Multimed Tools Appl (2017) 76:22195–22212
DOI 10.1007/s11042-017-5038-6

Breaking video into pieces for action recognition

Ying Zheng[1,2] · Hongxun Yao[1] · Xiaoshuai Sun[1] · Xuesong Jiang[1,2] · Fatih Porikli[2]

Received: 30 November 2016 / Revised: 10 July 2017 / Accepted: 14 July 2017 /
Published online: 2 August 2017

Abstract We present a simple yet effective approach for human action recognition. Most of the existing solutions based on multi-class action classification aim to assign a class label for the input video. However, the variety and complexity of real-life videos make it very challenging to achieve high classification accuracy. To address this problem, we propose to partition the input video into small clips and formulate action recognition as a joint decision-making task. First, we partition all videos into two equal segments that are processed in the same manner. We repeat this procedure to obtain three layers of video subsegments, which are then organized in a binary tree structure. We train separate classifiers for each layer. By applying the corresponding classifiers to video subsegments, we obtain a decision value matrix (DVM). Then, we construct an aggregated representation for the original full-length video by integrating the elements of the DVM. Finally, we train a new action recognition classifier based on the DVM representation. Our extensive experimental evaluations demonstrate that the proposed method achieves significant performance improvement against several compared methods on two benchmark datasets.

✉ Hongxun Yao
h.yao@hit.edu.cn

Ying Zheng
zhengying@hit.edu.cn

Xiaoshuai Sun
xiaoshuaisun@hit.edu.cn

Xuesong Jiang
xsjiang@hit.edu.cn

Fatih Porikli
fatih.porikli@anu.edu.au

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

[2] Research School of Engineering, Australian National University, Canberra, Australia

Keywords Action recognition · Video partition · Video representation · Decision value matrix

1 Introduction

As an essential component of video semantic understanding, human activity recognition is an interdisciplinary research area traversing through computer vision, informatics, and cognitive sciences. It aims to identify the underlying actions in a given video automatically. To this end, it extracts inherent information related to human actions, determines the relation between low-level visual features and high-level semantics, and builds models for classification of video clips [2, 27]. Action recognition has an extensive list of applications including video surveillance, human-computer interfaces, augmented and virtual reality, and robotics [42, 55, 56].

Existing methods for multi-class action classification usually assign a class label to a given video as shown in the upper box drawn by the solid line of Fig. 1. Among many issues, viewpoint changes, intra-class variations, and background clutter make action recognition a very challenging task. To tackle these problems, many alternative approaches have been proposed. One branch of previous methods employs the bag of visual words (BoVW) with local spatio-temporal features; space-time interest points (STIP) [22] and improved dense trajectories (IDT) [44]. Others such as mid-level action element [21] and action bank [33] use intermediate or high-level video representations. These techniques report acceptable performance on some datasets yet fail to generalize and provide robust solutions.

Recent developments in the field of human activity prediction have shown that inferring ongoing activities from videos only containing parts of the activities is an achievable task [8, 18, 31]. When only a half of the video is observed, the model proposed by Xu et al. [51] obtains 91.67% accuracy on UT-Interaction Set #1 [32]. Inspired by their idea, we propose to improve the performance of action recognition by breaking the video into parts and

