ACTION RECOGNITION BASED ON SEMANTIC FEATURE DESCRIPTION AND CROSS CLASSIFICATION

Yang Zhao[1,3], *Qi Wang*[2,*], *Yuan Yuan*[1]

[1]Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.

[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

[3]Graduate University of the Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049, P. R. China.

ABSTRACT

Action recognition is a challenging topic in computer vision. In this work, we present a novel method for action recognition which is based on two claimed contributions: semantic feature description and cross classification. The designed descriptor is combined by several local 3D-SIFT and is informative and distinctive, reflecting the spatio-temporal clues of the video. The cross classification effectively combines the feature localization and action categorization together. The proposed method is justified on a popular dateset named UCF50 and the experimental results demonstrate that our method outperforms the state-of-the-art competitors.

Index Terms— action recognition, semantic feature, 3D-SIFT, cross classification

1. INTRODUCTION

Action recognition is a hot topic in compute vision, which focuses on identifying specific action types occurring in a video sequence. Recently, more and more people turn their attentions to this field and try the best to design an automatic recognition algorithm. The developed algorithms can facilitate many useful applications in real life. For instance, Microsoft has put forward a novel peripheral named Kinect which aims to interact with people. One of its critical core is the action recognition function that is responsible for understanding the human behavior. Now, Kinect has been widely used in motion sensing games and shows us a new world of exciting experience[1]. Therefore, it's not hard to find the action recognition technology is playing an increasingly more important role in many fields of people's daily life.

However, human action is extremely complicated because of the large variations in human appearance, posture and body size within the same class [2]. Various factors such as cluttered background, occlusion, camera movement and illumination change also have a powerful effect on action recognition [3] [4]. Under these conditions, to accurately recognize an action from others is still a big challenge.

Traditional action recognition technology consists of two major parts: feature description and feature classification. Feature description is the process of describing the features of an action in a mathematical way. Feature classification then sorts them into different categories. Though many methods have been proposed recently, there are still limitations mainly focusing on the descriptor construction and classification strategy. Motivated by these two points, this work proposes a method based on semantic feature description and cross classification.

1.1. Previous Work

Many methods have been presented to advance the action recognition performance in the past decades. There are generally two stages for the processing, feature description and feature classification. As the first step, feature description always determine whether the result is good or not. Recent researches pay more attention to local features which show a better performance than global features. For example, the SIFT descriptor by Lowe [5], the SURF descriptor by Bay *et al.* [6], the Saliency by Wang *et al.* [7] [8] and the HOG descriptor by Dalal and Triggs [9] are popular ones that have been proved to be effective. However, to describe an action without considering time is unreasonable. Therefore, more literatures focus on making descriptors with space-time correlation. For this purpose, SURF feature is extended to spatio-temporal domain by Willems *et al.* [10]. Laptev *et al.* [11] combine local HOG and temporal HOF descriptors. Scovanner *et al.*[12] propose 3D-SIFT to apply the SIFT in video cubes. However, most of these methods ignore the semantic features such as the information of human body structure, which will alternatively be very important for action recognition.

After extracting the features from an observed video, peo-

This work is supported by the State Key Program of National Natural Science of China (Grant No. 61232010), and the National Natural Science Foundation of China (Grant No. 61172143, 61379094 and 61105012).

ple always deal with human action recognition as a classification problem. Traditional classifiers such as SVM [13], RVM [14] and boosting [15] have been successfully used here. For example, Wang *et al.* [16] generate the one-against-rest classification model via non-linear support vector machine to distinguish one action from others. However, these methods only consider the feature difference between actions, but it should be noted that the features at different human parts might contribute unlikely to the action distinction.

Inspired largely by the two limitations of the aforementioned feature description and classification processes, we try to construct a novel descriptor which can describe the space-time property in specific semantic areas. At the same time, we try to locate these features and classify them through semantic cross classification, instead of treating them equally.

1.2. Overview of The Proposed Method

We propose a novel action recognition method taking advantages of a semantic feature descriptor and a cross classification model. For the feature description, each human body is divided into three semantic parts, each of which is then described separately. After that, two sets of classification models are learned for the further cross classification. The first set is the vertical classifiers for identifying the specific semantic part within one type of action video. The second set is the horizontal classifiers for distinguishing the same kind of semantic part among different action types.

