Supplementary material for STP-Net: Spatio-Temporal Polarization Network for action recognition using polarimetric videos

R. Krishna Kanth
TCS Research
Bangalore, India
rokkamkrishna.kanth@tcs.com

Akshaya Ramaswamy
TCS Research
Chennai, India
akshaya.ramaswamy@tcs.com

A. Anil Kumar
TCS Research
Bangalore, India
achannaanil.kumar@tcs.com

Jayavardhana Gubbi
TCS Research
Bangalore, India
j.gubbi@tcs.com

Balamuralidhar P
TCS Research
Bangalore, India
balamurali.p@tcs.com

1. Visualization of mis-classification

In order to analyse the mis-classification using the proposed architecture and make a comparison between polarized and unpolarized inputs, we visualize the confusion matrices, as shown in figures 1 and 2. Here are some of the observations:

1. The overall confusion with multiple classes is more in the case of unpolarized input, and is prominent on the FGA dataset.

2. On the HumanAct12 dataset, for classes *warm-up* and *jump*, the videos get classified into multiple incorect classes with unpolarized input. The general classification trend is the same with both inputs, but it is to be noted that without the polarimetric information, there is confusion even between classes that are quite dissimilar in the way they are performed, for example, *jump* is confused with *throw* action, and *drink* action is mis-classified as *boxing*.

3. On the FGA dataset, a vertical pattern is seen with unpolarized input, wherein videos from almost all the classes get wrongly predicted as *moving from left to right* or *placing something on the table*. It is evident that the network is unable to learn the features necessary to discriminate between these classes using the unpolarized information. In contrast, this is not observed in the matrix on the right in figure 2.

4. A key observation in figure 2 is that 67% of the *right to left* video get mis-classified as *left to right*, whereas this does not happen when polarimetric properties are captured in the input.

2. Visualization of class activation maps

The above observations are further validated by the visualization of the class activation maps for different actions in the two datasets, as shown in figures 3 and 4. The key inferences from these are as follows:

1. In the HumanAct12 dataset, it is observed that with polarized input, the network is able localize the correct action regions in atleast one frame for all the action classes. In contrast, without the polarization information, the network does not clearly highlight any regions such as in *sit* and *lift dumbell* actions, or the network focuses on other regions such as keypoints on the knees or chair visible in *talk on phone* action. These activation regions can be associated with multiple classes and shows that the network is not able to accurately localize the discriminative regions.

2. With the FGA dataset, the difference is significant. This is due to two reasons 1) the dataset is captured in a constrained environment, so with minimum external factors, the benefit of using the polarimetric modality is clearly validated, 2) the dataset captures fine-grained actions which are challenging to discriminate if the motion is not captured by the network. It is evident that with polarized input, the network exactly captures the motion regions. Without the polarization input, the network fails to localize the action, and in general seems to just focus on the centre of the frame. These also confirms the inferences from the quantitative evaluations and the class-wise plots that highlight the confusion with multiple classes and poor generalization across action classes.

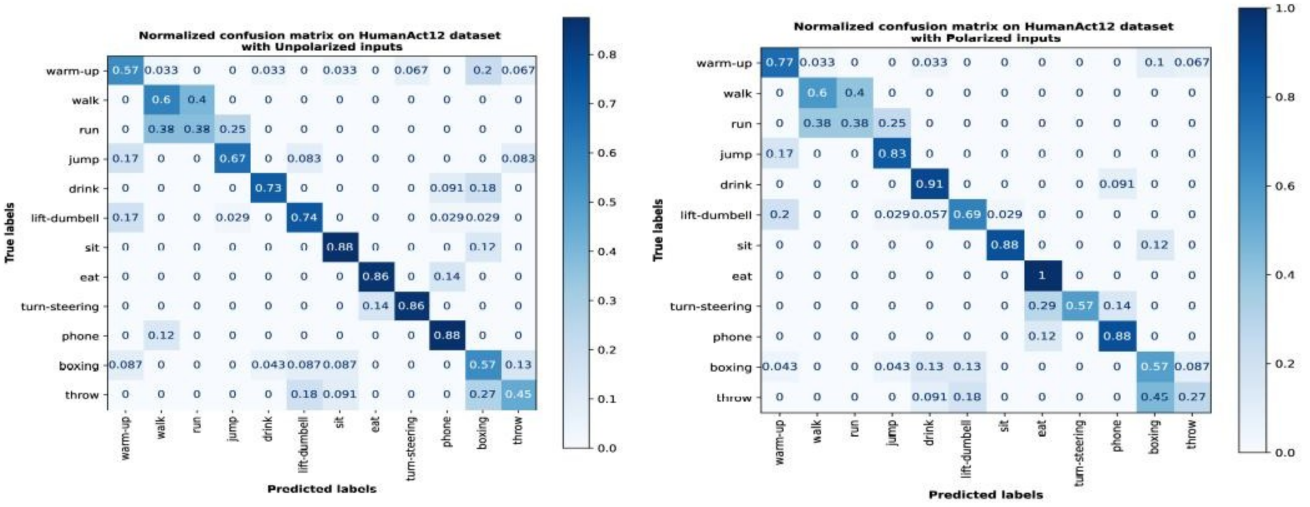

Figure 1. Normalized confusion matrix using STP-Net trained using unpolarized input (left) and polarized input (right) on the HumanAct12 dataset

