Action and Gesture Temporal Spotting with Super Vector Representation

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

[1]Southwest Jiaotong University, Chengdu, China
[2]Department of Information Engineering, The Chinese University of Hong Kong
[3]Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, CAS

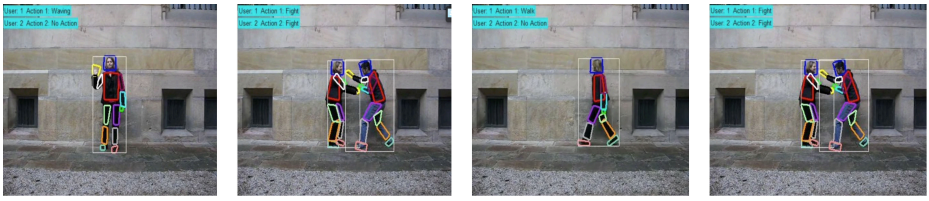
Abstract. This paper focuses on describing our method designed for both track 2 and track 3 at Looking at People (LAP) challenging [1]. We propose an action and gesture spotting system, which is mainly composed of three steps: (i) temporal segmentation, (ii) clip classification, and (iii) post processing. For track 2, we resort to a simple sliding window method to divide each video sequence into clips, while for track 3, we design a segmentation method based on the motion analysis of human hands. Then, for each clip, we choose a kind of super vector representation with dense features. Based on this representation, we train a linear SVM to conduct action and gesture recognition. Finally, we use some post processing techniques to void the detection of false positives. We demonstrate the effectiveness of our proposed method by participating the contests of both track 2 and track 3. We obtain the best performance on track 2 and rank 4^{th} on track 3, which indicates that the designed system is effective for action and gesture recognition.

Keywords: Action recognition, gesture recognition, temporal spotting, super vector

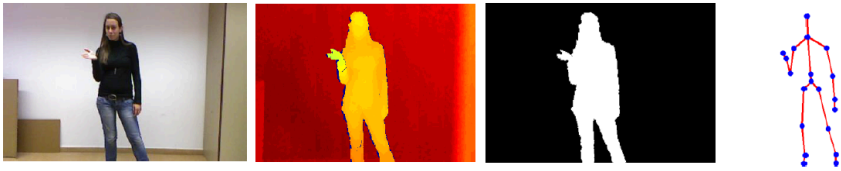
1 Introduction

Action and gesture recognition [2, 3] for a short video clip has become an important area in computer vision, whose aim is to classify the ongoing action or gesture into a predefined category. It has wide applications in our daily life such as human computer interaction, content based video retrieval, and sports video analysis. However, most of the existing research works focus on the action and gesture dataset, where the videos have been manually trimmed to bound the action interest, such as HMDB51 [4] and UCF101 [5]. These datasets have a limitation in measuring the effectiveness of proposed method in practical settings. Instead, in this paper, we try to address a more difficult problem, namely *temporal spotting* of action and gesture. We are given a continuous video stream and we need to recognize and temporally localize the ongoing action in the video sequence simultaneously.

We mainly describe our method designed for Track 2 and Track 3 of Looking at People (LAP) challenge [1] organized by ChaLearn in conjunction with the



(a) Track 2: Action/Interaction Recognition



(b) Track 3: Gesture Recognition

Fig. 1. Some examples of track 2 and track 3 of Looking at People Challenging. For track 2, we show several action instances such as waving, walk, and fight. For track 3, we show the different modalities provided for gesture recognition. From left to right: RGB information, depth information, user segmentation mask, and skeleton. Note that both figures are from the contest webpage.

ECCV 2014 conference. Track 2 focuses on action (interaction) recognition from RGB data, with 11 action classes, such as wave, point, clap, and couch [6]. Track 3 is about the gesture recognition from multi-modal data, including: color, depth, skeleton, and user mask [7]. This competition task aims to learn a vocabulary of gestures corresponding to 20 Italian cultural signs, such as vattene, vieniqui, perfettor, and ok. Some video examples of both tracks are shown in Figure 1. It is worth noting that, in test phase of both tracks, we are given a temporally untrimmed video which may contain multiple action and gesture instances. Our method need to temporally localize and recognize these ongoing instances.

