Open access • Proceedings Article • DOI:10.1109/DICTA.2018.8615804

Similar Gesture Recognition using Hierarchical Classification Approach in RGB Videos — Source link

Di Wu, Nabin Sharma, Michael Blumenstein

Institutions: University of Technology, Sydney

Published on: 01 Dec 2018 - Digital Image Computing: Techniques and Applications

Topics: Gesture recognition, Deep learning, Gesture, Convolutional neural network and Binary classification

Similar Gesture Recognition Using Hierarchical Classification Approach in RGB Videos

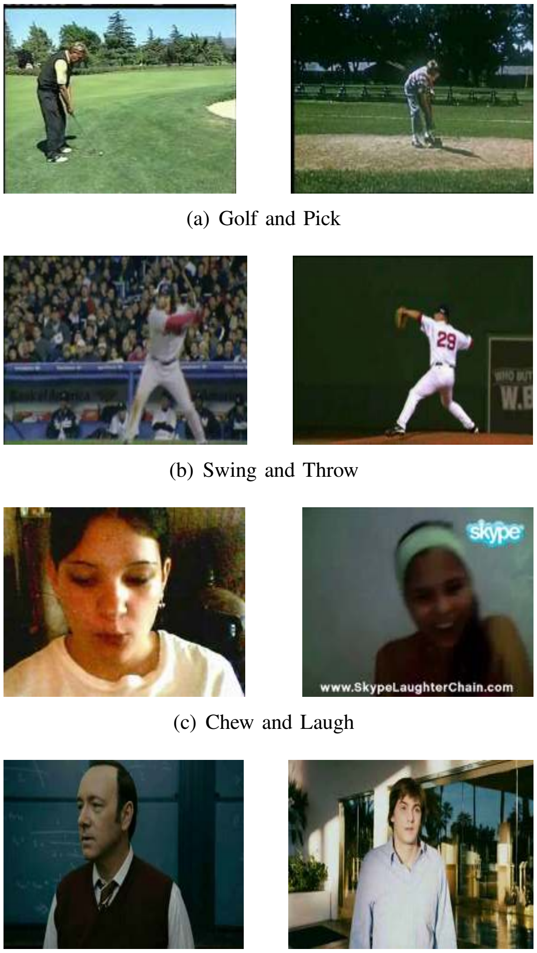
Di Wu, Nabin Sharma and Michael Blumenstein
School of Software, Centre for Artificial Intelligence
University of Technology, Sydney
Ultimo, New South Wales, Australia 2007
Email: Di.Wu-16@student.uts.edu.au, (Nabin.Sharma,Michael.Blumenstein)@uts.edu.au

***Abstract*—Recognizing human actions from the video streams has become one of the very popular research areas in computer vision and deep learning in the recent years. Action recognition is wildly used in different scenarios in real life, such as surveillance, robotics, healthcare, video indexing and human-computer interaction. The challenges and complexity involved in developing a video-based human action recognition system are manifold. In particular, recognizing actions with similar gestures and describing complex actions is a very challenging problem. To address these issues, we study the problem of classifying human actions using Convolutional Neural Networks (CNN) and develop a hierarchical 3DCNN architecture for similar gesture recognition. The proposed model firstly combines similar gesture pairs into one class, and classify them along with all other class, as a stage-1 classification. In stage-2, similar gesture pairs are classified individually, which reduces the problem to binary classification. We apply and evaluate the developed models to recognize the similar human actions on the HMDB51 dataset. The result shows that the proposed model can achieve high performance in comparison to the state-of-the-art methods.**

