Encoding Multi-resolution Two-Stream CNNs for Action Recognition

Weichen Xue, Haohua Zhao, and Liqing Zhang(✉)

Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, China
{xueweuchen,haoh.zhao,lqzhang}@sjtu.edu.cn

Abstract. This paper deals with automatic human action recognition in videos. Rather than considering traditional hand-craft features such as HOG, HOF and MBH, we explore how to learn both static and motion features from CNNs trained on large-scale datasets such as ImagNet and UCF101. We propose a novel method named multi-resolution latent concept descriptor (mLCD) to encode two-stream CNNs. Entensive experiments are conducted to demonstrate the performance of the proposed model. By combining our mLCD features with the improved dense trajectory features, we can achieve comparable performance with state-of-the-art algorithms on both Hollywood2 and Olympic Sports datasets.

Keywords: Deep learning · CNN · Action recognition

1 Introduction

Automatic action recognition is an important problem in computer vision and surveillance systems and has drawn significant attention in recent years. Recent research focuses on realistic datasets from movies, web videos such as Hollywood2 [6], UCF101 [11], and Olympic Sports [7]. The state-of-the-art performance of action recognition is given by a bag-of-words (BoW) representation of local features like HOG, HOF and MBH [12]. Recently, Convolutional neural networks (CNNs) are also introduced into action recognition task [10]. In some challenging datasets like UCF101, CNNs [10] have reported better performance than traditional local features [12]. However, CNNs require a huge amount of annotated training data. For some small size datasets such as Hollywood2 and Olympic Sports, we lack of sufficient training samples to train CNNs adequately. Therefore, there are a large number of works [5,13] exploring how to utilize CNNs trained on ImageNet to extract visual features.

In this work, we propose a novel algorithm to better employ the ImageNet trained CNNs. Motivated by the popularity of spatial pyramids in image classification [9], we propose multi-resolution latent concept descriptor (mLCD) features. By encoding the LCD features from multiple scales, the final video feature is able to give a better representation. On the other hand, when transfer

A. Hirose et al. (Eds.): ICONIP 2016, Part III, LNCS 9949, pp. 564–571, 2016.
DOI: 10.1007/978-3-319-46675-0_62

ImageNet trained CNNs, mLCD is only used to encode last convolution layer features of spatial networks. We extend our mLCD to temporal networks to capture motion features. The main contributions of this paper are summarized as follows:

1. We propose a multi-resolution extension to LCD [13] named mLCD, which extracts visual features from video frames at multiple scales.
2. We combine our mLCD with the two-stream CNNs [10], that is, using mLCD to encode features from temporal networks. As we know, this is the first work which encodes the features from last convolution layer of temporal networks.
3. We combine our mLCD features with the traditional improved dense trajectory [12] features, and conduct experiments on Hollywood2 and Olympic Sports datasets. The experimental results of the proposed algorithm achieves state-of-the-art performance.

2 Method

In this section, we first provide a description of our action recognition framework. Then the details of the proposed mLCD method are further elaborated. A discussion of the Fisher Vector and VLAD is given at the end of the section.

2.1 Action Recognition Pipeline

Our action recognition framework is shown in Fig. 1, it mainly consists of three parts: feature extraction, feature encoding and classification. Our proposed framework combines the hand-craft local features and the learned deep local features, encoding them with different methods and finally combining them with late fusion.

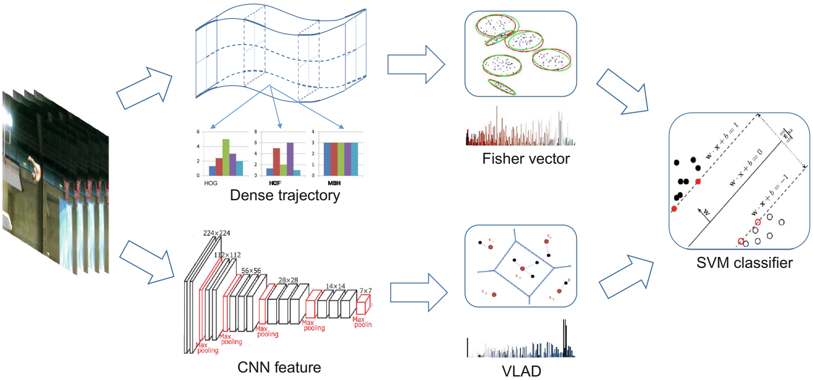


Fig. 1. Action recognition pipeline

For the hand-craft local descriptors, we adopt the improved dense trajectory features [12]. The densely sampled corner points are tracked to form dense trajectories. For each trajectory, different low-level features are computed in its spatial-temporal volume. In our framework, we compute histogram-based features including HOG, HOF and MBH. Normalization and PCA are applied to the local descriptors as mentioned in [12].

