TRAJECTORIES-BASED MOTION NEIGHBORHOOD FEATURE FOR HUMAN ACTION RECOGNITION

Xiang Xiao, Haifeng Hu, and Weixuan Wang

School of Electronics and Information Engineering, Sun Yat-sen University, Guangzhou, China

ABSTRACT

Recently, a common and popular method that produces competitive accuracy is to employ dense trajectories to identity human action. However, computing descriptors of dense trajectories may spend lots of time, and many trajectories which belong to the background trajectories may not be useful for the recognition. Moreover, the relationship between trajectories is always ignored. In this paper, we propose a trajectories-based motion neighborhood feature (TMNF) method for action recognition. We first select the trajectories of central particular region at the original video resolution to reduce the computation as well as the background trajectories. A new descriptor, which is referred to as TMNF, is proposed to explore the orientation and motion relationship between different trajectories. Finally, an improved vector of locally aggregated descriptors (IVLAD) method is used to represent videos and linear SVM is applied for classification. Experiments on the YouTube dataset demonstrate that our approach achieves superior performance.

Index Terms—Action Recognition, Dense Trajectories, Improved VLAD, linear SVM

1. INTRODUCTION

Human action recognition has caused wide attention in the area of computer vision due to its practical application in human-computer interaction, video surveillance and more. To date, it still imposes significant challenges such as background occlusions, view point changes, irregular motion, camera motion, *etc*. Many local space-time visual representations have been proposed to overcome these issues in the action recognition task. Laptev [1] detected sparse space-time interest points and computed histograms of the detected local points. Kliper-Gross *et al.* [2] proposed the motion interchange patterns (MIP) method to acquire local changes of motion directions by using local trinary pattern (LTP). Wang *et al.* demonstrated in [3] that dense sampling outperforms sparse space-time interest points.

In recent years, many approaches based on trajectories explore the underlying temporal motion [4]-[7]. Wang *et al.* [4] detected dense trajectories by tracking dense points in the optical flow field. Jain *et al.* [5] proposed to decompose visual motion of trajectories into dominant and residual motions, which could cancel the camera motion effectively. Murthy *et al.* [6] selected few trajectories to generate a new trajectories set termed 'ordered trajectories'. Jiang *et al.* [7] presented to use global and local reference points to capture motion information of trajectories and overcome the global motion. Although these methods perform good accuracy, the relationship of relative location and motion information of trajectories are discarded. In this paper, we select the trajectories which are located in the central particular region, to reduce the calculation as well as decrease the background trajectories to some extent. A new trajectories-based motion neighborhood feature (TMNF) is proposed to represent the selected trajectories. Compared with the classical methods, TMNF could capture the location and motion relationship between central trajectory and its nearest several trajectories. The results on human actions datasets demonstrate that increasing the TMNF descriptor is helpful for recognition. The motion boundary histogram (MBH) is widely used as the trajectory-aligned descriptor [8], which represents the gradient of the optical flow so that it could keep only the motion information about changes. In the process of MBH feature calculating, the vertical (y) and horizontal (x) components of optical flow are computed independently. In our model, we use these two components separately as two different features, which are MBHX and MBHY [9]. They may capture different aspects of motion information. We combine them with the MBH feature in order to obtain better result since they have complementary advantages.

How to build the bridge between low-level features and high-level action classes is a significant problem. Mid-level representation such as bag-of-words (BOW), fisher vector (FV) and vector of locally aggregated descriptors (VLAD) could be more efficient and compact than low-level representation [10]. According to [5]-[6], VLAD slightly outperforms BOW and FV under the same conditions. This paper proposes a video-level representation approach based

This work was supported in part by the NSFC, under Grant no. 61673402, NSF of Guangdong (grant nos. 2014A030313173 and 2016B010109002), the Fundamental Research Funds for the Central Universities of China and the Science and Technology Program of Guangzhou (grant no. 201704020180). Corresponding author: Haifeng Hu.

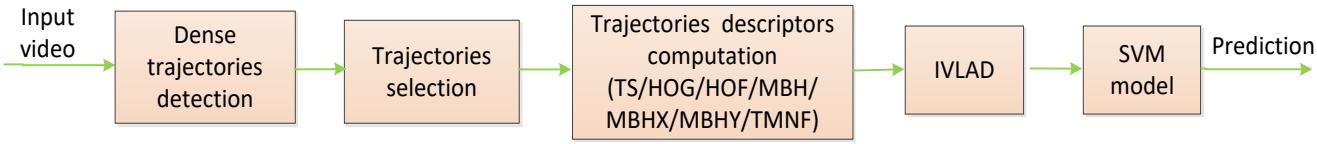

Fig. 1. Flowchart of the proposed method

on the original VLAD, which is referred to as improved VLAD (IVLAD).