As shown in Figure 2, our method is mainly composed of three steps: (i) temporal segmentation, (ii) clip classification, and (iii) post processing. Firstly, we temporally divide the untrimmed video sequence into several clips. We resort to a simple sliding window scheme for track 2 of action recognition, while we design a temporal segmentation method based on hand motion for track 3 of gesture recognition. Then, for each short video clip, we extract dense trajectories with four kinds of descriptors: HOG, HOF, MBHx, and MBHy [8]. We choose the Fisher Vector [9] as encoding method to obtain the global representation for each video clip and train a linear SVM for classification. Finally, we use some post-processing techniques to eliminate the false positive detections. For example, during our training phase, we train a classifier for the background class. This background classifier will enable us to eliminate some detections corresponding to the background class.

We will provide a detailed description about our method for both Track 2 and Track 3 in Section 2. Then, we will report the performance of our method on both Tracks in Section 3. Finally, we conclude our paper in Section 4.

2 Method

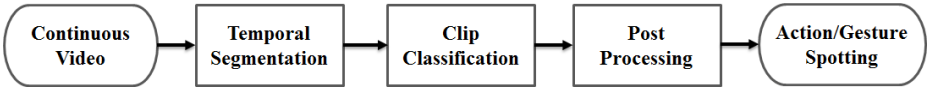


Fig. 2. The pipeline of our method for action and gesture temporal spotting. We firstly conduct video temporal segmentation. Then for each video clip, we extract super vector representation and perform clip classification. Finally, we use post processing techniques to eliminate the false detections.

We propose an temporal spotting system for the Track 2 and Track 3 of Looking at People (LAP) challenge as shown in Figure 2. Our system is mainly composed of three steps: (i) temporal segmentation, (ii) clip classification, and (iii) post processing. We will give a detailed description of these steps in the remainder of this section.

2.1 Temporal Segmentation

In this section, we mainly describe the method of dividing a continuous video into short clips, each of which may contain the action and gesture of interest. We design two different methods for track 2 of action recognition and track 3 of gesture recognition respectively.

Action Recognition. For the track 2 of action (interaction) recognition, it gives a continuous video stream, which contains a sequence of action instances. We need to firstly localize these instances and then recognize their corresponding action classes. We resort to a temporal sliding scheme to conduct action localization. Based on the observation on training dataset, we set the window length as 15-frames and scanning step as 5-frames. To speed up the sliding window process, we firstly extract the low-level features and encode these descriptors as described in the next section. Then we design a temporal integration histogram, with which we can efficiently calculate the feature histogram for the sub-window of any location and any length.

Gesture Recognition. Unlike Track 2, the track 3 of gesture recognition provides several data modalities such as RGB information, depth information, user mask, and skeleton information. We observe that the motion trajectory of human hand is an important cue for gesture spotting in a continuous video stream. With a reasonable assumption that the hands of an actor are almost

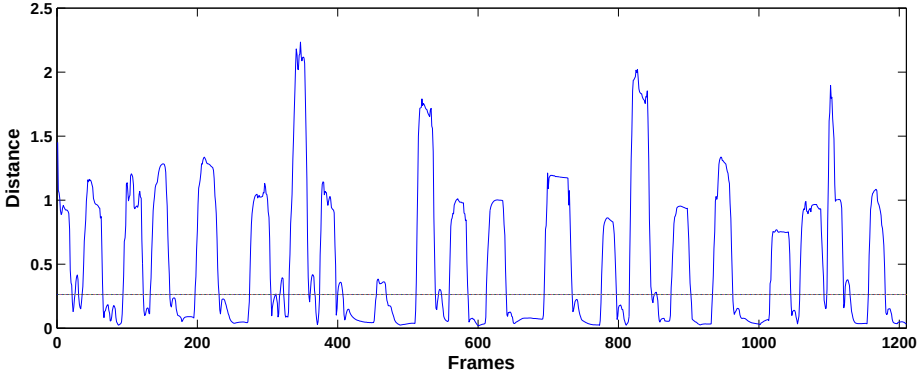


Fig. 3. A example of temporal segmentation for gesture recognition based on hand motion analysis.

in the same position when he is not performing a gesture, we propose temporal segmentation method based the analysis of hand trajectory.

