Action Recognition with Discriminative Mid-Level Features

Cuiwei Liu, Yu Kong, Xinxiao Wu, Yunde Jia
Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, China
{*liucuiwei, kongyu, wuxinxiao, jiayunde*}*@bit.edu.cn*

Abstract

Most of the existing action recognition approaches employ low-level features(e.g., local features and global features) to represent an action video. However, algorithms based on low-level features are not robust to complex environments such as cluttered background, camera movement and illumination change. In this paper, we present a novel random forest learning framework to construct a discriminative and informative mid-level feature from low-level features. Since a single low-level feature based representation is not enough to capture the variations of human appearance, multiple low-level features (i.e., optical flow and histogram of gradient 3D features) are fused to further improve recognition performance. This mid-level feature is employed by a random forest classifier for robust action recognition. Experiments on two publicly available action datasets demonstrate that using both the mid-level feature and the fusion of multiple low-level features leads to a superior performance over previous methods.

1. Introduction

How to represent an action video with discriminative features is very important in human action recognition. An expressive representation can provide sufficiently discriminative information to the classifier, and significantly improves the recognition performance. Video representations of an action can be roughly divided into two categories: global representation and local representation. Global descriptors [3, 10] extract visual information from videos to describe the entire human figure. The large-scale global features are powerful and abundant since much of the information is encoded, but they are sensitive to noise. Local representations [5, 15] describe a video by using a set of primitive cuboids, each of which is calculated around a detected interest point. Compared with global representations, local

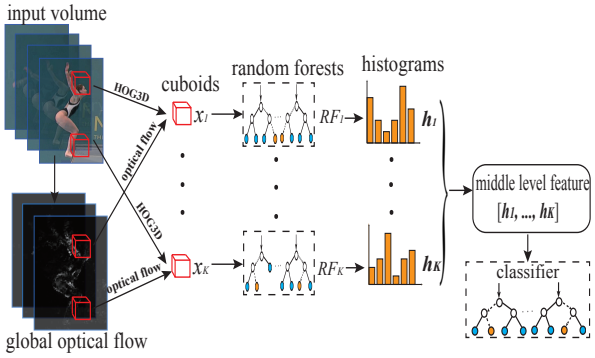

Figure 1. Overview of our method.

representations are more robust to partial occlusion and global deformation, but they depend on the reliable detection of accurate and informative interest points. Recent work [13, 7] have shown that the fusion of multiple features (e.g., local feature and global feature) is robust to video variations and benefits classification problems, but it still relies on low-level features. How to bridge semantic gap between low-level features and high-level action categories is a key problem in action recognition.

In this paper, we propose a novel random forest learning framework to build a compact mid-level representation based on multiple low-level features for action recognition. Our method starts from 3D cuboids of input action videos, and characterizes each cuboid by optical flow and histogram of gradient 3D (HOG3D). Low-level optical flow captures the global motion information and low-level HOG3D descriptor describes the local appearance and motion. Then, each cuboid is classified by the corresponding random forest, and its posterior probabilities of all categories are normalized to a histogram. The histograms of all the cuboids are concatenated to generate a compact mid-level representation. At last, random forest is utilized again to train our final classifier. An overview of the proposed approach is illustrated in Fig. 1.

The proposed action recognition algorithm has several appealing features. Firstly, we propose a novel discriminative mid-level feature, and action recognition based on this mid-level feature achieves excellent per-

formance on two datasets. Secondly, multiple low-level features are fused by a random forest based learning framework to describe human action. Lastly, different from [4] which is designed for two-class problem, our method can handle the multi-class problem naturally.

2. Our Method

2.1. Low-Level Feature

Two types of low-level features(i.e., optical flow [3] and HOG3D [5]) are merged to describe the motion and appearance of the interested human. Optical flow is calculated from the entire human figure to capture global motion information. This motion descriptor shows favorable performance with noise, so it can tolerate the jitter of human figures caused by human detector or tracker. HOG3D descriptor is extracted from each cuboid to characterize the local motion and appearance information. The large-scale global motion feature is powerful and abundant by encoding much of the information, whereas the local spatial-temporal feature can deal with noise and partial occlusion. Therefore the representation based on both of them is robust to variations and beneficial to action recognition.

