Human Action Recognition based on Motion Capture Information using Fuzzy Convolution Neural Networks

Earnest Paul Ijjina
Ph.D Research Scholar
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, India 502205
Email: cs12p1002@iith.ac.in

C Krishna Mohan
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad
Telangana, India 502205
Email: ckm@iith.ac.in

***Abstract*—In this paper, we propose a novel approach for human action recognition based on motion capture (MOCAP) information using a Fuzzy convolutional neural network. The MOCAP tracking information of human joints is used to compute the temporal variation of displacement between joints during the execution of an action. Fuzzy membership functions designed to emphasize the discriminative pose associated with each action are considered for feature extraction. The temporal variation of membership values associated with these fuzzy membership functions is considered as the feature representation for action recognition. A convolutional neural network (CNN) capable of recognizing local patterns in input data is trained to recognize human actions from the local patterns in the feature representation. Experimental evaluation on Berkeley MHAD dataset demonstrates the effectiveness of the proposed approach.**

***Keywords*—fuzzy convolutional neural network; human action recognition; motion capture (MOCAP) information;**

I. Introduction

Human behavior analysis is an important component in semantic analysis of multimedia content. Motion capture (MOCAP) refers to a broad range of techniques used to record the motion of subjects accurately. The MOCAP techniques range from tracking of wearable markers to human-skeletal prediction from depth and multi-view videos [1] [2]. Due to the accuracy and low noise of tracking information, it is widely used to capture ground-truth information in fields like sports, medicine and military. MOCAP information is also used to validate computer vision algorithms. Due to the availability of low-cost, high-mobility depth sensors like Microsoft Kinect with support for skeletal tracking in the last decade, there has been a significant increase in their use for human computer interaction (HCI) and marker-less motion tracking. Research on the effectiveness of MOCAP information for human action recognition [3] suggests that tracking information of certain joints provide more discriminative information for recognizing some actions. The experimental studies [4] also indicate that high-level pose features can outperform mid-level and low-level features for human action recognition. In this paper, we consider the MOCAP information of only three human joints to recognize eleven human actions.

978-1-4799-7458-0/15/$31.00

Over the last few decades, a broad range of human action recognition techniques were proposed that can recognize individual, group actions and the interaction of human-subjects with objects using different modalities. Action-bank features extracted from visual information in videos is used to learn discriminative dictionaries by Zhuolin Jiang *et al.* [5] to recognize human actions using 'label consistent K-SVD' algorithm. Weiyao Lin *et al.* [6] modeled human trajectories as heat sources to recognize group activities from the similarity of heat-maps. Alessandro Prest *et al.* [7] combined human detection, object detection and tracking techniques to recognize actions involving human object interaction. Shuiwang *et al.* [8] considered gray-level, gradient and optical-flow information of RGB video frames as inputs to a 3D convolutional neural network for human action recognition. Yu Zhu *et al.* [9] used deep neural network model to recognize human actions from spatial-location, temporal-differences and normalized motion trajectories of human joints. A broad range of joint, plane and velocity features were evaluated by Yun *et al.* [10] to recognize interaction among humans using support vector machines. Lu Xia *et al.* [11] recognize human action by modeling the temporal evolution of pose associated with action by a hidden markov model.

Some of the major factors influencing the effectiveness of human action recognition approaches are: 1) the features considered and 2) the computational complexity of the approach. In addition, human action recognition becomes more challenging due to: a) the existence of alternative limb movements for an action b) inconsistency in speed of execution of an action and c) the lack of alignment of movements across recordings for an action. In this paper, we address these issues by considering MOCAP skeleton information of only three human joints for feature extraction and a convolutional neural network architecture for classification. The reminder of this paper is organized as follows: In section 2, the proposed approach for human action recognition, feature extraction from MOCAP information and convolutional neural network (CNN) classifier are discussed. Experimental results were discussed in section 3. The last section gives conclusions of this work.

