2014 13th International Conference on Machine Learning and Applications

Human action recognition based on MOCAP information using 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*—Human action recognition is an important component in semantic analysis of human behavior. In this paper, we propose an approach for human action recognition based on motion capture (MOCAP) information using convolutional neural networks (CNN). Distance based metrics computed from MOCAP information of only three human joints are used in the computation of features. The range and temporal variation of these distance metrics are considered in the design of features which are discriminative for action recognition. A convolutional neural network capable of recognizing local patterns is used to identify human actions from the temporal variation of these features, which are distorted due to the inconsistency in the execution of actions across observations and subjects. Experiments conducted on Berkeley MHAD dataset demonstrate the effectiveness of the proposed approach.**

***Keywords*-convolutional neural networks (CNN); motion capture (MOCAP);**

I. Introduction

Human action recognition is a widely studied problem in computer vision. Action recognition algorithms use motion capture (MOCAP) information to learn and recognize the sequence of motions for each action. MOCAP refers to a broad range of techniques used to capture the motion of objects and people. The MOCAP techniques to capture human motion vary from tracking wearable markers to pose prediction from depth/multi-view videos [1] [2]. Human motion capture using MOCAP techniques is widely used to collect ground truth information for the validation of computer vision algorithms and for applications in sports, medicine and military. Over the years, the availability of low cost, high mobility depth sensors like Kinect led to the development and use of marker less motion tracking techniques in human computer interaction (HCI) systems and in entertainment. Experimental studies were conducted to identify the significance of pose information in recognizing human actions [3] and the most informative human joints for recognizing an action [4]. These studies suggests that high level pose features can greatly outperform mid and low level features for recognizing human actions and that human actions can be recognized with high accuracy from MOCAP information of a few human joints. In this paper, we consider the MOCAP information of only three human joints to recognize eleven human actions.

A wide range of techniques were proposed to recognize individual, group actions and interaction among humans or with objects by utilizing different modality. Zhuolin Jiang *et al.* [5] considered action bank features generated from RGB videos to learn discriminative dictionaries for human action recognition using 'label consistent K-SVD' algorithm. Human trajectories are modeled as heat sources by Weiyao Lin *et al.* [6] to recognize group activities from the similarity of heat maps. Techniques for human detection, object detection and tracking are combined by Alessandro Prest *et al.* [7] to model and localize human object interaction, for recognizing actions involving human object interaction. The gray level, gradient and optical flow of RGB video frames are used as features to train a 3D convolutional neural network for human action recognition [8]. The pairwise skeletal distance and spatio-temporal motion information are combined using random forest learning for 3D action recognition [9]. The spatial location, temporal differences and normalized motion trajectories of human joints are considered to recognize human actions from skeletal information using deep neural network model [10]. A broad range of joint, plane and velocity features were evaluated by Yun *et al.* [11] to recognize interaction among humans using support vector machines. Lu Xia *et al.* [12] represented human pose using histogram of 3D joint locations and recognized human actions by modeling the the temporal evolution of pose using HMM.

Some of the issues with existing algorithms are: 1) dependence on features with uncertain discriminative information 2) use of features with redundant information and 3) high computation complexity. In addition, human action recognition becomes more difficult due to the following reasons: a) 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 the proposed approach, we address these issues by considering MOCAP skeleton information of only three

978-1-4799-7415-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMLA.2014.30

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 an approach for human action recognition based on the features derived from MOCAP information using convolutional neural networks. The MOCAP information of three human joints is used to compute four distance metrics. The nature and range of variation of these distance based metrics for each action is used to construct the necessary discriminative features. The temporal variation of these features is given as input to the convolutional neural network for action recognition. Further details are provided in the following sections.

A. MOCAP distance metrics

The human MOCAP information contains the location of each human joint in 3D space over time. This joint tracking information can be used to compute distance, angle, velocity and other metrics/features for human action recognition. The distance based metrics considered in this paper for the computation of features are shown in the MOCAP skeletal structure in Figure 1(a).

