Action and Gesture Temporal Spotting with Super Vector Representation

Team member: Limin Wang[1,3], Xiaojiang Peng[2,3], Zhuowei Cai[3], Yu Qiao[3]

[1]The Chinese University of Hong Kong, Hong Kong
[2]Southwest Jiaotong University, Chengdu, China
[3]Shenzhen Institutes of Advanced Technology, CAS, China

September 6, 2014

Outline

1. Introduction

2. Method

3. Results

4. Conclusion

Outline

1 Introduction

2 Method

3 Results

4 Conclusion

Background

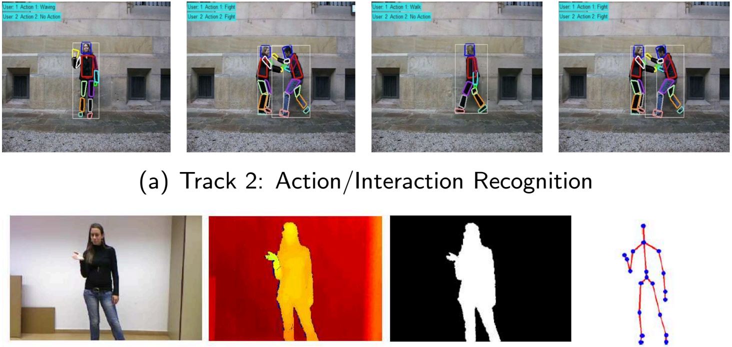
(a) HMDB51

(b) UCF 50 and UCF101

- Action recognition from a short video clip is widely studied in computer vision research (e.g. dataset of HMDB51 and UCF101).
- Action temporal spotting from a continuous video stream is a more difficult problem and less studied.
- It needs to recognize and temporally localize the ongoing action in the long video sequence simultaneously.

Challenging Tasks

Our team takes part in both track 2 and track 3 at Looking at People challenge:



(a) Track 2: Action/Interaction Recognition

(b) Track 3: Gesture Recognition

Figure: For track 2, we show several action instances such as waving and fight. For track 3, we show the different modalities (RGB, depth, mask, skeleton).

Outline

1 Introduction

2 Method

3 Results

4 Conclusion

Overview of Method

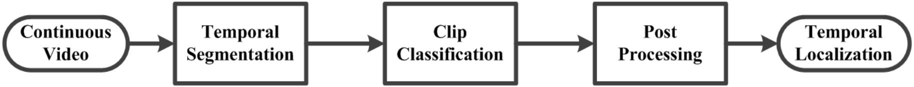


Figure: Pipeline of our method.

Our method for both tasks can be divided into three steps:

- (i) temporal segmentation (ii) clip classification (iii) post processing
- The methods for both tasks only differ in the step of temporal segmentation.

Temporal Segmentation for Track 2

- We resort to a temporal sliding window method to firstly divide continuous video stream into short clips.
- According to the observation on training dataset, we set the window duration as 15-frames and the scanning step as 5-frames.
- We firstly extract the low-level features for the whole video stream.
- We then design a temporal integration histogram to efficiently calculate the feature histogram for the sub-window of any location and any duration.

Temporal Segmentation for Track 3

- Unlike Track 2, Track 3 provides other data modalities such as skeleton, depth, and mask, which can provide extra information for temporal segmentation.
- We observe that the hands of an actor are almost at the same position when he is not performing the gesture.
- With this assumption, we propose a temporal segmentation algorithm based on the analysis of human hand trajectory.

Temporal Segmentation for Track 3

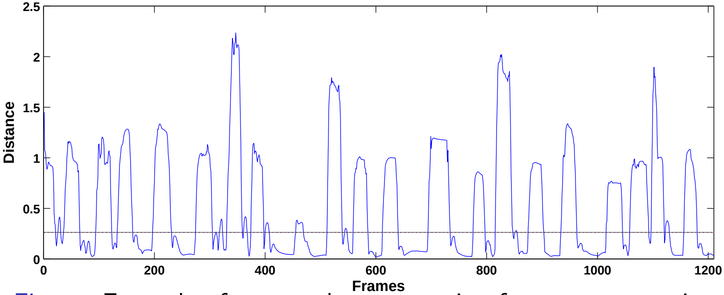


Figure: Example of temporal segmentation for gesture spotting.

- We use the histogram with 100 bins to estimate the spatial distribution of hand position in the whole video.
- We choose the center of bin with highest frequency as the hand position when the actor is static.
- With the distance of current hand position to its static one, we use a single threshold τ to determine whether actor is performing a gesture.