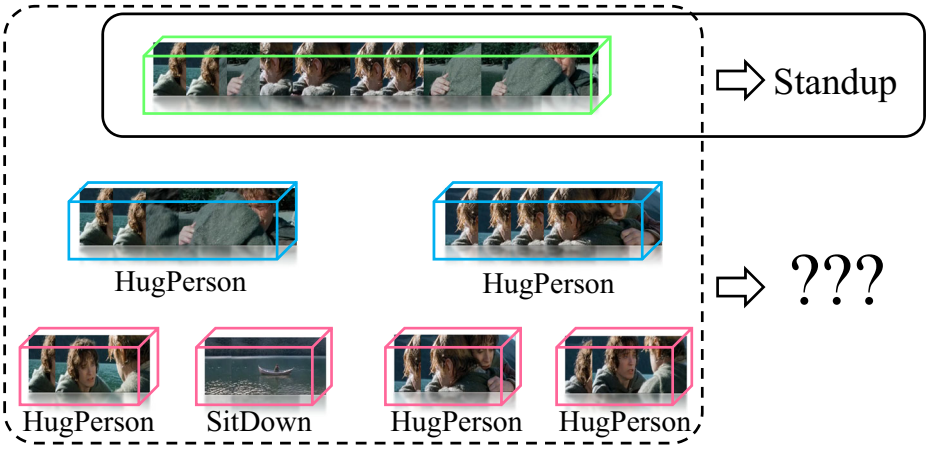

Fig. 1 An illustration of two different ways to recognize action in video. The sample clip, taken from the Hollywood human action dataset [23], shows the action "HugPerson" done by two actors of the movie "The Lord of the Rings: The Fellowship of the Ring"

combining the useful recognition information of these parts with the original full video. Actually, an activity composed of several sub-actions cannot afford more pressure from missing parts than simple action. In other words, it should be easier to recognize the segmented parts of simple action than human activity prediction from partially observed video. As the dashed box in Fig. 1 shows, we apply the classifier to the child clips described by improved dense trajectories [44] and present the predicted class labels of action below each clip. We can see that some of the predicted class labels are right while the label generated on the original video is wrong. It is important to explain that the example does not mean that the recognition performance from parts is higher than full video. What we want to show here is that it can make up for the shortcomings of general methods which output a class label for input video directly.

Based on the observation above, we propose a novel method based on video partition for action recognition. We first cut the original action video into two child parts from the middle and do this on child parts too. Similar to three-level binary tree, we can obtain seven action clips included with the original video as Fig. 1 shows. As the STIP [22] and IDT [44] with the BoVW framework are the most popular and successful methods, we use it to represent action clips. Then we train three classifiers for different layers of the binary tree and use the corresponding classifier to recognize action clips in each layer. After that, we construct a new representation by integrating elements in the decision value matrix (DVM) generated in the last step. Finally, a new classifier will be trained among the proposed DVM representation to fulfill the task of action recognition. The experimental results conducted on two public human action datasets demonstrate the performance of our method. The proposed method has extensive applicability and can be adopted by any feature descriptors, video representation approaches and classifiers.

The remainder of this paper is organized as follows. Section 2 reviews the works related to our research. Section 3 describes details of the proposed approach for action recognition. Section 4 presents the experimental results, gives some reasonable analyses and discussions about the additional results. Finally, we make a conclusion for this paper in Section 5.

2 Related work

Action recognition is a very active field which has been attracting increasing interest in the past decade [54]. Numerous methods have been proposed to tackle this problem from different views. Among these methods, the bag of visual words model (BoVW) with local features has been widely accepted by many researchers. Peng et al. [27] recently provide a comprehensive study of BoVW and three kinds of fusion methods. Our method based on video partition can be seen as one kind of score fusion, which is performed in the video segments level. It is different from most existing works of score fusion, as they fuse the scores from different descriptors on the video level [40]. To the best of our knowledge, this is the first work that explores the ability of video partition to improve the performance of action recognition. Actually because reviewing a large body of such existing works is beyond the scope of this paper, we refer the interested readers to insightful surveys and comparison of methods for action recognition [2, 5, 28, 49]. In this section, we will only introduce some works closely related to our research.

Extracting of space-time interest points is the most popular component in the process of action recognition. Interest points always appear in the locations that action suddenly changes in space and time, which also can be considered as the places containing useful

information to describe the action. There are a great number of methods to extract space-time interest points. The Harris3D detector of space-time interest points (STIP) proposed by Laptev et al. [22] is the most classical one. Because of the severe constraint on detected points, the number of interest points is generally small. Dollar et al. [10] apply the Gabor filter and change the scale of the neighborhood in both space and time to increase the number of detected points. Scovanner et al. [35] extend the bag of words paradigm and introduce a 3D SIFT descriptor for action recognition. Similarly, Klaser et al. [17] propose the HOG3D based on the histograms of oriented 3D spatio-temporal gradients.