***Index Terms*—Action Recognition, Neural Networks, Deep Learning, Computer Vision**

I. Introduction

Human action recognition is one of the most popular research area in computer vision. Diverse applications are designed based on the human action recognition technology such as, surveillance, video indexing, human-computer interaction, customer behaviour monitoring and analysis, etc across multiple domains. However, recognizing human actions accurately from video stream is a challenging task due to occlusion, low resolution, cluttered backgrounds and viewpoint variations, etc. [1] [2] [3]. Unlike action recognition from still images, videos include temporal information and genetic data augmentation which is essential to the classify actions/gestures more accurately. In early stages, researchers made assumptions on certain scale or fixed viewpoint when the video was captured. However, those assumptions doesn't reflect the real-world environment. Besides, early research also followed the two-steps approach to design the system. First, the hand-craft features are extracted from the video frames, followed by the design of classifiers based on the extracted features. Thus, most of the early research works calculate the motion and texture descriptors using spatio-temporal interest points which are built manually. In the real-world scenario, the performance of these hand-crafted features is low as they are highly problem-depended and lacks generalisization. Especially, for human action recognition, different actions may correspond to totally different patterns due to the environment changes and motion patterns.



(a) Golf and Pick

(b) Swing and Throw

(c) Chew and Laugh

(d) Turn and Walk

Fig. 1. Different classes of human activities with similar gestures [4]

Deep learning models [5] [6] [7] have become a priority choice to deal with the computer vision problems due their impressive performance in various computer vision related tasks. These models have the advantage of learning features from

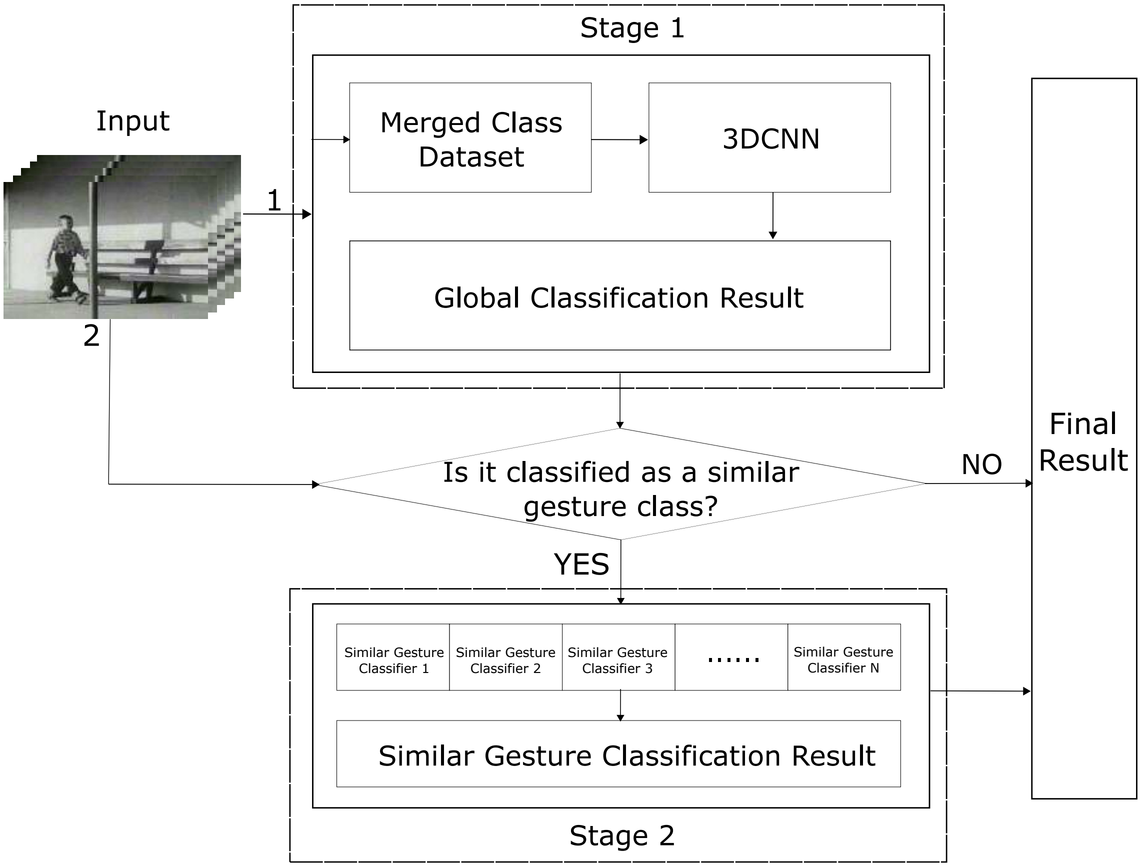

Stage 1
Input
Merged Class Dataset
3DCNN
1
Global Classification Result
2
Final Result
Is it classified as a similar gesture class?
NO
YES
Similar Gesture Classifier 1
Similar Gesture Classifier 2
Similar Gesture Classifier 3
......
Similar Gesture Classifier N
Similar Gesture Classification Result
Stage 2