2.2. Mid-Level Feature

In this section, we describe the proposed random forest learning framework which combines multiple low-level features to construct a mid-level representation with more discriminative power. Random forest has garnered growing interest in the computer vision community, and has been successfully applied to recognition and classification tasks [8, 16]. A random forest is an ensemble of decision trees, and classifies samples by majority voting over the outputs of all trees. In this paper, random forest is used to transform the low-level features into a discriminative mid-level representation.

In this work, a human body detector or tracker is required to extract the human-centered figures from a video. Each video is segmented to a set of fixed-size sub-volumes for further processing. Suppose that each sub-volume generates a group of cuboids, denoted as $\{x_k = (x_k^{OF}, x_k^{HOG3D}), k = 1 : K\}$, where x_k^{OF} is the optical flow feature of the local cuboid x_k sampled from the global optical flow, and x_k^{HOG3D} represents the HOG3D feature calculated in the local cuboid x_k. At each dense sampling position, a collection of cuboids is extracted from all the training data, then a random forest is built to get the mid-level representations of them.

Each tree in the random forest is independently grown from a bootstrap training set obtained by random sampling in the original training set, using CART methodology [2]. The training samples are dropped down a tree from the root. In order to split a node and the training data reaching the node, a binary number δ is randomly generated to decide which kind of low-level feature (i.e.,optical flow or HOG3D) is used for node split, and then feature candidates are randomly sampled from the selected type of feature. The node splits into two children nodes according to the feature candidate with the largest information gain, and each data at this node is sent to one of the children nodes. A node stops splitting if a pre-specified maximum tree depth has been reached, or the number of samples assigned to this node is small, or all of these samples belong to the same class. And then this node is treated as a leaf. Instead of marking the leaf node by one of the class labels [2], we use a vector to store the distribution, $P = [p^1, p^2, ...p^M]$, where M is the number of classes, and p^i is the posterior probability of data at this leaf node belonging to class i. The probability p_n^i of leaf node n is given by

$$p_n^i = p_n(y(x) = i) = \frac{N_i}{\sum_{j=1}^{M} N_j}, \quad (1)$$

where N_j represents the number of samples of class j arriving at node n. Posterior probability p_n^i is evaluated as the ratio of the number of samples of class i and the total number of samples that reach this node.

During testing, a sample x is passed down to each of the T trees. Suppose that sample x reaches leaf node $l(t, x)$ of tree t , then x is classified to the class with the largest average of the posterior probabilities of all trees:

$$\overline{y}(x) = \arg\max_i P_i(x) = \arg\max_i \frac{1}{T} \sum_{t=1}^{T} p_{l(t,x)}^i, \quad (2)$$

where $P_i(x)$ is the average of the posterior probabilities of class i and constitutes a good local descriptor. For a testing sample x , the posterior probabilities of all classes are normalized to a histogram: $h(x) = [\overline{P}_1(x), \overline{P}_2(x), ...\overline{P}_M(x)]$. However, there exists an overfitting problem for training samples that posterior probability of the true class is very close to 1. Hence, mid-level representations of training samples constructed in this manner are quite different from those of testing samples, and a classifier built on these mid-level representations would not achieve good performance.

In this paper, we introduce the out-of-bag estimate to construct the mid-level features for training samples. Out-of-bag estimate is a performance evaluation criterion characteristic of random forest. To build a tree, about 1/3 of the training samples are left out of the bootstrap training set. For a training sample x, the ensemble of trees not trained on x constitutes a new classifier, called the out-of-bag classifier. Then the out-of-bag estimate for the generalization error is the error rate of the

out-of-bag classifier on the training set. According to Breiman [2], empirical evidence shows that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. Therefore, for a training sample x, the out-of-bag classifier substitutes for the whole forest to get a posterior probability histogram $h(x)$.

