Spatio-temporal Video Re-localization by Warp LSTM - Supplementary Material

Yang Feng[♯*] Lin Ma[♮†] Wei Liu[♮] Jiebo Luo[♯]
[♮]Tencent AI Lab [♯]University of Rochester
{yfeng23,jluo}@cs.rochester.edu forest.linma@gmail.com wl2223@columbia.edu

In this supplementary, we first provide more information about our constructed dataset for the spatio-termpoal video re-localization task. Afterwards, more qualitative results are provided.

1. Dataset Statistics

Table 1 illustrates the number of combined labels in the training set according to the number of atomic actions within a combined label. Specifically, there are 36 combined labels containing only one atomic action. Most of the combined labels contain more than one atomic actions.

Table 1. Number of combined labels in the training set with respect to the number of atomic actions within a combined label.

Number of atomic actions	Number of combined labels
1	36
2	317
3	743
4	576
5	299
6	24
total	1925

The distribution of the number of samples of each combined label in the training set is shown in Figure 1. It can be observed that the number of tubelet samples of each combined action label follows a long-tail distribution. We also illustrate the combined labels with the largest number of tubelets in Table 2.

Moreover, Figure 2 shows the number of tubelets in the training set according to the tubelet length. It can be observed that the number of tubelets drops dramatically when the length increases.

[*]This work was done while Yang Feng was a Research Intern with Tencent AI Lab.

[†]Corresponding author.

Table 2. Top 10 combined labels with the largest number of tubelets.

Combined label	Number of tubelets
stand + listen to + watch	13875
stand + talk to + watch	11405
stand	9859
stand + talk_to	9157
stand + watch	9073
walk	6515
stand + listen_to	6186
sit	6176
sit + talk to	6167
sit + listen to + watch	5611

2. More Visualizations

Figure 3 shows the five links learned by TrajLSTM [1] for the two videos. It can be observed that the learned links seems to be static connections, while warp LSTM is able to learn the motions in the videos.

Two more visualization results are shown in Figure 4. The combined label of the first and second queries are "sit + hold + talk to + watch" and "stand + listen to + talk to", respectively. The combined label of the man in the first reference video is "sit + listen to + watch". "warp LSTM" correctly localizes the woman in the first reference video, while the other methods make mistakes in some video segments. There are two ground-truth tubelets in the second reference video. The combined label of the left man changes from "stand + listen to + talk to" to "stand + talk to". Most methods detect the tubelets correctly except that the "clip" makes a mistake in the fourth segment.

References

[1] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *NIPS*, 2017.

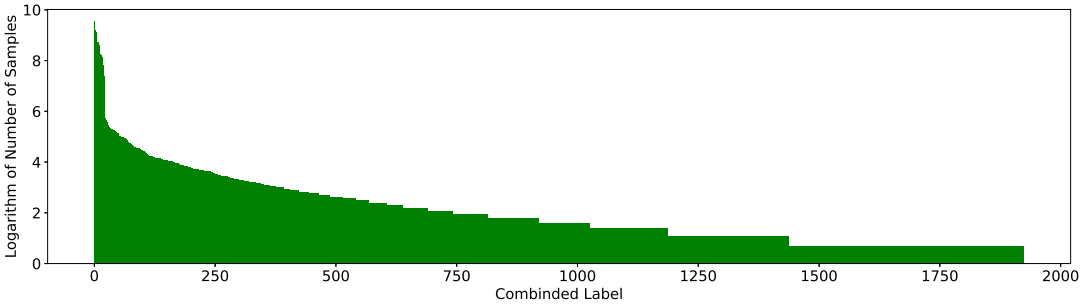

Figure 1. The distribution of the number of samples for each combined label in the training set.