Fig. 2. Proposed the Hierarchical 3DCNN Architecture

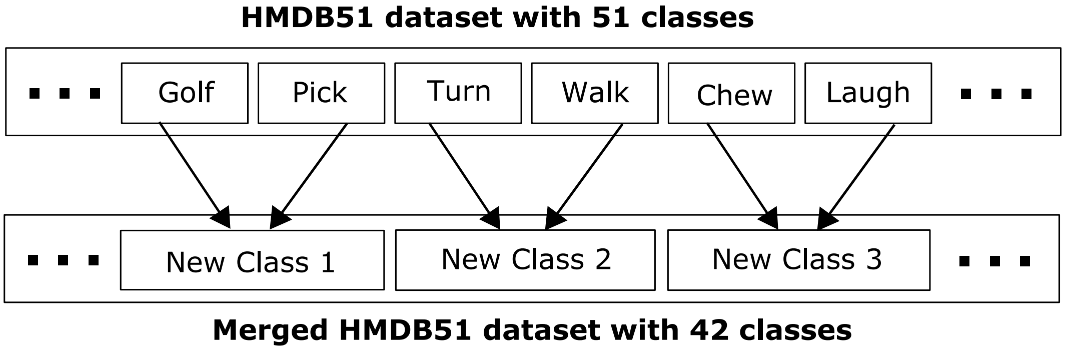

HMDB51 dataset with 51 classes
Golf
Pick
Turn
Walk
Chew
Laugh
New Class 1
New Class 2
New Class 3
Merged HMDB51 dataset with 42 classes

Fig. 3. Class merge progress

hierarchical neural network layers and automatically build the high-level representation from the raw video inputs. Hence, unlike the traditional hand crafted feature extraction methods, the CNN based feature extraction and classification process is embedded in an end-to-end pipeline. In short, a deep learning model applies multiple techniques such as local perception, weight sharing, multi-convolution kernel, down-pooling, etc. to study the features from the image or frames. The classifiers can be trained by either supervised or unsupervised methods, and the final result can be generated by the ensembling the results of multiple network layers. Deep learning techniques are widely used in visual object detection and tracking [8], handwriting and signature recognition [9], natural language processing [10], human action recognition [11], and image segmentation [12], etc. Convolutional Neural Network (CNN) is one of the popular deep learning models in computer vision research area. Convolutional neural networks are a type of deep models which include an input layer and an output layer. Between the two layers, there are multiple convolutional layers, pooling or sub-sampling layers, fully connected layers and normalization layers, which can be termed as hidden layers. Many research works have been done and showed that, with a well trained CNN model [13], the classifier could achieve high performance on object detection and recognition.

CNN has been wildly used for processing still images, because of its ability on feature construction through the different deep layer models. In this paper, we discover the use of the CNN models on video-based human action recognition. A simple way to apply the CNN on videos will be in the following steps. First, extract the frames from a video. Then, treat each frame as individual images and apply CNN models to recognize human actions at the image level. Thus, the approaches with the above strategy have been used in the early research works to analyze the human actions in videos [14]. However, the early works have the drawbacks such as they did not consider the temporal and motion information in the video frames. To adequately address this problem, A 3DCNN architecture [15] has been proposed by Ji et al. In the proposed method, the video will be analyzed by the multiple convolutional layers with 3D convolution and both the spatial and the temporal features are captured from three adjacent frames. Therefore, the motion and temporal information can be analyzed simultaneously.