II. Proposed Approach

In this paper, we propose a fuzzy convolutional neural network i.e., a convolutional neural network with fuzzy inputs for human action recognition based on the features extracted from motion capture (MOCAP) information. The tracking information of three human joints during the execution of an action are used to compute four distance measures. The nature and range of variation of these distance measures for each action are used to construct the fuzzy membership functions that can emphasize the discriminative pose associated with each action. The temporal variation of membership values of these fuzzy membership functions is used as the discriminative feature for human action recognition. A convolutional neural network capable of recognizing local patters in input data is used to recognizing human actions. Further details are provided in the following sections.

A. Computation of MOCAP Distance Measures

The MOCAP information for an action contains the 3D tracking information of human joints during the execution of the action. This tracking information can be used to compute distance, angle, velocity and other features for human action recognition. In this paper, the tracking information of only three human joints is used to compute four distance measures as shown in the MOCAP skeletal structure of Fig. 1.

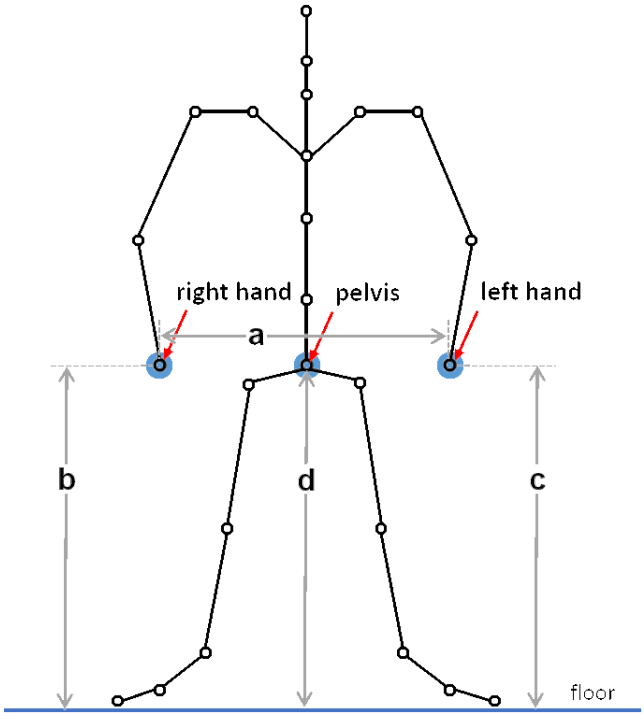


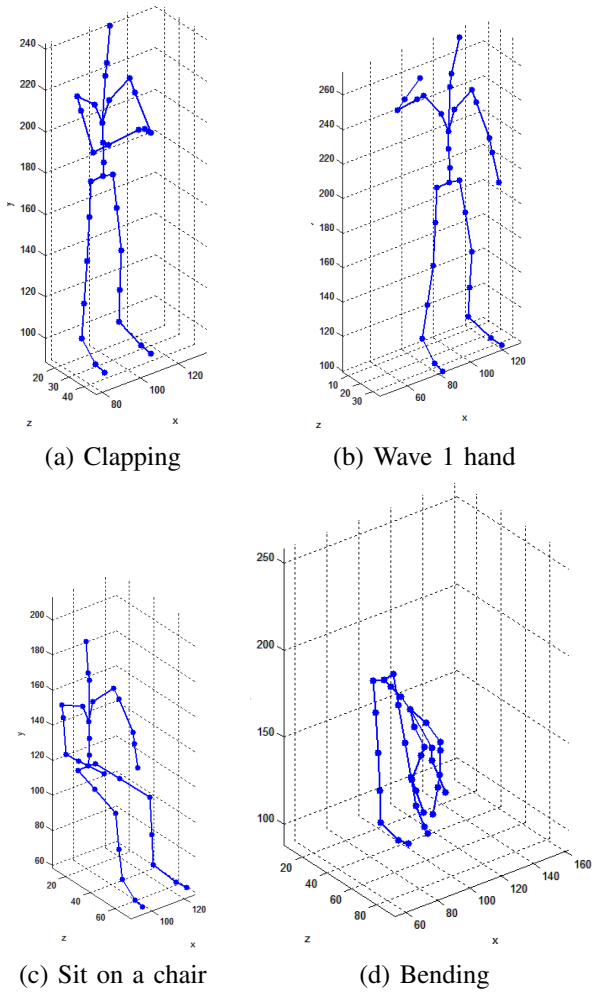
Fig. 1. MOCAP skeletal structure depicting the distance measures considered

