Pattern Recognition Letters 000 (2016) 1–10

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Classification of human actions using pose-based features and stacked auto encoder☆

Earnest Paul Ijjina*, Krishna Mohan C

Visual Learning and Intelligence Group (VIGIL), Department of Computer Science & Engineering, Indian Institute of Technology Hyderabad, Telangana 502285, India

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Human action recognition
Stacked auto encoder
Pose-based features
Fuzzy membership functions

ABSTRACT

In this paper, we propose a method for classification of human actions using pose based features. We demonstrate that statistical information of key movements of actions can be utilized in designing an efficient input representation, using fuzzy membership functions. The ability of stacked auto encoder to learn the underlying features of input data is exploited to recognize human actions. The efficacy of the proposed approach is demonstrated on CMU MOCAP and Berkeley MHAD datasets.

1. Introduction

Human action recognition is one of the major areas of research in computer vision due to its wide range of applications in video surveillance, ambient assisted living, robot vision, augmented reality, video indexing and retrieval, to name a few. The introduction of 3D cameras like Kinect and motion capture (MOCAP) systems in the last decade widened the modalities available for human action recognition. Even though Kinect was a low-cost RGB-D camera by Microsoft for their XBox gaming platform, it triggered a huge interest in the computer vision research community [6] to use RGB-D and pose based information for action and gesture recognition. The most common approach used for motion capture is by tracking a wearable marker or by using a skeletal prediction framework utilizing RGB-D video stream. Recent studies conducted to identify the relative importance of skeletal joints by Ofli et al. [14] suggest that some joints provide better discriminative information to recognize actions. Also, experimental study by Jhuang et al. [7] suggests that better discrimination among human actions can be achieved using high-level pose features compared to mid-level and low-level features. This gives motivation for using the pose based features for human action recognition.

A majority of pose based action recognition approaches use tracking information of various skeletal joints to compute features for action recognition. Features based on joint distance and joint motion are evaluated by Yun et al. [23] to recognize human interaction using support vector machine and multiple instance learning. It is observed that features computed from distance between pair of joints outperform other geometric features. A local view-invariant skeletal descriptor, skeletal quads is proposed by Evangelidis et al. [5]. A Gaussian mixture model (GMM) learnt on the training data is used to encode the quad as a fisher vector which is inturn used by the support vector machine (SVM) for classification. To capture the joint shape motion ques in a depth image, HON4D, a descriptor for activity recognition using depth videos is proposed by Oreifej and Liu [15] using SVM for classification. Models with inhomogeneous symmetric bias are trained with examples from an action domain in [20] and [21] for correcting the estimated human-pose. A framework for correcting human pose estimated from Kinect depth images, by combining the outputs of a random forest regression model and a pose prior model learned on motion capture data (using von Mises–Fisher distribution) is proposed by Shen et al. [19]. A hierarchical recurrent network fusing the pose information from five parts of the skeletal structure is proposed by Du et al. [4] to recognize actions from the temporally accumulated output. For action recognition in RGB videos, action-bank features extracted from visual information are used to train discriminative dictionaries using 'label consistent K-SVD' algorithm by Jiang et al. [9] to recognize human actions. Lin et al. [10] modeled human trajectories as heat sources to recognize group activities from the similarity of heat-maps. Prest et al. [18] combined human detection, object detection and tracking techniques to recognize human-human and human-object interactions. A 3D shape retrieval model using auto encoder for learning features from 2D projections of 3D shapes is proposed by Zhu et al. [25]. Ji et al. [8] used gray-level, gradient and optical-flow information of RGB videos as inputs to a 3D convolutional neural network for recognizing human actions. Xia et al. [22] recognized human actions by modeling the temporal evolution of pose associated with an action by a hidden Markov model (HMM).

☆ This paper has been recommended for acceptance by Xiang Bai.

* Corresponding author. Tel.: +91 9494466490.
E-mail address: cs12p1002@iith.ac.in, earnest.prof@gmail.com (E.P. Ijjina).

http://dx.doi.org/10.1016/j.patrec.2016.03.021
0167-8655/

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

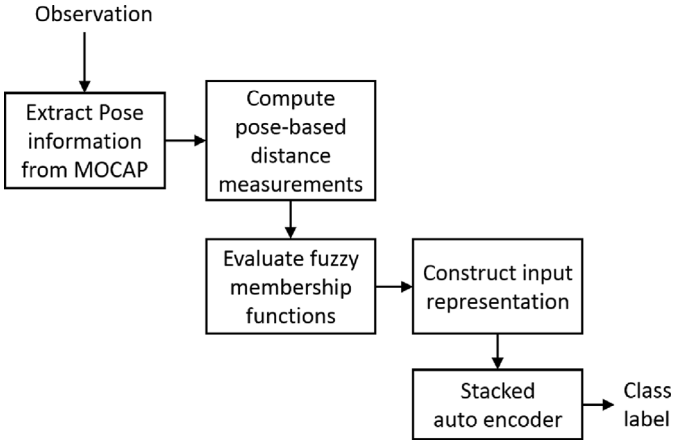


Fig. 1. Block diagram of the proposed approach for human action recognition.

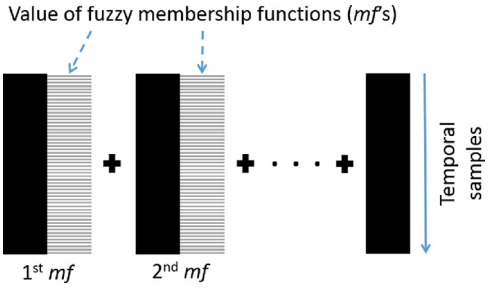


Fig. 2. Computation of input representation from values of fuzzy membership functions.

The effectiveness of human action recognition approaches is significantly affected by the features used for recognition and the computational complexity involved. Further, human action recognition poses additional challenges due to: (a) the existence of alternative limb movements for actions, and (b) the lack of synchronization of movements involved in an action. In this paper, we address these issues by considering MOCAP skeleton information of a small number of joints for feature extraction and a stacked auto encoder (to learn the underlying features) for classification. The reminder of this paper is organized as follows: Section 2 describes the proposed approach and the steps involved in the computation of input representation utilizing the domain knowledge of the actions. Section 3 elaborates the utilization of domain knowledge through statistical analysis for representing actions of MOCAP datasets. The classification results and analysis of features learned by the stacked auto encoder are also presented. Finally, Section 4 gives concluding remarks and the future work.