Indeed, the 3DCNN approaches improved the performance of the action recognition. However, human actions in videos are not as simple as static objects. With the different actions, the body parts will follow different sequence of gestures listed Figure 1. The gestures will be very similar in the most of videos frames when the people perform certain actions. For instance, playing golf is very similar as picking up something, because in the most frames people are supposed to bend their back which is very similar as in the Figure 1 (a). Similar situations will happen incase of ”Swing and Throw”(Figure 1(b)), ”Chew and Laugh” (Figure 1(c)) and ”Turn and Walk” (Figure 1(d)). Hence, the drawback of CNN in videos are obvious, as CNN will generate almost the similar features on some of the actions with the similar gestures.

Thus, the performance of the classifier will be decreased by the mis-classified classes. To analyze the similar actions effectively and accurately, we propose a hierarchical classification model, in which the first layer classifies multiple classes, whereas, the second layer focus on classifying similar gestures. Specifically, in the first layer, confusing/similar gesture pairs are merged to form single classes. Hence, the problem space for first level of classification is reduced to less number of classes and higher accuracy can be achieved. In the second level of classification, the merged pair of classes are handled explicitly. In the second level of the classification the problem space is reduced to two classes. A binary classifier is applied to the respective merged pair of classes in order to resolve the confusion. The overall performance is measured by combining the first and second layer results.

We applied the proposed method on the HMDB51 dataset, which consist of 51 different actions recorded by Serre Lab from Brown University. We ensemble the actions contains the similar gestures (i.e., Turn and Walk, etc.) into single actions/classes as the input. The proposed system achieved high performance compared with the baseline CNN models. Our experiment also shows that the developed hierarchical model outperforms other baseline models on the similar actions.

TABLE I
MERGING THE SIMILAR GESTURE CLASSES

Classes	Accuracy reported in [16]	**Merged Classes**
$Jump$	**0.38**(low)	New Class 1
$Catch$	1.00	
$Kick\ Ball$	**0.31**(low)	New Class 2
$Punch$	0.51	
$Laugh$	0.41	New Class 3
$Chew$	0.47	
$Pick$	**0.27**(low)	New Class 4
$Golf$	1.00	
Sit	**0.39**(low)	New Class 5
$Stand$	**0.27**(low)	
$Throw$	**0.16**(low)	New Class 6
$Swing\ Baseball$	**0.16**(low)	
$Turn$	**0.222**(low)	New Class 7
$Walk$	**0.38**(low)	
$Wave$	**0.14**(low)	New Class 8
$Shake\ Hands$	0.82	
$Sword$	**0.13**(low)	New Class 9
$Sword\ Exercise$	0.42	

The major contributions of this work can be summarised as follows:

- We concentrate on mis-classification problem on similar gestures, instead of focusing on the whole dataset to improve the classification performance.
- We propose to ensemble the results from a hierarchical 3DCNN architecture (H3DCNN) to boost the performance of the final output. The performance of the classifier on similar actions will increase the combined global results and binary classifier results.
- We evaluate the hierarchical models on the HMDB51 dataset in comparison to the baseline CNN methods. Experimental results show that the proposed method outperforms other baseline methods on similar gesture actions, and also on the overall accuracy.

The rest of this paper is organized as follows: We introduce some related work for action recognition in Section II. The dataset preparation and hierarchical 3DCNN architecture will be discussed in Section III. The experiment result has been reported in Section IV. The discussion and conclusion are in the Section V and VI respectively.

II. Related Work

In this section, we will briefly review the recent works related to our proposed model including 3DCNN methods and motion-related methods.

The basic idea of the 3DCNN is to perform the 3D convolution on videos which was proposed by Ji et al. [15]. The 3DCNN architecture generates the features of grey, gradient and optical flow by the hardwired layer from adjacent frames as different channels. Then, it applies convolution and subsampling on multiple channels. The final feature representation will be combined from all the channels. Based on the 3DCNN architecture, Tran et al. [17] proposed an optimized temporal kernel length for 3DCNN with a small $3 \times 3 \times 3$ kernel and built a new 3DCNN network with VGG-style. The new

3DCNN network named as C3D, contains eights convolutional layers, five pooling layers, and two fully connected layers, which could generate generic, efficient and compact features. The approaches mentioned above were trying to obtain temporal information from 3 to 16 video clips, respectively. To get a stabilized temporal information, Varol et al. [18] introduce a long-term temporal convolution (LTC) networks. Unlike 3DCNN, the LTC require more extended video clips with the length of 60 to 100 frames, which could demonstrate high-quality optical flow as the input.