Laptev et al. [23] represent each video sequence by histograms of visual word occurrences over a space-time volume. After that, the bag of visual words (BoVW) model with local spatio-temporal features soon became the leading framework and achieve excellent performance on some public datasets [3, 19, 36]. As there are lots of local spatio-temporal features, Wang et al. [45] evaluate and compare some feature detectors and descriptors under the standard BoVW framework. Among all these hand-crafted descriptors, dense trajectory [43] is one of the most popular methods. They combine dense sampling with feature tracking for multiple spatial scales. Taking into account the camera motion, Wang et al. [44] propose the improved dense trajectories (IDT) with fisher vector encoding and obtain the state-of-the-art performance for action recognition. Xu et al. [50] survey three kinds of aggregating methods with dense trajectories for action recognition. In consideration of the high computation cost of IDT, Kantorov et al. [15] design a new motion-based local descriptor which could drastically improves the speed of video feature extraction, feature encoding, and action classification by two orders of magnitude at the cost of a minor decrease in the recognition accuracy.

As one kind of features between local and global feature, the mid-level feature is first applied to the task of image classification and achieve good performance [6]. Soon after that, mid-level based video representation is introduced into the area of action recognition [24, 26]. These types of methods model the middle parts of action that may correspond to key frames and spatio-temporal cubes which can best describe the action or just some parts of object related to the action. Based on low-level representation like local spatio-temporal descriptors and dense trajectories, Raptis et al. [29] propose a mid-level action model learned by treating part-cluster assignments as latent variables and using a graphical model to study the relations between mid-level parts. Jain et al. [13] present a method for video representation based on mid-level discriminative spatio-temporal patches, which can be mined by exemplar-based clustering approach. Zhu et al. [59] introduce a two-layer representation of videos for action recognition, named ‘acton’ which is learned via a max-margin multi-channel multiple instances learning framework.

Unlike previous related work, Zhang et al. [57] propose a strongly-supervised approach which models action as a composition of volumetric patches discovered in a data-driven training process. To represent and recognize complex actions, Wang et al. [46] propose motion atom and phrase which are respectively designed for describing the motion information of short and long temporal scale. Zhou et al. [58] present a new approach by mining discriminative mid-level human-object interaction parts for fine-grained action recognition. Recently, Lan et al. [21] present a hierarchical mid-level action element (MAE) representation for action recognition. These elements are discovered by a discriminative clustering algorithm automatically and encoded in spatio-temporal segments ranging from entire action sequence to action parts. Compared to local descriptors which contain lots of boundaries and corners, mid-level based representation may be more suitable to describe complicated actions for the reason that it includes more semantic information.

It is worth mentioning that deep learning based methods are becoming more mainstream in the field of action recognition with the rapid development of deep convolutional neural network (CNN). In the last few years, deep learning based methods have achieved excellent performance in most of the image related tasks, such as image classification and face recognition. Along with this trend, highlands in the area of video research have been occupied one after another [14, 48, 52], with no exception for action recognition [4, 39]. In general, the training process of CNN model needs a huge number of examples. In view of the facts that current most used datasets for action recognition, such as KTH [34], YouTube action [25], or even larger available datasets like HMDB51 [20], UCF50 [30] and UCF101 [38], are relatively limited to the number of instances and their variety, Karpathy et al. [16] constructed a big video dataset named Sports-1M with amazing million orders of magnitude. In addition to the Sports-1M, there are other large-scale video datasets such as Activitynet [7] and Youtube8M [1].

Simonyan et al. [37] propose a two-stream deep convolutional network which incorporates separate spatial and temporal recognition streams. They obtain competitive performance with the state of the art and made the two-stream ConvNets model accepted by lots of researchers. Wang et al. [47] exploit the improved dense trajectories [44] and two-stream ConvNets [37] to build a new action representation, called trajectory-pooled deep-convolutional descriptor (TDD). Building upon the two-stream architecture but made