2. Proposed approach

In this work, we propose the use of pose based features computed from motion capture (MOCAP) information for human action recognition using stacked auto encoder. The block diagram of the proposed approach is shown in Fig. 1. The MOCAP information corresponding to the input observation is used to extract the temporal variation of the subject's pose during the execution of the action. This pose information is used to compute pose-based distances from a small number of skeletal joints. These distances are evaluated by a set of fuzzy membership functions, that are designed to emphasize the unique motion pattern of each action. The membership values from these functions is concatenated (with zero data) to generate the input representation of the action, as shown in Fig. 2. In the figure, '+' represents concatenation, black strips represent zero data and the width of each vertical strip is 2 pixels (i.e., values are duplicated). This representation of actions is given as input to a stacked auto encoder for action recognition. The architecture of the stacked auto encoder used for classification is shown in Fig 3. The first and second layers consists of 100 and 50 neurons, respectively. The last layer consists on n neurons, where n represents the number of action types (classes) to be recognized. The weights of the first two layers are initialized through pre-training. The last layer is a soft-max layer trained to generate a binary output with 1 for the predicted class and 0's for the rest. The next section covers the experimental evaluation of this approach on MOCAP datasets.

3. Experimental study

As explained in the previous section, distances computed from MOCAP skeletal-joints are used in the computation of input representation. The joints considered in this study are shown in Fig. 4 with red color. Following are the distances computed from these joints.

a: displacement between the left and right hand
b: height of right-hand above the ground
c: height of left-hand above the ground
d: height of pelvis above the ground
p: height of right-leg above the ground
q: height of left-leg above the ground

The intuition behind considering these joints and distances will be discussed in the experimental setup for each dataset. As MOCAP information contains the tracking information of human-joints over time (t), the value of these distance variables changes with time. If x represents a distance variable, then its value at time t is represented by $x(t)$. The maximum, minimum and range of x in an observation are denoted by x_{max}, x_{min} and x_r, respectively. The next section covers the experimental evaluation of the proposed approach on CMU MOCAP dataset.

3.1. CMU MOCAP dataset

The CMU MOCAP dataset [1] consists of motion capture (MOCAP) information corresponding to locomotion actions performed by various subjects. The locomotion activities and their variations considered is this evaluation are given in Table 1. The next section describes the utilization of domain knowledge about these actions in the design of an effective input representation.

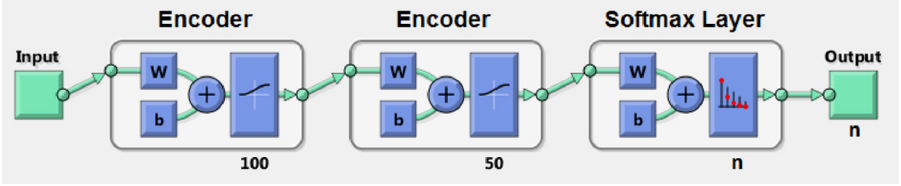


Fig. 3. Architecture of the stacked auto encoder used for classification.

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

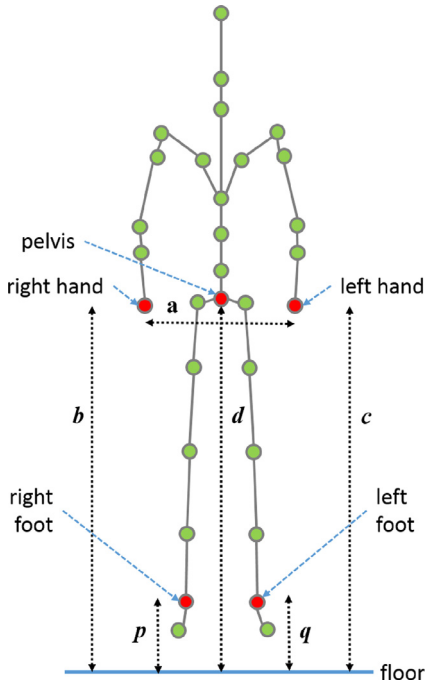


Fig. 4. MOCAP skeletal structure depicting the pose based distances considered for action recognition.

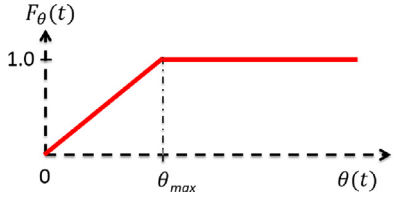


Fig. 5. Graphical representation of the membership function $F_\theta(t)$.

Table 1
Variations in locomotion actions considered for evaluation.

Action	Variations
Jump	Jump
	Forward jump
	High jump
	Jump up and down, hop on one foot
Run	Run
	Run, sudden-stop
	Run, veer left/right
	Run, 90-degree left/right turn
	Run around in a circle
Walk	Walk
	Slow walk
	Walk, exaggerated stride
	Navigate-walk forward, backward, sideways
	Walk/wander
	Walk, veer left/right
	Walk, 90-degrees left/right turn
	Slow-walk, stop
	Walk with anger, frustration
	Walk stealthily
	Walk/ hobble
	Whistle, walk jauntily
	Muscular, heavyset persons walk
	Walk forward, turn around, walk backward
	Walk around, frequent turns,
	Cycling walk along a line
	Navigate-walk forward, backward on a diagonal
	Navigate-walk forward, backward,
	Sideways on a diagonal
	Walk around

Table 2
Observations considered to generate the training dataset.

Action	Sample	P2P	# Of window
	(subject:trial)	Distance	Slides
Jump	13:40	36	9
Run	09:01	92	23
Walk	39:03	128	32

Table 3
Confusion matrix of the proposed approach for CMU MOCAP dataset.

		Predicted class label		
		Jump	*Run*	*Walk*
Actual class label	*Jump*	17		
	Run		41	5
	Walk			135

Table 4
Normalization of distance variables *a*, *b*, *c* and *d*.