In this paper, the tracking information of the right-hand, left-hand and the pelvis are considered to compute the following displacement measures: *a*) displacement between the left and right hands *b*) height of right-hand above the ground *c*) height of the left-hand above the ground and *d*) the height of pelvis above the ground. As some of these measures are dependent on the skeletal characteristics of the subject executing the action, these measures are normalized using the T-pose of the subject executing the action. The T-pose is a reference pose used to capture the basic skeletal characteristics of a subject. The distance measures $\{a, b, c, d\}$ are normalized across subjects to ensure consistency in their nature and range of variation for an action across subjects, as explained in Table I.

TABLE I. Normalization of measures *a*, *b*, *c* and *d*

Measure	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

As a result of normalization, *b* takes a value above 1 when the left hand is above the left shoulder and *c* takes a value above 1 when the right hand is above the right shoulder. Some of the actions that can be recognized using these measures are shown in Fig. 2.



(a) Clapping (b) Wave 1 hand

(c) Sit on a chair (d) Bending

Fig. 2. MOCAP skeletal structure representing the key pose for some human actions

Some of the observations from Fig. 2 are 1) *clapping* action can be recognized from the distance between the hands $\{a\}$ 2) *wave one hand* action can be identified from the height of hands above the ground $\{b, c\}$ 3) *sitting on a chair* action can

be detected from the height of pelvis above the ground $\{d\}$ and 4) *bending* action can be recognized from the height of hands above the ground $\{b, c\}$. It can be observed that $\{b, c\}$ can be used to recognize both *wave one hand* and *bending* actions but the range of variation of b and c is different for these actions. Thus, some of the measures can be used to recognize more than one action, depending upon their range of variation. The process of feature extraction exploiting the nature and range of variation of the distances $\{a, b, c, d\}$ for each action, is explained in the following section.

B. Feature Extraction

As explained in the previous section, more than one action can be recognized from the nature and range of variation of these distance measures. The typical nature and range of variation of these four normalized distance measures over time for some actions is shown in Fig. 3.

By grouping actions based on limb movements, {*jump*, *jumping jack*, *sit-down then stand-up*, *sit-down* and *stand-up*} actions are lower-body actions and {*bending, punching, wave two hands, wave one hand, clapping, throwing*} actions are the upper-body actions. Some of the observation from Fig. 3 are:

- d takes a value below zero for *jump*, *jumping jack*, *sit-down then stand-up*, *sit-down* and *stand-up* actions.
- d takes a value below -0.31 for *sit-down then stand-up*, *sit-down* and *stand-up* actions only
- d takes a value above 0.05 only for *jump* and *jumping jack* actions
- the nature (frequency) of variation of metric d below -0.31 is different for *jump* and *jumping jack* actions
- the above properties can be used to detect lower-body actions without any false-positives from upper-body actions due to the difference in the range of variation of d for these two types of actions.

Six membership functions $F_1(t), F_2(t), F_3(t), F_4(t), F_5(t), F_6(t)$ defined by the equations (1), (2), (3), (4), (5), (6) are designed to exploit these observations for action recognition. The plot of these six membership functions is shown in Fig. 4. Where, $x(t), x_r,\ x_{min}\ and\ x_{max}$ represents the *temporal variation, range, minimum* and *maximum* values of the distance variable x during the execution of an action and t denotes time i.e., temporal-sample.

$$F_1(t) = \begin{cases} 1, & \text{if } a(t) < 0 \\ 1 - \frac{a(t)}{2}, & \text{if } 0 \le a(t) \le 0.5 \\ 0, & \text{if } a(t) \ge 0.5 \end{cases} \tag{1}$$