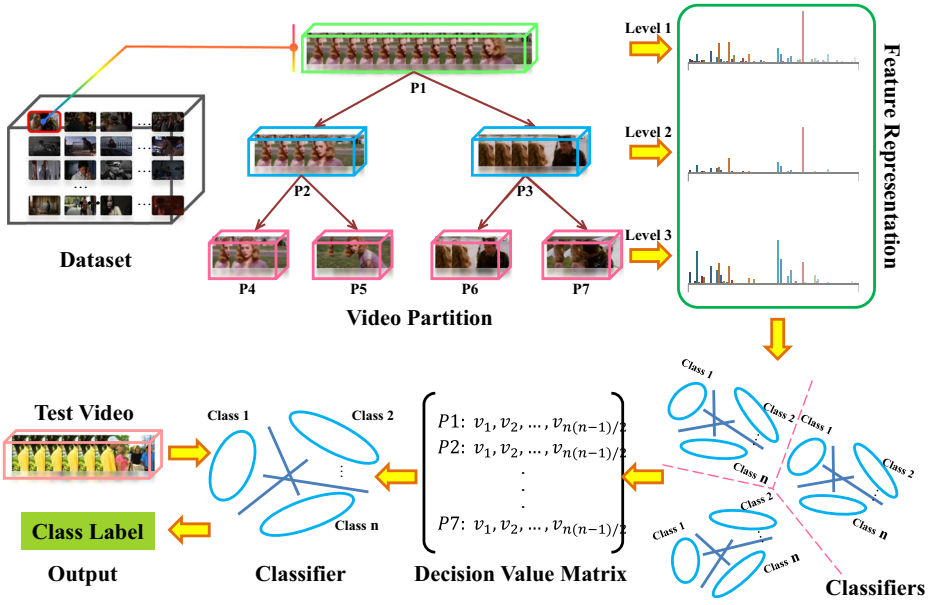

Fig. 2 The pipeline of our method. For each video in the dataset, we first do the process of video partition and obtain two child parts and four grandchild parts which form a three-level binary tree. All the parts are coded by an existing feature representation method. For each level of the binary tree, we train a classifier among the training video clips produced at that level. After that, the classifiers are applied to recognize video clips at the corresponding level and we can get a decision value matrix. Based on the matrix, we construct a new representation for the original video. Eventually, we can obtain a final classifier trained on the new representation and use it to recognize the test video

up for its drawbacks, Feichtenhofer et al. [11] investigate several fusion methods and propose a new ConvNet architecture of spatio-temporal fusion for human action recognition. Tran et al. [41] learn spatio-temporal features using 3D ConvNets trained on large-scale video datasets. Besides, the recurrent neural network architecture embedded with Long Short-Term Memory (LSTM) [12] also has attracted the attention of researchers [53].

3 Method

In this section, we describe the video partition based method for action recognition. As our goal is to discover a generally applicable way to improve the performance of existing approaches, we first give a glimpse of video representation which is the basis of action recognition. Then we present our method of segmenting video into small clips. After that, we present how the video partition can be beneficial to the problem of action recognition. Figure 2 shows the overall pipeline of our method.

3.1 Video representation

The proposed action recognition method is not limited to specified features, representations or classifiers. It can be considered as a general framework that can be implemented by different solutions. The proposed method aims to achieve a higher performance for action recognition based on existing solutions. According to practical experience, we choose the most popular BoVW pipeline embedded with two widely-used local features, name space-time interest points (STIP) [22] with HOG, HOF descriptors, and improved dense trajectories (IDT) [44] with HOG, HOF, MBHx, MBHy descriptors. It should be noted that other approaches like action bank [33] and two-stream ConvNets [37] are also good choices.

Space-time interest points (STIP) Interest points are detected by a space-time extension of the Harris operator for a fixed set of multiple spatio-temporal scales as illustrated in Fig. 3 for the action HandShake. For each interest point, it computes the histograms of oriented gradient (HOG) descriptors and histograms of optical flow (HOF) descriptors of the associated space-time patch. The local descriptors concatenate several histograms from a space-time grid defined on the patch and generalize SIFT descriptor to the space-time domain.