There are mainly three steps for the proposed method. First, three semantic features are defined and located to distinguish actions effectively. Each feature is actually a semantic region on the human body (e.g., head or leg) and is enough to highlight the difference between the actions. Second, in order to effectively describe it, we employ the space-time information simultaneously. Instead of treating the feature as a whole, we describe it as a concatenation of a series of sub-part descriptions. All the sub-parts are expressed by 3D-SIFT respectively and then they are arranged with a specific order. At last, we utilize the vertical classifiers to select the specific semantic features for the examined action type. Then for the same type of semantic feature within one video sequence, horizontal classifiers are employed to classify the action. Fig. 1 shows the detail of our method.

1.3. Contribution

Although the local space-time feature has been used in many methods and most of them have good performance, the proposed one in this paper is distinguished with them in the following two aspects, which also make the main contributions of this paper.

- Design a novel semantic feature descriptor. Most of the traditional local space-time feature descriptors ignore the structural information between parts. The descriptor we proposed is sensitive to the relative location of human parts and has certain semantic clues within it. With this descriptor we can distinguish one action from another via semantic analysis. Our experimental comparison with several popular methods utilizing other descriptors (such as SIFT, SURF) proves that the presented description is more effective.

- Propose an accurate cross classification strategy. The state-of-the-art classification methods mainly focus on classifying the feature descriptors without locating the feature areas. In this paper, we choose the most valuable semantic feature areas through the vertical classifiers and then the actions are recognized with the horizontal classifiers. The two classification processing ensures the quite valuable semantic features and the highly effective recognition performance. This strategy is novel and has never been used before. Experiments show that this framework can bring an averaged 12.71% accuracy increase to the competitors.

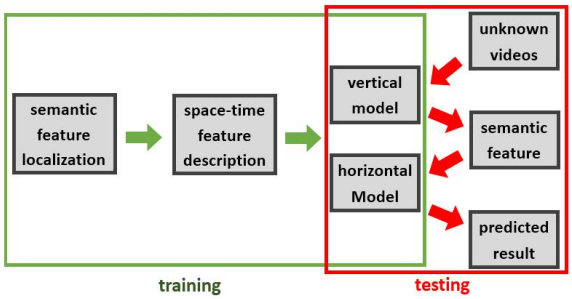

Fig. 1. Flowchart of the proposed method.

The rest of the paper is organized as follows. Section 2 presents our semantic feature descriptor which is based on local space-time exploration. Section 3 describes the action recognition with two novel classification models. Section 4 gives the experimental results to justify the effectiveness of the proposed method. Section 5 concludes the paper.

2. SEMANTIC FEATURE DESCRIPTOR

The description of feature plays an important role in action recognition. In this work, we try to choose the location which is valuable and rich in semantics, and then construct a descriptor to store the semantic information of action.The flowchart of this procedure is illustrated in Fig. 2 and the detailed procedure is presented as follows.

2.1. Locate the semantic feature area

In order to describe a human action with the most valuable characteristics, we use three semantic features which are manually defined in the training stage and automatically detected

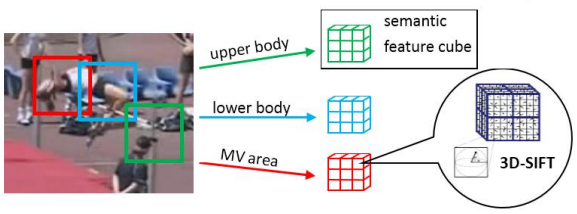

Fig. 2. Illustration of the semantic feature definition. We define three semantic features to describe one action, such as upper body, lower body and MV area in this work. Each feature is a spatio-temporal cube consisting of 3×3 sub-cubes, which are described by a series of 3D-SIFT. The final description is a concatenation of the 9 3D-SIFT descriptors.

in the testing stage. They are the upper body, lower body and the most valuable (MV) area of the activity. For example, riding a bicycle can be divided into handing on the handlebars, pumping the pedals and turning a wheel. In fact, any kinds of semantic features are reasonable, only if they are able to distinguish from the other actions. The semantic features between two kinds of actions can be different, but in the same kind of action, we must define them under the same principle. Once the semantic features of a specific action are decided on the examined t^{th} frame, several subsequent frames are also attached to this frame. The obtained 3D cube is denoted by the center of the t^{th} frame $C_i^t = (cx_i^t, cy_i^t)$. The whole cube is set as $60 \times 60 \times 15$ in pixels and we sample the features every 15 frames for an input video. This parameter setting is a empirical choice that balances computational feasibility and performance accuracy.

