2014 13th International Conference on Machine Learning and Applications

One-shot periodic activity recognition using Convolutional 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*—Activities capture vital facts for the semantic analysis of human behavior. In this paper, we propose a method for recognizing human activities based on periodic actions from a single instance using convolutional neural networks (CNN). The height of the foot above the ground is considered as features to discriminate human locomotion activities. The periodic nature of actions in these activities is exploited to generate the training cases from a single instance using a sliding window. Also, the capability of a convolutional neural network to learn local visual patterns is exploited for human activity recognition. Experiments on Carnegie Mellon University (CMU) Mocap dataset demonstrate the effectiveness of the proposed approach.**

***Keywords*-convolutional neural networks (CNN); human activity recognition;**

I. Introduction

Activity recognition is a major area of research in Computer Vision as they embody rich semantic information. The two major sources of data for activity recognition are video and motion capture data (Mocap). Mocap refers to a wide range of methods used for recording motions of objects and people. They are widely used to collect ground truth information for validating computer vision algorithms and for applications in sports, medicine, military and entertainment. Wearable markers attached near joints were tracked to estimate the motion of human joints. Experimental studies on the significance of pose information for human activity recognition [1] suggests that high level pose feature greatly outperform low/mid-level features. The most informative joints for human action recognition were analyzed [2] to design the best discriminative features. The availability of low cost, high mobility sensors such as Kinect and multi-view video sequence processing opened new possibilities in marker-less motion capture [3][4]. We have considered Mocap skeleton information for evaluating our approach due to the inadequacy of accurate pose estimation algorithms.

Group activities can be identified from the similarity of their heap maps [5] by modelling human trajectories as a series of heat sources. A discriminative dictionary learning algorithm 'label consistent K-SVD' [6], that associates labels for each dictionary item is used for action recognition using the action bank features. The grey value, gradient, optical flow along x and y axis are used as features to train 3D convolutional neural networks for action recognition[7]. The pairwise skeleton distance and spatio-temporal motion information is combined using random forests learning method for 3D action recognition[8]. Deep Neural Network model [9] using position of joints, temporal differences and normalized trajectories of motion are used for action recognition on skeleton data. A diverse range of joint, plane and velocity features are evaluated using support vector machines for two person interaction detection [10].

Some of the issues with existing algorithms are: 1) dependence on features with uncertain discriminative information 2) need for larger training dataset for better accuracy and 3) high computation complexity. The proposed method tries to address these issues by considering Mocap skeleton information for feature extraction on a single training instance and a convolutional neural network architecture for classification. The reminder of this paper is organized as follows: In Section 2, the proposed method for human activity recognition, feature extraction and convolutional neural network (CNN) classifier are discussed. Experimental results were discussed in Section 3. Section 4 gives conclusion of this work.

II. Proposed Approach

A method for human activity recognition for activities based on periodic actions using a convolutional neural network architecture for classification is proposed. The local patterns in the temporal variation of features for each activity is captured and utilized by the CNN to classify human activities. The features considered in this method are elaborated in the following section.

A. The features and training dataset

The temporal variation of height of feet above the ground are the features considered to classify the locomotion activities: jump, run and walk. The variation of foot height against the sample count for jump, run and walk activities is shown in Fig. 1. These features are normalized in each trial, to vary in the range of zero to one. The CMU Mocap dataset subject-trials used for feature analysis and peak to

978-1-4799-7415-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICMLA.2014.69

peak (P2P) distance computation for locomotion activities is shown in Table I.

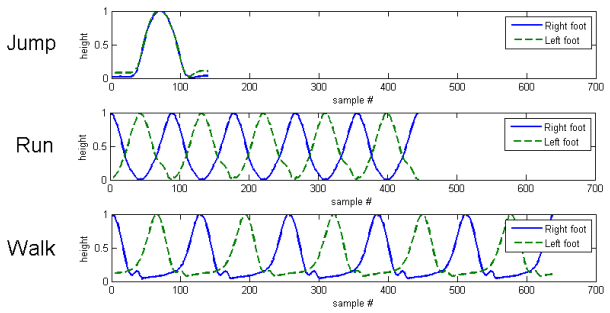


Figure 1. Temporal variation of height of feet above ground for jump, run and walk activities.