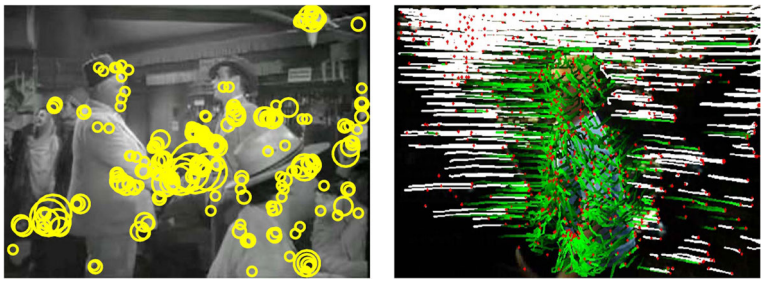

Fig. 3 Examples of two kinds of local features. *Left*: space-time interest points in multiple scales detected for the video frame contained the action HandShake. *Right*: improved dense trajectories by removing trajectories in white color due to camera motion

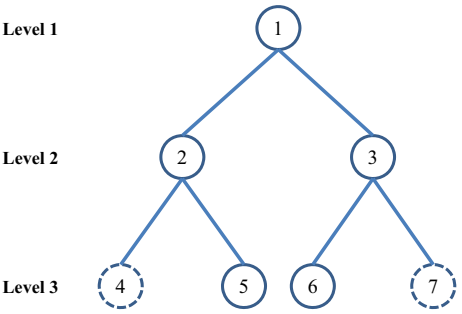

Fig. 4 Illustration of video partition for action recognition. The child parts and grandchild parts combined with the original video construct a three-level binary tree. Each *circle* refers to a video part. The *dashed circle* means that some kind of feature extraction method probably doesn't find descriptors in that part because of no motion happened

Improved dense trajectories (IDT) Dense trajectories are obtained by sampling dense points in different spatial scales from each video frame and tracking these points based on the displacement information from a dense optical flow field. Although the method of dense trajectories has achieved great success in action recognition, it has performance drawback due to poor consideration of camera motion. To improve the performance, Wang et al. [44] propose the improved dense trajectories which explicitly estimates camera motion and takes it into account to correct them. The right one of Fig. 3 gives an example of improved dense trajectories by removing trajectories due to camera motion.

In the experiments, we use the implementation released on the website of Laptev[1] for STIP and Wang[2] for IDT. Following the standard BoVW pipeline, we build a visual vocabulary by k-means clustering among the extracted descriptors and assign each interest point or trajectory to a visual word label. Then each video clip is represented by histograms of visual word occurrences. After that, we normalize each attribute by scaling it to [0,1].

3.2 Video partition for action recognition

Sometimes it is very difficult for existing methods to recognize the action in a video. With the idea of video partition, our goal is to improve the performance of action recognition by designing a new approach which can be generalized and applied to most of the existing solutions. Our method is different with the space-time pyramid which divides video into some spatial and temporal grids [23]. The method concatenates multiple descriptors generated from these grids into a single descriptor. It can be viewed as one kind of descriptor level fusion [27]. The video partition we proposed is not designed for fusion of descriptors. We apply it to discover more useful information from child parts of original video at the decision level which should be helpful to overcome the drawback of direct recognition from the given full video.

For a video $P1$ in human action datasets, we first segment it into two equal parts $P2$ and $P3$ from the middle. Then we do the same process on the obtained child parts and get four grandchild parts, $P4$, $P5$, $P6$ and $P7$. In consideration of the video length in most of human action datasets, a further segmentation of the video will produce lots of very small clips from which we cannot extract enough meaningful features for action recognition. So we only choose three-level segmentation in this paper. Included with the original video $P1$, we totally get seven parts distributed in three levels as shown in Fig. 4. After that, three action classifiers are trained for each level by cross validation. Give the trained classifiers,

[1] https://www.di.ens.fr/laptev/download.html

[2] https://lear.inrialpes.fr/people/wang/improved_trajectories

we use it to classify every action parts in the corresponding level and get the decision value matrix (DVM) shown as follows,

$$\begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \dots & \dots & \dots & \dots \\ v_{71} & v_{72} & \dots & v_{7m} \end{bmatrix} \tag{1}$$

where each row indicates the decision values output by the corresponding classifier for video part P_i, $i = 1, 2, ..., 7$. The m is the total number of decision values which should be determined by the selected classifier. In particular for classification methods based on support vector machine (SVM), $m = n(n-1)/2$ for one against one strategy and $m = n$ for one against all strategy, in which n is the number of classes of that human action dataset. As we select the default classification strategy of LIBSVM, the parameter m equals to $n(n-1)/2$.