Compared to the CNN based approaches; many works applied the motion related information as an input to CNN, such as, optical flow and motion vectors to incorporate temporal information. The two-stream model became a popular and important method for action recognition. Simonyan et al. [19] proposed an architecture to apply the optical flow as the input to obtain the motion information. The temporal and spatial information was processed in parallel, and fused with the softmax scores from the two streams. Feichtenhofer et al. [20] considered both spatial fusion and temporal fusion and proposed an improved two-stream model with bilinear fusion and 3D pooling. Adel et al. [16] aggregated the temporal coherent descriptors such as Histogram of oriented gradients (HOG), histogram of optical flow (HOF), motion boundary histogram (MBH) and fisher vectors (FVs) into a multiple kernel learning (MKL) algorithm which performs the optimal kernel and parameters from a large set of kernels to reduce the bias. In this paper, we use the joint sequence to represent high-level motion information which is more unique to specific actions than the optical flow. Also, we propose to fuse the two streams with a long-term convolutional network to achieve high accuracy on similar actions.

III. Methodology

This section describes the proposed architecture used to perform the task of action recognition on similar gestures. This section details about the dataset preparation method and presents the proposed hierarchical classification architecture.

A. Data preparation

Experiments are conducted on HMDB51 dataset, which is a state-of-art dataset to evaluate the proposed architecture (Figure 2). HMDB51 is a large and generic available public dataset for real-world actions collected by SERRE LAB from Brown University and firstly released on ICCV 2011 [4]. The videos of this dataset were collected from the Youtube and some movies which include a variety of actions with different human gestures including human body movements, body and objects interactions and some facial actions. It contains 7000 video clips distributed across 51 action classes, in which each class has around 100 video clips. It is a challenging dataset because the video clips of each class has different person performing the same gesture. Each subject performing the same action on different gestures and viewpoints have been recorded into 4 to 6 video clips. The proposed architecture is capable of handling the mis-classified actions which have similar gestures.

The most important process is how the similar gesture classes are merged to form a single class. To determine which classes to merge, we define two rules:

- Rule 1: Choose the classes with highest mis-classification rate, and
- Rule 2: Choose two classes which have similar gestures and have maximum confusion.

In order to identify the similar and confusing gesture classes, the overall performance of the state-of-the-art method [16] reported recently, was considered. Table I provides details about the performance of the similar and most confusing gestures, and also provides the information about the gesture pairs merged together to form single class. Similar gesture actions such as, ”Jump & Catch”, ”Pick & Golf”, ”Laugh & Chew” and ”Sit & Stand” etc. are chosen and merged into one class as shown in Figure 3. After the merging the classes, the number classes in the complete dataset (HMDB51) will reduce from 51 classes to 42 classes. Moreover, the size/number of samples in the complete dataset remains the same. This process will decrease the mis-classification rate and improve the overall accuracy of the dataset, as the dataset now has unique gestures.

B. Architecture Description

Figure 2 presents the proposed action recognition architecture. The hierarchical structure have two stages. The proposed architecture doesn't depend on any particular dataset and is generic. It can be applied to model real-world scenarios for gesture recognition. For the current work, HMDB51 dataset was considered for experiments and validating the proposed hierarchical architecture. The input data from HMDB51 dataset has 51 classes initially. After merging the similar gesture class pairs based on the rules defined in the previous section, 42 classes were formed.

The first stage of the proposed hierarchical classification model focuses on classifying the generic classes (complete dataset), whereas, the second stage resolves the similar/confusing gestures. Once an input video is classified to one of the similar/confusing gesture class by the first stage, the sample is passed to the second stage for further classification. The second stage comprises of a different binary classifiers, one for each of the confusing gesture pairs. The target binary classifier as selected automatically based on the first stage classification results. Additionally, in the first stage if a sample video is not classified as one of the similar gesture class, the sample video is not passed to the second stage and the predicted result is considered as the final result.