From Fig. 1, it can be observed that both feet are at the same height above the floor level for jump activity and only one foot is above the floor-level for run, walk activities. The foot-up and foot-down actions in run and walk are periodic, with same periodicity for both foot. The peak to peak (P2P) distance of a feature is indicative of the duration the corresponding foot is above the ground and is smaller for run than for walk. Thus, for a given time period, the number for foot-up and foot-down actions in run activity will be more than the once in walk activity.

Activity	sample (subject:trial)	P2P distance
jump	13:40	36
run	09:01	92
walk	39:03	128

Table I
SAMPLES CONSIDERED TO GENERATE TRAINING DATASET

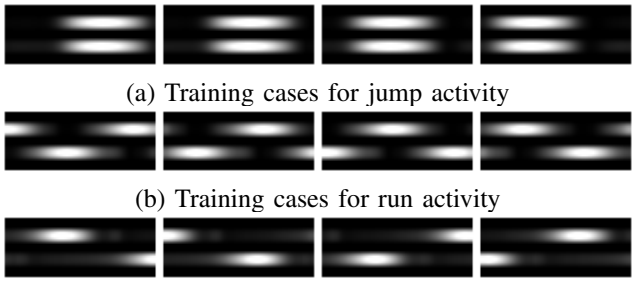
In case of jump, due to the absence of periodic actions a P2P distance of 34 samples is considered by default and a window of 138 samples is considered to represent the signal. The window is positioned on the signal to align the peak with the center of the window as shown in Fig. 1. A sliding window with a length of 104 samples is run between the two peaks of the feature associated with right foot shown in Fig. 1 with a shift size of four samples, to generate the training cases. Instead of repeating this process between the two peaks of the feature associated with left foot, as the actions of left and right foot have same periodicity for run and walk activities, the vertical reflections of the data generated for right foot are added to the training dataset as shown in Table II.

The temporal variation of these features is down sampled by 4 to construct a 2D representation of size 1026. A margin of two elements is left on the top, bottom and between the features. The features are duplicated along the vertical axis for better performance. Some of the training cases generated

P2P distance	# of window slides	# of training cases
36	9	18
92	23	46
128	32	64

Table II
OF TRAINING CASES PER ACTIVITY

using the sliding window for jump, run and walk activities are shown in Fig 2(a), 2(b) and 2(c) respectively. From



(a) Training cases for jump activity

(b) Training cases for run activity

(c) Training cases for walk activity

Figure 2. Training cases for jump, run and walk activities

the 2D representations in Fig. 2 it can be observed that there is a distinctive visual pattern for these three activities. A rich training dataset capturing all possible feature patters is generated by using a fixed-length sliding window on the training instances. The classification scheme to capture and classify the activities from their local patterns, is elaborated in the following section.

B. Activity classification using CNN

A convolutional neural network (CNN) [11] is capable of recognizing local patterns with some degree of shift and distortion. This characteristic is exploited to classify activities from the local patterns in their features. A typical CNN architecture used as a classifier [12] consists of an alternating sequence of convolution and subsampling layers followed by a neural network for classification. The architecture used in the proposed method is shown in Fig. 3 and the 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.

The 2D representation of activities mentioned in section 2.1 is presented as input to the classifier. The CNN is trained using back-propagation algorithm for 5000 epochs on the training dataset to classify activities from their 2D representation. The test dataset consisting of subject-trials excluding the ones used for training is used for performance

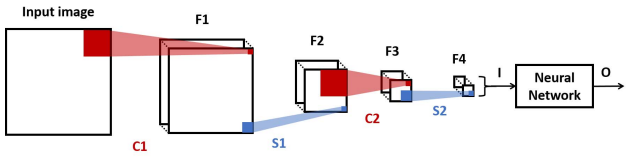


Figure 3. CNN architecture considered in the proposed approach

C1:	3×3 templates	F1:	10, 8×24 feature maps
S1:	2×2 templates	F2:	10, 4×12 feature maps
C2:	3×3 templates	F3:	20, 2×10 feature maps
S2:	2×2 templates	F4:	20, 1×5 feature maps
I:	100 vector	O:	3 outputs

Table III
CNN CONFIGURATION

evaluation of the proposed approach. The experimental setup and results are discussed in the following section.

III. EXPERIMENTAL RESULTS

The proposed approach for activity recognition is evaluated on locomotion activities: jump, run and walk in CMU Mocap dataset [13]. The variations in these activities considered for evaluation are shown in Table IV.