In our approach, we consider the tracking information of the right-hand, left-hand and the pelvis to compute the following distance based metrics: **a**) distance 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. The nature and range of variation of these distance metrics are used for action recognition. Some of the actions that can be recognized using these metrics are shown in Figure 1(b) to (e).

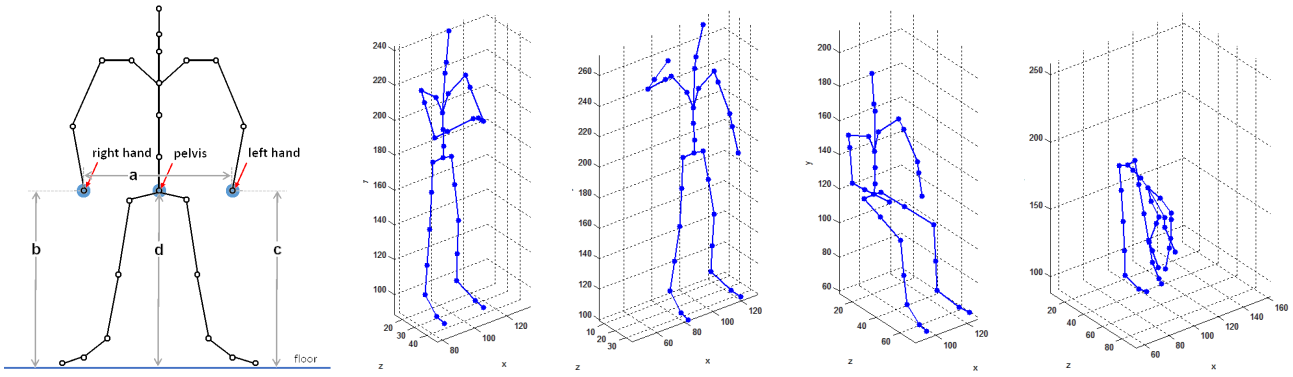
Some of the observations from Figure 1 are 1) *clapping* action can be recognized from distance between the hands (metric **a**) 2) *wave one hand* action can be identified from the height of hands above the ground (metrics **b**, **c**) 3) *sitting on a chair* action can be detected from the height of pelvis above the ground (metric **d**) and 4) *bending* action can be recognized from the height of hands above the ground (metrics **b**, **c**). It can be observed that metrics **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 metrics can be used to recognize more than one action, depending on their range of variation. The process of feature extraction from distance metrics, exploiting their nature and range of variation for each action, is explained in the following section.

B. Feature extraction

The four distance metrics computed from MOCAP information are normalized for each subject using their T-pose information to ensure consistency in the nature and range of variation of these metrics across subjects with varying physique. This is accomplished by normalizing metric **a** by dividing it with the distance between the hips. Metric **d** by subtracting and dividing it with the value of **d** in T-pose. Metrics **b**, **c** are normalized by dividing them with the height of left, right shoulders respectively to ensure consistency in their range of variation for subjects with long/short hands. Thus, metric **b** takes a value above 1 when the left hand is above the left shoulder. Similarly, metric **c** takes a value above 1 when the right hand is above the right shoulder. As some of these metrics can be used to recognize more than one action, the nature and range of variation of these metrics for each action needs to be considered for feature extraction. The typical nature and range of variation of these four normalized distance metrics over time for some actions is shown in Figure 2.