We firstly estimate the position of actor hands when he is not performing a gesture. With the given pose information, we use a 2D histogram of 100×100 cells overlapped with a pixel grid to estimate the spatial distribution of hand position in the whole video. Then we choose the cell with highest frequency as the static hand position. Based on the static hand position, for each frame, we calculate the distance of current hand position to the static position. Finally, according to this distance, we use a single threshold τ to determine whether the actor is performing a gesture or not. An example of temporal segmentation is shown in Figure 3. With this simple yet effective method, we split a untrimmed video into several clips, each of which can be used for gesture recognition in the next section.

2.2 Clip Classification

In this section, we give a detailed description of video clip representation and classification. Note that, for both track 2 and track 3, we resort to the same video representation, and we just use the RGB modality for gesture recognition of track 3.

The key element of clip classification is the visual representation of video data. Following the success of Dense Trajectories (IDTs) [1] in action recognition on the wild videos [8], we choose the IDTs as our low-level features of RGB data. Specifically, we use the public code released on the project page. We set the length of trajectory as 9 and extract four kinds of descriptors, namely HOG, HOF, MBHx, and MBHy. Some examples of extracted dense trajectories are shown in Figure 4. From these examples, we observe that these extracted dense

[1] https://lear.inrialpes.fr/people/wang/dense_trajectories

Fig. 4. Examples of extracted dense trajectories for videos from both Track 2 of action recognition and Track 3 of gesture recognition.

trajectories focus on the foreground regions with high motion saliency. The four kinds of descriptors correspond to different views of video data such as static appearance, dynamic motion, and motion boundary. These different aspects are of importance for action and gesture recognition.

With these low-level descriptors, we adopt the Bag of Visual Words [10] model to obtain the global representation. According to the recent study works [11, 12], super vector based encoding methods are very effective by aggregating different order statistics in a high-dimensional feature representation. Specifically, we choose Fisher Vector [9] as the encoding method using the implementation of vlfeat [13]. Given a set of local descriptors $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_N] \in \mathbb{R}^{M \times N}$, we first use the PCA and Whiten technique to remove the correlations among different dimensions and normalize the variance. Then, based on the transformed feature descriptors $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ where $\mathbf{x}_i = \Lambda \mathbf{f}_i$ and $\Lambda \in \mathbb{R}^{D \times M}$ is the transform matrix of PCA and Whiten processing, we learn a generative Gaussian Mixture Model (GMM):

$$p(\mathbf{x}; \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k), \tag{1}$$

where K is mixture number, and $\theta = \{\pi_1, \mu_1, \Sigma_1, \cdots, \pi_K, \mu_K, \Sigma_K\}$ are model parameters. $\mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$ is D-dimensional Gaussian distribution.

With some reasonable assumptions that the posterior probability is sharply peaked on a single value of k for any descriptor $\mathbf{x}$, the Fisher Information Matrix (FIM) is a diagonal matrix [9]. Then Fisher vector is derived from Fisher Kernel [14] as follows:

$$\mathcal{G}_{\pi_k}^{\mathbf{X}} = \frac{1}{\sqrt{\pi_k}} \sum_{i=1}^{N} (\gamma_k(\mathbf{x}_i) - \pi_k), \tag{2}$$

$$\mathcal{G}_{\mu_k}^{\mathbf{X}} = \frac{1}{\sqrt{\pi_k}} \sum_{i=1}^{N} \gamma_k(\mathbf{x}_i) \left(\frac{\mathbf{x}_i - \mu_k}{\sigma_k} \right), \tag{3}$$

$$\mathcal{G}_{\sigma_k}^{\mathbf{X}} = \frac{1}{\sqrt{2\pi_k}} \sum_{i=1}^{N} \gamma_k(\mathbf{x}_i) \left[\frac{(\mathbf{x}_i - \mu_k)^2}{\sigma_k^2} - 1 \right], \tag{4}$$

where $\gamma_k(\mathbf{x})$ is the posteriori probability of local descriptor $\mathbf{x}$ assigned to k^{th} Gaussian Mixture:

$$\gamma_k(\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\mathbf{x}; \mu_i, \Sigma_i)}. \tag{5}$$

In our current implementation, we choose the first order and second order super vector with power ℓ_2-normalization ($\alpha = 0.5$) as our super vector representation $\mathcal{S}$:

$$\mathcal{S} = [\mathcal{G}_{\mu_1}^{\mathbf{X}}, \mathcal{G}_{\sigma_1}^{\mathbf{X}}, \cdots, \mathcal{G}_{\mu_K}^{\mathbf{X}}, \mathcal{G}_{\sigma_K}^{\mathbf{X}}], \quad \mathcal{S} = \frac{\text{sign}(\mathcal{S})\sqrt{|\mathcal{S}|}}{\|\sqrt{\mathcal{S}}\|_2}. \tag{6}$$