Measurement	Normalization
a	Divided by the distance between the hips
b	Divided by the height of left shoulder
c	Divided by the height of right shoulder
d	Subtract and divide by the value of *d* in T-pose

3.1.1. Input representation

The domain knowledge for *jump*, *run* and *walk* actions suggests that these actions involve periodic movement of feet (lift-up and put-down the foot) and they only differ in the duration and relative (simultaneous or alternative) motion of the feet. For *walk* and *run* actions, the left and right foot move alternatively, whereas for *jump* action, the feet move at the same time. Similarly, the duration between successive foot movement is different for *walk* and *run* actions. This domain knowledge suggests that the height of the feet above ground i.e., distances p and q could be used in the computation of an input representation, suitable for discrimination. As the height of the feet above the ground may change between observations, they need to be normalized within each observation. The normalization of these variables is achieved using the fuzzy membership function $F_\theta(t)$ given in Eq. (1), whose plot is shown in Fig. 5. Here, θ represents the variable p or q. The plot of $F_p(t)$ and $F_q(t)$ for a typical *jump*, *run* and *walk* action is given in Fig. 6. The plots suggest that the nature of variation of these variables (i.e., synchronous or asynchronous change and the frequency) is different for these actions. It also suggests the periodic nature of these actions i.e., re-occurrence of the same foot movement. The membership values are concatenated to generate the input representation.

$$F_\theta(t) = \begin{cases} 0 & \text{if, } \theta(t) \le 0 \\ \frac{\theta(t)}{\theta_{max}} & \text{if, } 0 < \theta(t) < \theta_{max} \\ 1 & \text{if, } \theta(t) \ge \theta_{max} \end{cases} \quad (1)$$

The periodic nature of these locomotion actions can be used in the design of a one-shot training model in which a single observation per class is used to train the model. We considered a temporal window of 104 samples and shift it 4 samples at a time on the training observation, to obtain the training cases. The details of the training cases generated from each training observation is given in Table 2. As *jump* action is performed only once, we considered 34 (α-cut at 0.7) as the P2P distance. Some of the training cases generated for these actions is shown in Fig 7. All the observations of actions described in the second column of Table 1, excluding the 3

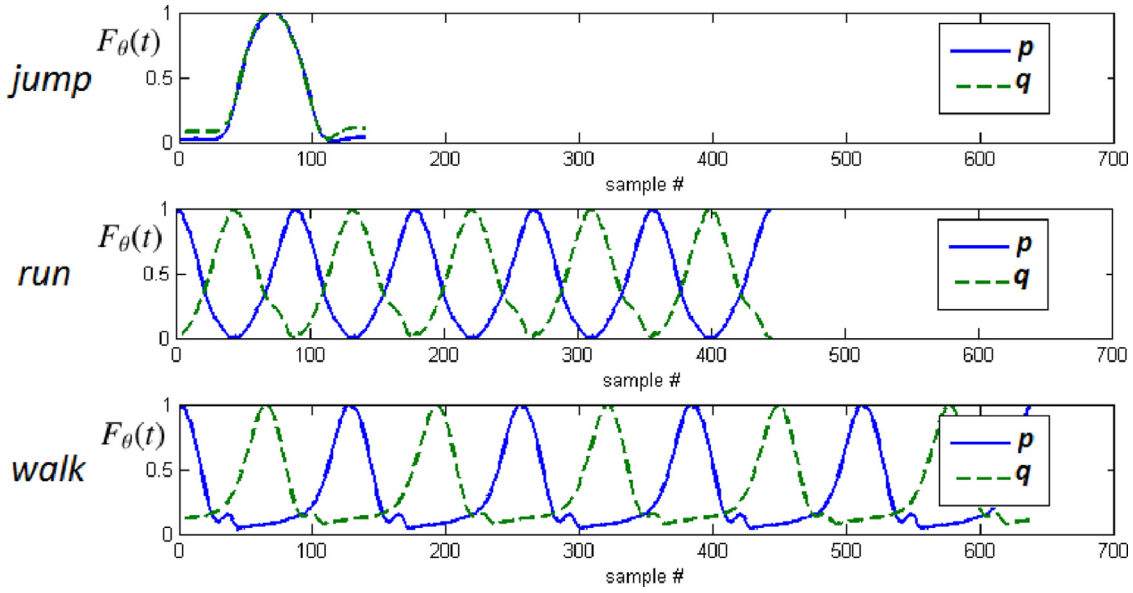


Fig. 6. Temporal variation of height of feet above ground for *jump*, *run* and *walk* actions.

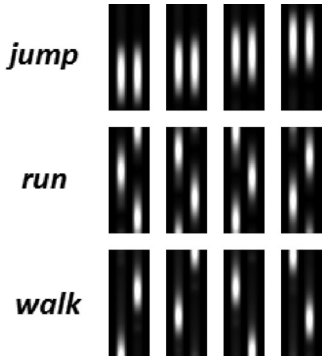


Fig. 7. Training cases for jump, run and walk actions.

observations used in training, are used as the test dataset. The first 104 temporal samples of observations is used in the computation of input representation. The next section covers the experimental results and analysis.

3.1.2. Results and analysis

The representation discussed in the previous section is down-sampled to a 26 × 10 representation and used as input to a stacked auto encoder (SAE) with three layers for action recognition. The number of neurons in first and second layer are 100 and 50, respectively. The last layer is a soft-max layer trained to generate a binary output with 1 for the predicted class and 0's for the rest. The model is trained for 500 epochs using conjugate gradient descent algorithm and evaluated on test dataset. The confusion matrix of the proposed approach is given in Table 3. The misclassification of 5 out of 198 observations results in a recognition accuracy of 97.47%. We have also conducted experiments of the proposed action representation using convolutional neural network (CNN) classifier. The recognition performance using CNN classifier is 94.44%, which is less than the performance of the proposed approach. The features learned by the 100 neurons in the first layer of SAE classifier are shown in Fig. 8. The high recognition accuracy of the proposed approach suggests an efficient input representation and an effective recognition model. The next section covers the experimental evaluation of the proposed approach on MHAD dataset.

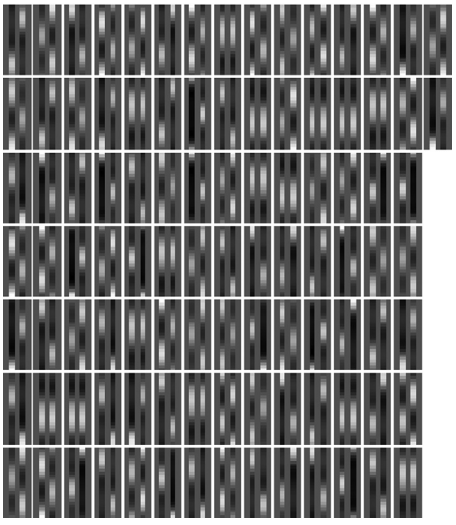


Fig. 8. Features learned by the first layer of stacked auto encoder for CMU MOCAP dataset.

