Part-based Graph Convolutional Network for Action Recognition: Supplementary Material

Kalpit Thakkar
kalpit.thakkar@research.iiit.ac.in
P J Narayanan
pjn@iiit.ac.in

Center for Visual Information Technology (CVIT), Kohli Center for Intelligent Systems (KCIS), IIIT Hyderabad, India

In this document, we present findings from further quantitative analysis on the action recognition results. Specifically, we compute the confusion matrices of the performance of different models and explain the useful model properties based on our observations. We find that graph-based models can understand actions which involve more motion better than those where skeleton motion is very less and contains object interactions. We also show the importance of using geometric and kinematic features instead of 3D joint locations by performing an experiment on graph-based model of Yan *et al*. [2].

1 Quantitative Analysis

We compute the confusion matrices for performance of our part-based graph model, graph model using only one part and Yan's graph model [2]. We did not include Li's graph model [1] as no code has been provided by the authors to reproduce the results. The performance for cross subject (CS) evaluation protocol is considered as it is more challenging than the cross view (CV) evaluation protocol. The confusion matrices for different models are shown in Figure 1 (model-1), 2 (model-2) and 3 (model-3). The recognition accuracy for each of these models for cross subject (CS) evaluations is 85.6, 87.5 and 81.5 respectively. The model corresponding to Figure 1 is a one-part graph model which does not divide the skeleton graph into parts and it takes a combination of relative joint coordinates $\mathbf{D}_R$ and temporal displacements $\mathbf{D}_T$ as input. The model corresponding to 2 is our four-part graph model with $\mathbf{D}_R$ and $\mathbf{D}_T$ as input. Finally, Figure 3 corresponds to graph-based model introduced in Yan *et al*. [2] for skeleton action recognition. We proceed to identifying the action classes for which the recognition performance is bad, explain what the reasons are for such performance, propose a possible solution and then compare performance across different classes for models with respect to model-2.

1.1 Commonly confused classes

The confusion matrices have boxes marked around certain values. These boxes represent the confused classes which are consistent across all models. For example, one of the boxes is around action classes 11 & 12, which correspond to "reading" and "writing" actions. These actions are mostly confused amongst each other and also with actions such as "playing with the phone / tablet" or "typing on a keyboard" (actions 29 & 30 present in the other marked box) which is clear from the confusion matrices. In all these actions, there is almost no skeleton motion and the differences are manifested in the form of interaction with different