The main contributions of this paper are summarized as follows:

1. A new descriptor (TMNF) is proposed to describe trajectories by capturing their orientation relationship and motion neighborhood information.
2. The central region is applied to reduce background trajectories, which decreases the computation remarkably.
3 k-means++ method is used for clustering and MBHX/MBHY features are combined with MBH to achieve rich description of human activities.
4. A video representation algorithm is utilized by applying improved VLAD technique.

2. OUR APPROACH

The proposed method is based on dense trajectories by Wang *et al.* [4]. Dense points are sampled from every frame and tracked in the dense optical flow field to reshape dense trajectories. We follow [4] and sample points with a grid step size of 5 pixels in 8 different space scales. Tracking is achieved by median filtering. Specifically, a feature point $P_{t+1}=(x_{t+1},y_{t+1})$ in the $t+1$ frame is tracked from the previous frame by:

$$P_{t+1}=(x_{t+1},y_{t+1})=(x_t,y_t)+(K\times w)|_{(\bar{x}_t,\bar{y}_t)} \tag{1}$$

where $w=(u,v)$ is the dense optical flow field, K denotes the median filtering kernel and $(\bar{x}_t,\bar{y}_t)$ denotes the rounded position of (x_t,y_t). A maximum value of trajectory length L is set to 15. Four types of features which are trajectory shape (TS), HOG, HOF and motion boundary histogram (MBH) are extracted to describe trajectories. However, we argue that there are many irrelevant trajectories in the background, which may increase the complexity of computation. Furthermore, the relationship between trajectories is ignored. In this paper, we propose a novel model to handle with these issues. Figure 1 illustrates the overall flowchart of proposed method.

2.1. Trajectories-based Motion Neighborhood Feature

In order to reduce the computation as well as irrelevant background trajectories, we first apply a pre-processing technique. We choose the central region (CR) of video frame at the original video resolution. Specifically, the width and height of the CR are selected at two-thirds of the original frame. For each trajectory T_u, it should be kept or not according to (2):

$$T_u=\begin{cases}\text{keep}; & \text{if } (\bar{x}_u,\bar{y}_u)\subset \text{CR}\\ \text{discard}; & \text{if } (\bar{x}_u,\bar{y}_u)\notin \text{CR}\end{cases} \tag{2}$$

where $(\bar{x}_u,\bar{y}_u)$ is the mean value of x and y coordinates of T_u. In this way, the total numbers of trajectories are dropping and the main part of the dropped trajectories is irrelevant, since the object of interest often occupies the center region [11]. We cluster the MBH feature of selected trajectories by using k-means++ method [12]. so that each MBH feature of a trajectory can be mapped to a visual word.

Fig. 2 shows the proposed motion neighborhood feature of trajectories (TMNF). For each T_u, we define NT_u as a neighbor trajectory of T_u, which consider the relative orientation locations and the MBH visual words of NT_u to T_u. We choose MBH feature since it captures more motion information than other trajectories features. Specifically, eight bins are used to quantize the spatial relative locations between NT_u and T_u. They could be depicted as follows:

$$RL(T_u,NT_u)=P_{T_u}-P_{NT_u}=(\bar{x}_{T_u}-\bar{x}_{NT_u},\bar{y}_{T_u}-\bar{y}_{NT_u}) \tag{3}$$

Thus the space coordinate is divided into 8 bins from 0° to 360°, each bin corresponds to a angle interval of 45°. We combine the orientation relationship with MBH visual words of NT_u. Assume that the MBH features of training trajectories are clustered to k clustering centers by *k*-means++ algorithm, so that there are $8\times k$ possible location-word combinations. The TMNF descriptor of a trajectory is formed by creating a matrix of size $N\times 8k$: the *i*-th row is a cumulative histogram corresponding to the first i ranked neighbor trajectories of T_u. The matrix is reshaped to a single $8kN$ dimensional vector to yield a final TMNF representation. Comparing with the number of $8k$, N is relatively small. It indicates that the TMNF representation is a sparse vector. We decrease the dimensionality subsequently with PCA. MBHX and MBHY are applied separately in our approach as two separate different features, since they may capture different aspects of motion information of trajectories.

2.2. Improved VLAD (IVLAD)

(a) (b)

Fig. 2. An illustration of the trajectories-based motion neighborhood feature, named as TMNF. (a) A trajectory T_u and its neighbor trajectory NT_u. Red points denote the mean spatial location of T_u and NT_u. Blue arrows represents the relative location orientation between T_u and NT_u. (b) The quantization scheme of orientation for generating the 8 bins representation.

Vector of locally aggregated descriptors (VLAD) is a feature representation approach that aggregates the descriptors based on a locality criterion in the feature space. Jegou *et al.* [13] first proposed VLAD method to address the problem of large-scale image search. Jain *et al.* [5] applied VLAD for action recognition.