After obtaining the decision value matrix, we concatenate all the decision values in the matrix belonging to the same video into a long vector by the following formula,

$$X = [v_{11}\ v_{12}\ ...\ v_{1m},\ \dots,\ v_{71}\ v_{72}\ ...\ v_{7m}]. \tag{2}$$

Then we normalize the vector and take X' as the representation of original video. Based on the new DVM representation, a final classifier is trained to recognize the action in videos. We choose the radial basis function (RBF) with the following formulation as the kernel of SVM.

$$K(x, x_i) = exp\left(-\frac{||x - x_i||^2}{\delta^2}\right). \tag{3}$$

According to the kernel function, the training vectors x_i are mapped into a higher dimensional space. Then the SVM will find a linear separating hyperplane with the maximal margin in this higher dimensional space. It should be noted that other classification models like Bayesian and neural network can also be applied to our method. But for the sake of simplicity, we only choose the SVM classifier to evaluate the effectiveness of our method.

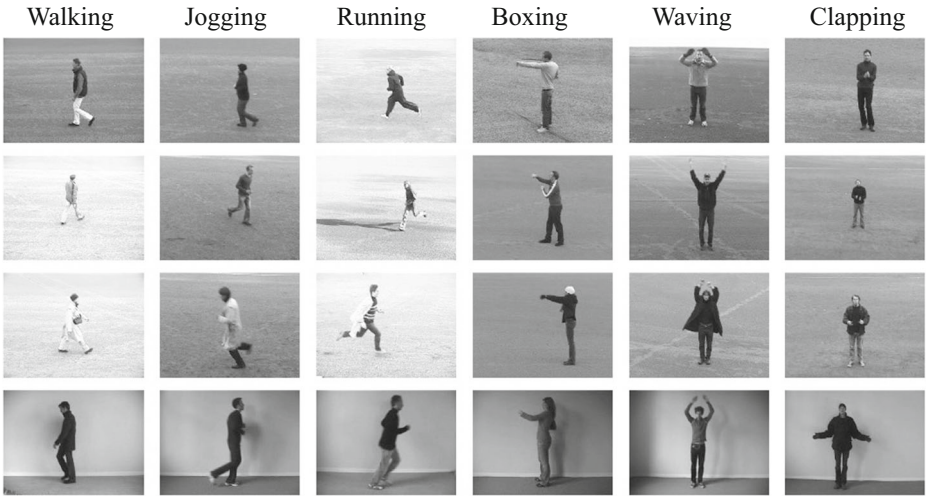

Fig. 5 Sample frames from the KTH human action dataset with six classes (*columns*) and four scenarios (*rows*) presented

Fig. 6 Sample frames of each action class from the Hollywood dataset. *From left to right, the first row*: Kiss, GetOutCar, SitDown, HugPerson. *Second row*: HandShake, StandUp, AnswerPhone, SitUp

4 Experiments

In this section, we describe the detailed experimental settings and show the results on two public human action datasets. We first introduce the datasets used for evaluation and their corresponding experimental setups. Then we present implementation details of our experiments. After that, we evaluate the performance of our method for action recognition and explore different factors that may impact on the final recognition accuracy. Finally, we make a further discussion about our work.

4.1 Datasets

We conduct experiments on two public datasets, KTH [34] and Hollywood human action dataset [23]. The KTH dataset is relatively simple while the Hollywood dataset is more complicated as it is collected from real movies. Some examples of video frames from the two action datasets are illustrated in Figs. 5 and 6.

The KTH dataset consists of 600 video files in total and each class has 100 videos which have a uniform resolutions of 160×120 pixels.[3] The videos are collected from 4 different scenarios and evenly divided into 6 types of actions: walking, jogging, running, boxing, hand waving and hand clapping. Furthermore, each video contains about four subsequences used as a sequence in the experiments. There are altogether 2391 sequences in the dataset. We train models on the training + validation set ($8+8$ people) and report average accuracy for evaluation on the test set (9 people).