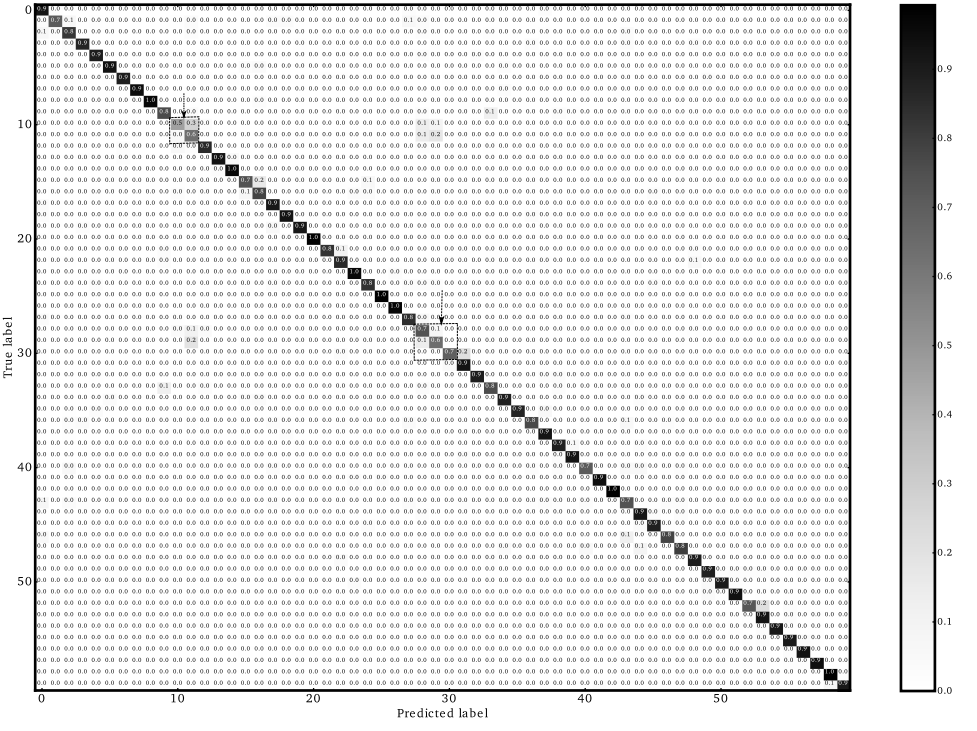

Figure 1: Confusion matrix for model with one part and combined geometric + kinematic features as input.

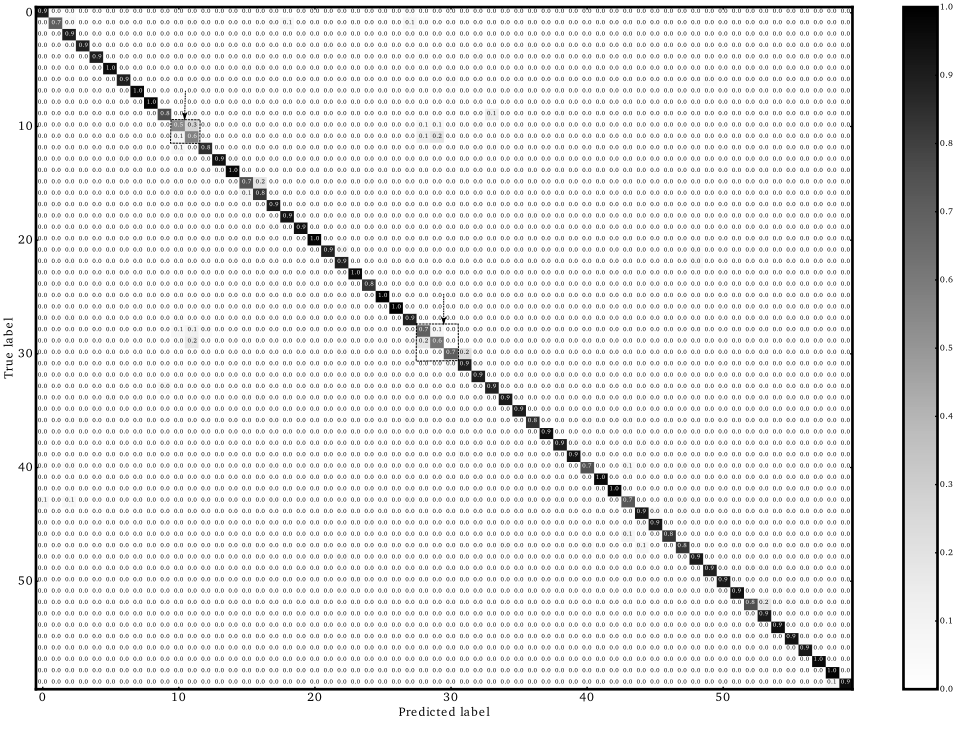

Figure 2: Confusion matrix for model with four parts and combined geometric + kinematic features as input.