The final results are calculated by combining the global classification result from Stage 1 and similar gesture classification result from Stage 2. To obtain the final result of the original dataset, we will remove the result of the similar gesture classes in Stage 1 and consider the result from the binary classification in Stage 2.

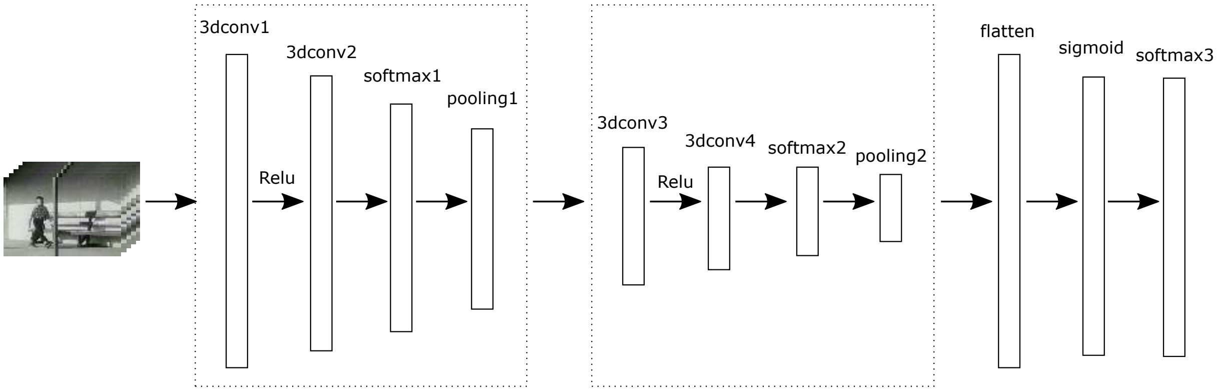


Fig. 4. Architecture of 3DCNN

C. Experiments setup

We use the Tensorflow [21] and Keras [22] frameworks to construct and train the neural networks. These frameworks are able to assist us to design the neural network architectures and algorithms for executing on GPUs. In our setup, we use NVIDIA P6000 and CUDA 8 platform to complete the experiments. We also applied a two level 3DCNN neural network with same kernel size 3×3 in Tensorflow and Keras for both global classification and similar gesture classification as shown in Figure 4. During the experiment, 60% and 30% of the whole dataset will be set as training and testing set respectively, and the rest of 10% will be set as validation set. We use the original RGB frames as the input and the result will be the benchmark to compare the performance between the system include and exclude the proposed hierarchical architecture.

IV. Evaluation

In this section, our proposed methodology is evaluated on the HMDB51 dataset. The accuracy (ACC) is used as an evaluation metric. The proposed 3DCNN architecture achieved the accuracy of 0.46 as reported in Table II. The resultant low accuracy classes is grouped into the pair of new classes based on the similar gestures. The total classes number of classes after grouping reduced from 51 to 42. It is demonstrated that classification accuracy increased to 0.52 globally after new classes. However, the reported increased classification result does not represent the performance on the whole dataset. Therefore, to assess the performance on the whole dataset, binary classification is applied to the new pair of classes and finally extending it to the classification result for all the classes in the dataset. The inclusion of binary classifiers in the hierarchical architecture further boost the performance to 0.632 accuracy.

TABLE II
Recognition accuracy on the HMDB51 dataset

Method	Accuracy
3DCNN on original dataset	0.46
H3DCNN on merged dataset	0.52
H3DCNN with binary classification	**0.632**

The average accuracy for the newly paired classes is reported in Table III. After the low accuracy performance of 18 classes, the classes have been merged into 9 classes. The results show that the average accuracy of each pair is overwhelming the result [16]. A significant increase in accuracy can be seen in the classes (Jump & Catch) from 0.69 to 0.82. It is also demonstrated that huge improvement of 0.16 to 0.82 can be seen for the classes Throw & Swing Baseball.