3.2. Berkeley MHAD dataset

The Berkeley MHAD dataset [13] consists of MOCAP information of 11 actions performed by 12 subjects. We considered three joints to compute 4 distances which are later used in input representation. As these distances are dependent on the height of the subject and the length of limbs, they are normalized using the reference pose (T-pose) of the subject as explained in Table 4. All future references of *a*, *b*, *c* and *d* refer to these normalized distance variables. The change in pose (MOCAP skeletal structure) for the actions in Berkeley MHAD dataset along with the variation of the four distance variables *a*, *b*, *c* and *d* is shown in Figs 9 and 10. The front-view and side-view of some of the key poses that appear during the execution of these actions are shown in these figures in blue

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

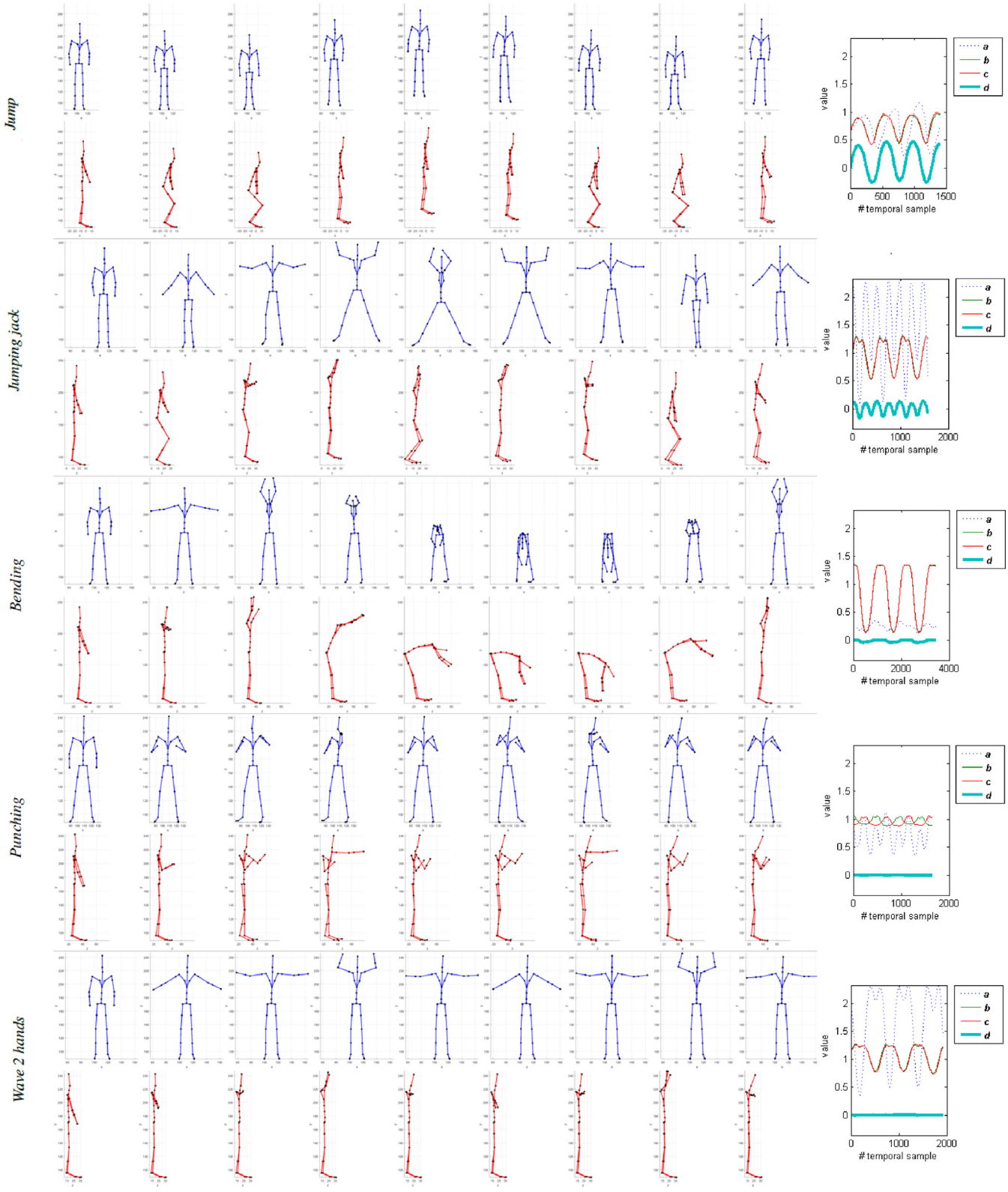


Fig. 9. Variation in MOCAP skeletal structure and distance variables {*a*, *b*, *c*, *d*} when performing *jump*, *jumping jack*, *bending*, *punching* and *wave 2 hands* actions.

and red color, respectively. The next section explains how these four distances are used in the design of an input representation.

3.2.1. Input representation

As explained in the previous section, four distance measurements are used to recognize 11 actions in the MHAD dataset. From Figs. 9 and 10, it can be observed that these actions differ in the nature and range of limb (hand and leg) movement. Thus, representing actions by their unique motion may result in an action representation, more suitable for discrimination. The domain knowledge about these action suggests that, jump related actions like *jumping* and *jumping jack* can be recognized by the height

