Human Action Recognition Based On Multi-level Feature Fusion

Y.Y. Xu, G.Q. Xiao, X.Q. Tang
Southwest University
Chongqing, China

***Abstract*—an efficient multi-level feature fusion descriptor for human action recognition is introduced in the paper. The descriptor is built by the low-level features, which include three trajectory features, HOF and SIFT combination with the mid-level class correlation feature. Inspired by the recent popularity of dense trajectories in image recognition, they have been utilized to represent actions. It is favorable to extract scene information for action recognition, since human actions have the tightly affinity on specific natural scenes. In addition, noting that different action classes may often share similar motion patterns, we introduce the mid-level class correlation feature to describe relationships among different video classes. Finally, to achieve the better recognition results, bag-of-word model is employed to describe the video by sets of visual words. The average accuracy of the proposed method for action recognition is up to 92.6% on UCF sports dataset.**

Keywords-human action recognition; multi-level feature; bag-of-words

I. INTRODUCTION

Human action recognition based on videos is a hot topic in the field of computer vision, and it has extensive applications and potential economic values in the area of human computer interaction and video indexing. The main task of action recognition is to process and analyze the original image sequences, to learn and understand human action or behavior. The crucial issues of action recognition are video feature extraction and description, which will affect the results of the action classification significantly. Many feature extraction methods have been proposed, in which features can be summarized as global features and local features [4]. Local features are a more popular way for representing human actions, which achieve the advanced results for action recognition when combined with a bag-of-features description [1]. For example, Wang et al. [1] gained 88.2% accuracy in UCF sports dataset. However, they still have limitations on real-world video owing to complex backgrounds and high intra-class variances.

In this paper, an efficient multi-level feature fusion method for action recognition is introduced and the flowchart is shown in Figure 1.

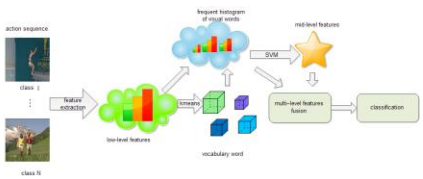

FIGURE I. FLOWCHART OF MULTI-LEVEL FEATURES FUSION FOR ACTION RECOGNITION.

Firstly we extract low-level features such as trajectories and HOF (histograms of optical flow). The trajectories are gained by tracking densely sampled points using optical flow fields. Since HOG (histograms of oriented gradients) and MBH (motion boundary histograms) are the effective descriptors on a variety of datasets, they will be used to describe our dense trajectories.

In many cases, human actions can be recognized not only by the moving human, but also the properties of the specific natural scenes, which can give us associatively contextual information about the action scenes. In our work, the action scenes are employed and modeled using the SIFT feature from randomly selected frames in the videos.

Due to different action classes may often share similar motion patterns of a part of the body and such class correlations among different action classes can be used to distinguish different actions. Therefore, probabilities that the video belongs to each action class are directly applied to represent the mid-level class correlation feature.

A multi-level feature fusion method is utilized to achieve higher recognition accuracies, and the fusion feature vector is formed by concatenating feature vectors of trajectories, SIFT, HOF and the mid-level class correlation.

This article is organized as follows. Section 2 reviews related works on human action recognition. In section 3 we introduce multilevel features for action description. Finally, experimental setups and results are presented in scetion4.

II. RELATED WORK

Human action recognition plays a key part in computer vision in recent years and how to extract identifiable and valid features have become a crucial problem. As mentioned previously, features could be mainly divided into two categories: global features and local features.

The global features such as silhouette-based tracking [2] or a motion shape template [6] cannot effectively describe actions which contain complex and vibrational backgrounds. To solve these problems, local features have been used to represent actions. Laptev and Lindeberg [3] proposed space-time interest points alike to the Harris detector. Ever since, other space-time interest points based methods [4-5] have extensively used for action recognition. In addition, local features in combination with a bag-of-features representation achieve effective results for action recognition [7-9]. Wang et al. [1] proposed an efficient approach to extract dense trajectories, and then they utilized bag-of-features to describe trajectories. Kovashka and

Grauman [8] exploited multiple bag-of-words models to represent the layers of space-time cubes at different scales.