The Hollywood dataset has 8 action classes:[4] AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, StandUp. All videos are obtained from 32 Hollywood movies. The dataset is divided into two training set and a test set. We choose the clean training set to train models and the test set to evaluate performance of methods. The clean training set contains 219 action samples that were manually verified to have 231 correct labels. The test set consists of 211 manually annotated action samples with 217 labels. The scenes of Hollywood movies are very complex such as background and viewpoint change, so it is a very difficult benchmark for action recognition. We perform evaluation according

[3] http://www.nada.kth.se/cvap/actions

[4] http://lear.inrialpes.fr/people/marszalek/data/hoha

Table 1 The classification accuracy of different methods with four kinds of local features on the KTH dataset and the performance improvements compared to original method

	HOG	HOF	HOGHOF	IDT
div1	76.2	89.1	87.2	95.1
div2	67.7	84.9	82.8	93.9
div4	65.3	77.7	78.7	91.5
ours	**79.2**	**89.7**	**89.8**	**95.8**
	+3.0	+0.6	+2.6	+0.7

to the splits of clean training and test as described in [23] and present exhaustive results on the dataset.

4.2 Implementation details

For STIP based features, the HOG and HOF descriptors are computed on a 3D video patch with $3 \times 3 \times 2$ spatio-temporal blocks in the neighborhood of each detected STIP. 4-bin HOG and 5-bin HOF descriptors are then computed for all blocks and are concatenated into a 72-element and 90-element descriptors respectively. We choose the STIP with HOG, HOF, the combination of HOG and HOF descriptors to evaluate our method. For IDT based features, we choose the combined descriptors (Trajectory+HOG+HOF+MBH) with default parameter settings. For simplicity, we use HOG, HOF, HOGHOF and IDT to denote the four kinds of local features. Regarding codebook generation, the k-means algorithm is adopted to cluster a subset of features sampled from the training videos. The number of clusters is set to $k = 4000$ as described in [23], which has shown empirically to give good results and is consistent with the values used for static image classification.

In the experiments, we take the most widely used SVM as the action recognition classifier. Specifically, we use the code of LIBSVM implemented by Chang et al. [9] released on their website.[5] Besides, we apply 5-fold cross-validation to find the best parameters c and g for multi-class classification.

4.3 Results of action recognition

We first evaluate the performance of our method for action recognition on KTH and Hollywood human action datasets. The results are shown in Tables 1 and 2. The abbreviations, "div1", "div2", "div4" and "ours", correspond to the classifiers trained under the following conditions: trained on the original training data, training data consists of child parts, training data consists of grandchild parts and our method. In experiments, class labels of original videos are taken as the labels of corresponding segmented parts. We present classification accuracy of the methods and shows the performance improvements compared to "div1" which is the original method. It can be seen that the performance of classifiers trained after video partition has shown a downward trend but still maintains a relatively high accuracy. Our method achieves the best performance compared to the original method on both action datasets. Take the IDT descriptor as an example, there is a modest improvement 0.7% for the KTH dataset. While on the Hollywood dataset, we obtain a very substantial improvement that reaches 7.6%. Besides, we give all the best parameters c and g selected for SVM in Tables 3 and 4.

[5] http://www.csie.ntu.edu.tw/cjlin/libsvm

Table 2 The classification accuracy of different methods with four kinds of local features on the Hollywood dataset and the performance improvements compared to original method

	HOG	HOF	HOGHOF	IDT
div1	29.0	27.6	32.9	33.6
div2	25.2	25.7	22.1	31.3
div4	23.5	21.9	24.1	27.6
ours	**32.9**	**32.9**	**33.8**	**41.2**
	+3.9	+5.3	+0.9	+7.6

Table 3 The best parameters *c* and *g* selected for SVM model on the KTH dataset

	HOG		HOF		HOGHOF		IDT	
	Cost	Gamma	Cost	Gamma	Cost	Gamma	Cost	Gamma
div1	32	0.0078125	2048	0.0001221	2048	0.0004883	32	0.0078125
div2	8	0.125	32	0.125	32	0.125	32	0.0078125
div4	8	0.125	8	0.125	8	0.125	8	0.03125
ours	0.25	0.5	16	0.0001221	0.5	1	16	0.03125

Table 4 The best parameters *c* and *g* selected for SVM model on the Hollywood dataset