From Figure 2, it can be observed that the height of pelvis above the ground (metric **d**) takes a value below zero for *jump*, *jumping jack*, *sit-down then stand-up*, *sit-down* and *stand-up* actions. These actions require significant movement in the lower body and can be considered as lower body actions. The remaining actions can be considered as the upper-body actions, which do not involve significant movement in the lower body. As metric **d** takes a value below -0.31 for *sit-down then stand-up*, *sit-down* and *stand-up* actions only the variation of **d** below -0.31 is considered as a feature. Similarly, metric **d** takes a value above 0.05 only for *jump* and *jumping jack* actions due to which the variation of metric **d** above 0.05 is considered as another feature. The nature (frequency) of variation of metric **d** below -0.31 is different for *jump* and *jumping jack* actions, thereby acting as a discriminative feature for these two actions. The two features considered so far (using metric **d**) can recognize five lower body actions without any false positives from other actions as the operating range of metric **d** for upper body actions do not overlap with these features.

As metrics **b** and **c** take a value below 0.35 for *bending* action, variation of **b** below 0.35 is considered as a feature to recognize *bending* action. The variation of metrics **b** and **c** above 0.95 are also considered as features as they can recognize *punching*, *wave one hand* and *wave two hands* actions. For *punching* and *wave two hands* actions metrics **b** and **c** follow a sine wave but are out of sync for *punching* and in-sync for *wave two hands* action. Only one of the metrics **b**, **c** follow a sine wave for *wave one hand* action. The nature of variation of **b** with respect to **c** provides the necessary discriminative information for recognizing these three actions. The features considered so far avoid any false positives in recognizing *punching*, *wave one hand* and *wave*



(a) MOCAP distance metrics (b) Clapping (c) Wave 1 hand (d) Sit on a chair (e) Bending

Figure 1. MOCAP skeletal structure depicting the distance metrics considered and the key pose for some human actions

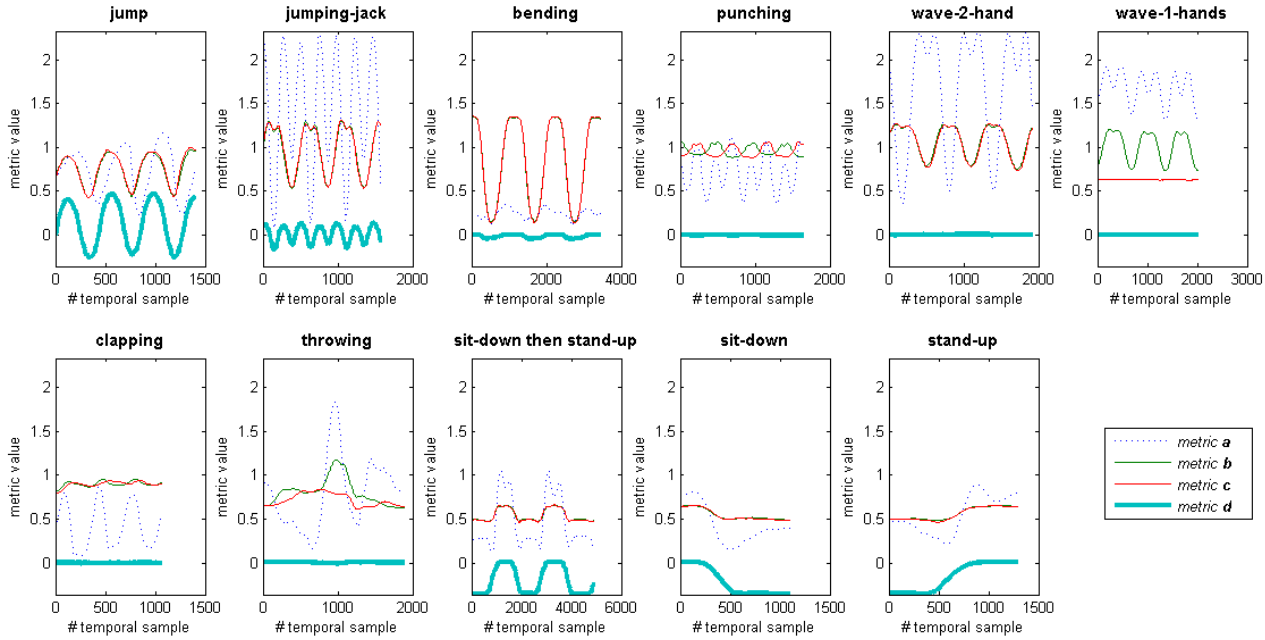


Figure 2. Plot of variation of normalized distance metrics over time for some actions