For the learned deep features, to capture both the semantic high-level features and motion features, we utilize the two-stream CNNs [10] to extract video features. In our framework, we combine LCD [13] (latent concept descriptor) and two-stream CNNs. We also extend LCD to its multi-resolution version, which is described in Sect. 2.2.

Once the improved dense trajectory features and multi-resolution LCD features are extracted, we encode the improved dense trajectory features with Fisher Vector and encode mLCD features with VLAD. The choice of encoding methods is discussed in Sects. 2.3 and 3.3. The two kinds of encoded features are combined using late fusion, and the concatenated video-level features are feed to SVM classifier to obtain the final classification result.

2.2 Multi-resolution LCD

LCD [13] (latent concept descriptor) is a method to encode the CNN extracted features by traditional BoW methods. LCD extracts the pooling layer feature rather than the full-connected layers, which contains spatial information. Specifically, for VGG16 architecture, the dimension of last convolution layer is $7 \times 7 \times 512$, which can be viewed as 49 local features of dimension 512. These local descriptors can be encoded by any BoW methods including Fisher Vector and VLAD. We propose two major improvements of the original LCD method.

Firstly, we combine LCD with the two-stream CNNs. Traditional LCD only utilizes the feature from the spatial network, thus, it can merely capture the static semantic information. To capture the motion feature, we propose to embed LCD into the temporal network, that is, the last convolution layer of temporal network is also viewed as local features and encoded by the same way as spatial network. The local descriptors from the two networks are encoded independently and combined with late fusion.

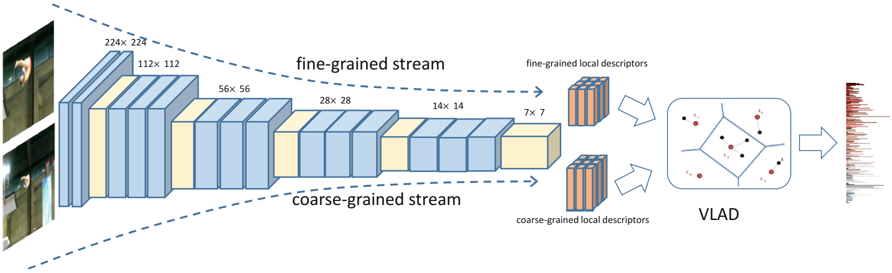


Fig. 2. Multi-resolution latent concept descriptor

The second important improvement of LCD is that we extend the LCD to its multi-resolution version, which is named mLCD. Our intuition is motivated by the success of spatial pyramid in both image classification [9] and action recognition [12]. For both the spatial network and temporal network, as shown in Fig. 2, we provide two kinds of input to the networks. The entire images (or optical flow) are feed into the network to gain the coarse-grained local descriptors. Fine-grained local descriptors are obtained by feeding the central crop of image into the same network. We call the two procedure coarse-grained stream and fine-grained stream respectively. We encoding the coarse-grained local features and fine-grained local features together to generate the feature of a video.

2.3 Fisher Vector or VLAD

Once the multi-resolution local descriptors are extracted by our networks, we can employ Fisher Vector [9] or VLAD [1] to encode the local descriptors into a video descriptor. Either Fisher Vector or VLAD can be viewed as an alternative of bag-of-words encoding, but both of them have shown better performance than traditional BoW encoding methods in image classification [9] and action recognition [8].

Fisher Vector [9] encoding, derived from Fisher Kernel, is the gradient of the log-likelihood with respect to a parameter. Generally, we fit the data with a Gaussian Mixture Model (GMM) with diagonal covariance matrix, so given a single local descriptor x, the gradient vector of log-likelihood respect to the model parameter is as follows:

$$\mathcal{G}^{x}_{\mu,k} = \frac{1}{\sqrt{\pi_k}}\gamma_k\left(\frac{x-\mu_k}{\sigma_k}\right) \tag{1}$$

$$\mathcal{G}^{x}_{\sigma,k} = \frac{1}{2\sqrt{\pi_k}}\gamma_k\left[\left(\frac{x-\mu_k}{\sigma_k}\right)^2 - 1\right] \tag{2}$$

where γ_k is the weight of local descriptor x to the k-th gaussian component, and is calculated by $\gamma_k = \frac{\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(x;\mu_k,\Sigma_k)}$

The Fisher Vector of one descriptor x is the concatenation of these gradients, $S_x = [\mathcal{G}^{x}_{\mu,1}, \mathcal{G}^{x}_{\sigma,1}, ..., \mathcal{G}^{x}_{\sigma,K}, \mathcal{G}^{x}_{\sigma,K}]$. The final Fisher Vector of one video is the sum of the Fisher Vectors of local features, $S = \sum S_x$.

