Protocol 11: Implementation Plan and Data Preprocessing

14.03.2019

Network Architecture for Lameness Detection

For behavior understanding, the dynamics of geometric body configuration may be more important than appearance. Similar to action recognition, in which the position of joints can effectively indicate the human behaviors, it is also true for lameness detection, in which the cow behaviors are of interest. Thus, skeleton-based methods would be promising for lameness detection. Although this type of methods require pose estimation as the preprocessing step, this step can be done by open source pose estimators. Two deep neural networks used for a similar task, namely action recognition, are shown below and will be applied to lameness detection.

1. Spatial-Temporal Graph Convolutional Networks (ST-GCN) [1]

- **Reference:**

  Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

- **Input data:** Joint coordinate vectors

- **Architecture:** 9 layers of spatial temporal graph convolution operators (ST-GCN units) with a softmax classifier at the end. The Resnet mechanism is applied on each ST-GCN unit.

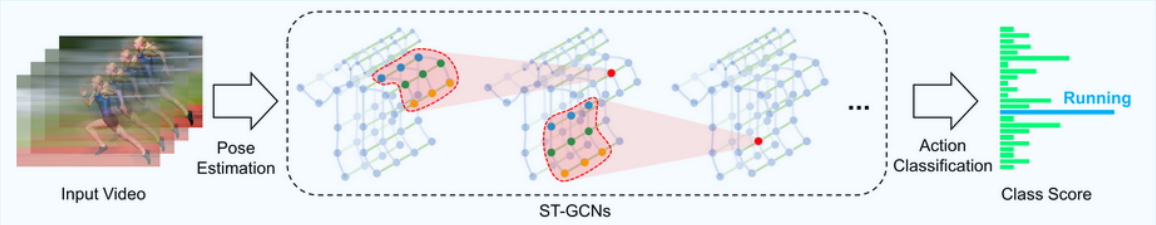

Figure 1: The spatial-temporal graph convolution network (ST-GCN) framework.

- **Arguments:**
  - Skeleton sequence can be represented by spatial-temporal graph, making graph neural networks great candidates.
  - The network has learnable edge importance weighting, which can recognize the joint importance to improve the recognition performance.
  - The authors provide open source with pre-trained model.

- **Source:** https://github.com/yysijie/st-gcn

2. Two-Stream Convolutional Network [2]

- **Reference:**

  Li, Chao, et al. "Skeleton-based action recognition with convolutional neural networks." 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017.

- **Input data:**

  - Skeleton sequence
  - Skeleton motion (temporal difference)

- **Architecture:** Two-stream CNN: 7 layers (3 convolutional and 4 fully-connected layers)

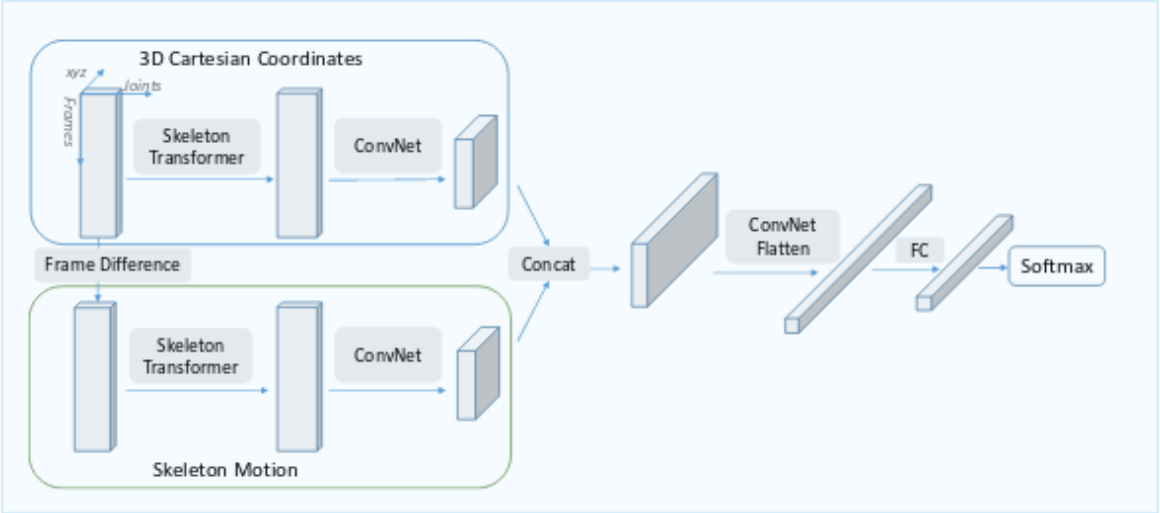

Figure 2: Two-stream architecture for action classification.

- **Arguments:**
  - Compared to the original two-stream network [3], which uses raw images and stacked optical flows as input, this network is fed with joint coordinates. With smaller dimension of input data, the computational cost is thus less.
  - The network is comparatively simple with fewer parameters to learn.
  - There are more studies adopted CNN to learn skeleton features and achieved impressive performance [4].
  - The authors improved their results by proposing a similar network architecture [4] that learns hierarchical co-occurrence features from skeleton sequences automatically.

- **Source:**
  - https://github.com/XiaoCode-er/Two-Stream-CNN
  - https://github.com/huguyuehuhu/HCN-pytorch

Data Preprocessing

To improve the performance of cow detection, 100 images of cows with different lighting conditions were used to fine-tune the TensorFlow object detection API. Faster R-CNN with Inception v2 was used as the model. The overall processing time is 3hrs 46min 8sec, resulting in 501 video clips.

Bibliography

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 597–600, IEEE, 2017.

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, pp. 568–576, 2014.

[4] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.