Improved Dense Trajectories

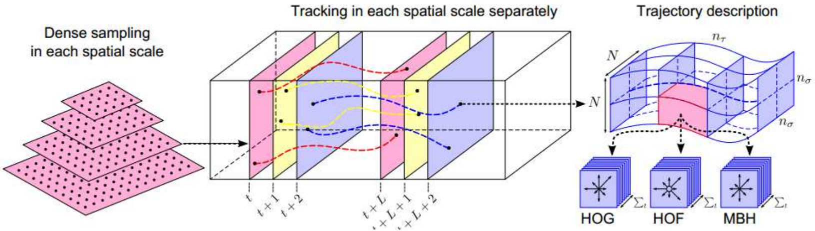


Figure: Illustration of dense trajectory extraction.

We extract dense trajectory features and use four kinds of descriptors: HOG, HOF, MBHx, and MBHy.

Heng Wang and Cordelia Schmid, *Action Recognition with Improved Trajectories*. in ICCV, 2013.

Some Examples

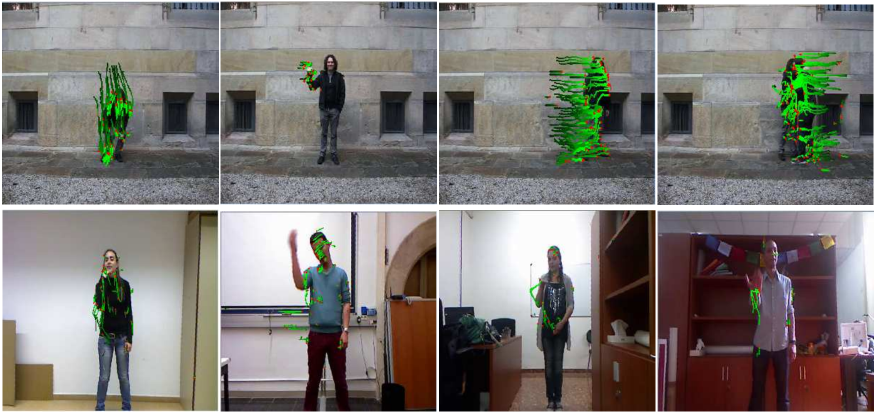
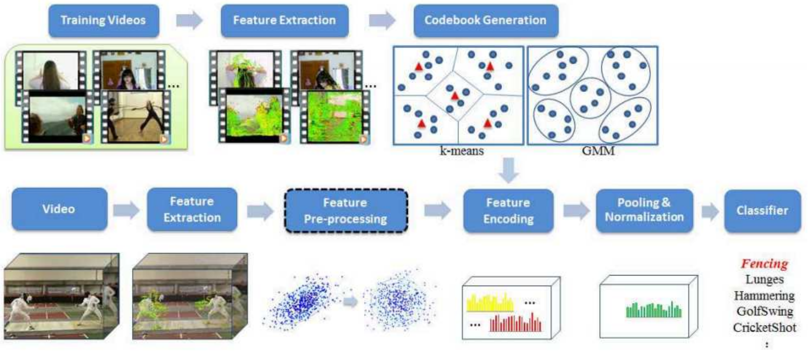


Figure: Examples of extracted dense trajectories for videos from both Track 2 of action recognition and Track 3 of gesture recognition.

Bags of Visual Words



- Bag of Visual Word (BoVW) has many choices for each step.
- Super vector encoding obtains higher performance.

Xiaojiang Peng, Limin Wang, Xingxing Wang, Yu Qiao, *Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice*. in CoRR abs/1405.4506, 2014.

Fisher Vector

- Given a set of transformed descriptors: $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$, we learn a generative GMM: $p(\mathbf{x}; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$
- where K is mixture number, and $\theta = \{\pi_1, \mu_1, \Sigma_1, \cdots, \pi_K, \mu_K, \Sigma_K\}$ are model parameters.
- Then Fisher vector for a clip is derived as follows:

$$\mathcal{G}_{\mu_k}^{\mathbf{X}} = \frac{1}{\sqrt{\pi_k}} \sum_{i=1}^{N} \gamma_k(\mathbf{x}_i) \left(\frac{\mathbf{x}_i - \mu_k}{\sigma_k} \right),$$

$$\mathcal{G}_{\sigma_k}^{\mathbf{X}} = \frac{1}{\sqrt{2\pi_k}} \sum_{i=1}^{N} \gamma_k(\mathbf{x}_i) \left[\frac{(\mathbf{x}_i - \mu_k)^2}{\sigma_k^2} - 1 \right],$$

- In our current implementation, we choose power ℓ_2-normalization ($\alpha = 0.5$) to obtain final representation $\mathcal{S}$:

$$\mathcal{S} = [\mathcal{G}_{\mu_1}^{\mathbf{X}}, \mathcal{G}_{\sigma_1}^{\mathbf{X}}, \cdots, \mathcal{G}_{\mu_K}^{\mathbf{X}}, \mathcal{G}_{\sigma_K}^{\mathbf{X}}], \quad \mathcal{S} = \frac{\mathrm{sign}(\mathcal{S})\sqrt{|\mathcal{S}|}}{\|\sqrt{\mathcal{S}}\|_2}.$$

Classifier

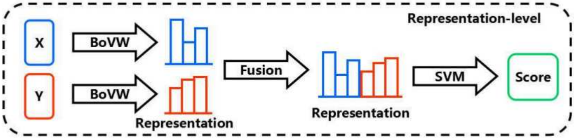


Figure: Feature fusion and classifier training.

