Review：Video Based Human Action Recognition Via Deep Learning Algorithm

Jianyang Xie, Main Supervisor: Yalin Zheng, Second Supervisor: Nguyen, Anh Industry Supervisor: Xiaoyun Yang(Remark AI)

Background

- Human action recognition (HAR) has a wide range of applications, e.g., intelligent video surveillance, robot vision, human-computer interaction, game control, and so on.
- Accurate action recognition in videos is still a challenging task because of the complexity of the visual data e.g., due to varying camera viewpoints, occlusions, and abrupt changes in lighting conditions.

Purpose

- Developing a robust machine learning-based model for the accurate and effective recognition of human action via surveillance video.
- Investigating a self-supervised architecture to alleviate data annotation dependencies, improving its generalization in industry application.
- Developing a lightweight model and making it easy to be integrated into industry applications.

Method：Literature Search

CNN-based methods

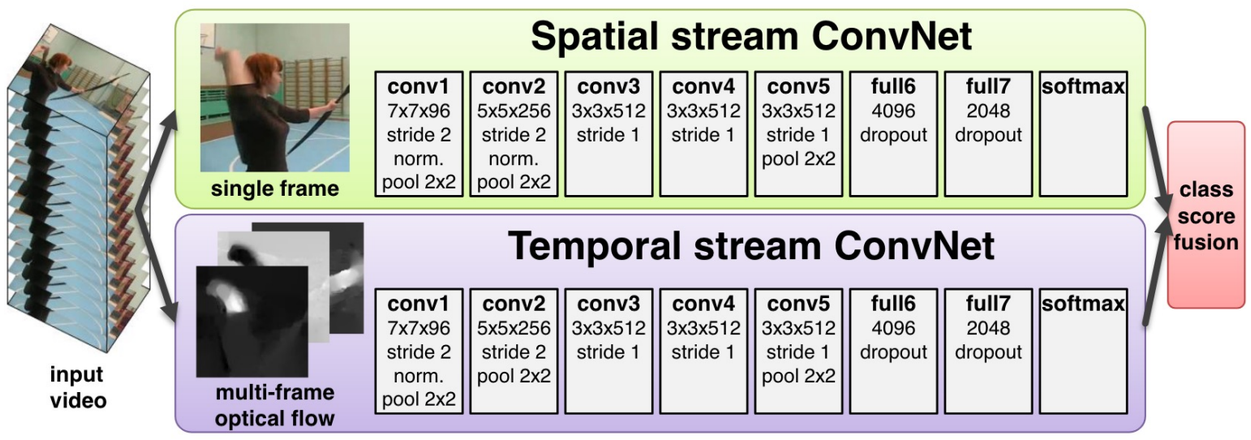

Figure.1 two stream architecture for video-based action recognition

Figure.1 shows a classical CNN-based human action recognition method, Spatial Stream ConvNet is utilized for appearance feature extraction, Temporal Stream ConvNet is used to process time dimension information.
2D and 3D CNNs were utilized to extract feature for action classification
There are three optimization directions were investigated:

- Multi-stream information such optical flow was combined as input for human action recognition
- Varieties of fusion methods were proposed to evaluate the interaction between spatial and temporal stream
- 3D CNN was utilized and optimized to extract spatial and temporal information at the same time

Skeleton-based methods

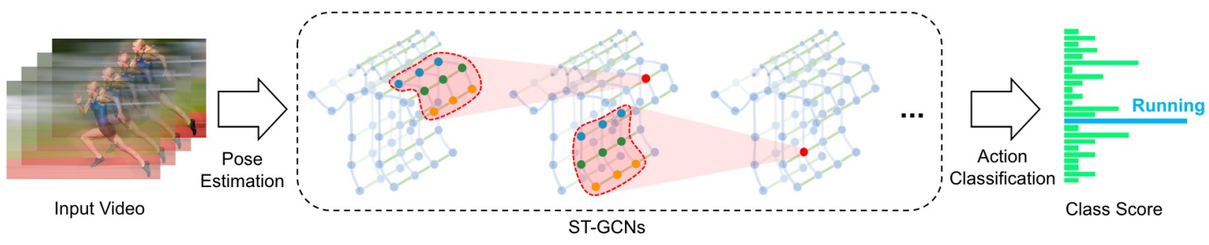

Figure.2 Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition

Figure.2 shows a basic framework of skeleton-based action recognition method, body joint was estimated by utilizing pose estimation methods, then the skeleton graph generated based on the body joint, finally, Graph Convolutional Networks were applied for classification. There are some optimized directions were investigated to overcome drawbacks:

- Self-Adaptive graph generation methods were proposed to realize a flexible representation of body joint skeleton
- Combined the CNN-based methods with skeleton information through utilizing multi-stream strategy

Conclusion

Although Deep learning-based human action recognition has achieved great success, there are some challenges::

- CNN-based methods are easily affected by circumstance changes and illumination variations in videos
- Skeleton-based method ignored the sematic information of graph node and edge, which limited the representation ability of the skeleton.
- Both types of methods heavily rely on a large amount of annotation, which may cause overfitting.
- Models are too heavy to be utilized in industry application

Project Progress

Problem: Skeleton-based methods consider all body joints as the same type and ignore the semantic information, which caused misclassification between similar actions. For example, the Skeleton-based method classified the phone call action into clapping action. Both of these two actions is the interaction o two part of our body.

Solution: Encoding the semantic information such as body joint type into a graph by generating a ***heterogeneous graph***. Compared with the ***homogeneous graphs*** which are widely utilized in existing skeleton-based methods, the heterogeneous graph considers node type as important property, and thus power for complex systems modelling. The pipeline of ***heterogeneous graph*** generation is shown in Figure 3.

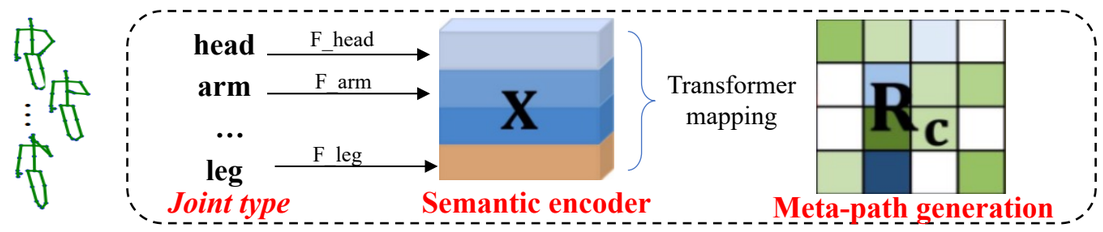

Figure.3 The pipeline of heterogeneous graph generation

Reference

[1] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition." Proceedings of the Neural Information Processing Systems (NIPS). 2015.
[2] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. “Spatial temporal graph convolutional networks for skeleton-based action recognition.” AAAI conference on artificial intelligence. 2018.

Engineering and Physical Sciences Research Council
UNIVERSITY OF LIVERPOOL
REMARK AI