Supplementary materials for
Intra- and Inter-Action Understanding via Temporal Action Parsing

Dian Shao Yue Zhao Bo Dai Dahua Lin
CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
{sd017, zy317, bdai, dhlin}@ie.cuhk.edu.hk

1. Implementation Details

Every video is evenly divided into 25 snippets, from each of which a frame is selected. We use the BN-Inception [1] pre-trained from ImageNet [3] as a backbone network to obtain the 1024-d feature for every frame. The network is trained with Adam [2] with a learning rate of 0.001. For the TCN-based method, the sequential model contains two layers of temporal convolution with kernel size of 5 and channel number of 256, followed by a non-linear activation of ReLU. For ISBA, we first use K-means to pre-group the frame-level features into K clusters, where $K = 64$. The rest follows the official implementation. For CTM, we maximize the log likelihoods for all possible labelings. We provide Figure 1 to demonstrate the difference of the optimization objectives between CTM, TCN and TransParser.

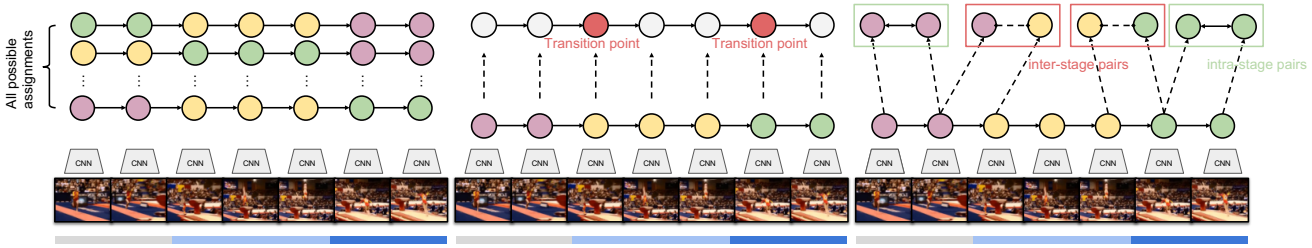


Figure 1: Illustration of the methods that we use in the experiment section. **Left**: The CTM method maximizes the sum of log likelihoods for all possible assignments; **Middle**: The TCN method predicts the probability of sub-action transition for every frame; **Right**: Our TransParser tries to enlarge the difference between adjacent sub-actions while encouraging semantic consistency within the same sub-actions.

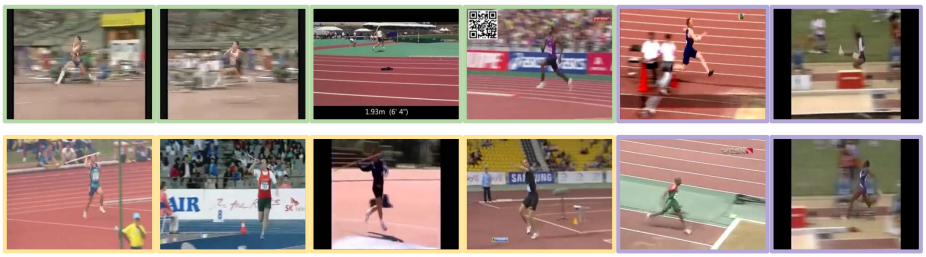
2. More Qualitative Results

Figure 4 illustrates the histogram of the most distinctive pattern that the miner ϕ discovers across different action classes. We can see that only a few number of rows of the miner are useful for every action. Further, certain patterns repeatedly occur at different classes. This means that some sub-actions are shared across different actions. We show some representative patterns that are mined by TransParser, along with the retrieved video frames in Figure 2. For example, approach running is commonly seen in the action of long jump, triple jump, tumbling, and pommel horse; Shot put, discus throw and hammer throw share similar appearance of standing after throwing. In contrast, some patterns only occur at one or two classes. They mainly focus on the visual cues which are specific for that action. As shown in Figure 3, ϕ_{13} depicts the action of squatting at the beginning of weightlifting; ϕ_{22} depicts walking or standing still which is commonly seen at the beginning of diving. ϕ_{15} and ϕ_{55} depict a person standing on a pommel horse or a balance beam respectively. Some falsely retrieved samples (with a red box) are visually similar.