Furthermore, the class correlation feature has attracted many researchers of computer vision. Xu and Chang [10] introduced concept score to represent mid-level semantic features. Liu et al. [11] utilized both data-driven and human specified attributes to characterize human actions via semantic concepts. Parikh and Grauman [12] explored relative attributes to obtain multiple semantic relationships which have more abundant information for video recognition.

III. MULTILEVEL FEATURES FOR ACTION DESCRIPTION

Motion information can't be described precisely by only using a kind of specific feature, thus multi-level feature fusion becomes popularly. In this paper, the mid-level class correlation feature and the low-level features which include three trajectory features, SIFT and HOF are extracted.

A. Low-level Features

There are many methods to extract trajectories. For example, Messing et al. [7] described the information of trajectories by using Lucas-Kanade optical flow to track Harris 3D interest points. Wang et al. [1] firstly sampled feature points in dense grids, and then the trajectories were obtained by tracking them using densely optical flow. In our paper, the method proposed by Wang et al. [1] is employed to extract trajectories.

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 discretising the 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. Figure 2 shows dense trajectory description.

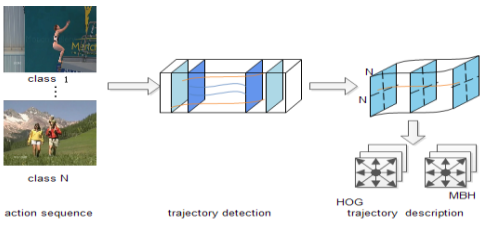

FIGURE II. DENSE TRAJECTORY DESCRIPTION.

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.

Furthermore, SIFT features are utilized to represent action scenes. To reduce the computational complexity, we extract SIFT from randomly selected 20% frames instead of all frames in the video.

B. Bag-of-features

The vector dimensions of some features are different for different video, which is disadvantageous for classification using SVM. To address this issue, bag-of-features is utilized to describe the actions by a set of visual words.

Firstly, to build a vocabulary for each descriptor (trajectories, HOG, MBH, SIFT), we cluster features using k-means and regard each cluster as a visual word of the vocabulary. Then the features are classified by selecting closest to the vocabulary word. Finally the video descriptors are represented by histograms of visual vocabulary word occurrences.

$$H^l = \left[h_1^l, h_2^l, \mathrm{L}\ , h_n^l\right] \tag{1}$$

Where n denotes the number of visual vocabulary words and h_i^l is the frequency of ith vocabulary word occurrences in lth descriptor.

C. Mid-level Class Correlation Feature

The recognition capability of the low-level visual features is limited. Therefore, we introduce the mid-level class correlation feature which represented by relationships among different action classes. Since different action classes may often share similar motion patterns of the human body parts such as walking and running, and such class correlations among different classes can be used to distinguish different actions. The mid-level class correlation feature is described by a group of probabilities of the videos belonging to each action class which are determined by the pre-learned classifiers of the corresponding action classes.

To capture class correlation feature, an independent SVM, which is used for producing probability of the video belongs to each action class, is learned for each action class using the low-level features. The mid-level class correlation feature is then described by $M^l = \left[m_1^l, m_2^l, \mathrm{L}\ , m_C^l\right]$, where m_i^l represents probability of the video belonging to ith action class from the lth kind of low-level feature and C is the number of action classes.

In order to achieve higher accuracies for action recognition, we employ multi-level feature fusion method. The fusion feature vector can be formed by concatenating all features, when all the low-level features and the mid-level class correlation feature are learned.

IV. EXPERIMENT

In this section, we will evaluate our method for human action recognition on the UCF sports dataset which contains 150 videos with 10 different types of sport actions.