2.2.1. VLAD method

In VLAD, the difference between each feature descriptor and its closest center is collected as residual vector. For each clustering center u_i, the according residual vectors are computed as a sub-vector v^i:

$$v^i = \sum_{x:q(x)=u_i} x - u_i \tag{4}$$

where $q(x)$ is the clustering label of feature x. Then concatenating all obtained sub-vectors to yield a D dimensional vector $v=[v^1 ... v^k]$, where $D = k \times d$, k is the number of centers and d is the dimension of feature.

Two-stage normalization is subsequently achieved. Firstly, the 'power-law normalization' is applied [14]. Each component $v_j, j = 1 \, to \, D$ is modified as $v_j = |v_j|^{\alpha} \times sign(v_j)$, where α is a parameter such that $\alpha \le 1$. The second stage is the *L*2-normalization, which yields the final VLAD vector.

2.2.2. Improved VLAD

Original VLAD approach forms the final representation by summing up all residual vectors. However, some components of a residual vector may change significantly. This may lead to that the individual descriptors will contribute unequally to the VLAD representation [15]. To tackle this problem, we use *L*2-normalization to the residual vectors so that all descriptors could contribute equally. Thus Eq. (4) is modified as:

$$v^i = \sum_{x:q(x)=u_i} \frac{x - u_i}{\|x - u_i\|} \tag{5}$$

Compared with original VLAD, the above operation can guarantee all descriptors equally contribute, while bursts are processed according to subsequent power-law normalization.

Fig. 3. Sample frames from the YouTube dataset

3. EXPERIMENTS

3.1. Dataset

We evaluate the performance of proposed method on a challenge human action dataset: YouTube [16].

YouTube dataset contains 11 action categories and there are 1168 videos in total. Fig. 3 shows several frames from the YouTube dataset. For each class, sequences are grouped into 25 folds with more than 4 action clips in each fold. We follow the protocol in [16] by using the leave-one-out cross validation scheme, which involves using one fold as the testing videos and remaining folds as the training samples.

3.2. Experimental setup

We made all experiments on the Tianhe-2A platform. It's ranked as the world's fastest supercomputer and could be used to speed up the computation remarkably. The number of neighbor trajectories N is set to 5, we cluster each type of feature to 256 centers ($k = 256$). The α of power-law normalization is set to 0.2. In order to handle with multi-class classification problem, we apply the LIBSVM with the linear kernel.

3.3. Experimental results

We conduct experiments on the YouTube dataset to test the proposed approach. Firstly, we compare the number of

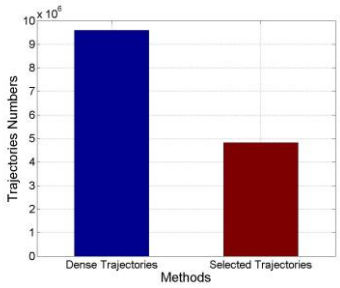

Fig. 4. Number of trajectories in different schemes for 300 simple videos from YouTube datasets.

Table 1. The performance of five methods on the YouTube dataset.

Method	Accuracy
Dense Trajectories (DT) + BOW [4]	84.2%
DT + MBHX + MBHY + BOW	84.5%
DT + TMNF +BOW	88.2%
DT + TMNF + VLAD	89.7%
Proposed method	**91.4%**

trajectories of our approach with original dense trajectories based method. Secondly, we make comparisons between several approaches to evaluate the performance of proposed TMNF and IVLAD. Thirdly, our approach compares with the state-of-the-art on YouTube dataset.

1) *Number of trajectories*: Fig. 4 shows the number of trajectories of two trajectories based methods on both datasets. As we can see, our operation produces the minimum number of trajectories, which is about 50% of the original dense trajectories method. This will reduce the computation of the subsequent operation and it does not result in significant loss of accuracy.

2) *Evaluation of the proposed method:* We evaluate the performance of proposed TMNF and IVLAD. Three approaches are applied for comparison. They are called 'DT+MBHX+MBHY+BOW', 'DT+TMNF+BOW' and 'DT+TMNF+VLAD', respectively. 'DT+MBHX+MBHY +BOW' only adds MBHX and MBHY features on the basis of dense trajectories (DT) method, while 'DT+ TMNF+BOW' adds the TMNF feature to the DT. Moreover, to validate the availability of IVLAD, we replace the IVLAD with the traditional VLAD method in the 'DT+TMNF+VLAD' approach. We summarize the results on the YouTube dataset in Table 1. By the comparisons of Table 1, we draw several conclusions. First, the 'DT+MBHX+MBHY+BOW' and 'DT+TMNF+BOW' methods outperform the dense trajectories (DT) method. It indicates the fact that extracting TMNF and MBHX/MBHY features are beneficial to capture the motion information for action recognition. Moreover, the accuracy of 'DT+TMNF+BOW' has greatly increased by nearly 4 percent, which shows that the proposed TMNF is a distinguished feature by comparison with other original features used by [4]. Second, our method achieves a higher performance than 'DT+TMNF+VLAD' (*e.g.*, from 89.7% to 91.4%), which indicates the efficiency of the MBHX/MBHY features and the IVLAD representation. Finally, the proposed method achieves the best results of 91.4%, which is higher than any other methods reported in Table 1. It means that the proposed TMNF and IVLAD representation are complementary to dense trajectories based method.