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

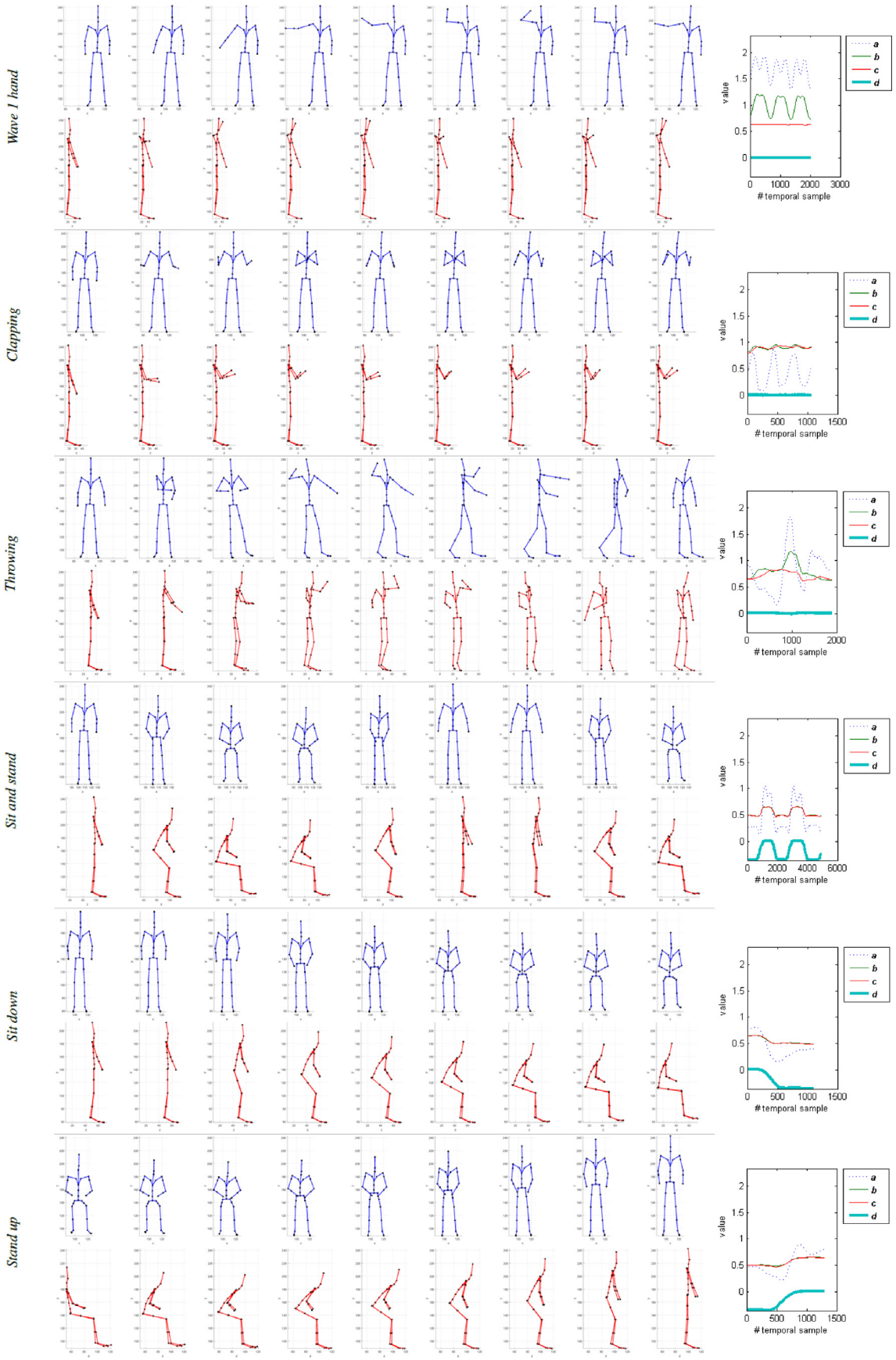


Fig. 10. Variation in MOCAP skeletal structure and distance variables {*a*, *b*, *c*, *d*} when performing *wave 1 hand*, *clapping*, *throwing*, *sit and stand*, *sit down* and *stand up* actions.

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

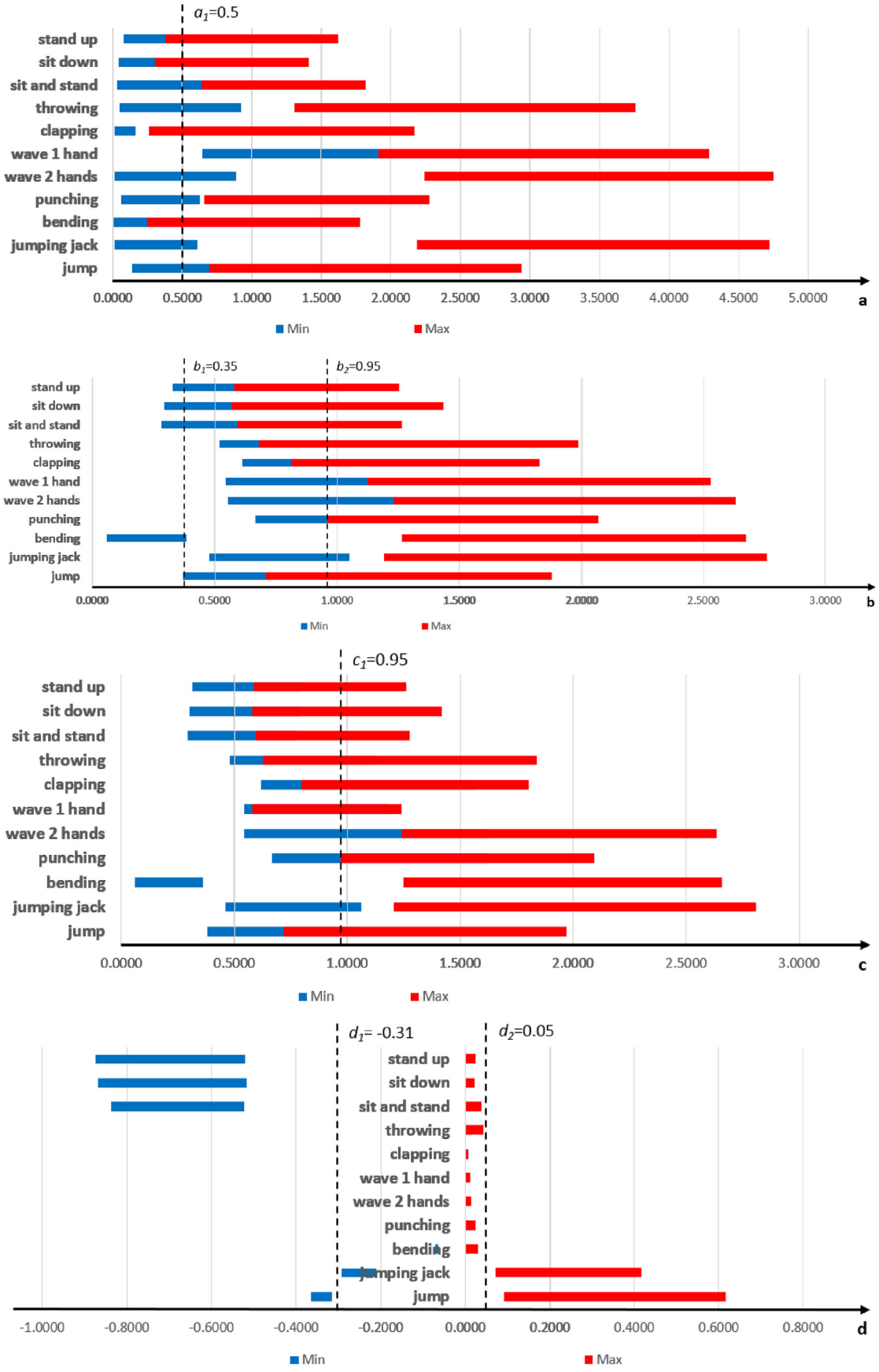


Fig. 11. The plot of range of minimum (in blue) and maximum (in red) values of the distance variables *a*, *b*, *c*, *d* for observations in MHAD dataset. (Best viewed in color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

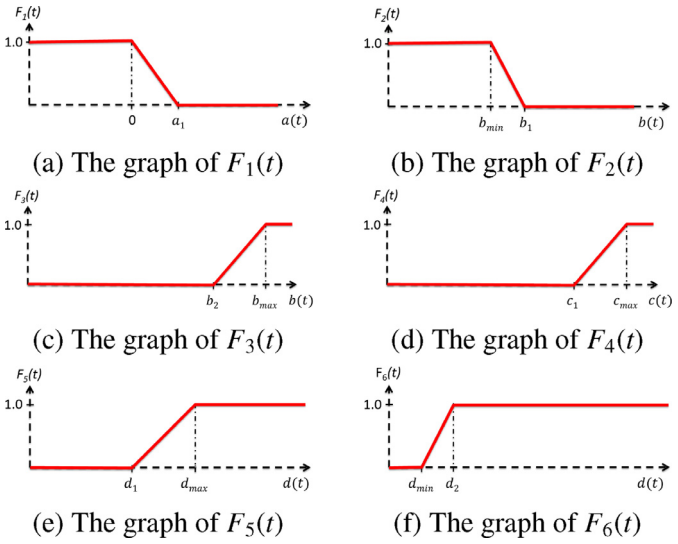


Fig. 12. Graphical representation of the six membership functions.