two hands actions. The variation of metrics **b** and **c** above 0.95 can also be used to recognize *throwing a ball* action from their nature of variation as shown in Figure 2. The variation of metric **a** below 0.5 is considered as a feature to recognize *clapping* action and all the features defined so far avoid any false positives during recognition. The summary of computation of features from distance metrics is given in Table I. The effective range of variation of a metric for each feature is determined emperically. The features #2 to #6 are locally scaled to vary between 0 and 1 when their range exceeds 0.02 and feature #2 is complemented during scaling. The features are scaled to normalize the variation in limb movements for actions. To normalize the variation in speed of execution of actions, the number of temporal samples of all actions are down sampled to 104 samples.

Feat. #	Distance metric used	range considered
1	dist. between hands (**a**)	less than 0.5
2	height of right hand (**b**)	less than 0.35
3	height of right hand (**b**)	greater than 0.95
4	height of left hand (**c**)	greater than 0.95
5	height of pelvis (**d**)	greater than 0.05
6	height of pelvis (**d**)	less than -0.31

Table I
COMPUTATION OF FEATURES FROM DISTANCE METRICS

The temporal variation of significant features for each

action is shown in Table II. Actions are analyzed by grouping them into two categories namely upper body and lower body actions considering the entropy of various human joints during these actions. The significant features for recognizing an action are shown in the third column of Table II, with optional features in square brackets and their temporal variation as a 2D representation in column four. The 2D representation is obtained by duplicating the features considering a two element margin from borders and between the features. The features are duplicated to emphasize the local patters for efficient classification.

Action	Type	Significant Feature	Variation
clapping	upper-body	1 [2 3 4]	
bending	upper-body	[1] 2 [3 4]	
punching	upper-body	[1 2] 3 4	
wave 1 hand	upper-body	[1 2] 3 [4]	
wave 2 hands	upper-body	[1 2] 3 4	
throwing a ball	upper-body	[1 2] 3 [4]	
jump	lower-body	5 [6]	
jumping jack	lower-body	5 [6]	
sit and stand	lower-body	[5] 6	
sit-down	lower-body	[5] 6	
stand-up	lower-body	[5] 6	

Table II
VARIATION OF SIGNIFICANT FEATURES FOR EACH ACTION

Thus the six features constructed from the four metrics provide the necessary discriminative information to classify these actions. The 2D representation of the typical temporal variation of the six features for some actions is shown in Figure 3. This 2D representation of the temporal variation of the six features provides the necessary discriminative information (as local patterns) to recognize each action. The use of a convolutional neural network to recognize human actions from the temporal variation of these features is elaborated in the following section.

C. Action recognition using CNN

A convolutional neural network (CNN) [13] is capable of recognizing local patterns with some degree of shift and distortion. This characteristic is exploited to classify human actions from the local patterns in the 2D representation of actions. A typical CNN classifier architecture used as a classifier [14] consists of an alternating sequence of convolution and subsampling layers followed by a neural network for classification. The architecture considered in the proposed approach is shown in Figure 4 whose configuration is listed in Table III. Here, {C1, C2} represents the convolution layers, {S1, S2} represent the subsampling layers, {F1, F2, F3, F4} represent the feature maps generated at the output of 1, 2, 3, 4 layers of the CNN classifier respectively and I, O denote the input, output of the Neural network. Sigmoid transfer function is used as the activation function in all the neurons and backpropagation algorithm in batch mode is used for training.

