Supplementary Material: Action Graphs: Weakly-supervised Action Localization with Graph Convolution Networks

Maheen Rashid
University of California, Davis
mhnrashid@ucdavis.edu

Hedvig Kjellström
KTH Royal Institute
of Technology
hedvig@kth.se

Yong Jae Lee
University of California, Davis
yongjaelee@ucdavis.edu

1. Number of Training Videos Per Graph

An interesting consequence of our graph based method is that it enables the network to learn from similarity relations between time segments from *multiple* videos. In other words **G** can be constructed from time segments from multiple videos during training which can allow our network to learn from instances of the same action from different videos, as well as cluster together background time segments that are common to videos of different classes. All results in the main paper used a graph video size of 1. In this section we explore the effect of varying the number of videos per graph during training. We continue to test on a single video at a time during test time. As a result, increasing graph video size creates a disparity between train and test settings as the network comes to expect features that are modified by graph edges going between time segments from different videos.

Figure 1 shows the effect of varying number of graph videos on performance. For THUMOS '14, increasing graph video size does not lead to an increase in performance, with performance following a downward trend as graph video size is increased. On the other hand, ActivityNet's performance is improved up till graph video size of 4 and then performance begins to decline as the gap between train and test settings increases.

One reason for this difference between datasets can be the typical number of action occurrences per video. While THUMOS '14 often features multiple short occurrences of an action in a video, ActivityNet often features a single long occurrence. As a result, using information from multiple videos is more beneficial to ActivityNet as it can use multiple videos to build a better background model. Automatically deciding an optimal graph size for a dataset is an interesting direction for future research.

2. Qualitative Results

Figure 2 shows additional qualitative results from the THUMOS '14 dataset. The ground truth is shown in blue, with our detections in green. Overall, our method is good at localizing all occurrences of an action.

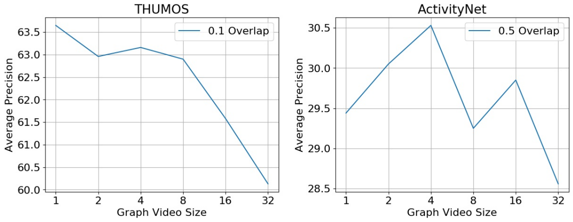

Figure 1. The effect of varying graph video size on THUMOS and ActivityNet. While performance decreases for THUMOS as graph video size is increased, for Activity performance improves before declining as train and test setting become very different.

The first row shows an example of a video with multiple 'Hammer Throw' occurrences during most of the video, followed by a few occurrences of 'Clean and Jerk'. Our method is able to localize almost all occurrences, however sometimes the localizations are too short in length, or broken in to multiple occurrences. On the other hand, in the second example of 'Soccer Penalty', our model provides localizations that are a little too long compared to the ground truth.

In Figure 3 we show some failure examples of our system. Multiple action occurrences that happen close in time are lumped together as a single detection in the first and second examples for the actions of 'Tennis Swing' and 'Cricket Bowling'. However, the network is able to distinguish multiple occurrences of both actions from longer segments of time when no action is happening, as indicated by the lack of false positives. While our network is able to localize almost all instances of 'Shot Put' in the third row, our detections do not span the full duration of the action and have poor overlap. Finally, our network fails completely in the last example of 'Volleyball Spiking', where it localizes the start and end of the spiking action rather than its actual duration.

