $\mathbf{A}$ is increased to $160 \times 160K$.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{160} \end{bmatrix} \xrightarrow{\text{After resizing}} \begin{bmatrix} \mathbf{a}_1 & & \\ & \ddots & \\ & & \mathbf{a}_{160} \end{bmatrix} \tag{2}$$

The above method is applied to the matrices of three features. The sizes of these matrices are changed as 160 × 8 → 160 × 1280, 160 × 40 → 160 × 6400 and 160 × 40 → 160 × 6400 for joint-joint angle, joint-joint vertical and horizontal distances respectively. When we apply SVD to the new matrices, we obtain 160 singular values from each. These values are combined to create the final descriptor with size 1 × 480 which will be used for action recognition.

3. **Experimental Results.** In this paper, we evaluate the performance of our descriptor and compare it with the state of the art on KTH [17] and Weizmann [18] datasets by using SVM classifier. We use radial basis function (RBF) kernel with parameters $\gamma = 0.01$ and $C = 100$. KTH dataset consists of six different human activities: running, walking, jogging, hand waving, hand-clapping, and boxing. The activities are performed by 25 subjects in four different environments. These are indoors, outdoors, outdoors with scale variation, and outdoor with different clothes. The dataset contains 2,391 sequences with homogeneous backgrounds. The size of images is 160 × 120. We use the same settings given in [19]. The accuracy corresponds to the average of the recognition rates for each

Table 1. Selected distance features

Start joint	End joint (V)	End joint (H)
neck	right shoulder	right shoulder
	right elbow	right elbow
	spine base	spine base
	left hip	left hip
left shoulder	right elbow	right elbow
	right wrist	right wrist
	left hip	left hip
left elbow	left knee	left knee
	left ankle	left ankle
	right hip	right hip
	right knee	right knee
	right ankle	right ankle
	right shoulder	right shoulder
	right elbow	right elbow
	left hip	left hip
	spine base	–
left wrist	left knee	left knee
	left ankle	left ankle
	right knee	right knee
	right ankle	right ankle
	right elbow	right elbow
	right wrist	right wrist
	spine base	–
	right hip	–
right shoulder	left knee	left knee
	right hip	right hip
	left hip	left hip
	–	right knee
right elbow	left knee	left knee
	left ankle	left ankle
	right hip	right hip
	left hip	left hip
	spine base	spine base
	–	right wrist
right wrist	left hip	left hip
	left ankle	left ankle
	right hip	right knee
spine base	right hip	right hip
	right knee	right knee
left hip	right knee	right knee
	right ankle	right ankle
left knee	right ankle	right ankle
nose	right wrist	–

human action class. Table 2 shows the recognition results of our method and state-of-the-art studies. The recognition accuracy of our model is 96.51%. Although our method could not outperform all of the studies, it showed a comparable recognition performance. We took only the recent studies of the last three years into the benchmark. In [19], authors use a novel bag of visual words encoding scheme which achieves 98.4% accuracy rate. In [13]

Table 2. Average recognition accuracy of our method and state-of-the-art methods for KTH dataset

Method	Average accuracy (%)
Naveed et al. [20]	92.30
Cho and Byun [21]	94.55
Khare et al. [22]	95.72
Aslan et al. [14]	96.14
Our method	**96.51**
Cortes et al. [19]	98.40
Sharif et al. [13]	99.90

authors employ a novel weighted segmentation method and also a rank correlation-based feature selection approach which gives out 99.9% accuracy rate.

Our method was also tested on the Weizmann dataset. This dataset contains 90 low-resolution videos. There are 10 different activities which are running, jumping in place, jumping forward, bending, waving with one hand, jumping jack, jumping sideways, jumping on one leg, walking, and waving with two hands. Each activity is performed by 9 different subjects. The experimental setup in this dataset is based on leave-one-person scheme. Then, for each subject, there are 10 videos, which correspond to the testing set and 80 to the training set.

Table 3 shows that our method outperforms all of the state-of-the-art results by giving a 100% recognition rate. The accuracy is the average of the recognition rates of 10 activities. Similar to the results of KTH dataset, the highest two performances belong to [13] and [19].

Table 3. Average recognition accuracy of our method and state-of-the-art methods for Weizmann dataset

Method	Average accuracy (%)
Aslan et al. [14]	91.11
Naveed et al. [20]	92.70
Xiao and Song [23]	96.50
Cortes et al. [19]	96.60
Sharif et al. [13]	98.12
Our method	**100**

4. **Conclusion.** We propose an efficient model for human activity recognition in 2D videos. Our model applies SVD on a biometric feature set which consists of joint-joint angle, joint-joint horizontal, and joint-joint vertical distances between selected joints. The output of SVD is used as the main descriptor. The experiments show that our model outperformed state-of-the-art studies in Weizmann and also obtained comparable results in the KTH dataset. For future studies, we are aiming to adopt our model for activity recognition in real-time videos.