The above process is done for each sampling position separately. Histograms of all the K cuboids extracted from an input volume v are concatenated to construct the mid-level feature of v : $f(v) = [h(x_1), h(x_2), ..., h(x_K)]$.

2.3. Final Classifier

With the mid-level feature $f(v)$, random forest is utilized again to train our final classifier because it is able to learn multiple classes discriminatively. As described in Section3.2, random forest classifies a sample according to the average of the posterior probabilities of all trees. The random selection of the feature type is eliminated, since only mid-level feature is employed.

3. Experiments

Our method is evaluated on the Weizmann action dataset [1] and the UCF sports dataset [9]. The Weizmann dataset is composed of 90 videos acted by 9 different people, each performing 10 actions. We employ leave-one-person-out cross-validation scheme, and report the average accuracy. The UCF sports dataset contains 149 sports sequences of 10 actions. To increase the amount of data samples, we add a horizontally flipped version of each sequence to the dataset following the work of [12]. We totally have 149 pairs of videos, 127 pairs of which are randomly selected for training, leaving the rest 22 pairs for testing. We randomly split the dataset for seven times and report the average results.

In this paper, we extract cuboids ($16 \times 16 \times 9$) from the fixed-size figure volume ($80 \times 40 \times 10$) at regular positions, and the stride between two adjacent cuboids is 5 pixels. Due to the randomness of random forest, we run the final classifier for thirty times with a group of fixed parameters, and report the average results, and report the average results.

3.1. Results on Two Datasets

Table 1 lists the action recognition results of our method compared with the previous work. Our method gets the accuracy of 100% on the Weizmann dataset with comparable results with other state-of-the-art methods. For the UCF dataset, our method outperforms the previous work, and the standard deviation of our results is 1.14%. In addition, all methods use different testing schemes. Evaluations are done with a five-fold cross-validation in [16], while leave-one-out cross-validation scheme is adopted in [11, 6, 12, 9]. In order to reduce the computation complexity, we use 85% of the dataset for training and the rest 15% for testing. Hence our model is trained on less data than that of [11, 6, 12, 9], yet obtaining better results.

Table 1. Recognition accuracy.

Method	Weizmann	UCF
Wang et al. [11]	–	88.2%
Kovashka et al. [6]	–	87.27%
Yao et al. [16]	95.60%	86.60%
Wang et al. [12]	–	85.60%
Rodriguez et al. [9]	–	69.20%
Wu et al. [14]	98.9%	–
Fathi and Mori [4]	100%	–
Our method	100%	90.10%

3.2 Evaluations of the Mid-Level Feature

The proposed method can be considered as a two-stage random forest framework, in which the first stage fuses low-level features into the mid-level feature and the second stage takes the mid-level feature as the input of final classifier. To investigate the discriminative power of the mid-level feature, we compare our two-stage method with the one-stage method only using low-level features on the Weizmann dataset, and the results are listed in Table 2. Our two-stage method achieves higher performance than the one-stage method, which demonstrates the efficacy of the mid-level feature.

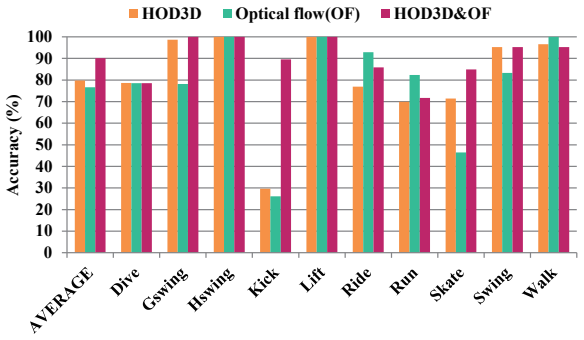

Figure 2. Comparison of feature fusion and single feature on the UCF dataset.