It is worth noting that we separately construct super vector representation $\mathcal{S}_i$ for the i^{th} kind of descriptor according to the above description. We then concatenate the four kind of super vector representation as a whole one $\mathcal{S} = [\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4]$. Using this concatenated representation, we train a linear SVM for each action and gesture class using the implementation of LIBSVM [15]. For multiclass classification, we use the one-vs-all training scheme and choose the prediction with the highest score as its predicted label.

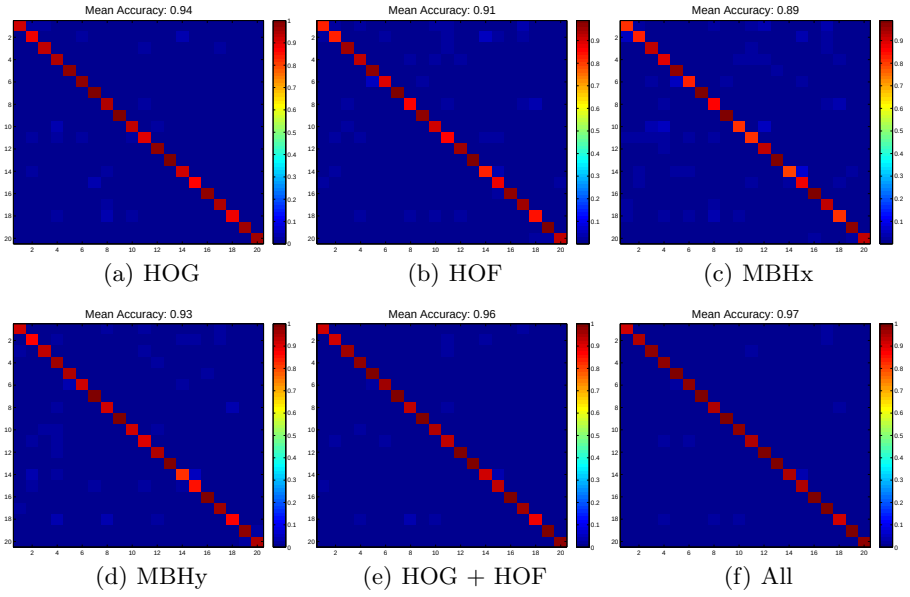
2.3 Post Processing

In order to avoid the false detections, which may correspond to the irrelevant background action classes, we design a post processing technique. Firstly, during training phase of action and gesture recognition, we mine some instances of static background or noisy motion. We then use these instances to train a classifier which represents the background class. During test phase, if a video sub-window or clip is predicted as the background class, we will remove this detection. Secondly, we use a single threshold -0.8 to eliminate those detections with low confidence score. In our evaluation, we find this post processing step is very effective for removing those false positive detections and improving the performance of action and gesture spotting.

3 Evaluation

In this section, we present the experimental results for track 2 of action recognition and track 3 of gesture recognition at Looking at People (LAP) challenge [1]. For both tracks, we set the number of GMM mixture as 256.

Dataset and Evaluation Measurement. For track 2 of action recognition, it has 11 action categories such as wave, point, clap and so on. For training data, it has 5 video sequences, containing 135 action instances in total. For validation



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

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

Fig. 5. Results of different descriptors (HOG, HOF, MBHx, and MBHy) and their combinations (HOG+HOF, All) on the gesture recognition.

data, there are 2 video sequences including 44 action instances. There are 2 video sequences with 52 action instances for final evaluation in test phase. We firstly train our model on the training dataset and adjust the parameters according to its performance on the validation dataset. Finally, in order to increase the training data, we retrain our model on both the training and validation dataset, and verify its performance on the testing dataset.

For track 3 of gesture recognition, there are totally 20 gesture categories of Italian signs. For training dataset, there are 470 video sequences, corresponding to $6,830$ gesture instances. For validation dataset, it has 230 video sequences and $3,454$ gesture instances. For final evaluation, there are 240 video sequences for testing.