REFERENCES

[1] M. T. Uddin and M. A. Uddin, Human activity recognition from wearable sensors using extremely randomized trees, *Int. Conf. Electr. Eng. Inf. Commun. Technology*, pp.769-778, 2015.

[2] A. Jalal et al., Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home, *Indoor Built Environment*, vol.22, no.1, pp.271-279, 2013.

[3] Y. Zhan and T. J. Kuroda, Wearable sensor-based human activity recognition from environmental background sounds, *Journal of Ambient Intelligence and Humanized Computing*, vol.5, no.1, pp.77-89, 2014.

[4] A. Jalal and M. A. Zeb, Collaboration achievement along with performance maintenance in video streaming, *Int. Conf. Comput. Inf. Technology*, Dhaka, Bangladesh, pp.369-374, 2007.

[5] A. Jalal, S. Kim and B. J. Yun, Assembled algorithm in the real-time H.263 codec for advanced performance, *Int. Workshop Enterprise Netw. Comput. Healthcare Industry*, pp.295-298, 2005.

[6] A. Jalal and S. Kim, Algorithmic implementation and efficiency maintenance of real-time environment using low-bitrate wireless communication, *IEEE Workshop Softw. Technol. Future Embedded Ubiquitous Syst.*, pp.81-88, 2006.

[7] A. Jalal, M. Z. Uddin and T. Kim, Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home, *IEEE Transactions on Consumer Electronics*, vol.58, no.3, pp.863-871, 2012.

[8] S. Kamal, A. Jalal and D. Kim, Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM, *Journal of Electrical Engineering and Technology*, vol.11, no.6, pp.1857-1862, 2016.

[9] A. Jalal, Y. Kim and D. Kim, Ridge body parts features for human pose estimation and recognition from RGB-D video data, *Int. Conf. Comput., Commun. Netw. Technol.*, pp.1-6, 2014.

[10] A. Jalal et al., Human activity recognition via the features of labeled depth body parts, *Lecture Notes in Comput. Sci.*, vol.7251, pp.246-249, 2012.

[11] A. Jalal, T. K. Jeong and T. S. Kim, Development of a life logging system via depth imaging-based human activity recognition for smart homes, *Int. Symp. Sustainable Healthy Buildings*, pp.91-95, 2012.

[12] A. Jalal and S. Kamal, Real-time life logging via a depth silhouette-based human activity recognition system for smart home services, *Int. Conf. Adv. Video Signal Based Surveillance*, pp.74-80, 2014.

[13] M. Sharif, M. A. Khan, F. Zahid, J. H. Shah and T. Akram, Human action recognition: A framework of statistical weighted segmentation and rank correlation-based selection, *Pattern Analysis and Applications*, vol.23, pp.281-294, 2020.

[14] M. F. Aslan, A. Durdu and K. Sabanci, Human action recognition with bag of visual words using different machine learning methods and hyperparameter optimization, *Neural Computing and Applications*, DOI: 10.1007/s00521-019-04365-9, 2019.

[15] O. F. Ince, I. F. Ince, M. E. Yildirim, J. S. Park, J. K. Song and B. W. Yoon, Human activity recognition with analysis of angles between skeletal joints using a RGB-depth sensor, *ETRI Journal*, vol.42, pp.78-89, 2020.

[16] S. Arciniegas-Alarcón, M. García-Peña and W. J. Krzanowski, Imputation using the singular value decomposition: Variants of existing methods, proposed and assessed, *International Journal of Innovative Computing, Information and Control*, vol.16, no.5, pp.1681-1696, 2020.

[17] C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: A local SVM approach, *International Conf. Pattern Recognition (ICPR)*, pp.32-36, 2004.

[18] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, Actions as space-time shapes, *International Conf. Computer Vision (ICCV)*, pp.1395-1402, 2005.

[19] X. Cortes, D. Conte and H. Cardot, A new bag of visual words encoding method for human action recognition, *International Conf. Pattern Recognition (ICPR)*, pp.2480-2485, 2018.

[20] H. Naveed, G. Khan, A. U. Khan, A. Siddiqi and M. U. G. Khan, Human activity recognition using mixture of heterogeneous features and sequential minimal optimization, *International Journal of Machine Learning and Cybernatics*, vol.9, pp.2329-2340, 2019.

[21] S. Y. Cho and H. R. Byun, Human activity recognition using overlapping multi-feature descriptor, *Electronics Letters*, vol.47, pp.1275-1277, 2011.

[22] M. Khare, J. Gwak and M. Jeon, Complex wavelet transform-based approach for human action recognition in video, *International Conference on Control, Automation and Information Sciences (ICCAIS)*, pp.157-162, 2017.

[23] Q. Xiao and R. Song, Action recognition based on hierarchical dynamic Bayesian network, *Multimedia Tools and Applications*, vol.77, pp.6955-6968, 2018.