ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Action Recognition Based on Weight BOVW

Taizhe Tan, Chuhong Li [a)]

Guangdong University of Technology, Guangdong, 510006, China

Corresponding author:253298502@qq.com

Abstract. To improve the accuracy and robustness of human action recognition, a kind of human action recognition algorithm based on the weighing bad-of-word model was put forward. The features extracted from video sequences through existing algorithms contained HOG feature, HOF feature, and MBH feature, and relevant visual vocabulary table was acquired through the K-means clustering method. Statistics of the word frequency of all visual words of each category was made, and the words that rank at the front in the word frequency were selected to conduct the weighted normalization processing and acquire the weight of the words. The test samples conducted weighing expressions on the words whose frequency ranked at the front positions during the generalization process of bad of words, and it conducted classification and recognition of the characteristics of the bag of words through SVM. The result of UCF Sports DS experiment showed that the new algorithm had a good recognition capacity.

Key words: bag of words; action recognition; SVM.

INTRODUCTION

The broad application prospect of human action recognition in intelligent monitoring, human-computer interaction and other fields triggered attention by more and more researchers. However, due to the complexity and variety of human actions and the change of camera view, human behavior recognition remains a difficult and hot issue in the field of computer vision field [1].

Thanks to the researchers of their study, a large number of human action recognition algorithms have been raised, which mainly include: method on account of template matching [2], method on account of optical flow, [3] and method on account of features[4-5].The reference[6] mentioned a human behavior recognition method on account of dense trajectories, that is, tracking the feature points of optical flow field to obtain the trajectory, and calculating the trajectory displacement vector and Histogram of Oriented Gradient (HOG) of the locus neutron space-time block, Histograms of Oriented Optical Flow(HOF) and Motion Boundary Histograms(MBH) [7] as the underlying local feature descriptors of video sequences. These local feature descriptors are then used as bags of visual word model (BOVW)[8] inputs to gain video overall expression. At last, the overall video expression will merge together new features for category recognition with support of vector machine then come out with better recognition result. However, the expression of BOVW only contains the characteristics of single action, which means when there are large differences among the same actions, the differences between the expressions of words bags will also be large, then will affect the classification accuracy. At the same time, there are great differences between the common characteristics of the same class and other classes, causing less expression of the original word bag model.

In view of the above mentioned problems, this paper proposes a weighing idea to enhance the visual feature representation of action classes. Common action class features in the same kind of scene often appear on some words which are in front of the word frequency comparison, the combination of which can form the discriminative feature among classes. If these words are weighted in the process of generating the word bag, the test class will add the common features of the class to improve the classification accuracy.

FEATURE EXTRACTION

In order to construct the BOVW model, we first need to extract the behavior feature. In our paper, the method proposed by Wang et al. [1] is employed to extract trajectories. Firstly sampled feature points in dense grids, and then the trajectories were obtained by tracking them using densely optical flow.

To describe the motion information of dense trajectories, two descriptors (HOG, MBH) within 3D space-time field of size $N \times N \times L$ around the trajectory are computed. To remain structure information, the space-time field is segmented into $n_\sigma \times n_\sigma \times n_\tau$ space-time grid. The HOG of a grid is constructed by discrete gradients into eight bins. The MBH descriptor divides optical flow into horizontal and vertical components, and then Histograms of oriented gradients are computed for each of them. Compared with optical flow information, The MBH has shown more robust and more suitable for action classification, since the MBH expresses the gradient of the optical flow.

Except for trajectory features, HOF descriptor is also extracted. To construct HOF, The angles and magnitudes of optical flow are computed and then divided into 8 bins and 72 bins, respectively. Figure 1 shows dense trajectory description.

