**Investigation on Human Action Recognition**

**Miaoran Chen**

**What is action recognition:**

Identify human’s action based on a clip of a clip of video. The video is usually short and contains on main human action. Thus action recognition could be seen as a classification problem, whose input is video and output is the label of human action.

**Problems：**

1. variation between the same action expressed by different people.
2. environmental obstruction.
3. Human act in different speed. Thus it is hard to determine the start and end of an action.

**Common Dataset:**

1. [The HMDB-51 dataset(2011)](http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#Downloads)

6849 samples and 51 categories, size: 320\*240. Mostly come from movies.

1. UCF-101(2012)

13320 samples and 101 categories, size: 320\*240. All come from Youtube. This dataset contains 5 action types.

1. Sport-1M(2014)

110m samples and 487 categories. The labels of this dataset was generated by prediction, not accurate enough.

1. Kinetics(2017)

50m samples and 600 categories. Manual annotation was used to rectify this dataset repeatedly. Kinetics contains the widest range of human action among these dataset.

**Common algorithm:**

**optic flow:**

Main idea: Estimating an object’s motion based on video frames. Calculating relative pixels’ motion in video frames.

Assumed condition:

1. Brightness constancy: the brightness of a point would not change in the video from stem to stern. This is the basic requirement of the input videos.
2. Objects’ motion are not drastic: the position of a point would not change drastically.

Common optic flow algorithm type:

1. Gradient-based

Using the differential of image grayscale value to calculate objects’ speed. Horn-Schunck algorithm and LK algorithm are representative ones.

1. Matching-based
   1. Locating and tracking the main part of an object through the video. This method applies to those objects with big size.
   2. Locating the similar areas in different frames and calculating the displacement among them.
2. Energy-based

Based on frequency actually. Using spatiotemporal filtering to process image. This method is calculating-expansive.

1. Phase-based

Calculating an object’s speed based on the shift of phase.

1. neural network-based

Integrating human’s visual mechanism and deep network.

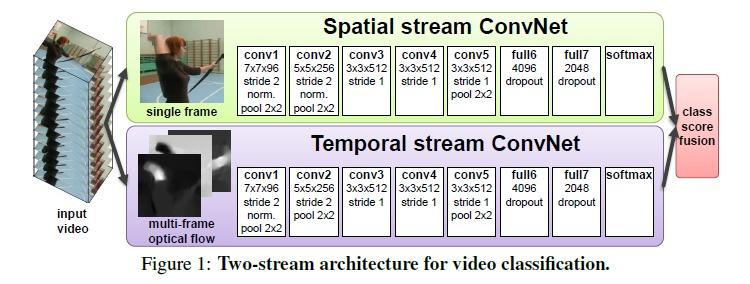
**iDT: improved dense trajectories(2013):**

Main idea: Extracting the trajectories in a video sequence and extracting 4 features: HOF, HOG, MBH, trajectory based on the trajectories. HOF used to calculate grayscale and others used to calculate dense optic flow. This features are encoded with the used of the Fisher Vector. The results are generated from SVM.

This algorithm performed best in the field of action recognition before neural network was employed. Although deep learning-based methods have outperformed iDT, **the integration of iDT and neural networks** can always gain promotion more or less.

**Two-stream:**

Main idea: Training two CNN and then fusing them. One based on RGB image and the other based on optic flow image. The RGB image is generated randomly from the video. The optic flow image is generated by fusing several optic flow of continuous frames.



TSN:

An advanced two-stream. Since two-stream could not work well on those actions with precedence order. Researchers advanced it by cutting the input video into 3 parts. Employing two-stream on each video respectively.

**C3D:**

Main idea: Adding temporal information to traditional 2D convolutional network to Integrate the video information in a short time series.

**Primary research directions:**

**Structure:**

Mainstream networks are mostly generated from two-stream and C3D.

TSN: [Temporal Segment Networks: Towards Good Practices for Deep Action Recognition](https://arxiv.org/abs/1608.00859" \t "_blank)

TSN + different weight of video parts: [Deep Local Video Feature for Action Recognition](http://openaccess.thecvf.com/content_cvpr_2017_workshops/w14/papers/Lan_Deep_Local_Video_CVPR_2017_paper.pdf" \t "_blank)

Two-Stream I3D : [Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset](https://arxiv.org/abs/1705.07750)

T3D: [Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification](https://arxiv.org/abs/1711.08200)

[Temporal Pyramid Pooling](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1711.04161)：[End-to-end Video-level Representation Learning for Action Recognition](https://arxiv.org/abs/1711.04161)

**Connection:**

Contains 2 types: connection within one network between 2 different layers, connection between 2 networks(mainly in two-stream network).

[Spatiotemporal Multiplier Networks for Video Action Recognition](https://link.zhihu.com/?target=http%3A//openaccess.thecvf.com/content_cvpr_2017/papers/Feichtenhofer_Spatiotemporal_Multiplier_Networks_CVPR_2017_paper.pdf)

[Spatiotemporal Pyramid Network for Video Action Recognition](https://link.zhihu.com/?target=http%3A//openaccess.thecvf.com/content_cvpr_2017/papers/Wang_Spatiotemporal_Pyramid_Network_CVPR_2017_paper.pdf)

[Attentional pooling for action recognition](https://link.zhihu.com/?target=https%3A//papers.nips.cc/paper/6609-attentional-pooling-for-action-recognition.pdf)

[ActionVLAD for video action classification](https://link.zhihu.com/?target=http%3A//openaccess.thecvf.com/content_cvpr_2017/papers/Girdhar_ActionVLAD_Learning_Spatio-Temporal_CVPR_2017_paper.pdf)

[Deep Convolutional Neural Networks with Merge-and-Run Mappings](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1611.07718)

**My review:**

Optic flow is an indispensable part of action recognition algorithms. Most of the top performing action recognition methods use optical flow as a "black box" input(*On the Integration of Optical Flow and Action Recognition 2017*). Because optic flow is invariant to appearance. How to encode motion effectively is the most important problem in the field of optic flow right now. So how to integrate optic flow to neural network and how to make it encode more effectively may be a meaningful research direction. Furthermore, the different impact proportion of video frames in C3D may works. Since different video frames contribute different to the result of classification.

Among the 3 main research directions mentioned above, each of them has problems remained to solve. For structure, I think the major problem is how to extract spatial-temporal properties as much as possible from the input video. For connection, the major problem is how to distribute weights to different video parts properly and how to interrelate temporal information and spatial information effectively. More information about the field of action recognition needs more reading.