Supplementary Material: Multi-View Action Recognition using Contrastive Learning

Ketul Shah[1] Anshul Shah[1] Chun Pong Lau[1] Celso M. de Melo[2] Rama Chellappa[1]
[1]Johns Hopkins University [2]DEVCOM Army Research Laboratory
{kshah33, ashah95, clau13, rchella4}@jhu.edu celso.miguel.de.melo@gmail.com

1. Data Augmentations

Video augmentations were applied similar to [4]. For each clip, spatial augmentations are applied to all frames consistently. Specifically we use, random crop, random horizontal flip with flip probability 0.5, gaussian blur with a standard deviation chosen randomly from $[0.1, 2]$ and color jitter the following parameter values: 0.4 for brightness, contrast, saturation and 0.1 for hue. For temporal augmentation, we sample clips from random time in video. At test time, we use the center crop with no augmentations.

2. 1-crop results

We present our single (center) crop test results on NTU-60, NTU-120 and NUMA datasets in Table 1, along with results of baseline methods which report 1-crop test results. Our method shows significant improvement in performance over previous methods.

	NTU-60		NTU-120		NUMA
	xview	xsub	xset	xsub	xview
DMCL [3]	-	-	84.3	-	-
Glimpse Clouds [1]	93.2	86.6	-	-	-
Vyas *et al*. [5]	86.3	82.3	-	-	83.1
Ours	97.6	91.3	86.4	85.4	89.1

Table 1: 1-crop test results on NTU-60, NTU-120 and NUMA.

3. More details on RoCoG experiments

RoCoG [2] is a gesture recognition dataset consisting of synthetic and real videos from seven gestures captured from multiple viewpoints.

Each video in the original dataset contains multiple instances of a person performing different activities. We preprocess the data by temporally splitting each video into individual instances containing a single activity. This results in 9912 synthetic videos and 970 real videos, each with a corresponding gesture label.

For RoCoG experiments, we use sixteen frame clips with skip rate of two as input for each method. Label smoothing of 0.2 is used for classifier targets. Temperature is set to 0.07. We choose the inception I3D network as the feature encoder with 16 frame input clips.

4. Societal Impact

In our work we propose a novel approach for multi-view action recognition. As such, our contribution is on a more fundamental level and we do not anticipate any harms through the method itself. But, given that we are working with videos it is essential to ensure that we obtain required consent for the people in the video. Further, we also show experiments using synthetic data for training which eliminates privacy concerns.

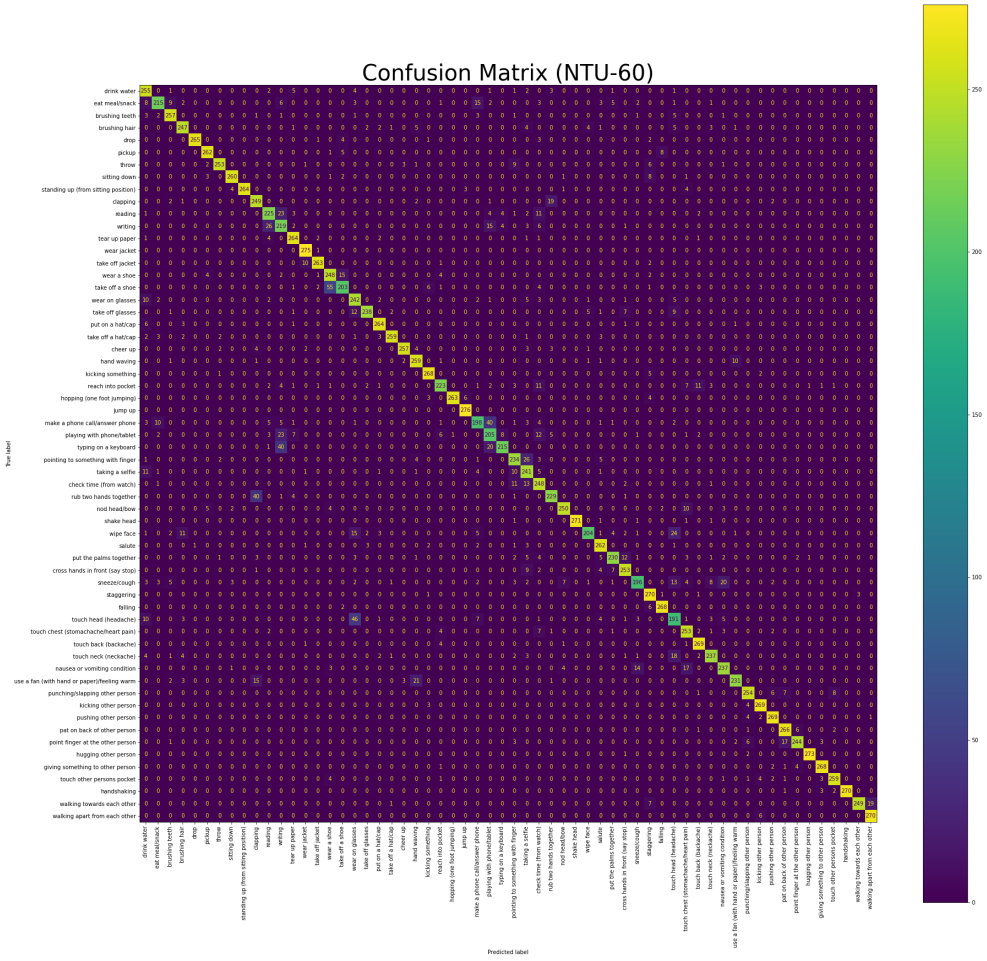

Figure 1: Confusion Matrix for NTU-60 test set.

References

[1] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Celso M de Melo, Brandon Rothrock, Prudhvi Gurram, Oytun Ulutan, and BS Manjunath. Vision-based gesture recognition in human-robot teams using synthetic data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10278–10284. IEEE, 2020.

[3] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2755–2764, 2021.

[4] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in NIPS*, 33:5679–5690, 2020.

[5] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *European Conference on Computer Vision*, pages 427–444. Springer, 2020.