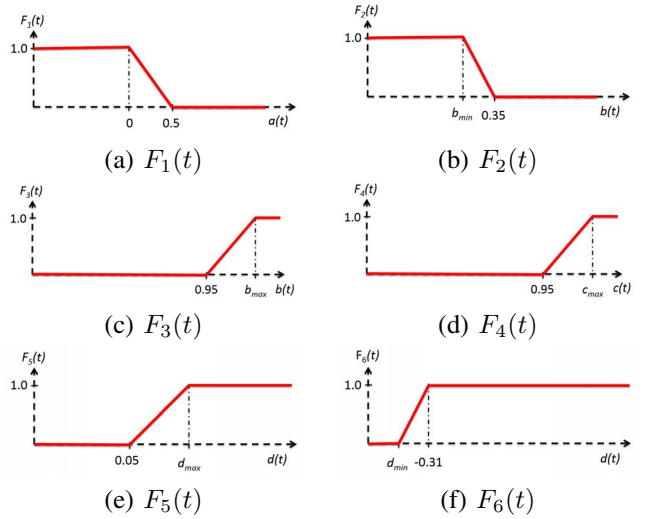
$$F_2(t) = \begin{cases} 0, & \text{if } b_r < 0.02 \\ 1, & \text{if } b_r \ge 0.02, b(t) \le b_{min} \\ \frac{(0.35 - b(t))}{(0.35 - b_{min})}, & \text{if } b_r \ge 0.02, b_{min} < b(t) < 0.35 \\ 0, & \text{if } b_r \ge 0.02, b(t) \ge 0.35 \end{cases} \tag{2}$$

$$F_3(t) = \begin{cases} 0, & \text{if } b_r < 0.02 \\ 0, & \text{if } b_r \ge 0.02, b(t) \le 0.95 \\ \frac{(b(t) - 0.95)}{(b_{max} - 0.95)}, & \text{if } b_r \ge 0.02, 0.95 < b(t) < b_{max} \\ 1, & \text{if } b_r \ge 0.02, b(t) \ge b_{max} \end{cases} \tag{3}$$

$$F_4(t) = \begin{cases} 0, & \text{if } c_r < 0.02 \\ 0, & \text{if } c_r \ge 0.02, c(t) \le 0.95 \\ \frac{(c(t) - 0.95)}{(c_{max} - 0.95)}, & \text{if } c_r \ge 0.02, 0.95 < c(t) < c_{max} \\ 1, & \text{if } c_r \ge 0.02, c(t) \ge c_{max} \end{cases} \tag{4}$$

$$F_5(t) = \begin{cases} 0, & \text{if } d_r < 0.02 \\ 0, & \text{if } d_r \ge 0.02, d(t) \le 0.05 \\ \frac{d(t) - 0.05}{d_{max} - 0.05}, & \text{if } d_r \ge 0.02, 0.05 < d(t) < d_{max} \\ 1, & \text{if } d_r \ge 0.02, d(t) \ge d_{max} \end{cases} \tag{5}$$

$$F_6(t) = \begin{cases} 0, & \text{if } d_r < 0.02 \\ 0, & \text{if } d_r \ge 0.02, d(t) \le d_{min} \\ \frac{d(t) - d_{min}}{-0.32 - d_{min}}, & \text{if } d_r \ge 0.02, d_{min} < d(t) < -0.31 \\ 1, & \text{if } d_r \ge 0.02, d(t) \ge -0.31 \end{cases} \tag{6}$$



(a) $F_1(t)$ (b) $F_2(t)$ (c) $F_3(t)$ (d) $F_4(t)$ (e) $F_5(t)$ (f) $F_6(t)$

Fig. 4. Plot of membership functions

The parameters used in the membership functions are determined empirically and the use of *minimum* and *maximum* values of distance variables for each action normalizes the movements for the action across multiple recordings of the same action. The middle 60% of the recording is used to compute the temporal variation of the distance variables i.e., *a(t), b(t), c(t), d(t)* to exclude the movements involved in beginning and ending an action. To normalize the variation in speed of execution of actions, the number of temporal samples in the 60% recording are down-sampled to 26 temporal samples. The values of the six fuzzy membership functions associated with these 26 temporal samples is computed and appended to form a 6×26 matrix, that is used as the feature representation