Table 2. Comparison of our method with the state-of-the-art approaches on the YouTube dataset.

Method	Accuracy
Wang *et al.* [4]	84.2%
Simon *et al.* [18]	89.0%
Le *et al.* [19]	75.8%
Wang *et al.* [17]	89.8%
Brendel *et al.* [20]	77.8%
Our method	**91.4%**

3) *Comparison to the state-of-the-art*: Table 2 compares our result to other works on YouTube dataset. The recognition performance reported by Jones *et al*. [18] is 89.0% using multigraph representation called Feature Grouped Spectral Multigraph (FGSM). Wang *et al.* [17] improve the dense trajectories method by explicitly estimating camera motion and produce good performance. [19] applies Laplacian Eigenmaps (LE) on histogram and uses RBF kernel for classification. The result for [20] is produced by using an exemplar-based approach, which represents activities as time series of a few snapshots of human-body parts. As can be seen in Table 2, our method outperforms all the other methods. Furthermore, the proposed method achieves result in an easier and effective manner because the TMNF is described by sparse representation. Since we combine the neighborhood motion information and location information, our method can detect accurately the actions with great changes of motion. In future, it may be possible to achieve even better results by detecting the interest region automatically.

4. CONCLUSIONS

In this paper, we propose a novel TMNF-based method for action recognition. We select the trajectories in the central region, and propose a new TMNF descriptor to represent trajectories. IVLAD is presented to obtain the final video representation and linear SVM is used for classification. Experimental results verify the effectiveness of the proposed method.

5. REFERENCES

[1] I. Laptev, "On space-time interest points," in *Proc. International Journal of Compute Vision*, pp. 107-123, 2005.

[2] O. Kliper-Gross, Y. Gurovich, and T. Hassner, "Motion interchange patterns for action recognition in unconstrained videos," in *Proc. European Conf. on Computer Vision*, pp. 256-269, 2012.

[3] H. Wang, M.M. Ullah, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. British Machine Vision Conference*., pp. 127-138, 2009.

[4] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3169-3176, 2011.

[5] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 2555–2562, 2013.

[6] O.V. Ramana Murthy and R. Goecke, "Ordered trajectories for large scale human action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Works.*, pp. 412–419, 2013.

[7] Y.G. Jiang, Q. Dai, W. Liu, X.Y. Xue, and C.W. Ngo, "Human action recognition in unconstrained videos by explicit motion modeling," *IEEE Trans. on Image Processing*, pp. 3781–3795, 2015.

[8] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. European Conf. on Computer Vision*, pp. 428–441, 2006.

[9] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," in *Proc. International Journal of Compute Vision*, pp. 60-79, 2013.

[10] C. Liu, Y. Kong, X. Wu, and Y. Jia, "Action recognition with discriminative mid-level features," in *Proc. IEEE Int. Conf. on Pattern Recognit.*, pp. 3366–3369, 2012.

[11] A. Karpathy, G. Toderici, S. Shetty, and T. Leung, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, pp. 1725–1732, 2014.

[12] D. Arthur and S. Vassilvitskii, "The advantages of careful seeing," in *Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 1027-1035, 2007.

[13] H. Jegou, F. Perronnin, M. Douz*e, and* J. Sanchez*,* "Aggregating local image descriptors into compact codes," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* pp. 1704-1716, 2012.

[14] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. European Conf. on Computer Vision*, pp. 143-156, 2010.

[15] J. Delhumeau, P.H. Gosselin, and H. Jegou, "Revisiting the VLAD image representation," in *Proc. of the 21st ACM Int. Conf. on Multimedia*, pp. 653-656, 2013.

[16] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1996-2003.

[17] H. Wang, and C. Schmid, "Action recognition with improved trajectories," in: *Int. Conf. on Comput. Vision*, pp. 3551-3558, 2013.

[18] S. Jones, and L. Shao, "A multigraph representation for improved unsupervised/semi-supervised learning of human actions," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, pp. 820-826, 2014.

[19] M. Belkin, and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," in: Neural Computation, pp. 1373-1396, 2002.

[20] W. Brendel, and S. Todorovic, "Activities as time series of human postures," in *Proc. European Conf. on Computer Vision*, pp. 721-734, 2010.