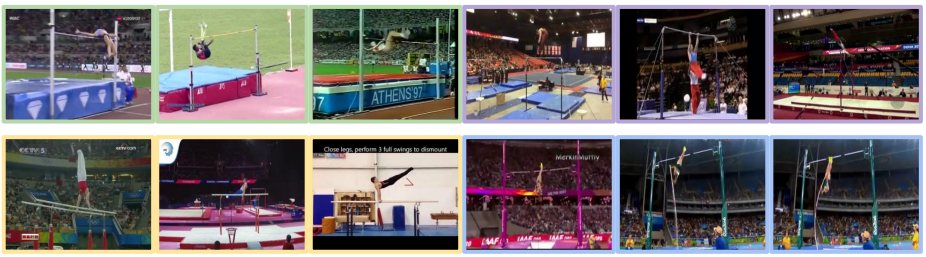
A video demo is also provided to illustrate the dynamics of sub-actions given a query video.

Pattern-3: approach running

Pattern-16: approach running (high jump) / bounding (long jump)



Pattern-30: circling



Pattern-52: (after) throwing

Figure 2: Illustration of some representative video frames retrieved by a certain ϕ_k. Frames in the same color belong to the same action label. The figure is best viewed in color.

Pattern-13: squatting

Pattern-22: walking

Pattern-15: pommel horse

Pattern-55: beam

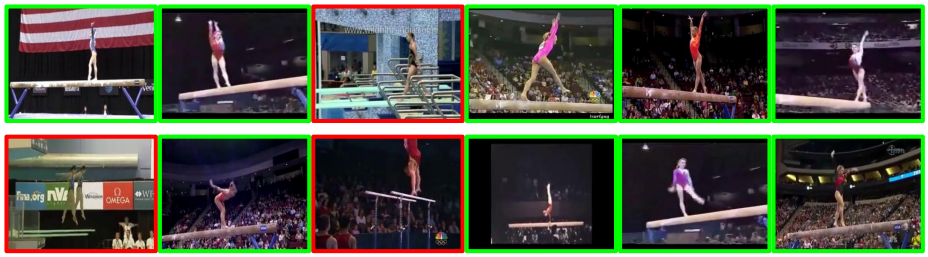


Figure 3: Illustration of some representative video frames retrieved by a certain ϕ_k. The figure is best viewed in color.

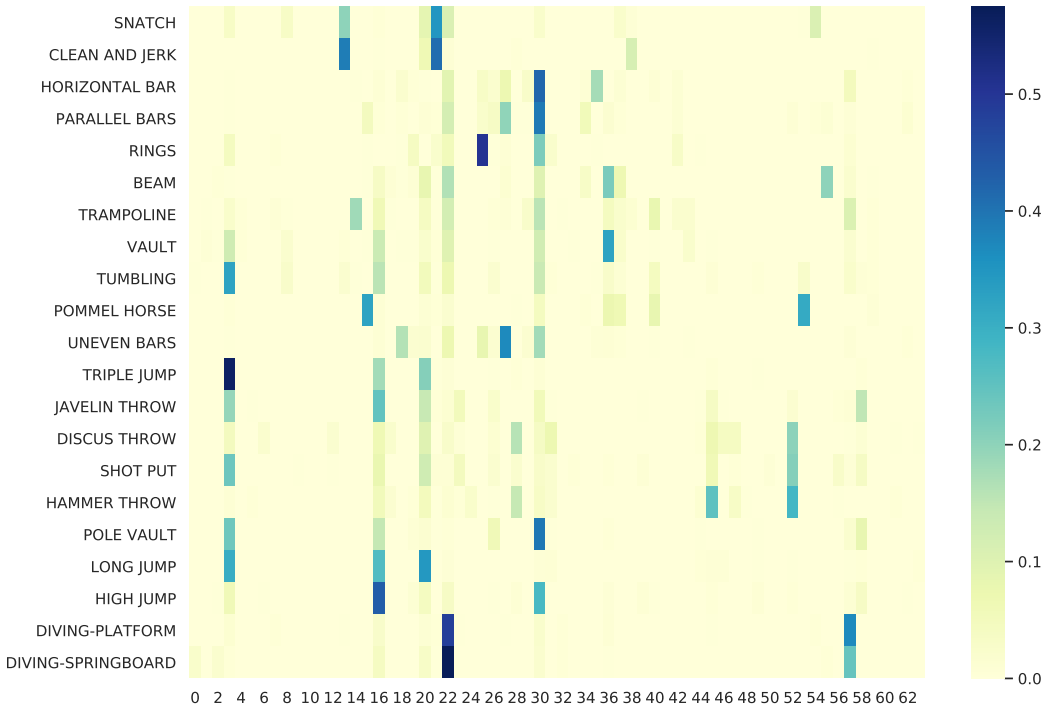


Figure 4: Histogram of occurrences of the mined patterns ϕ_k across different action classes.

References

[1] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. JMLR, 2015. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1