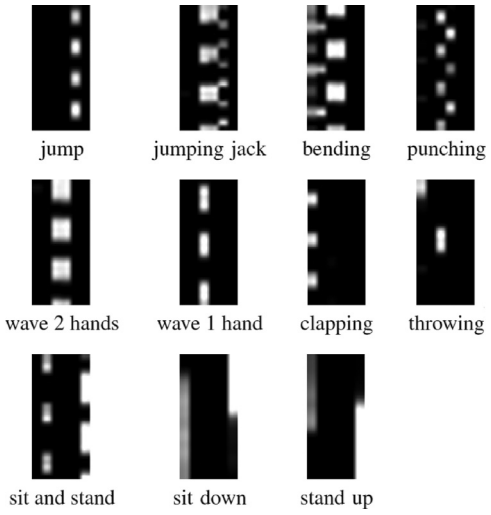


Fig. 13. Visualization of input representation for the 11 MHAD actions. (The representation is scaled-up by 3 time for better visibility).

of hands above the ground as it increases during the execution of these actions. Similarly, the hand-waving actions like *wave 1 hand* and *wave 2 hands* can be recognized by the height of the hands above the ground. It is critical to notice that both jump and hand-waving actions can be recognized from the height of the hands, but the range of variation of height is different for these two categories. Similarly, *clapping* and *jumping jack* can be discriminated from the range of variation of the distance between hands (a). The sitting based actions are characterized by the movement in the lower body and thereby can be recognized using the height of pelvis (d). This interpretation can be extended to recognize a single action from the nature and range of variation of multiple distance variables. The movements involved in these actions are statistically analyzed by plotting the range of minimum and maximum values of these four distance variables for the 11 MHAD actions in Fig. 11. For a distance variable x and an action class c, the region in blue represents the range of minimum value of x for all observations of class c. Similarly, the region in red represents the range of maximum value of x for all observations of class c i.e., the maximum value of the distance variable x, $\max(x(t))$ for any observation of class c will fall within the range corresponding to the red region. From the figure, it can be observed that the range of variation of these variables is not identical for these actions. This analysis is utilized in the design of six membership functions $F_1(t)$, $F_2(t)$, $F_3(t)$, $F_4(t)$, $F_5(t)$, $F_6(t)$ whose equations are given in Eq. (2) to Eq. (7), respectively, and their corresponding graphical representation is shown in Fig 12. The optimum value of the constants a_1, b_1, b_2, c_1, d_1 and d_2 is empirically determined to be 0.5, 0.35, 0.95, 0.95, −0.31 and 0.05, respectively. From the equations, it can be observed that the core function gets executed only when there is a significant change in the value of the variable. During the execution of these gestures, some additional movements may be needed during the beginning and ending of the action, which are not part of the key movements associated with an action. To overcome this practical constraint, we omit the first and last 20% of temporal samples of observations. The temporal variation of the membership value of these 6 fuzzy membership functions is concatenated and down-sampled to 26 temporal samples, resulting in an input representation of size 26 × 13 (i.e., the width of each vertical strip in Fig. 2 is 1 pixel). The typical input representation of the 11 gestures is shown in Fig 13. The next section describes how this temporal representation is used for action recognition using a stacked auto encoder.

The membership functions are:

$$F_1(t) = \begin{cases} 1 & \text{if, } a(t) < 0 \\ 1 - \frac{a(t)}{2} & \text{if, } 0 < a(t) \le a_1 \\ 0 & \text{if, } a(t) \ge a_1 \end{cases} \quad (2)$$

$$F_2(t) = \begin{cases} 0 & \text{if, } b_r < 0.02 \\ 1 & \text{if, } b_r \ge 0.02, b(t) \le b_{min} \\ \frac{(b_1 - b(t))}{(b_1 - b_{min})} & \text{if, } b_r \ge 0.02, b_{min} < b(t) < b_1 \\ 0 & \text{if, } b_r \ge 0.02, b(t) \ge b_1 \end{cases} \quad (3)$$

$$F_3(t) = \begin{cases} 0 & \text{if, } b_r < 0.02 \\ 0 & \text{if, } b_r \ge 0.02, b(t) \le b_2 \\ \frac{(b(t) - b_2)}{(b_{max} - b_2)} & \text{if, } b_r \ge 0.02, b_2 < b(t) < b_{max} \\ 1 & \text{if, } b_r \ge 0.02, b(t) \ge b_{max} \end{cases} \quad (4)$$

$$F_4(t) = \begin{cases} 0 & \text{if, } c_r < 0.02 \\ 0 & \text{if, } c_r \ge 0.02, c(t) \le c_1 \\ \frac{(c(t) - c_1)}{(c_{max} - c_1)} & \text{if, } c_r \ge 0.02, c_1 < c(t) < c_{max} \\ 1 & \text{if, } c_r \ge 0.02, c(t) \ge c_{max} \end{cases} \quad (5)$$

$$F_5(t) = \begin{cases} 0 & \text{if, } d_r < 0.02 \\ 0 & \text{if, } d_r \ge 0.02, d(t) \le d_1 \\ \frac{d(t) - d_1}{d_{max} - d_1} & \text{if, } d_r \ge 0.02, d_1 < d(t) < d_{max} \\ 1 & \text{if, } d_r \ge 0.02, d(t) \ge d_{max} \end{cases} \quad (6)$$

$$F_6(t) = \begin{cases} 0 & \text{if, } d_r < 0.02 \\ 0 & \text{if, } d_r \ge 0.02, d(t) \le d_{min} \\ \frac{d(t) - d_{min}}{d_2 - d_{min}} & \text{if, } d_r \ge 0.02, d_{min} < d(t) < d_2 \\ 1 & \text{if, } d_r \ge 0.02, d(t) \ge d_2 \end{cases} \quad (7)$$

3.2.2. Results and analysis

The representation described in the previous section is given as input to a stacked auto encoder to recognize the human actions. The first hidden layer of the stacked auto encoder is trained to learn the features necessary to reconstruct the input at its output. A second hidden layer is trained to reconstruct the features learned by the first hidden layer and a soft max layer at the end is used for classification. The proposed approach is evaluated on MHAD dataset using 5-fold cross validation. The stacked auto encoder is trained using conjugate gradient descent algorithm [12] for 400

Table 5
Confusion matrix of the proposed approach for 5-fold cross validation on Berkeley MHAD dataset.