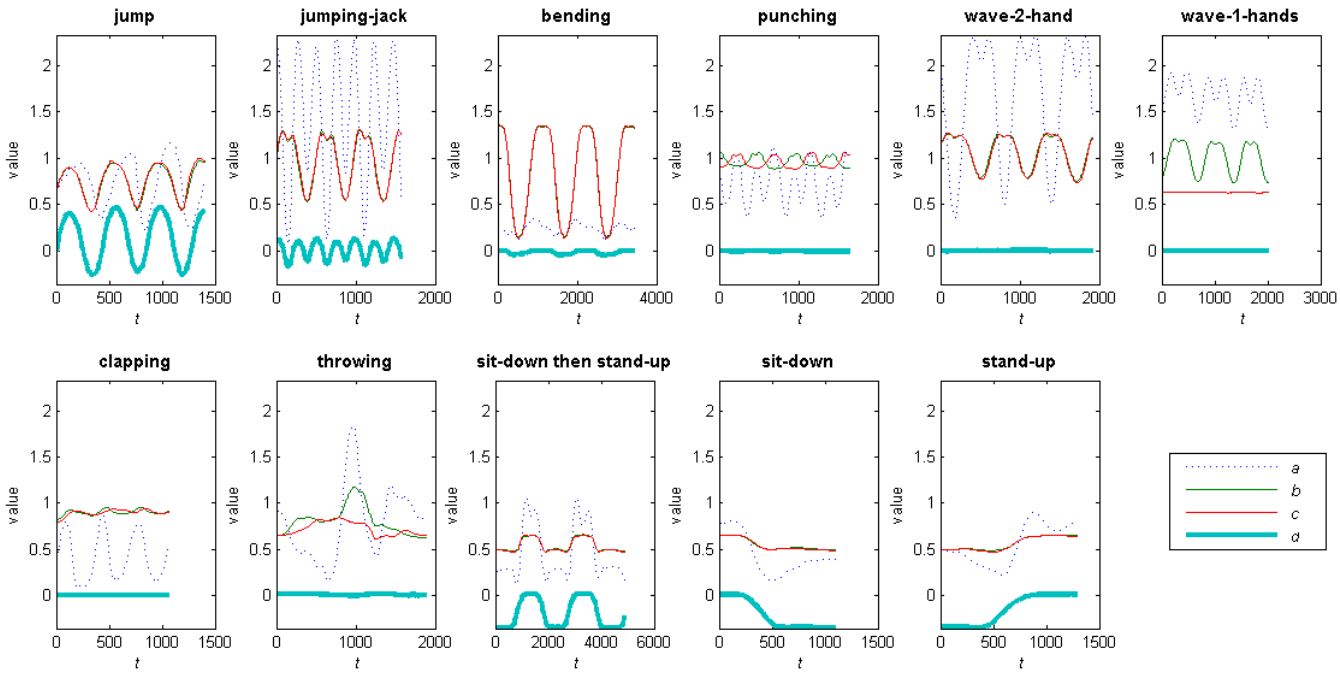


Fig. 3. Plot of temporal variation of normalized displacement measures for some actions

associated with actions. The typical feature representation associated with each action is shown in Fig.5. For better visualization, a down-sampling of 104 temporal samples is used instead of 26 temporal samples and a padding of one line is used between the temporal variations of the values of membership functions during appending.

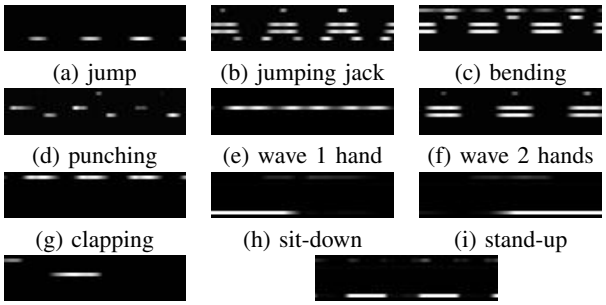


Fig. 5. Typical feature representation for various human actions.