Regarding the evaluation measurement, for both track 2 and track 3, it uses the **Jaccard Index** to evaluate the performance of action and gesture spotting. Specifically, the Jaccard Index is defined as follows:

$$J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}}, \tag{7}$$

where $A_{s,n}$ is the ground truth of action (or gesture) n at sequence s, and $B_{s,n}$ is the prediction for such action (or gesture) at sequence s. For final evaluation, our method is evaluated based on the mean Jaccard Index among all action (or gesture) classes. For track 2, the number of action classes is 12, and there are 20 gesture classes for track 3.

Gesture Recognition. As the clip representation using iDTs with Fisher Vector are adapted from the action recognition tasks [12], we firstly conduct an exploration experiment to verify its performance on gesture recognition. Specifically, we use the $6,830$ gesture instances from the training dataset for SVM training. We demonstrate the effectiveness of this representation on the $3,454$ gesture instances from the validation dataset. The experimental results are shown in Figure 5 and we observe that this representation is very effective for capturing the visual information of gesture. Combining all the descriptors, we obtain a mean accuracy of 0.97 on the validation dataset.

Contest Result. We report the action and gesture spotting performance for both track 2 and track 3 of Looking at People (LAP) challenge, and the results are shown in Table 1 and Table 2. Our spotting system obtains the best performance for track 2 and ranks the 4^{th} for track 3. It is worth noting that our system only use the modality of RGB for gesture recognition, while other top performers use other modalities such as skeleton and depth. The top performance on both tracks demonstrate that our designed temporal spotting system is very effective for action and gesture recognition.

Table 1. Contest results for track 2 of action recognition.

Rank	Team	Score
1	**Ours**	0.507173
2	Pei et al. [16]	0.501164
3	Shu [17]	0.441405

Table 2. Contest results for track 3 of gesture recognition.

Rank	Team	Modalities	Score
1	Neverova et al. [18]	Skeleton, depth, RGB	0.849987
2	Monnier et al. [19]	Skeleton, depth, RGB	0.833904
3	Chang [20]	Skeleton, RGB	0.826799
4	**Ours**	RGB	0.791933

4 Conclusion

We have presented our method designed for the contests of track 2 and track 3 at Looking at People (LAP) challenge. We firstly segment each video sequence into clips and then use a super vector representation to describe the clip visual content. The performance on both tracks demonstrate that our method is effective for action and gesture recognition. In the future, we may consider using more modalities to further improve the our spotting performance for gesture

recognition.

Acknowledgements: This work is partly supported by Natural Science Foundation of China (91320101, 61036008, 60972111), Shenzhen Basic Research Program (JC201005270350A, JCYJ20120903092050890, JCYJ20120617114614438), 100 Talents Program of CAS, Guangdong Innovative Research Team Program (201001D0104648280).

References

1. Escalera, S., Bar, X., Gonzlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: ECCV workshop. (2014)
2. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. **43**(3) (2011) 16
3. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C **37**(3) (2007) 311–324
4. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV. (2011) 2556–2563
5. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR **abs/1212.0402** (2012)
6. Sanchez, D., Bautista, M., Escalera, S.: Hupab 8k+: Dataset and ecoc-graphcut based segmentation of human limbs. Neurocomputing (2014)
7. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.J.: Multi-modal gesture recognition challenge 2013: dataset and results. In: ICMI. (2013) 445–452
8. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1) (2013) 60–79
9. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010) 143–156
10. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV. (2003) 1470–1477
11. Wang, X., Wang, L., Qiao, Y.: A comparative study of encoding, pooling and normalization methods for action recognition. In: ACCV. (2012) 572–585
12. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. CoRR **abs/1405.4506** (2014)
13. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)
14. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: NIPS. (1998) 487–493
15. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM TIST **2**(3) (2011) 27
16. Yong, P., Bingbing, N., Indriyati, A.: Mixture of heterogeneous attribute analyzers for human action detection. In: ECCV workshop. (2014)
17. Shu, Z.: Human action detection in videos with improved dense trajectories and sliding window. In: ECCV workshop. (2014)

18. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: Multi-scale deep learning for gesture detection and localization. In: ECCV workshop. (2014)
19. Monnier, C., German, S., Ost, A.: A multi-scale boosted detector for efficient and robust gesture recognition. In: ECCV workshop. (2014)
20. Chang, J.Y.: Nonparametric gesture labeling from multi-modal data. In: ECCV workshop. (2014)