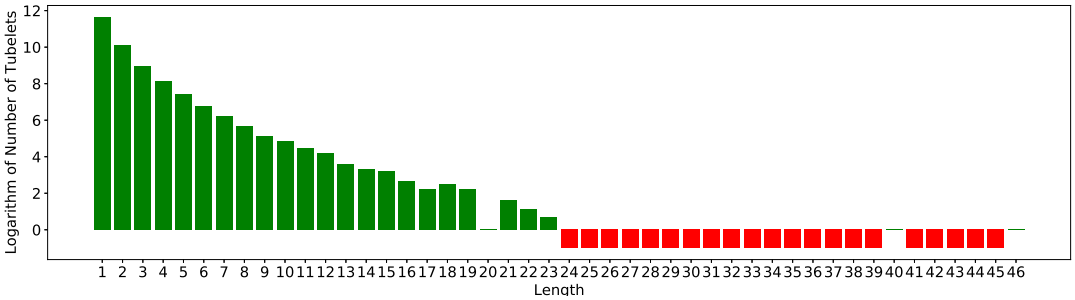

Figure 2. The logarithm of the number of tubelets in the training split according to length. Red bars mean that the number of tubelets with the corresponding length is ZERO.

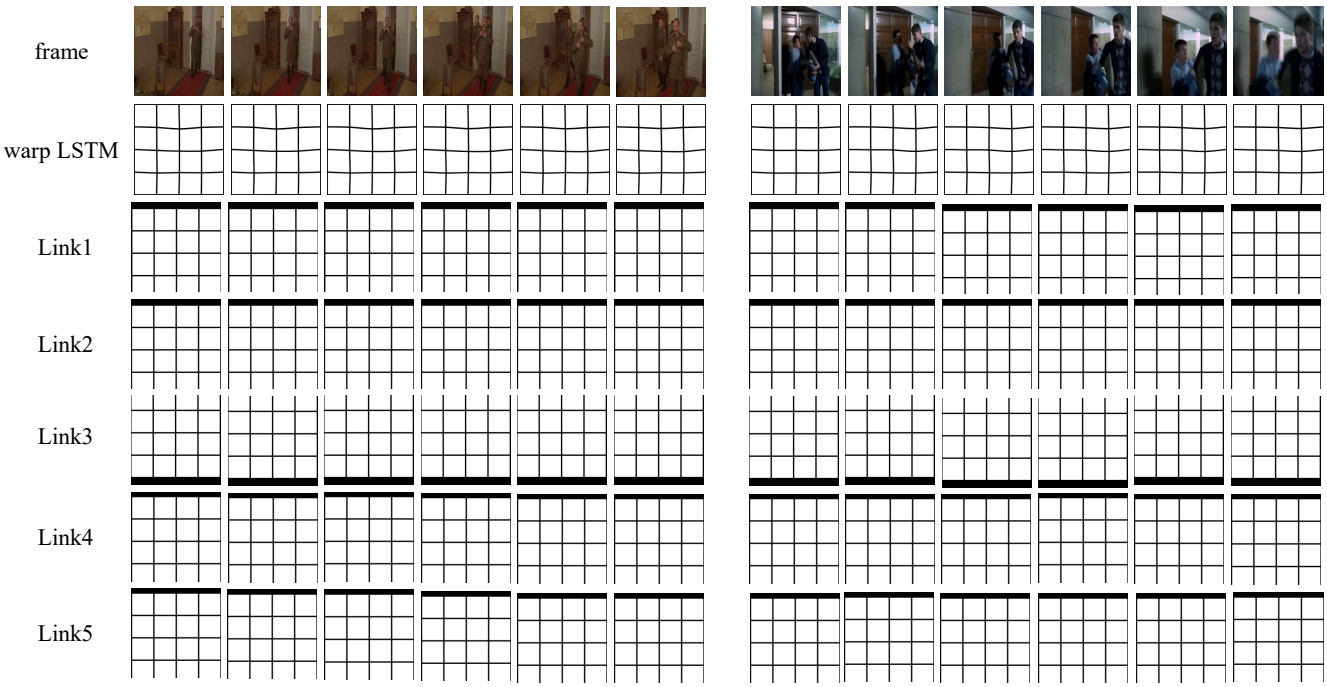

Figure 3. The visualization of the learned links in TrajLSTM [1].

Figure 4. More qualitative results of the spatio-temporal video re-localization.