The 6×26 representation of actions is used to recognize actions from the local patterns in this feature representation. The details of the CNN classifier considered for action recognition is explained in the following section.

C. Action Recognition using CNN

A convolutional neural network (CNN) [12] is a feed forward neural network capable of recognizing local patterns in input data with some degree of shift and distortion. This characteristic is exploited to classify human actions from the local patterns in the 6×26 representation of actions. A typical CNN classifier architecture[13] consists of an alternating sequence of convolution and subsampling layers followed by a neural network for classification. Fig. 6 presents the CNN architecture considered in the proposed approach and Table II lists the configuration considered. The representation uses, {C1, C2} to denote convolution layers, {S1, S2} to denote sub-sampling layers and {F1, F2, F3, F4} to denote the feature maps generated at the output of {1, 2, 3, 4} layers respectively. The input I of the last layer of the CNN classifier is obtained by the vector representation of the fourth feature map and O denotes output of last layer of the CNN classifier. The size and count of convolution, subsampling masks used for the evaluation of the CNN architecture in Fig. 6 is listed in Table II. The size and count of masks and feature-maps shown in Fig. 6 are not drawn per scale and Table II should be referred for details of the respective configuration.

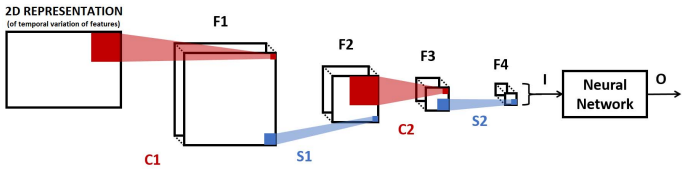


Fig. 6. Proposed CNN architecture for human action recognition

TABLE II. CNN CONFIGURATIONS USED FOR EVALUATION

Config.	C1	S1	C2	S2	I
1	(3×1)@5	1×1	(1×3)@10	1×1	520
2	(1×3)@5	1×2	(1×3)@10	1×2	300

The first CNN configuration tries to recognize vertical local patterns i.e., local patterns across the values of membership functions at a given time and the second configuration tries to recognize horizontal local patters in the temporal variation of

value of membership functions over time. The CNN classifier is trained using back-propagation algorithm in batch mode. The CNN classifier is trained to recognize the human actions from the 6×26 feature that is given as input to the CNN. The experimental setup and results are discussed in the following section.

III. Experimental results

The proposed approach for human action recognition from MOCAP information using fuzzy convolutional neural network is tested on Berkeley MHAD dataset [14] that consists of 11 actions performed by 12 subjects, repeating each action 5 times in every recording with 5 recordings per each action-subject pair. The statistics of actions in this dataset for one subject are shown in Table III.

TABLE III. Statistics of actions in Berkeley MHAD dataset for one subject

Action	# of repetitions/recording	# recordings	$\approx$ length
Jump	5	5	5 sec
Jumping jack	5	5	7 sec
Bending	5	5	12 sec
Punching	5	5	10 sec
Wave 2 hands	5	5	7 sec
Wave 1 hand	5	5	7 sec
Clapping	5	5	5 sec
Throwing	1	5	3 sec
Sit then stand	5	5	15 sec
Sit-down	1	5	2 sec
Stand-up	1	5	2 sec

The dataset captures the MOCAP information by tracking the 43 LED's worn by the subject during the execution of actions. Other multi-modal information like audio, depth and acceleration are also captured synchronously with MOCAP information for each action. Five-fold cross-validation mentioned in [14] is used to evaluate the performance of the proposed approach by training the CNN classifier using back-propagation algorithm in batch-mode with a batch-size of 4 for 500 epochs. The classification error for the 5-folds considering the two CNN configurations in Table II is shown in Table IV.