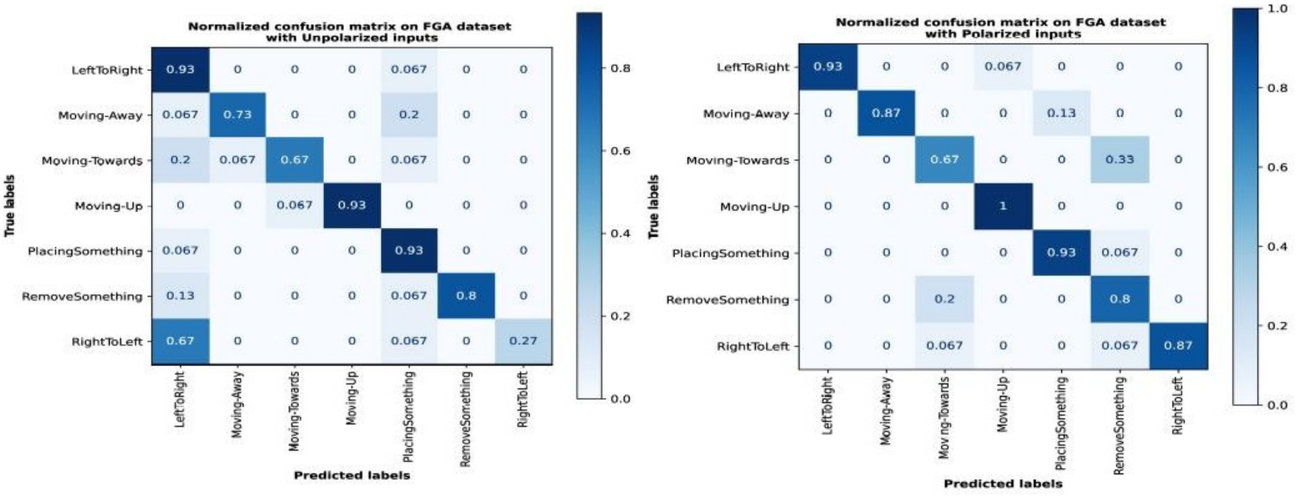

Figure 2. Normalized confusion matrix using STP-Net trained using unpolarized input (left) and polarized input (right) on the FGA dataset