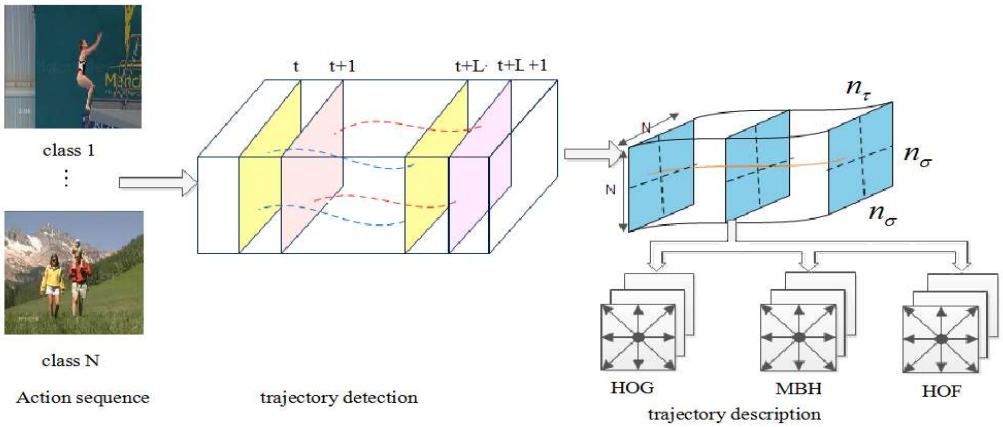

FIG.1 The process of trajectory extracted and described

WEIGHING BAG OF WORDS

Aimed at the problem that different videos would extract different interest points and quantities, the Bag of Visual Word could well solve this problem. Through K-means[9] clustering of all video features extracted above, including HOG feature, MBH feature and HOF feature, it acquired relevant visual vocabulary table and then express each video through the aggregation of visual words.

The realization of bad of word model was divided into 4 steps, as shown in Figure 2.

(1) Extract features and descriptions.

(2) Construct the visual vocabulary table. Respectively construct the visual vocabulary tables of each descriptor, that is, 3 visual vocabulary tables were acquired through 3 kinds of descriptions.

(3) Classification of features. Calculate the Eulidean distance between each feature and the visual word corresponding to it in the visual vocabulary table, namely the distance to the center of clustering, and affiliate it to the smallest crowd to complete the classification of the feature.

(4) Express each video with frequency histogram. Make statistics of classification of each feature in the video, and acquire the normalization of the histogram to express each video.

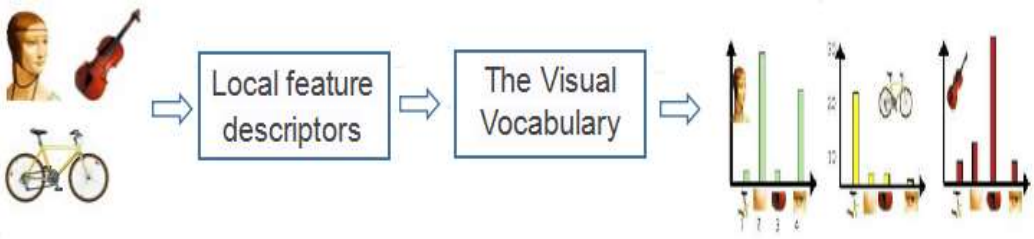

FIG.2 The process of realizing Bag of Visual words

Traditional Bag of Visual Word has been broadly used in the action recognition. While there are also some shortcomings, so a number of improved algorithms appeared. E.g. adding in spatial information into the visual representation[10] and improvement of clustering algorithm[11], etc., the improved algorithm of this thesis was mainly the strengthening of weights of important words to improve the expression of remarkable visual features of images. Detailed algorithm steps are as follows:

Expression of overall bad of words: make statistics of word frequency of each visual word in the bag of visual words of all video sequences in each category, then acquire the overall expression of the bag of words of each category.

(1) Make sure the weights of important words; take the frequencies of the first N words from the statistical results and conduct normalization processing on the frequency of the N words, show the relevant weights of the words with the reciprocal of the frequencies, and their values are respectively $\omega_i(0<i<N+1, i$ is an integer). Conduct normalization processing of those weights once again and make the weights within 0 and 1. Then add 1 to each weight, that is $\omega_i=\omega_i+1$. Then the weights are quantized within 1 and 2 to eliminate the bad influences caused by overlarge weights.

Select each category of important words to conduct weighing: after bag of words distribution of feature vectors of each video sequence, take $\omega_{i,}$ the weight of first N words in the category that the video sequences belong to and multiply by the frequency of the word corresponding to $\omega_{i,}$ the weight of single video frequency. Then the expression of weighing bag of visual words can be acquired.

Through the above processing, the acquired expression of weighing bag of words can sufficiently reflect the obvious features of each action. As the conversion flows of weighing bag of visual words shown in Figure 3.