Thus, an average classification accuracy of 99.248% can be obtained using five-fold group-wise cross validation using the second CNN configuration. The variation in limb movements for actions, the speed of execution of actions and the alignment of movements across recording of an action result in noisy signal with minor shift. Despite the noise, a high classification accuracy is obtained due to the tolerance of CNN to noisy input signal with minor shift. Table V shows reported results for action recognition using MOCAP information on Berkeley MHAD dataset. Ferda Ofli *et al.* [14] considered angles at 21 joints and Muhammad Shahzad Cheema *et al.* [15] used 3D position of all the 43 joints to recognize actions from MOCAP information.

Most of the existing MOCAP action recognition algorithms use 3D information of more than 20 joints and some of these approaches use computationally expensive models for recognition. In contrast to the existing approaches, our approach uses MOCAP information of only three human joints to attain better accuracy than most of the existing approaches on Berkeley MHAD dataset. The major research contributions of this work are: 1) the design of discriminative features from a small number (three) of informative joints using fuzzy membership functions and 2) the use of CNN classifier with linear masks for action recognition.

TABLE IV. CNN classification error of five folds for 5-fold cross validation, for various CNN configurations

Config\Fold	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg.
1	2.27%	1.51%	1.51%	1.51%	3.78%	2.116%
2	0.75%	0.75%	0.75%	0.0%	1.51%	0.752%

TABLE V. Performance of different human action recognition approaches on Berkeley MHAD dataset

Approach	# of joints considered	Accuracy (in %)
MOCAP with NN [14]	21	75.55
MOCAP with K-SVM [14]	21	79.93
Multi factor classification [15]	43	87.83
Single factor action [15]	43	89.85
Our approach (5-fold cross validation)	3	**99.248**

IV. Conclusions

An approach for human action recognition with features extracted from MOCAP information using fuzzy convolutional neural network architecture is presented. Experimental results suggests that a high classification accuracy can be achieved by extracting features using fuzzy membership functions and MOCAP information of only three human joints. The ability of a convolutional neural network to recognize local patterns with some degree of shift and noise is exploited to recognize actions from the local linear patterns in features. The future work involves extensive experimentation on other MOCAP datasets and datasets with predicted human joint information like JHMDB [4], to identify the set of features suitable for action recognition across multiple datasets.

Acknowledgment

This work was supported by the Deity, Govt of India (Grant No. 13(6)/2010CC&BT)

References

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review." *IEEE Transactions on Cybernetics*, vol. 43, no. 5, June 2013, pp. 1318–1334.

[2] K. Li, Q. Dai, and W. Xu, "Markerless shape and motion capture from multiview video sequences." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, March 2011, pp. 320–334.

[3] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition." in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2012, pp. 8–13.

[4] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition." in *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, December 2013, pp. 3192–3199.

[5] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, November 2013, pp. 2651–2664.

[6] W. Lin, H. Chu, J. Wu, B. Sheng, and Z. Chen, "A heat-map-based algorithm for recognizing group activities in videos." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, November 2013, pp. 1980–1992.

[7] A. Prest, V. Ferrari, and C. Schmid, "Explicit modeling of human-object interactions in realistic videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, April 2013, pp. 835–848.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, January 2013, pp. 221–231.

[9] K. Cho and X. Chen, "Classifying and visualizing motion capture sequences using deep neural networks," *Computer Research Repository (CoRR)*, vol. abs/1306.3874, 2013.

[10] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning." in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2012, pp. 28–35.

[11] L. Xia, C.-C. Chen, and 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.

[12] Y. Bengio, "Learning deep architectures for ai," *Foundation and Trends in Machine Learning*, vol. 2, no. 1, January 2009, pp. 1–127.

[13] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Master's thesis, Technical University of Denmark, Asmussens Alle, Denmark, 2012.

[14] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, January 2013, pp. 53–60.

[15] M. S. Cheema, A. Eweiwi, and C. Bauckhage, "Human activity recognition by separating style and content," *Pattern Recognition Letters*, In Press.