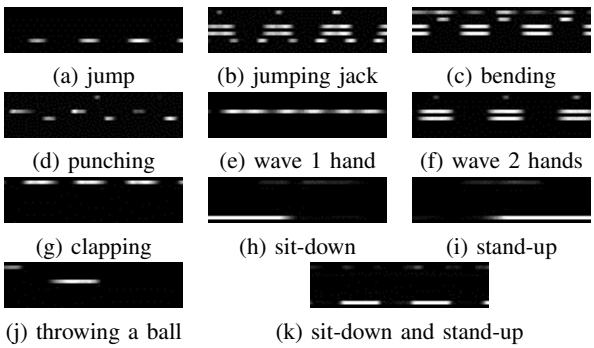


Figure 3. 2D representation of temporal variation of features for various human actions.

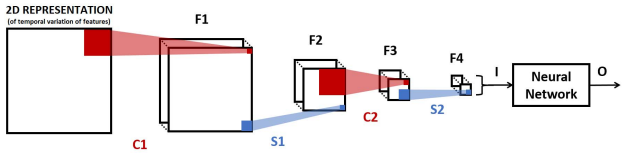


Figure 4. Proposed CNN architecture for human action recognition

Layer: Configuration		Feature map: #, config	
C1:	3×3 templates	F1:	4, 24×24 feature maps
S1:	2×2 templates	F2:	4, 12×12 feature maps
C2:	3×3 templates	F3:	8, 10×10 feature maps
S2:	2×2 templates	F4:	8, 5×5 feature maps
I:	200 vector	O:	11 outputs

Table III
CONFIGURATION OF THE PROPOSED CNN ARCHITECTURE

The 2D representation of actions elaborated in previous section is scaled to generate a 26×26 matrix, which is presented as input to the classifier. The CNN is trained to classify actions from their 2D representation. 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 convolutional neural network is tested on Berkeley MHAD dataset [15]. The dataset consists of 11 actions performed by 12 subjects repeating every action 5 times in each recording with 5 recordings per action-subject pair. The details of the human actions in this dataset for one subject are shown in Table IV.

Action	# of repetitions/recording	# recordings	≈ 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

Table IV
THE DESCRIPTION, # OF REPETITIONS PER RECORDING, # OF RECORDINGS AND DURATION OF ACTIONS IN BERKELEY MHAD DATASET FOR ONE SUBJECT

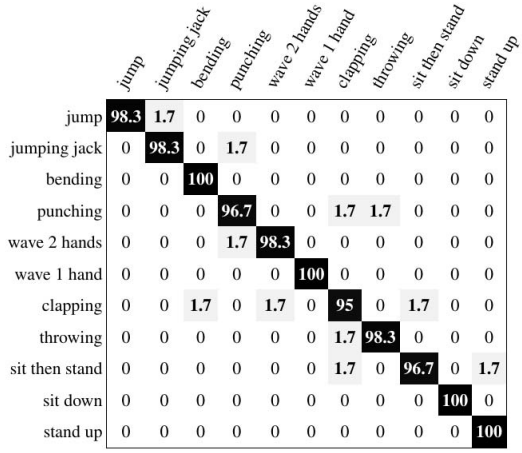


Figure 5. Confusion matrix of proposed approach using MOCAP information on Berkeley MHAD dataset