2.2. Describe the semantic feature area

Once the area of the semantic feature is decided, the next step is to construct a corresponding feature descriptor. To make the descriptor more informative and discriminative, the whole cube is further divided into $3 \times 3 \times 1$ sub-cubes, each of which contains $20 \times 20 \times 15$ pixels. The obtained sub-cubs are separately described by 3D-SIFT as $X_k = (x_1^k, x_2^k, ..., x_D^k), k \in [1, 9]$, where D is the dimensions of descriptor and then they are concatenated together $\mathbf{X} = [X_1, X_2, ..., X_9]$. Instead of taking the cub as a whole, this strategy can reflect the variability of the space-time cube and can provide more decisive clues for the classification.

3. CROSS CLASSIFICATION

In this part, the details of the proposed cross classification are introduced. There include two procedures as illustrated in Fig. 3, vertical classification and horizontal classification. The vertical classification model is defined for every kind of action and is used to detect the semantic features within each type. The horizontal model is also a multiclass classification

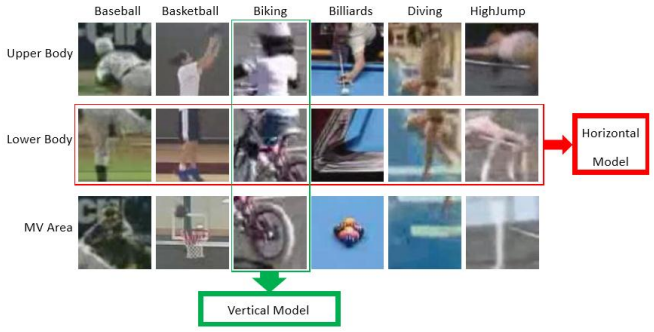

Fig. 3. Illustration of the cross classification model. Three semantic features are defined and trained to get the longitudinal models for one specific action. For one kind of semantic feature, one horizontal model is also learned to classify the actions.

procedure that utilizes the semantic features of one kind to distinguish the action types.

3.1. Vertical Model

The vertical model is the most distinguished part in our method. It uses the learned classifiers to identify different features for one action type. That means for each action type, 3 classification models are learned to distinguish the 3 features. In the training stage, the features are labeled manually and described by the concatenated 3D-SIFT. Then the classification model is learned by multiclass support vector machines (SVM)[17]. The SVM can be treated as an optimization problem showed in Eq. 1

$$\begin{aligned} \min_{\mathrm{w},b,\xi} \quad & \frac{1}{2}\mathrm{w}^T\mathrm{w} + C\sum_{i=1}^{l}\xi_i, \\ subject\ to \quad & y_i(\mathrm{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \end{aligned} \tag{1}$$

where $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_i)$ is the kernel function donoted by Eq. 2

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0. \tag{2}$$

In the testing stage, we first sample several seed points in the examined frame and then a corresponding 3D cube the same size as the previously defined semantic feature is constructed around each point. After describing the acquired cube, the previously learned vertical models are applied to determine the feature types. The detected three features with the highest probability is stored for the further action prediction. This process continuously repeated every 15 frames until the input video is finished.

Methods	tS	HGS	HFS	MBH	AB	MS	DT	proposed
Accuracy	67.2%	68.0%	68.2%	82.2%	76.4%	88.0%	85.6%	89.22%

Fig. 4. Accuracy comparison of the proposed method and the competitors.

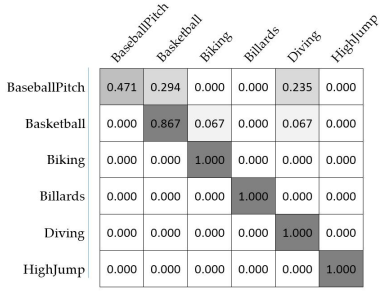

Fig. 5. Confusion matrix of the proposed method.

3.2. Horizontal Model

The horizontal model plays like traditional action recognition because it directly uses features to make prediction. But it is enforced by a three-round decision. Since there are three kinds of semantic features for every action, three horizontal models should be learned. The final result considers all these three results and returns the most likely label through a WTA (Winner-Take-All) voting strategy. The training and testing procedure is quite similar to the aforementioned step and no more introductions are given.

In general, the training process is conducted beforehand with a manually labeled ground truth. After that, when we get an unknown video, the vertical models are first employed to obtain the three kinds of semantic features. Then the horizontal models are applied to predict which kind of actions it belongs to. The prediction is repeated three times on the three feature types and the final decision is made by combining them with a WTA principle.