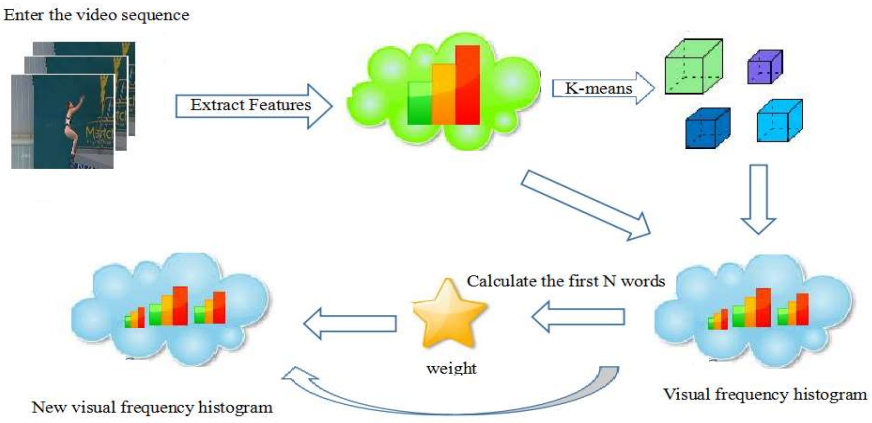

FIG.3 The process of transforming Weighted Bag of words

EXPERIMENT

In this section, we compared the recognition accuracy of different kernel function of BOVW firstly. And to determine the kernel function of the subsequent experiment. Figure 4 and figure 5 respectively show the average recognition accuracy of the improved BOVW method and the original BOVW method which use different kernel functions and codes. We can see from the figure 4 and figure 5 that Histogram Inter SVM is superior to Inter SVM and Linear SVM. In addition, when the code number is 1000, the effect of recognition is great. It is best to use the

improved BOVW algorithm parameter N for 25.Compared with the original BOVW algorithm, the results are shown in figure 6, which shows that the overall average accuracy of new BOVW algorithm improved is about 3%.

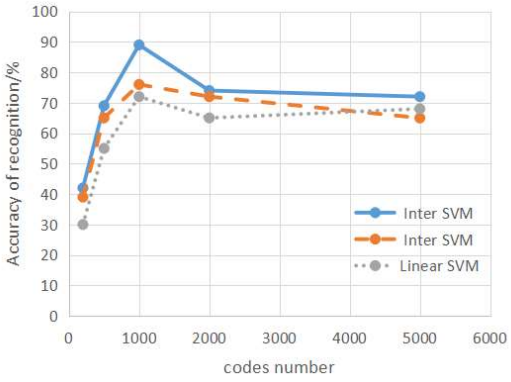

FIG.4 The overall average classification accuracy of our improved BOVW method

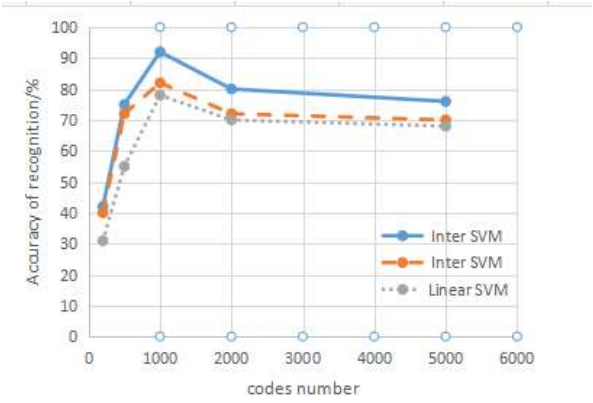

FIG.5 The overall average classification accuracy of ortginal BOVW method

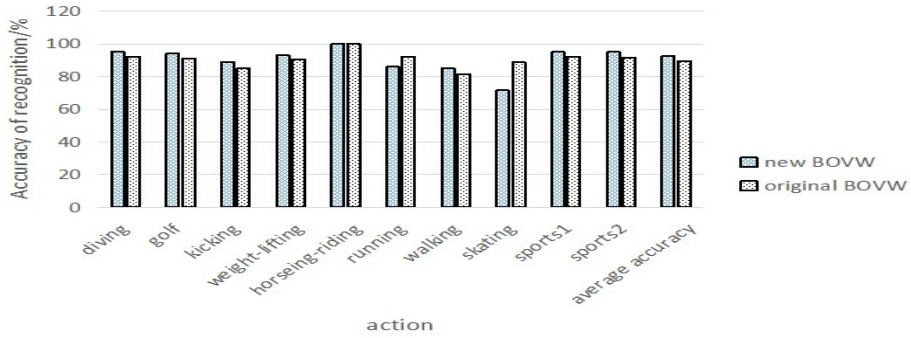

FIG.6 Two kinds of BOVW algorithm accuracy comparison chart

To further evaluate the performance of the proposed method, we compared the action recognition accuracy of the proposed method with those of existing methods and the results on UCF sports datasets are shown in Table 1.