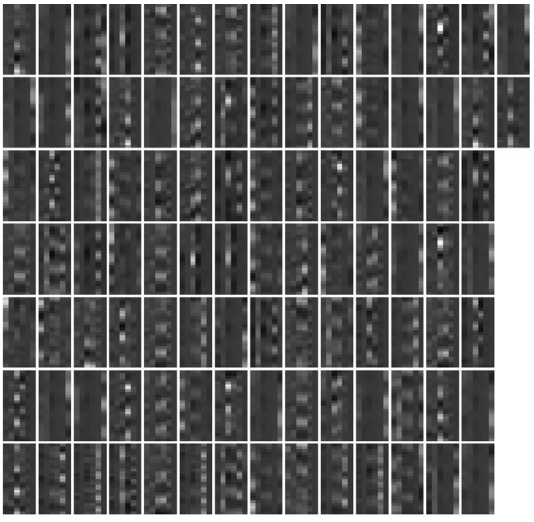
	Jump	Jumping-jack	Bending	Punching	Wave-2-hand	Wave-1-hands	Clapping	Throwing	Sit & stand	Sit-down	Stand-up
Jump	100										
Jumping-jack		96.7			3.3						
Bending		1.7	96.7		1.7						
Punching				100							
Wave-2-hand		1.7			98.3						
Wave-1-hands						100					
Clapping					1.7		95.0		1.7		1.7
Throwing				1.7				96.7	1.7		
Sit & stand-up									96.7	1.7	1.7
Sit-down										100	
Stand-up							1.7				98.3

Table 6
Performance of different human action recognition approaches on Berkeley MHAD dataset.

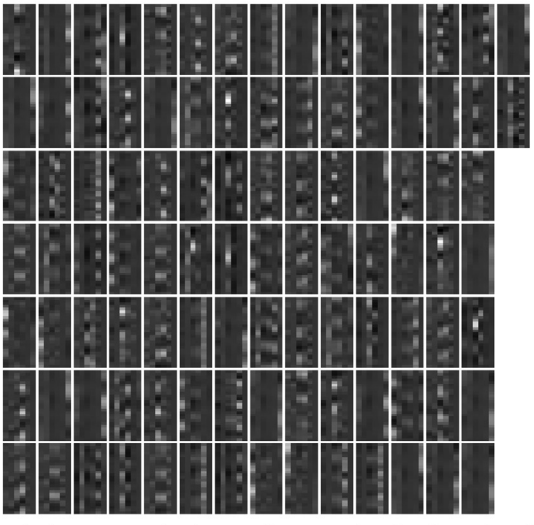
Approach	Accuracy
Bag of words by Foggia et al. [16]	72.9
Kernel SVM by Ofli et al. [13]	79.93
Modeling of styles by Cheema et al. [3]	89.85
Deep learning by Foggia et al. [17]	85.8
Conditional RBM by Mocanu et al. [11]	82.42
Cloud sequence by Zhang et al. [24]	85.7
Edit distance by Brun et al. [2]	87.1
SMIJ by Ofli et al. [14]	95.37
Proposed representation with CNN classifier	97.27
Proposed approach	**98.03**

epochs during unsupervised feature learning and fine-tuning of weights. An average classification accuracy of 98.03% is obtained using the proposed approach whose confusion matrix is given in Table 5. The performance comparison of various approaches for 5-fold cross validation on MHAD dataset is shown in Table 6. From this table, it can be observed that the performance of the proposed approach is better than many approaches even when MOCAP information of only 3 joints is considered in input representation. Even though the performance of proposed approach using CNN classifier is close to SAE, we prefer SAE classifier as its performance is less sensitive to the initial weights (due to its layer-wise pre-training), when compared to a CNN classifier. This indicates the effectiveness of input representations (utilizing statistical information about actions) and the stacked auto encoder classification framework.

The features learned by the first layer of stacked auto encoder for the first two splits in 5-fold cross validation on MHAD dataset is given in Fig. 14. From the figure, it can be observed that similar features are learned across both the folds. The input representation and the reconstructed input using the features learned by the first layer of stacked auto encoder, for some observations is given in Fig 15. From the figure, it can be observed that the reconstruction of input signal eliminates noise and normalizes the representation across observations, which could be the reason for the effectiveness of stacked auto encoder.



(a) Stacked auto encoder feature for first split of MHAD dataset



(b) Stacked auto encoder feature for second split of MHAD dataset

Fig. 14. Features learned by the first layer of stacked auto encoder for (a) the first split and (b) the second split, for 5-fold cross validation of MHAD dataset.

4. Conclusion

This paper presents an input representation based on pose features for human action recognition using stacked auto encoder. It is shown that an effective MOCAP action representation can be built by utilizing the domain knowledge about the key movements of the actions. The unsupervised feature learning capability of the stacked auto encoder is exploited to learn the discriminative features required for classification. The low misclassification error of the proposed approach for CMU MOCAP dataset and 5-fold cross validation of MHAD dataset suggests the effectiveness of the input representation and the stacked auto encoder classification framework. Future work will extend this approach to other multi-modal datasets containing other modalities of data like acceleration, audio etc.