4. EXPERIMENTS

We employ a popular dataset named UCF 50 provided by [18]. It has 50 action categories, consisting of real-world videos taken from the YouTube website. There are 25 to 30 groups in each type of action and the video clips in the same group are similar. All the videos in this dataset are resized to 960×720. In our test, we choose four video clips in each group to be training data and the others are used to test our method. Six actions are selected in UCF50: BaseballPitch, Basketball, Biking, Billards, Diving and HighJump. Three types of semantic features are defined in each kind of action, as well as three vertical models. Six horizontal models for inter-category classification are also constructed.

Firstly, we compare our method with several traditional methods [19] which are combined by the bag-of-features and support vector machine such as HOG + SVM (HGS), HOF + SVM(HFS) and trajectories + SVM(tS). And then several state-of-the-art methods are chosen to compare with our method. For example, Dalal *et al.* [20] proposed the motion boundary histograms descriptor(MBH) which computes derivatives for both horizontal and vertical components of optical flow. Wang and Schmid [21] propose an improved dense trajectory and it shows a good performance on many datasets(DT). Sadanand and Corso present a method of high-level action representation named action bank[22](AB). Wu *et al.* [23] use multilevel features and latent structural SVM to recognize the actions (MS).

Fig. 4 presents the recognition results of different methods evaluated by accuracy. We can see clearly that the traditional methods perform poorly with a low accuracy rate (no more than 70%), while the state-of-the-art methods are all above 75%. But the proposed method outperforms all the competitors, with an accuracy of nearly 90% and averaged increase of 12.71%. In order to see how the proposed method performs on the six actions of the dataset, we also present its confusion matrix in Fig. 5. It is manifest that almost all the recognition results achieve outstanding performance. This success is mainly due to the semantic feature description and cross classification. The semantic features and the vertical classification enable the obtained descriptor more distinctive and decisive, because they focus on specific semantic areas. The horizontal classification fuses the results from three separate feature types to make the final decision. This is more reasonable than with a single feature. All these factors together ensure the effectiveness of our method.

5. CONCLUSION

In this paper, we propose a novel semantic feature descriptor based on 3D-SIFT and a cross classification strategy of vertical and horizontal models. The descriptor is distinctive and informative, because of its focus on particular semantic areas. The cross classification strategy is also critical because it ensures the precisely localized semantic features are utilized for action prediction. Thanks to the contributions of the two steps, the action recognition performance has been advanced greatly. To be specific, the proposed method can predict the action types more accurately than state-of-the-art methods on a difficult dataset that is publicly available, with an averaged accuracy increase of 17%.

In the future, more advanced spatio-temporal connections will be considered in the action recognition process to properly describe the evolving characteristics of an action. Besides, the tracking process will be added in the training stage to facilitate the manual labeling.

6. REFERENCES

[1] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, 2014.

[2] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 489–496.

[3] Z. Jiang, Z. Lin, and L. S. Davis, "A unified tree-based framework for joint action localization, recognition and segmentation," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1345C1355, 2013.

[4] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE T. Cybernetics*, vol. 43, no. 2, pp. 660–672, 2013.

[5] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, 1999, pp. 1150–1157.

[6] H. Bay and A. Ess, "Surf: Speeded up robust features," vol. 110, no. 3, pp. 346–359, 2008.

[7] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 34–41, 2013.

[8] Q. Wang, Y. Yuan, and P. Yan, "Visual saliency by selective contrast," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 23, no. 7, pp. 1150–1155, 2013.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[10] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conference on Computer Vision*, 2008, pp. 650–663.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 708–721.

[12] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM Conference on Multimedia*, 2007, pp. 357–360.

[13] C. Schuldt, I. Laptev, , and B. Caputo, "Recognizing human actions: A local svm approach," in *International Conference on Pattern Recognition*, 2004, pp. 556–561.

[14] M. E. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems 12*, pp. 652–658.

[15] Y. Ke, R. Sukthankar, , and M. Hebert, "Efficient visual event detection using volumetric features," in *IEEE Conference on Computer Vision*, 2005, pp. 166–173.

[16] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009, pp. 124.1–124.11.

[17] C.-C. Chang and C.-J. Lin, "Libsvm : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2011.

[18] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," in *Machine Vision and Applications*, 2012, pp. 1–11.

[19] H. Wang, A. Kläer, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," in *International Journal of Computer Vision*, 2013, pp. 60–79.

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

[21] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, 2013.

[22] C. Liu, J. Yuen, and A. Torralba, "Action bank: A high-level representation of activity in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368–2382, 2011.

[23] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 8, pp. 1422–1431, 2013.