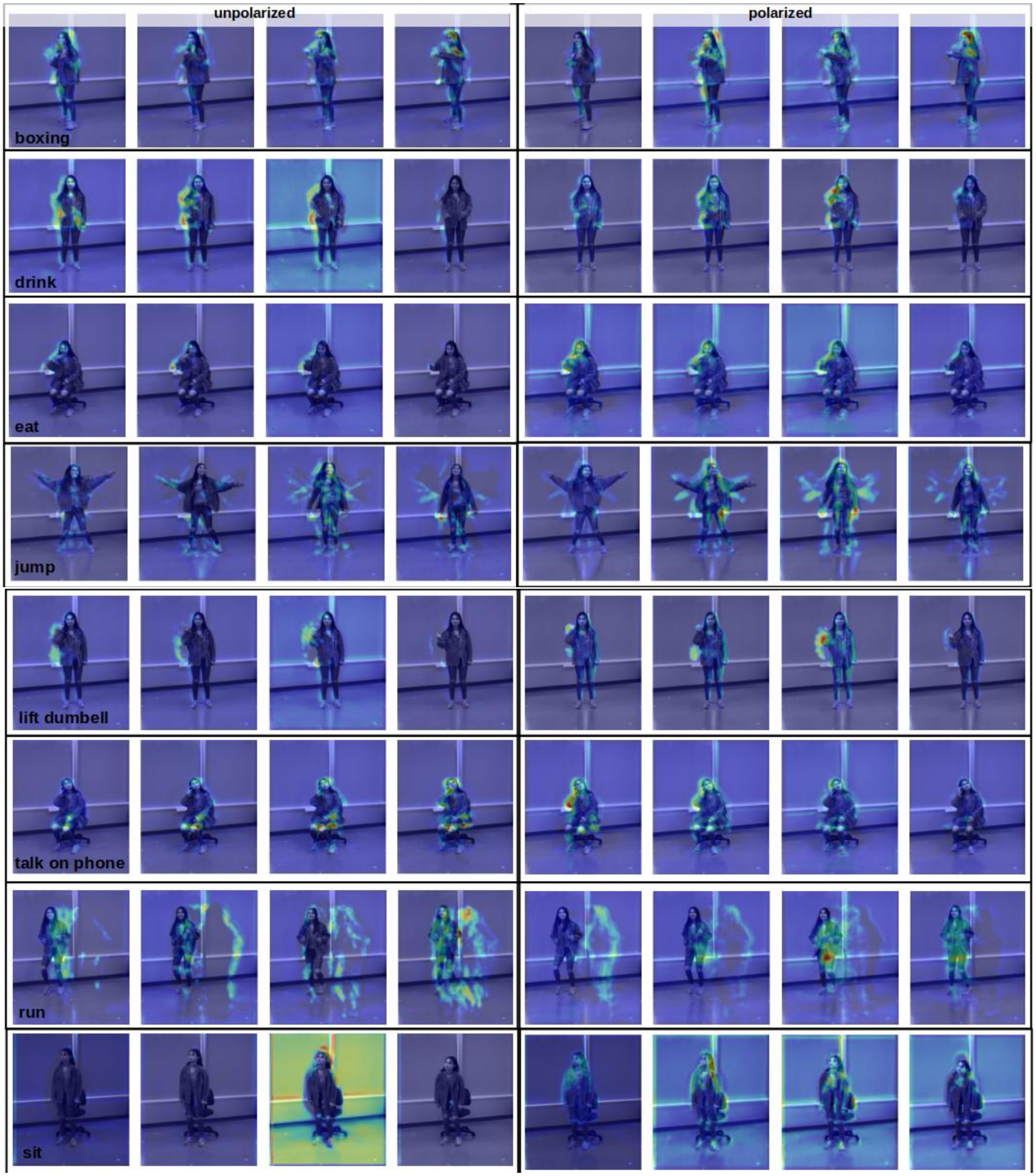

Figure 3. Class activation maps achieved using unpolarized input and polarized input for different actions on the HumanAct12 dataset

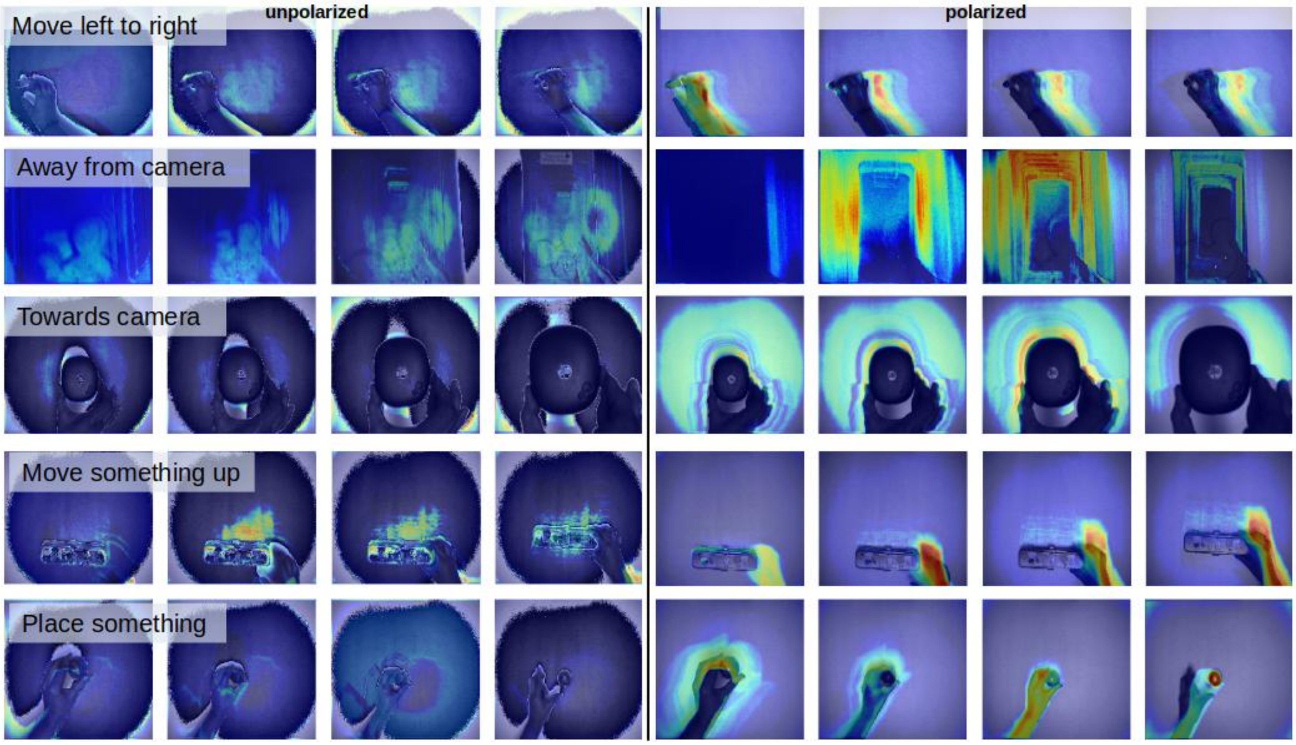

Figure 4. Class activation maps achieved using unpolarized input and polarized input for different actions on the FGA dataset