When computing trajectory descriptions, we set $N = 32$, $n_\sigma = 2$ and $n_\tau = 3$. Trajectory length and sampling step size are fixed to $L = 15$ and $W = 5$ respectively in our experiment. Table 1 lists the recognition accuracies employing different features including three trajectory features (trajectory, HOG, MBH), HOF, SIFT and fusion of five kinds of low-level visual features based on both low-level features and multi-level

features (i.e. combination of the low-level features and the mid-level class correlation feature).

TABLE I. ACTION RECOGNITION ACCURACIES USING DIFFERENT FEATURES.

features	Low-level	Multi-level
trajectory	68.3%	73.2%
HOG	75.6%	78%
MBH	75.6%	78%
HOF	65.85%	70.7%
SIFT	70.7%	70.7%
fusion features	89.51%	92.68%

From Table 1, we get the observations as follows.

(2) Although a kind of feature alone is less effective performance, the association of five types of features can universally achieve improved results.

(3) The low-level visual features, with combination to the mid-level class correlation feature can enhance the discriminating accuracies in most cases, which proves the effectiveness to improve recognition accuracies when using relationships among action classes.

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 dataset are shown in Table 2.

TABLE II. ACTION RECOGNITION ACCURACIES OF DIFFERENT METHODS ON THE UCF SPORTS DATASETS.

methods	Recognition accuracy
Proposed method	92.68%
Wang et al.[13]	85.6%
Wu et al. [9]	92.48%
Yu Kong et al.[14]	88.8%
Kovashka and Grauman[15]	87.27%
Le et al. [16]	86.5%

The average accuracy of the proposed method is higher than that of the approach in [1], which use the same dense trajectory features. Furthermore, the result proves that our method is comparable to the state-of-the-art methods.

V. CONCLUSION

This paper has proposed an efficient method for human action recognition involving feature extraction and description of the features before the classification. Five types of low-level visual features are extracted, and the bag-of-words model is used to describe features better. To employ the relationships among different classes for action recognition, we introduce the mid-level class correlation feature which is represented by using a group of probabilities of the videos belonging to each action class from the pre-learned classifiers of the corresponding action classes. In addition, multi-level features which concatenate the low-level visual features and the mid-level class correlation feature are used to SVM multiple classifier for human action recognition. The experimental results prove that the proposed approach in this paper is more excellent than other state-of-the-art methods.

ACKNOWLEDGEMENT

In this paper, the research was sponsored by National Key Technologies Research and Development Program of China (Project No. 2013BAD15B06).

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] Weinland, D. & Boyer, E., Action recognition using exemplar-based embedding. *Proc. of IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, INSPEC 10140035: 1-7

[3] Laptev, I., on space-time interest points. *International Journal of Computer Vision*, 64(2-3), pp. 107-123, 2005.

[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] Gorelick, L., Blank, M., Shechtman, E., Irani, M. & Basri, R., Actions as Space-Time Shapes. *IEEE trans. pattern analysis and machine intelligence*, 29(12), pp. 2247-2253, 2007.

[7] Messing, R., Pal, C., & Kautz, H., Activity recognition using the velocity histories of tracked keypoints. *Proc. of Int. Conf. on Computer Vision*, pp. 104-111, 2009.

[8] Kovashka, A. & Grauman, K., Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2046–2053, 2010.

[9] 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.

[10] Xu, D., & Chang, S.F., Video event recognition using kernel methods with multilevel temporal alignment. *IEEE trans. pattern anal. mach. intell.*, 30(11), pp. 1985-1997, 2008.

[11] Liu, J., Kuipers, B. & S. Savarese, Recognizing human actions by attributes. *Proc. of IEEE Conf. on CVPR*, pp. 3337–3344, 2011

[12] Parikh, D. & Grauman, K., Relative attributes. *Proc. of IEEE Conf. on International Conference on Computer Vision*, pp. 503–510, 2011.

[13] 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.

[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.