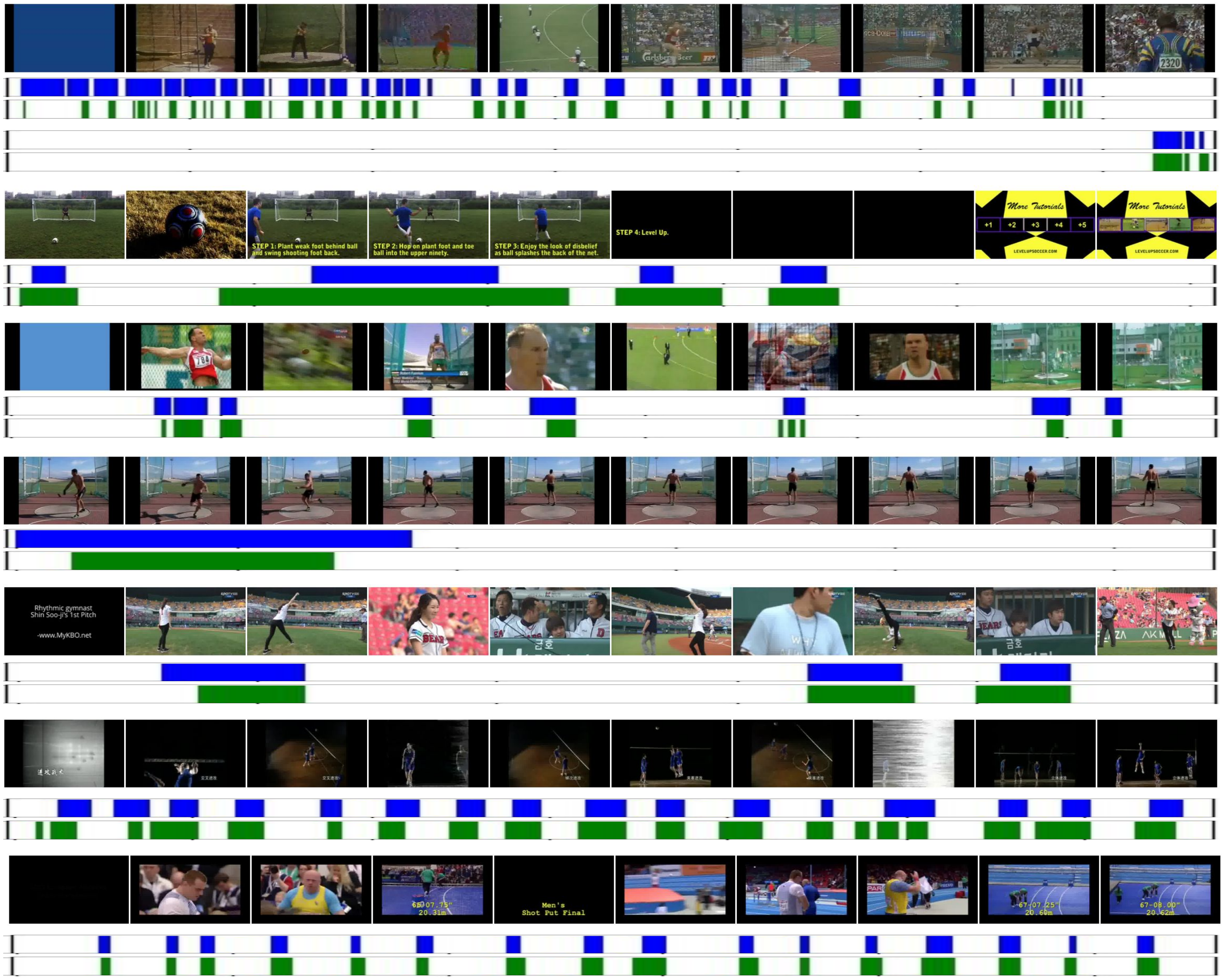
STEP 1: Plant weak foot behind ball and swing shooting foot back.
STEP 2: Hop on plant foot and toe ball into the upper ninety.
STEP 3: Enjoy the look of disbelief as ball splashes the back of the net.
STEP 4: Level Up.
More Tutorials
+1
+2
+3
+4
+5
LEVELUPSOCCER.COM
Rhythmic gymnast
Shin Soo-ji's 1st Pitch
-www.MyKBO.net
Men's
Shot Put Final

Figure 2. **Qualitative results:** The groundtruth is in blue, and our detections are in green.

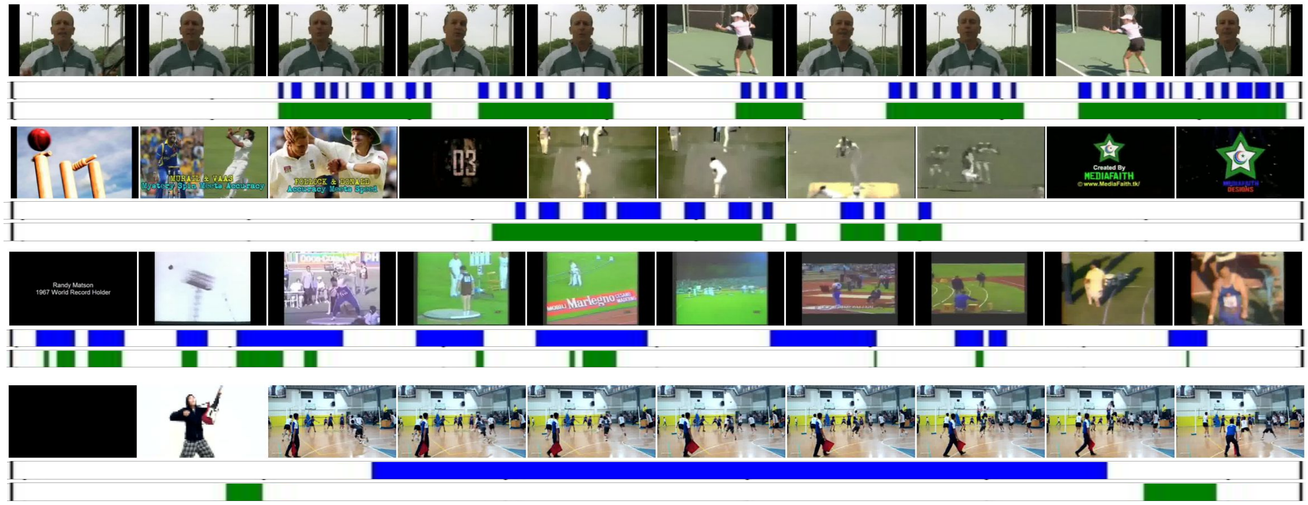

Figure 3. **Failure results:** The groundtruth is in blue, and our detections are in green.