TABLE.1 Average accuracy table

Methods	Recognition
Proposed method	92.71%
Wang et al.[12]	85.5%
Wu et al. [13]	92.48%
Yu Kong et al.[14]	88.6%
Kovashka and Grauman[15]	87.57%
Le et al. [16]	86.7%

CONCLUSION

The paper has proposed an efficient method for human action recognition involving feature extraction and description of the features before the classification. Three types of low-level visual features are extracted, and the bag-of-words model is used to describe features better. the expression of BOVW only contains the characteristics of single

action, which means when there are large differences among the same actions, the differences between the expressions of words bags will also be large, then will affect the classification accuracy. Therefore, this paper puts forward a kind of strengthening weight of feature to strengthen the significant visual feature representation of action class. There are more obvious differences between the classes, and the characteristics of the classes are more significant. The experimental results prove that the proposed approach in this paper is more excellent than other state-of-the-art methods.

REFERENCES

1. Wang, H., Klaer, A., Schmid, C. & Liu, C.-L., Action recognition by dense trajectories. Proc. of IEEE Conf. On Computer Vision and Pattern Recognition, pp. 3169–3176, 2011.
2. Ando H,Fujiyoshi H． Human-Area Segmentation by Selecting Similar Silhouette Images Based on Weak-Classifier Response / / Proc of the 20th International Conference on Pattern Recognition． Istanbul,Turkey,2010: 3444－3447.
3. Wang H, Yi Y. Tracking Salient Key points for Human Action Recognition[C]//IEEE International Conference on Systems,Cybernetics. IEEE, 2015:3048-3053.
4. Niebles, J.C, Wang, H. & Fei-Fei, L., Unsupervised learning of human action categories using spatial-temporal words. International journal of computer vision, 79(3), pp. 299-318, 2008.
5. Shechtman, E. & Irani, M., Space-time behavior based correlation. Proc. of Int. Conf. on Computer Vision and Pattern Recognition, pp. 405–412, 2005.
6. Wang Heng, Klaser A, Schmid C, et al. Action Recognition by Dense Trajectories [C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Washington D.C, USA IEEE Press 2011:3169-176.
7. Li Q, Cheng H, Zhou Y, et al. Human Action Recognition Using Improved Salient Dense Trajectories[J]. Computational Intelligence & Neuroscience, 2016, 2016(5):675-459.
8. Faraki M, Palhang M, Sanderson C. Log-Euclidean bag of words for human action recognition[J]. Iet Computer Vision, 2015, 9(3):331-339.
9. Yang Zunqi,Zhang Qiannan.Research on attention of microblog user based on K-means cluster analysis[J].Journal of Intelligence,2013,32(8):142-144.
10. Sheng Hua,Zhang Guizhu.A clustering method combining K-means and fast search algorithm of density peaks[J]. Computer Applications and Software,2016,33(10):260-269.
11. Yang Yi,Shawn N. Spatial pyraimid co-occurrence for image classification[C]//IEEE International Coference on Computer Vision. Barcelona, Spain:[s. n.], 2011:1465-1472.
12. Wang, H., Ullah, M. M., Klaer, A. Laptev, I. & Schmid, C. Evaluation of local spatio-temporal features for action recognition. Proc. of British Machine Vision Conference, pp. 1–11, 2009.
13. Wu, X., Xu, D. & Duan, L., Action recognition using multilevel features and latent structural SVM. IEEE transactions on circuits and systems for video technology, 23(8), pp. 1422-1431,2013.
14. Kong, Y. & Zhang, X.Q.. Adaptive learning codebook for action recognition. Pattern recognition letters, 32(8), pp. 1178-1186, 2011.
15. Kovashka, A. & Grauman, K., Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. Proc. of IEEE Conf. on CVPR, pp. 2046–2053, 2010.
16. Le, Q. V., Zhou, W. Y., Yeung, S. Y. & Ng, A. Y., Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3361–3368, 2011.