VLAD [1] can be viewed as a simplified version of Fisher Vector. VLAD employs k-means algorithm to obtain the codebook rather than fit data with GMM. Once the codebook $\{d_i : i = 1, 2, ..., K\}$ is calculated, for each local descriptor x, its VLAD encoding can be calculated as $S_x = [\omega_1(x - d_1), \omega_2(x - d_2), ..., \omega_K(x - d_k)]$.

For our mLCD local descriptors, we make some quantitative analyses to decide which encoding method should be adopted. In Fig. 3, we plot the value distribution of Fisher Vector and VLAD, both of which have been normalized. It is shown that for Fisher Vector, most of the values are positive but the values of VLAD are distributed more uniform. Therefore, it is natural to assume VLAD is a better choice, and the experimental result in Sect. 3 validates our assumption.

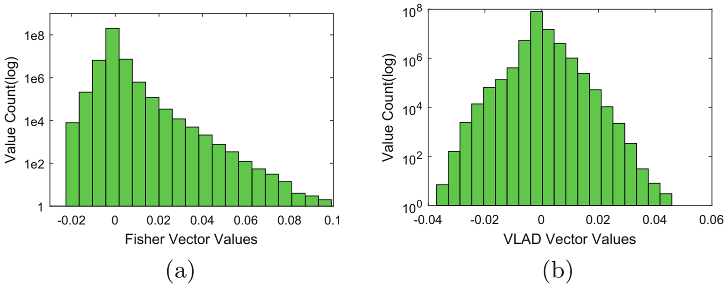


Fig. 3. The value distribution of Fisher Vector and VLAD

3 Experiments

In this section, we first introduce the datasets used in our experiments. Then we present the implement details of our algorithms. Some quantitative analyses are made to show the effectiveness of our method. Finally, a comparison with the state-of-the-art methods is given.

3.1 Datasets

Hollywood2. The Hollywood2 dataset [6] has been collected from 69 different Hollywood movies and includes 12 action classes. It contains 1,707 videos split into a training set (823 videos) and a test set (884 videos). The performance is measured by mean average precision (mAP) over all classes, as in [6].

Olympic Sports. The Olympic Sports dataset [7] consists of athletes practicing different sports collecting from YouTube. There are totally 16 sports actions (such as clean and jerk, bowling, basketball lay-up, discus throw), represented by a total of 783 video sequences. We use 649 sequences for training and 134 sequences for testing as recommended. mAP over all classes is reported as in [7].

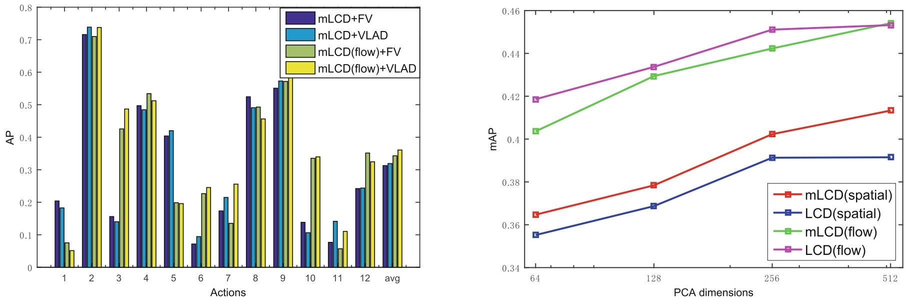
3.2 Implement Details

In our experiments, HOG, HOG and MBH descriptors form a 396-dimension vector (96+108+96+96), dimension of which is reduced to half with PCA. The dense trajectory feature are encoded by Fisher Vector and square root normalization and L2 normalization are both applied. For the learned deep feature, we adopt VGG16 for both spatial network and temporal network. The spatial network is trained on the ImageNet and the temporal network is pre-trained on ImageNet and is finetuned on UCF101. The mLCD features are encoded by VLAD. For each local feature, we search 5 nearest neighbors in the codebook to encode it. Square root and L2 normalization is also applied at last. For multi-class SVM, we adopt the one-vs-all method. The hyperparameters of each SVM is decided via cross validation.

3.3 Quantitative Analysis

We conduct rich quantitative analyses to demonstrate the effectiveness of our algorithm. In this section, all the experiments are conducted on the Hollywood2 dataset. First, we make an analysis of the choice of some hyperparameters in our method. In Fig. 4(a), we encoding mLCD local descriptors with Fisher Vector and VLAD separately. It is showed that VLAD can gain a slightly better performance than Fisher Vector, which is consistent with the analyses in Sect. 2.3.

