## A. Appendix

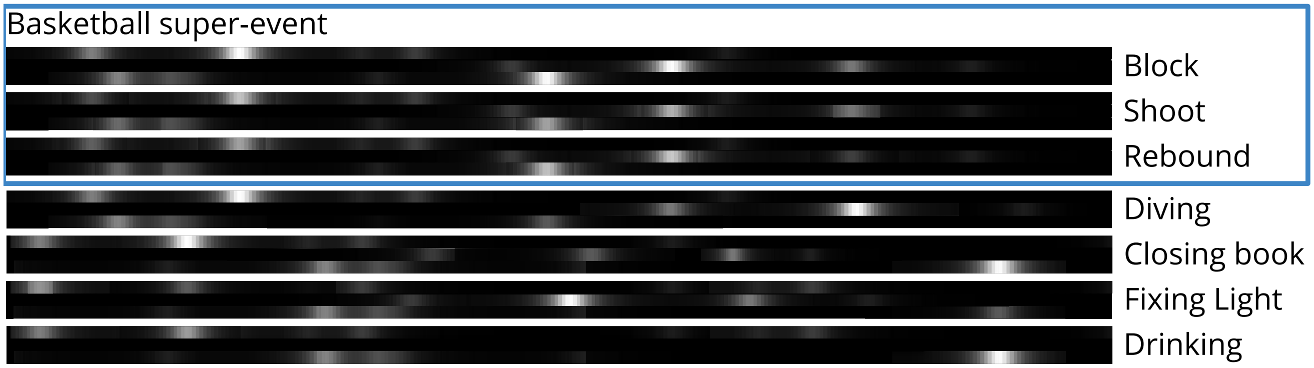

Figure 1. Comparison of the learned temporal structure filters on several different activity classes. The filters are visualized by combining our learned temporal structure filters using their soft-attention weights. We obtain 5 $T \times N$ filters, and sum them based on the learned attention weights. These are global super-event representations which select intervals from the entire video. These filters are scaled to match the length of the video by construction. As a result, even though the videos are continuous and have different lengths, these filters capture the temporal relationships/ordering between the activities assuming their overall relative locations within each video are similar. If such assumption does not hold, we can use the relative version of our super-event representation with more computation. This figure shows that the basketball activities end up learning a very similar global super-event structure (i.e., they share it), while unrelated activities learn different temporal structures.