Evaluations of feature fusion are done on the UCF dataset. Fig. 2 reports the recognition rate of each class combining multiple low-level features compared with that using a single low-level feature. As is shown in Fig. 2, optical flow does well in classifying actions with drastic motion such as "run", while HOG3D is good at distinguishing actions by appearance such as "Golf-swing". Taking "kick" for example, only one foot

Table 2. Comparison of one-stage method with our two-stage method on the Weizmann dataset.

Method	**Average**	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
one-stage	**88.89%**	100%	100%	66.67%	100%	77.78%	77.78%	77.78%	100%	100%	88.89%
two-stage	**100%**	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

moves saliently in this action and it is not distinguishable from "walk" just by shape, thereby algorithms only using optical flow or HOG3D cannot work well. However, our method performs better by merging the two types of low-level features.

3.3. Influence of Parameters

In this section, we investigate the effect of several parameters of random forest on the UCF dataset. Random forests are adopted twice in our method, once for the construction of mid-level feature, and the second time for the final classification.

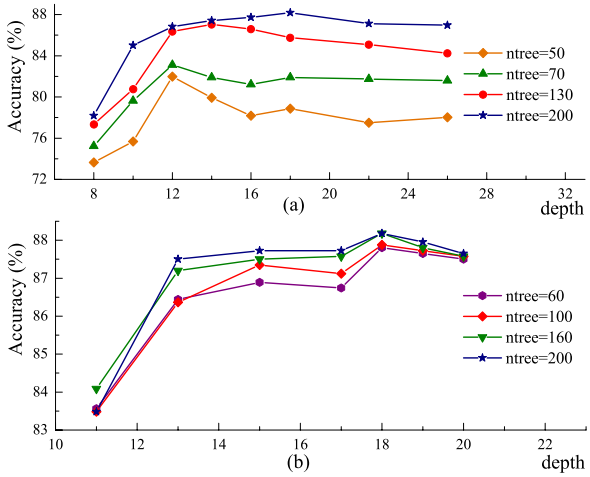

Figure 3. Influence of different parameters on the UCF dataset.

Fig. 3(a) depicts how the parameters of random forest used to build the mid-level feature affect the performance. As is shown in Fig. 3(a), the curves are first increasing and then decreasing due to the overfitting problem. Actually, the depth for which the forest is overfitting depends on the number of trees. For curve with larger number of trees, the maximum accuracy is obtained at deeper depth. Furthermore, it is interesting to observe that more trees can improve the performance.

In Fig. 3(b) we show the quality of the final random forest classifier with different number of trees and different depths. Experiment results are summarized as a function of depth with each curve representing a particular number of trees. We can observe that overfitting begins around depth 18 for all curves. Moreover, it is obvious that increasing the number of trees gradually improves the performance of the classifier.

4. Conclusions

We have presented a random forest based algorithm to learn a discriminative mid-level feature by fusing low-level optical flow and HOG3D features for action recognition. Evaluations on two datasets prove that our method is robust to variations of actions and achieves high performance. In addition, the proposed framework is general for merging any type of low-level features.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

[2] L. Breiman. Random forests. *Machine Learning, vol. 45, no. 1, pp. 5-32*, 2001.

[3] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[4] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.

[5] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[6] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[7] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *CVPR*, 2008.

[8] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Trans. Pattern Anal. Mach. Intell, vol.30, no.9, pp.1632-1646*, 2008.

[9] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[10] C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[11] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[12] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[13] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *CVPR*, 2009.

[14] X. Wu, Y. Jia, and W. Liang. Incremental discriminant-analysis of canonical correlations for action recognition. *Pattern Recognition, vol.43, no.12, pp.4190-4197*, 2010.

[15] X. Wu, D. Xu, L. Duan, and J. Luo. Action recognition using context and appearance distribution features. In *CVPR*, 2011.

[16] A. Yao, J. Gall, and L. V. Gool. A hough transform-based voting framework for action recognition. In *CVPR*, 2010.