Figure 4(b) shows the impact of PCA dimension reduction to our method. In this experiment, we fix the encoding method as VLAD, and apply PCA to mLCD and LCD local space-time features to show the impact of PCA. As presented in Fig. 4(b), the performance of both mLCD and LCD is largely damaged by PCA dimension reduction. It is also clear that our mLCD method is more sensitive to PCA. For the temporal network, when the local discriptors keep the original dimension (512), mLCD reports a mAP of 45.41 %, while LCD gains 45.32 %. When PCA dimension is smaller, LCD obtains a better performance compared with mLCD. Therefore, in the following experiments, we do not apply PCA to mLCD local features.



(a) (b)

Fig. 4. Quantitative analysis of hyperparameters choice. (a) The mAP of mLCD when encoded by Fisher Vector and VLAD. (b) The mAP of LCD and mLCD when applied PCA dimension reduction

Table 1 compares our mLCD with LCD [13] under different settings. We can draw mainly two conclusions from this table. First, it is clear that mLCD outperforms LCD on both Hollywood2 and Olympic Sports under different configurations. Second, spatial network performs better than temporal network on both datasets. There are several possible reasons leading to this result: it may be decided by the scenes of two datasets, which indicate that on Hollywood2 dataset and Olympic Sports dataset, motion feature is less import than static feature. Another reason is that temporal network is trained on UCF101, which is relative small compared with ImageNet, therefore, temporal network tends to be overfitting on UCF101.

Table 1. The mAP of our method with different configurations

Methods	Hollywood2	Olympic sports
LCD	0.3915	0.7739
mLCD	0.4133	0.8003
LCD (flow)	0.4532	0.7429
mLCD (flow)	0.4541	0.7554
LCD + LCD (flow)	0.5101	0.8296
mLCD + mLCD (flow)	0.5163	0.8530

3.4 Comparison with the State of the Art

Table 2 compares our method with the most recent results reported in literature of th two datasets. On Hollywood2 dataset, trajectory based methods [2,4,12] achieve great success. Wang et al. [12] report 64.3 % by combining dense trajectories, motion features and human detectors. Jain et al. [3] report 66.6 % by introduce a large scale of concept detector. Our method improves the state-of-the-art result by around 0.1 % by combining shallow and deep features.

Olympic Sports is a collection of sports videos. This dataset contains rich structure information and significant camera motion. Therefore, traditional trajectory based methods [2,4] does not perform well on this dataset. Wang et al. [12] introduce human detectors to remove the background trajectories and gain a mAP of 91.1 %. Our experiments show that without computational expensive detectors, we can also obtain a slightly better result of 91.4 %.

Table 2. The mAP of our method with different settings

Methods	Hollywood2	Olympic Sports
Jiang et al. [4]	0.595	0.806
Manan Jain et al. [2]	0.625	0.832
Wang et al. [12]	0.643	0.911
Mihir Jain et al. [3]	0.666	–
IDT + mLCD (Ours)	0.669	0.914

4 Conclusion

In this paper, we explore how to effectively utilize CNNs trained on ImageNet and UCF101 to improve the performance of action recognition. We introduce multi-resolution latent concept descriptors (mLCD) to encode both spatial and temporal network, and conduct experiment on Hollywood2 and Olympic Sports datasets. We report a better result compared with the current state-of-the-art methods.

Acknowledgements. The work was supported by the National Natural Science Foundation of China (61272251), the Key Basic Research Program of Shanghai Municipality, China (15JC1400103) and the National Basic Research Program of China (2015CB856004).

References

1. Arandjelovic, R., Zisserman, A.: All about VLAD. In: CVPR. pp, 1578–1585. IEEE (2013)
2. Jain, M., Jégou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: CVPR, pp. 2555–2562. IEEE (2013)
3. Jain, M., van Gemert, J.C., Snoek, C.G.: What do 15,000 object categories tell us about classifying and localizing actions? In: CVPR, pp. 46–55 (2015)
4. Jiang, Y.-G., Dai, Q., Xue, X., Liu, W., Ngo, C.-W.: Trajectory-based modeling of human actions with motion reference points. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 425–438. Springer, Heidelberg (2012)
5. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR, pp. 1725–1732. IEEE (2014)
6. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR, pp. 2929–2936. IEEE (2009)
7. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
8. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. arXiv preprint (2014). arXiv:1405.4506
9. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. IJCV **105**(3), 222–245 (2013)
10. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS, pp. 568–576 (2014)
11. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint (2012). arXiv:1212.0402
12. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV, pp. 3551–3558. IEEE (2013)
13. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative CNN video representation for event detection. In: CVPR, pp. 1798–1807 (2015)