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021

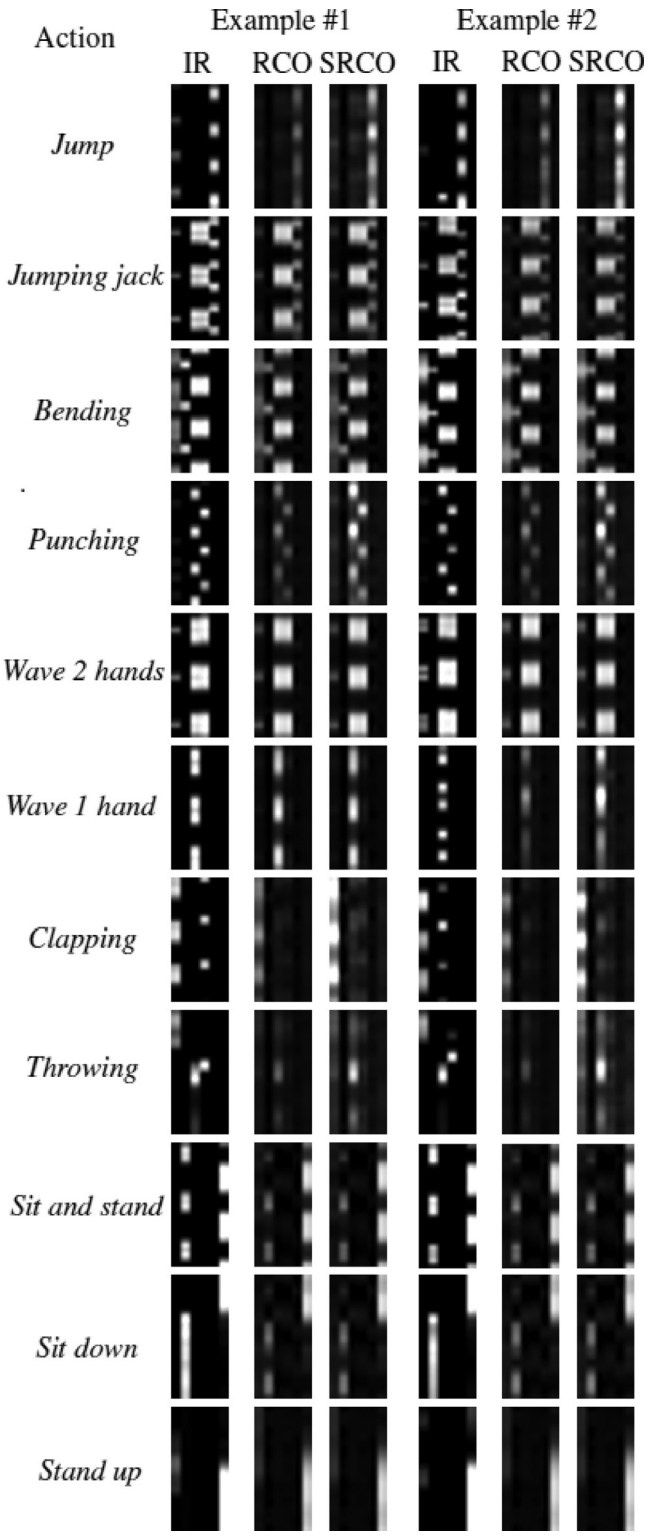


Fig. 15. Typical input representation and its reconstruction using the first layer of stacked auto encoder. (Here, IR is input representation, RCO is reconstructed output using the 1st layer of stacked auto encoder and SRCO is the scaled reconstructed output to to reach a maximum gray value of 255. (The representation is scaled-up for better visualization).

References

[1] Cmu human motion capture database, (http://mocap.cs.cmu.edu/). (accessed: 02.09.16).
[2] L. Brun, P. Foggia, A. Saggese, M. Vento, Recognition of human actions using edit distance on aclet strings, in: VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 2, Berlin, Germany, 11–14 March, 2015, 2015, pp. 97–103.
[3] M.S. Cheema, A. Eweiwi, C. Bauckhage, Human activity recognition by separating style and content., Pattern Recogn. Lett. 50 (2014) 130–138. Depth Image Analysis.
[4] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition., in: CVPR, IEEE, 2015, pp. 1110–1118.
[5] G. Evangelidis, G. Singh, R. Horaud, Skeletal quads: Human action recognition using joint quadruples, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2014, pp. 4513–4518.
[6] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: a review., IEEE Trans. Cybern. 43 (2013) 1318–1334.
[7] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M.J. Black, Towards understanding action recognition., in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 2013, pp. 3192–3199.
[8] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition., IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 221–231.
[9] Z. Jiang, Z. Lin, L.S. Davis, Label consistent k-svd: learning a discriminative dictionary for recognition., IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 2651–2664.
[10] W. Lin, H. Chu, J. Wu, B. Sheng, Z. Chen, A heat-map-based algorithm for recognizing group activities in videos., IEEE Trans. Circ. Syst. Video Technol. 23 (2013) 1980–1992.
[11] D.C. Mocanu, H.B. Ammar, D. Lowet, K. Driessens, A. Liotta, G. Weiss, K. Tuyls, Factored four way conditional restricted boltzmann machines for activity recognition, Pattern Recogn. Lett. 66 (2015) 100–108. Pattern Recognition in Human Computer Interaction.
[12] M.F. Møller, Original contribution: A scaled conjugate gradient algorithm for fast supervised learning, Neural Netw. 6 (1993) 525–533.
[13] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley mhad: a comprehensive multimodal human action database, in: Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), 2013, pp. 53–60.
[14] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (smij), J. Vis. Commun. Image Represent. 25 (2014) 24–38.
[15] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences., in: CVPR,IEEE, 2013, pp. 716–723.
[16] P. Foggia, G. Percannella, A.S.M. V., Recognizing human actions by a bag of visual words, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, IEEE SMC 2013, 2013.
[17] P. Foggia, A. Saggese, N.S.M. V., Exploiting the deep learning paradigm for recognizing human actions, in: IEEE (Ed.), Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS 2014), 2014.
[18] A. Prest, V. Ferrari, C. Schmid, Explicit modeling of human-object interactions in realistic videos, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 835–848.
[19] W. Shen, R. Lei, D. Zeng, Z. Zhang, Regularity guaranteed human pose correction, in: Asian Conference on Computer Vision, Springer, 2014, pp. 242–256.
[20] K.D. Wei Shen, X. Bai, T. Leyvand, B. Guo, Z. Tu, Exemplar-based human action pose correction and tagging, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1784–1791, doi:10.1109/CVPR.2012.6247875.
[21] K.D. Wei Shen, X. Bai, T. Leyvand, B. Guo, Z. Tu, Exemplar-based human action pose correction, IEEE Trans. Cybern. 44 (2014) 1053–1066.
[22] L. Xia, C.C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D), Rhode Island, USA, 2012, pp. 20–27.
[23] K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in: IEEE, 2012.
[24] H. Zhang, C. Reardon, C. Zhang, L.E. Parker, Adaptive human-centered representation for activity recognition of multiple individuals from 3d point cloud sequences, in: IEEE, 2015, pp. 1991–1998.
[25] Z. Zhu, X. Wang, S. Bai, C. Yao, X. Bai, Deep learning representation using autoencoder for 3d shape retrieval, in: Proceedings of the International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), IEEE, 2014, pp. 279–284.

Please cite this article as: E.P. Ijjina, K.M. C, Classification of human actions using pose-based features and stacked auto encoder, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.03.021