- We separately construct Fisher vector representation for each kind of descriptor.
- We use the representation-level fusion method to combine these different descriptors.
- We adopt one-vs-all training scheme and train a linear SVM for each action class.

Post Processing

- In order to avoid the false detections, we design a post processing step.
- During training phase, we mine some instances of static background or noisy motions, and then train a classifier corresponding to background class.
- If a video clip is predicted as background class, we will remove this detection.
- We also use a single threshold (-0.8) to eliminate those detections with low confidence score.

Outline

1. Introduction

2. Method

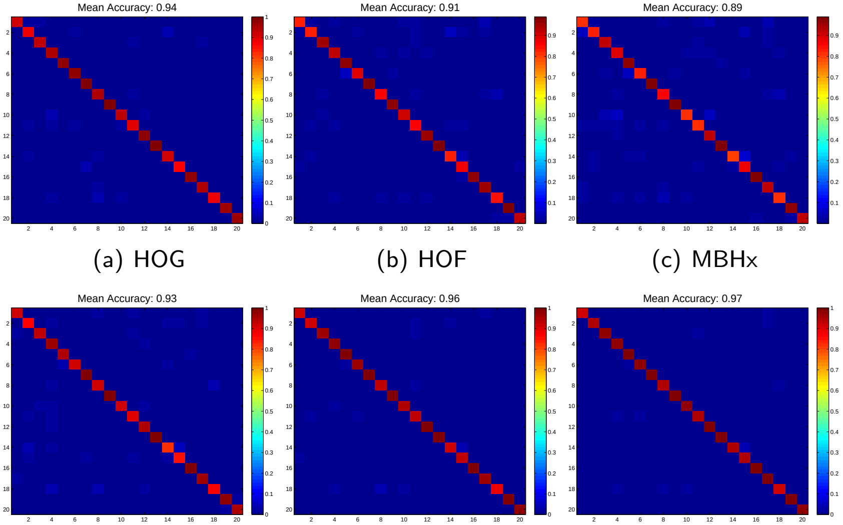
3. Results

4. Conclusion

Experimental Setup

- For both tracks, the descriptors of HOG, HOF, MBHx, and MBHy are firstly dimension reduced with PCA by a ratio 0.5 and then whitened to have unit variance in each dimension.
- We set the number of GMM mixture as 256.
- For track 2, there are 11 action classes, 5 video sequences with 135 action instances for training, 2 videos with 44 action instances and 2 videos with 52 instances for validation and test, respectively.
- For track 3, there are 20 gesture categories of Italian signs. There are 470 sequences with 6,830 gesture instances for training, 230 sequences with 3,454 instances for validation, and 240 sequences for testing.

Gesture Recognition



(a) HOG (b) HOF (c) MBHx

(d) MBHy (e) HOG + HOF (f) All

Action Temporal Spotting

Table: Contest results for track 2 of action recognition.

Rank	Team Name	Score
1	**Our Team**	0.507173
2	ADSC	0.501164
3	SBUVIS	0.441405
4	DonkeyBurger	0.342192
5	UC-T2	0.121565
6	MindLAB	0.008383

Gesture Temporal Spotting

Table: Contest results for track 3 of gesture recognition.

Rank	Team Name	Modalities	Score
1	liris	Skeleton, depth, RGB	0.84998
2	CraSPN	Skeleton, depth, RGB	0.83390
3	JY	Skeleton, RGB	0.82680
4	**Our Team**	RGB	0.79193
5	Lionel Pigou	Depth RGB	0.78880
6	stevenwudi	Skeleton, depth	0.78731

Outline

1 Introduction

2 Method

3 Results

4 Conclusion

Conclusions

- We have presented a simple yet effective method for temporal spotting of action and gesture from continuous video stream.
- Using this method, we obtain competitive results for both track 2 of action recognition and track 3 of gesture recognition.
- From recognition results, we observe that Fisher vector of dense trajectories is very effective for describing visual content and obtains high recognition performance (0.97).
- To further improve the spotting accuracy, we need to consider designing more effective segmentation algorithm and using more modalities such as depth and skeleton.

Thank you!

Email:lmwang.nju@gmail.com

Welcome to our ECCV poster presentation:

- Video Action Detection with Relational Dynamic-Poselets (Session 3B).
- Action Recognition with Stacked Fisher Vectors (Session 3B).
- Boosting VLAD with Supervised Dictionary Learning and High-Order Statistics (Session 2B).