Action Recognition in Videos Using Convolutional and Spatial-Temporal Interest Point Features

T. Tharmini and A. Ramanan
Department of Computer Science, Faculty of Science, University of Jaffna
tharmini7@gmail.com

Introduction

Recently human action recognition is an emerging topic of research in the field of computer. The factors such as occlusion of objects, camera parameters, scene clutter, illumination, body posture size and gender are increasing the complexity of action recognition. Recent trend in computer vision is the usage of convolutional neural network (CNN) which has been successful in image analysis like object recognition. On the other hand, action recognition using handcrafted features showed that space time interest point (STIP) performs better when compared to other local features. The proposed method in this study improves capturing spatial-temporal variation in human actions from videos using the STIP and convolutional features.

Objective

To improve the overall performance of action recognition task by using Convolution and STIP features.

Methodology

- Given a video, *fc7* features are computed for each video frame which is then mapped into a short binary code space using Iterative Quantization (ITQ) [4] , which is a method based on local sensitive hashing method (LSH).
- Key-frames are selected across the time space by picking the frames that their binary codes are different from their previous frames in a given video.
- A subset of key-frames (i.e., a snippet) is constructed using a fixed-sized window which is applied to the initial set of key-frames by striding the window with a constant factor. CNN flow is computed from the difference between the last key-frame and the first key-frame of a snippet.
- A vector representation is computed by stacking the CNN flows of each snippet.
- In the final step, all the videos are represented as Bag-of-Features (BoF) of temporal words.
- On the other hand, space-time interest points are searched in the video frames and feature descriptors are computed. These descriptors were also represented as BoF.
- Action-specific codebooks were constructed using K-means algorithm for BoF representation. The concatenated feature vector was then classified by a linear one-versus-all SVMs.

Methodology…

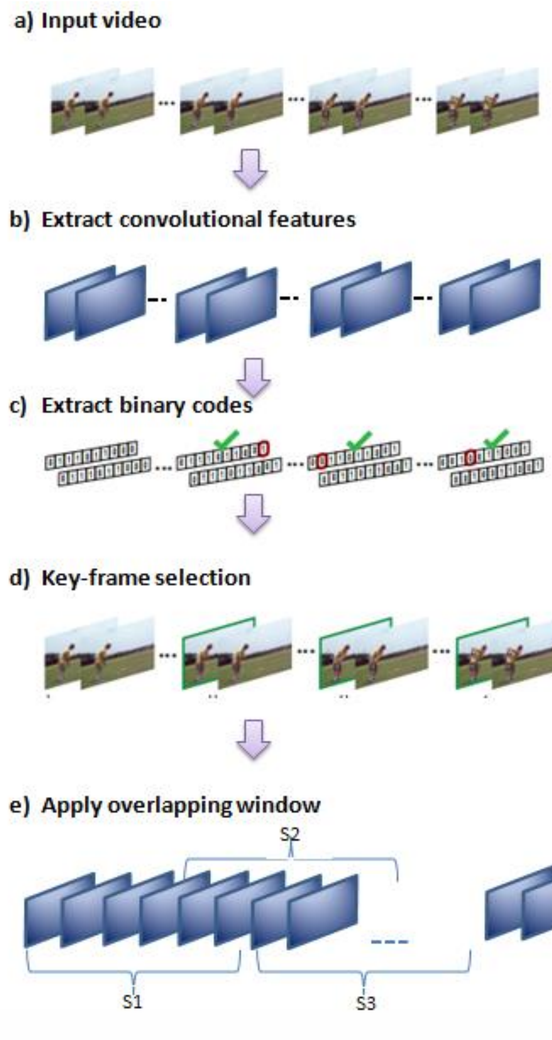

Figure 1: Snippet selection by selecting Key frames using LSH

Experimental setup

The proposed method was tested on KTH dataset [2]

- Testing included 9 subjects (2,3,5,6,7,10,22) and training set included rest of subject (16 subjects)
- Features: For CNN *fc7* features of VGG-F and for STIP HOG and HOF descriptors
- Codebook Construction: K-means algorithm
- Classifiers: Linear OVA-SVMs

Testing Result

Table 1: Comparison between action recognition rates of 24-bits key-frame selection with overlapping window size 8 (stride 4) with different K of K-means using convolutional features in classification

.

K=100	K=150	K=200
83.79%	**85.69%**	84.72%

Table 2: Comparison of STIP features in classification using action-specific and global codebook with different K of K-means

Action-specific Codebook		Global Codebook	
K=500	K=1000	K=500	K=1000
93.07%	**94.44%**	90.13%	91.67%

Finally, we combined STIP and CNN flow features for action classification by following the best parameter settings and type of codebook obtained in Table 1 and Table 2 as indicated in bold.
The combined feature set of STIP and CNN flow yields a classification rate of **94.91%** which is slightly better than the usage of an independent feature set.

Discussion

- Experiment reveals that how to recognize a video with feature representation. The overall performance of action classification has been improved when STIP and convolutional features are used together.
- This study evaluated the performance in two steps:

  (i) Finding key-frames and applying overlapping windows to extract CNN flow features that captures sub-actions in a video sequence
  (ii) Detecting local structures in space-time using STIP features.

- Each of the features was represented as bag-of-features.

Conclusion

Based on our experimental results, we conclude that the combined set of convolutional and STIP features perform better in action classification using the KTH dataset.

References

1. B. Banerjee and V. Murino, "Efficient pooling of image based CNN features for action recognition in videos," in Proceedings of IEEE International Conference on Accoust., Speech Signal Process. (ICASSP), pp. 2637–2641, Mar. 2017.
2. C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach in Pattern Recognition," in Proceedings of the 17th International Conference on Pattern Recognition, Aug 2004.
3. I. Laptev and T. Lindeberg, "Space-time interest points," in Proceeding on International Conference on Computer Vision, pp. 432–439, 2003.
4. Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in Proceedings of the International conference on CVPR, 2011.