The binary classification result for a new pair of classes is reported in Table IV. In comparison with [16], improved accuracy of sit action from 0.39 to 0.49, and the pick action have been improved from 0.27 to 0.94. Similar improvement in accuracy is noted for the classes (Wave, Throw, and Jump). The losses in Figure 5 shows that for the most confused pairs (Throw & Swing Baseball) and (Turn & Walk), the loss dramatically declined after 100 epochs. Although the performance of some of the actions may have a slight decrease in accuracy, with our proposed hierarchical approach, the global performance is increased.

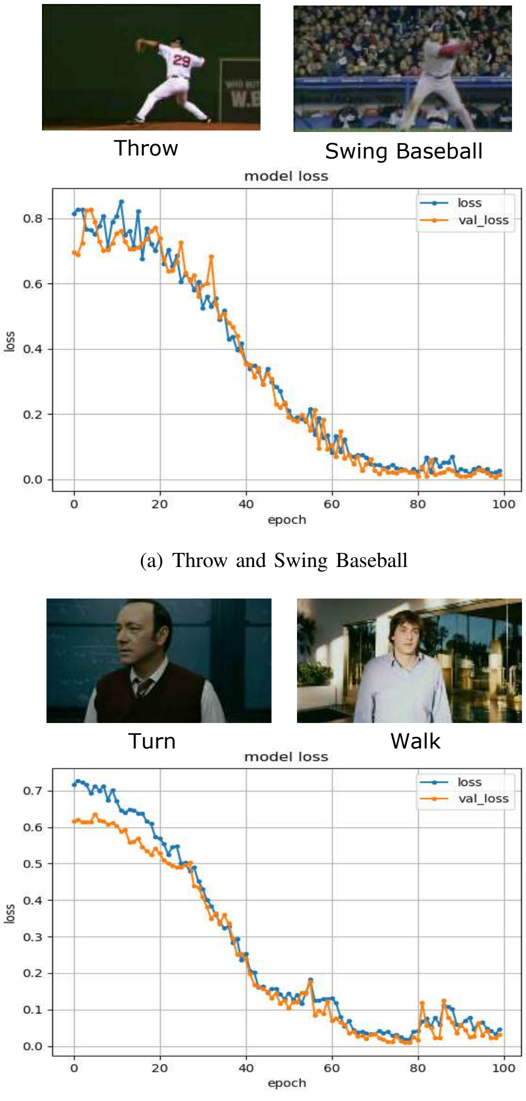
Table V shows the comparison between the proposed method and some of the state-of-the-art methods. In the HMDB51 dataset, we achieved an accuracy of 0.632. Thus our proposed architecture with two-stages can effectively classify actions on a global level, and similar gestures on the local level in the hierarchy thus outperforming the state-of-art methods [16].

V. Discussion

Our results show that the H3DCNN architecture indeed improves the performance of the classifiers. Although some of the unconfused classes can achieve a high classification accuracy around 90%, on average, we take an improvement on both globe accuracy and accuracy on similar gesture actions. We combine the convolutional and binary classification to achieve this improvement. This combination obtained better globe results and boost the results on similar classes compare with the state-of-arts works. Using 3DCNN combine with other methods such as LSTM does not achieve as same as

TABLE III
COMPARISON OF AVERAGE ACCURACY ON PAIRED CLASSES

Similar Gesture Pair	Accuracy reported in [16]	**Proposed Method**
Jump & Catch	0.69	**0.82**
Kick Ball & Punch	0.41	**0.95**
Laugh & Chew	0.44	**0.86**
Pick & Golf	0.64	**0.95**
Sit & Stand	0.33	**0.76**
Throw & Swing Baseball	0.16	**0.82**
Turn & Walk	0.3	**0.72**
Wave & Shake Hands	0.48	**0.8**
Sword & Sword Exercise	0.28	**0.84**



(a) Throw and Swing Baseball

(b) Turn and Walk

Fig. 5. The training loss and validation loss for the binary classification with the most confusing gestures