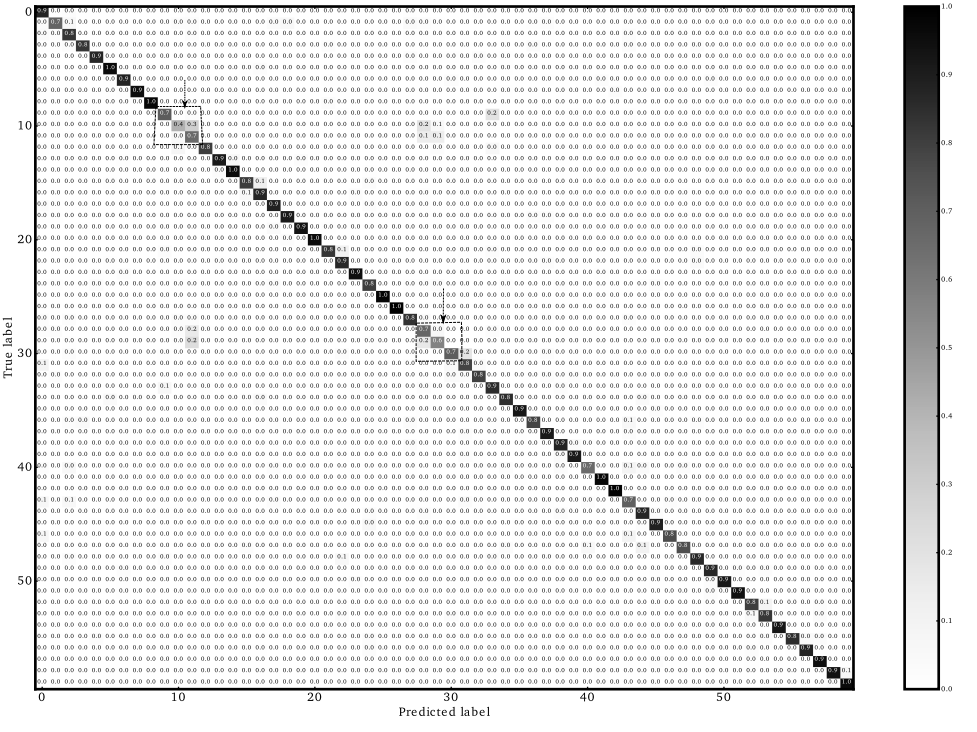

Figure 3: Confusion matrix for Yan's graph-based model [2] having 3D joint locations as input signals.

Model	Accuracy	
	CS	CV
Yan [2] (model-2)	81.5	88.3
Model-2 + $\mathbf{D}_R \| \mathbf{D}_T$	86.3	92.1

Table 1: Results on NTURGB+D for model-2 [2], with and without the combined signal $\mathbf{D}_R \| \mathbf{D}_T$ (relative coordinates and temporal displacements).

objects. Due to these properties, models using skeleton information for recognizing actions give lower performance for these action classes as they do not have access to object information. A possible approach to overcome this limitation on recognition potential is to use RGB information along with skeleton information in order to get information about objects as well.

1.2 Model-1 vs Model-2

Model-2 improves over Model-1 by using a part-based graph representation instead of considering the entire graph as one part. Model-2 achieves better recognition performance by improving over action classes such as "brushing teeth" (class 3), "cheer up" (class 22), "make a phone call/answer phone" (class 28), etc. These actions have a strong correlation with movement of both hands and legs. Due to this correlation, our part-based graph model is able to achieve better performance as it learns from these parts specifically and uses an intuitive way to divide the human body into parts. Being agnostic to parts in human skeleton helps in learning a global representation but learning importance of parts using such a model is difficult, compared to a part-based model.

1.3 Model-3 vs Model-2

Spatio-temporal model of Yan *et al*. [2] confuses the action of "clapping" as well along with the actions mentioned in section 1.1. The model proposed by Yan [2] partitions the edge set and uses the same vertex set for each partition of edge set. We believe that their model learns the importance of different edges in the skeleton graph and does not learn the importance of parts like our part-based graph model. In order to understand the influence of geometric and kinematic signals as input to a graph-based model, we use the signals on top of model-3 and we find that we get a boost in recognition performance for model-3. The recognition accuracy on NTURGB+D is shown in Table 1. This experiment shows that the signals help in improving recognition performance for different graph-models for skeleton action recognition.

2 Conclusion

Using a part-based model works better than using a model that does not partition the skeleton graph. However, using only skeletal data for action recognition is not enough as different actions might have similar dynamics of parts in the skeleton but different object interactions. In such cases, RGB information can be used to disambiguate interactions with objects. Providing the network with a cue that is known apriori to work well for the task at hand, viz. relative coordinates and temporal displacements for skeletal action recognition, can improve recognition performance by a large amount as we show in our experiment on previous state-of-the-art model for NTURGB+D [2].

References

[1] Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, and Jian Yang. Spatio-temporal graph convolution for skeleton based action recognition. *AAAI Conference on Artificial Intelligence*, 2018.

[2] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI Conference on Artificial Intelligence*, 2018.