Activity	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 IV
VARIATIONS IN ACTIVITIES CONSIDERED FOR EVALUATION

One trial from each of these classes is used to generate the training dataset as explained in section 2.1 and the remaining trials are used to generate the test dataset. For a given test trial, peak analysis is conducted on its features to identify the number of peaks in the signal. For a single peak signal, the sampling window is positioned to align the peak with the center of the window. For a signal with multiple peaks, consider the 104 samples after the first peak as the representative signal of the trial. The selected feature windows are scaled to produce 1026 feature maps, which are used for testing. The average classification error of the proposed approach is 5.56% and the confusion matrix for human activity recognition is shown in Fig. 4.

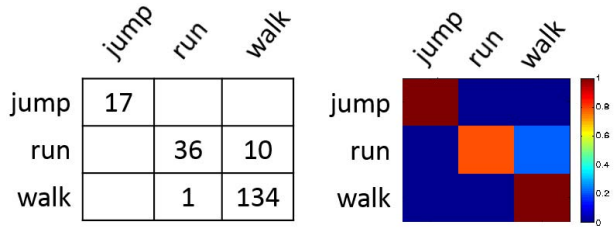


Figure 4. Confusion matrix for human activity recognition

The low misclassification error is due to the wide variations in the execution of activities. As a single window of 104 samples is used to represent a trail, the absence or interference of foot-up, foot-down actions with other actions during the window affects the accuracy of the proposed approach. Fig. 5 shows the temporal variation of height of feet above the ground for subject #16, trial #42 performing the activity run/jog, 90-degree left turn, which is misclassified. The samples between 22 and 125 are used as the representative signal for this trial. The overlap of run action with '90 degrees left turn' action could be the reason for its misclassification.

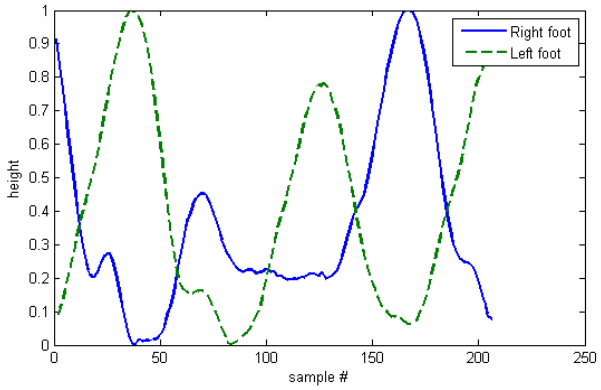


Figure 5. Temporal variation of height of feet above ground for the activity 'run/jog, 90-degree left turn'.

Experiments were conducted considering various CNN configurations in terms of number of layers, size of tem-

plates and number of feature maps. Experimental evaluation suggests that the configurations shown in Fig. 3 and listed in Table III is the optimal CNN configuration for this task. The variation in average classification error against training epochs is shown in Fig. 6.

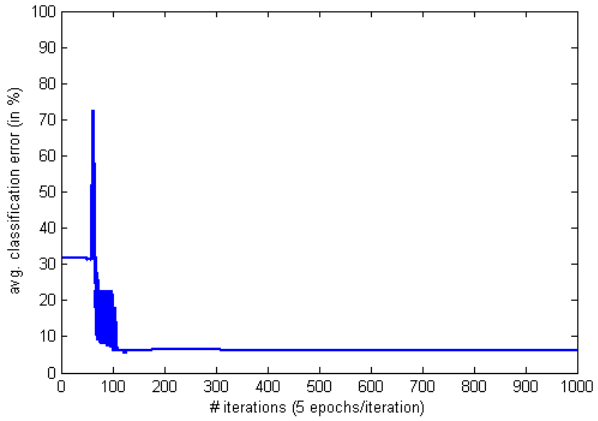


Figure 6. Average classification error against training epochs

It can be noticed that the classification error drops to 6.07% in 520 epoch. Experiments were conducted on a typical i5 machine without GPU's and the CNN classifier was trained for 5000 epochs in less than an hour.

IV. Conclusion

A classifier for human activities based on periodic actions, using height of feet above the ground as the feature and a CNN architecture for classification is presented. Experimental results on diverse variations of these activities from CMU Mocap dataset suggests low classification error and fast convergence. Interference of other actions in a trial with the foot-up and foot-down actions affects classification error. The future work includes the use of multiple windows per trial and late-fusion techniques for classification; additional features to sub-classify activities into their variants.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[12] 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.

[13] "CMU Human Motion Capture Database," http://mocap.cs.cmu.edu/.