TABLE IV
COMPARISON OF ACCURACY FOR EACH CLASS IN PAIRS AFTER BINARY CLASSIFICATION

Classes	Accuracy reported in [16]	**Proposed Method**
Jump	0.38	**0.95**
Catch	1.00	0.77
Kick Ball	0.31	**0.83**
Punch	0.51	**1.00**
Laugh	0.41	**0.93**
Chew	0.47	0.44
Pick	0.27	**0.94**
Golf	1.00	0.94
Sit	0.39	**0.49**
Stand	0.27	**0.66**
Throw	0.16	**0.7**
Swing Baseball	0.16	**0.93**
Turn	0.222	**0.57**
Walk	0.38	**0.81**
Wave	0.14	**0.55**
Shake Hands	0.82	**0.88**
Sword	0.13	**0.83**
Sword Exercise	0.42	**0.83**

TABLE V
COMPARISON OF RECOGNITION ACCURACY ON THE HMDB51 DATASET WITH STATE-OF-ART METHODS

Method	Accuracy
LSTM mode [23]	0.44
Two-stream CNN [19]	0.594
Learning to rank [24]	0.618
Coherence learning to rank with MKL [16]	0.62
Proposed Method	**0.632**

performance with our architecture, which can be explained with the advantage of binary classification.

Dynamic analysis and evaluation are also critical, in this work we only use 3DCNN as both globe classifier and binary classifier. There could be other classifiers can achieve a better result, which we will explore it in the future work. By joining other classifiers or methods, we could test different parameters which may improve the result as well. Also, we only test the result on the single HMDB51 dataset, which we obtain high performance. However, there are still many datasets contains actions with similar gestures. Future work will be dedicated into two parts. The first part is to test different methods or algorithms on multiple datasets, in which select and build the dataset with similar gestures will be considerable work. And the second part will design a system which can automatically pair the misclassified classes in the pre-processing stage.

The future work will be applying the proposed approach on the multiple datasets such as UCF101 and Youtube Action datasets. By evaluating the results, we will redesign the CNN network model in stage 2 and also design the end-to-end approach to make the system efficiency.

VI. Conclusion

In this paper, we design a new approach to handle the actions with similar gestures to improve the overall accuracy of a gesture recognition system. Analysis showed that a major reason for low performance is due to the confusion among the similar gestures. Hence, we focus on resolving the confusion among the class with similar gestures, in the current work. A generic hierarchical classification model is proposed in this work, which can be applied to any datasets/real-world application involving gesture recognition. The first stage classifies the individual class as well as the new class formed by merging the similar gestures. In the second stage, binary classification is used to resolve the confusion among the similar gesture classes. Experimental results indicate that the proposed approach outperforms not only other neural network architectures but also the methods which uses 3DCNN. Overall, our method achieves better performance on HMDB51 dataset, compared to the state-of-the-art action recognition approaches.

References

[1] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” *International Conference of Learning Representations*, 2016.

[2] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[3] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in *In European Conference on Computer Vision (ECCV)*, 2016.

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: a large video database for human motion recognition,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2556–2563.

[5] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.

[8] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, “A study on detecting drones using deep convolutional neural networks,” in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–5.

[9] X.-X. Niu and C. Y. Suen, “A novel hybrid cnn–svm classifier for recognizing handwritten digits,” *Pattern Recognition*, vol. 45, no. 4, pp. 1318–1325, 2012.

[10] R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[13] H. Mobahi, R. Collobert, and J. Weston, “Deep learning from temporal coherence in video,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 737–744.

[14] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.

[15] 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, 2013.

[16] A. Saleh, M. Abdel-Nasser, M. A. Garcia, and D. Puig, “Aggregating the temporal coherent descriptors in videos using multiple learning kernel for action recognition,” *Pattern Recognition Letters*, vol. 105, pp. 4–12, 2018.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[18] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[19] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in neural information processing systems*, 2014, pp. 568–576.

[20] C. Feichtenhofer, A. Pinz, and R. P. Wildes, “Spatiotemporal multiplier networks for video action recognition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 7445–7454.

[21] S. S. Girija, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” 2016.

[22] F. Chollet *et al.*, “Keras,” https://github.com/fchollet/keras, 2015.

[23] N. Srivastava, E. Mansimov, and R. Salakhudinov, “Unsupervised learning of video representations using lstms,” in *International conference on machine learning*, 2015, pp. 843–852.

[24] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5378–5387.