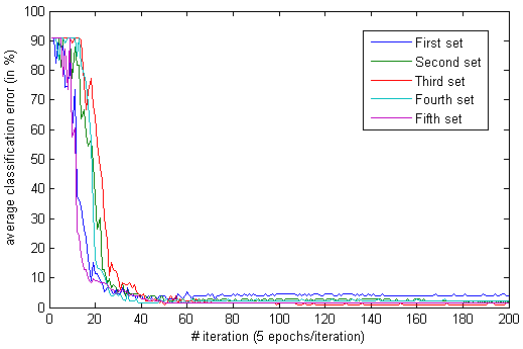


Figure 6. Plot of average classification error against training iterations

Each action is captured simultaneously by an optical motion capture system, four multi-view stereo vision camera arrays, two Microsoft Kinect cameras, six wireless accelerometers and four microphones. The motion capture data is acquired by tracking the 3D position of 43 LED markers placed at different body parts and at the 25 joints shown in Figure 1(a). The T-pose configuration of a subject contains the basic skeleton information of a subject, which is used in the proposed approach to normalize the MOCAP distance metrics. As each action is performed five times in each recording (except for three actions), the middle 60% of the recording is considered as the representative signal of each action upon which metric computation and features extraction are carried out. From the last column of Table IV, it can be observed that the length of the recording changes from action to action and across recordings. The representative signal length is scaled to generate a signal of 104 temporal samples which will be down-sampled further to 26 temporal samples before presenting it as the input to the CNN. The scaling of representative signal to 104 temporal samples results in the generation of feature vectors of same size for all actions, thereby normalizing the variation in speed of execution of an action across observations and subjects.

Five-fold cross-validation is used to evaluate the performance of the proposed approach by training the CNN classifier using back propagation algorithm with a batch size of 44 for 1000 epochs. The confusion matrix of the proposed approach on Berkeley MHAD dataset is shown in Figure 5.

An average classification accuracy of 98.38% is obtained using five-fold group-wise cross validation. 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 low classification error is obtained due to the tolerance of CNN to noisy input signal with minor shift. The plot of average classification error against training iteration for five-fold cross validation is shown in Figure 6.

Experiments were conducted considering various CNN configurations in terms of number of layers, size of templates and number of feature maps. The configuration shown in Figure 4 and listed in Table III is identified to be the best CNN configuration for this task. Table V shows reported results for action recognition using MOCAP information on Berkeley MHAD dataset. Ferda Ofli *et al.* [15] considered angles at 21 joints and Muhammad Shahzad Cheema *et al.* [16] 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. The current state of the art approach for action recognition on Berkeley MHAD dataset [17] considers all body part configurations (from MOCAP information of 28 joints) and temporal scales to attain 100% accuracy. 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 and

Approach	# of joints considered	Accuracy (in %)
MOCAP with NN [15]	21	75.55
MOCAP with K-SVM [15]	21	79.93
Multi factor classification [16]	43	87.83
Single factor action [16]	43	89.85
Disc. Hierarchy of LDSs [17]	28	100.00
Our approach (5-fold cross validation)	3	**98.38**

Table V
PERFORMANCE OF DIFFERENT HUMAN ACTION RECOGNITION APPROACHES ON BERKELEY MHAD DATASET

is comparable with the state of the art approach for action recognition from MOCAP information on Berkeley MHAD dataset. The major research contributions of this work are: 1) the design of discriminative features using a small number (three) of informative joints and 2) the representation of the features for classification by a CNN classifier.

IV. CONCLUSIONS

An approach for human action recognition with features extracted from MOCAP information using convolutional neural network architecture is presented. Experimental results suggests that a high classification accuracy can be achieved by considering features derived from only three MOCAP joints. The ability of a convolutional neural network to recognize local patterns with some degree of shift and noise is used to recognize actions from the nature and range of variation of features, in the presence of distortion due to variation in limb movements, speed of execution of actions and the alignment of movements across recordings. The future work involves extensive experimentation on other MOCAP datasets and datasets with predicted human joint information like JHMDB [3], to identify the set of features applicable for action recognition across multiple datasets.

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, pp. 1318–1334, June 2013.

[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, pp. 320–334, Mar. 2011.

[3] 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, Dec. 2013, pp. 3192–3199.

[4] 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.

[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, pp. 2651–2664, Nov. 2013.

[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, pp. 1980–1992, Nov. 2013.

[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, pp. 835–848, April 2013.

[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, pp. 221–231, Jan. 2013.

[9] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2013, pp. 486–491.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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)*, Jan 2013, pp. 53–60.

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

[17] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal, "Bio-inspired dynamic 3d discriminative skeletal features for human action recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2013, pp. 471–478.