	HOG		HOF		HOGHOF		IDT	
	Cost	Gamma	Cost	Gamma	Cost	Gamma	Cost	Gamma
div1	128	0.0019531	32	0.0004883	512	0.0000305	32	0.0078125
div2	128	0.001953	2048	0.0000305	2048	0.0001221	32	0.0078125
div4	32	0.0078125	8	0.03125	32.0	0.0078125	32	0.0078125
ours	0.125	0.0625	32	0.0001221	0.125	0.03125	0.5	0.25

Table 5 The comparison of classification accuracy between four types of kernel functions for SVM model on KTH dataset

	HOG	HOF	HOGHOF	IDT
linear	77.0	88.5	88.7	**95.8**
polynomial	76.1	87.8	87.2	93.9
sigmoid	78.7	88.9	88.9	95.1
rbf	**79.2**	**89.7**	**89.8**	**95.8**

Table 6 The comparison of classification accuracy between four types of kernel functions for SVM model on Hollywood dataset

	HOG	HOF	HOGHOF	IDT
linear	17.1	15.2	22.4	27.5
polynomial	23.8	23.8	23.8	23.7
sigmoid	29.0	31.9	30.5	35.1
rbf	**32.9**	**32.9**	**33.8**	**41.2**

Table 7 The classification accuracy for combination of ours and features in different dimension on KTH dataset

	ours	comb.	1024	512	256	128	64	32	16
HOG	79.2	80.2	80.2	80.4	80.4	**80.5**	80.4	79.9	80.2
HOF	89.7	90.4	90.4	90.3	90.4	**90.6**	90.5	90.1	89.9
HOGHOF	**89.8**	89.2	89.2	89.2	89.3	89.4	89.7	89.7	89.2
IDT	95.8	**96.2**	96.2	96.2	96.2	96.1	96.2	95.9	96.2

To find out which type of kernel function is best for SVM, we test four kernel functions on KTH and Hollywood datasets and show the comparison of classification accuracy in Tables 5 and 6. We can see that the radial basis function (RBF) kernel gets the best performance on both datasets. Especially on the Hollywood dataset, the accuracy of RBF kernel is far higher than other kernels. So we choose the RBF kernel for SVM in the experiments.

To further improve the performance, we combine our features with the features generated by the original method. In consideration of the dimension difference between two features, we also try the combination after dimension reduction by principal component analysis (PCA). It is worth noting that dimension reduction is not a required operation and many approaches ignore this step, such as vector of locally aggregated descriptor (VLAD) and sparse coding. In the experiments, we try different dimensions after PCA and want to find the best combination for action recognition. Tables 7 and 8 show the results on KTH and Hollywood datasets. The "comb." refers to combine our features with the original features directly, other numbers are the dimensions after PCA. From the tables, we find it is strange that the combination can get a better performance on the KTH dataset. But on Hollywood dataset, the combination is not effective. In contrast, the results of HOGHOF show that the combination can work on simple datasets like KTH. But for complicated datasets as Hollywood, applying our method alone will obtain a better performance.

4.4 Discussion

At the beginning of this paper, we make an assumption that it can give us a more accurate judgment of action by breaking the video into parts. The experimental results shown above have already demonstrated the effectiveness of our method. Now we want to make a further discussion about the upper bound of performance for action recognition. In Figs. 7 and 8, "rand2" and "rand4" indicate that the predict class label is selected randomly from two and four segmented parts of the original video. The "select2" and "select4" mean that the right class label is chosen from segmented parts by the human. That is to say, it is treated as a right prediction if one of them output a true label. From the Figures, it can be seen that our

Table 8 The classification accuracy for combination of ours and features in different dimension on Hollywood dataset

	ours	comb.	1024	512	256	128	64	32	16
HOG	**32.9**	30.5	30.5	30.5	30.0	30.5	31.0	31.4	31.0
HOF	**32.9**	32.9	32.9	32.9	32.4	32.4	32.4	32.4	32.9
HOGHOF	33.8	35.2	35.2	35.2	35.7	**36.2**	33.8	33.8	33.8
IDT	**41.2**	39.3	39.3	39.3	39.8	38.9	40.3	39.8	40.3

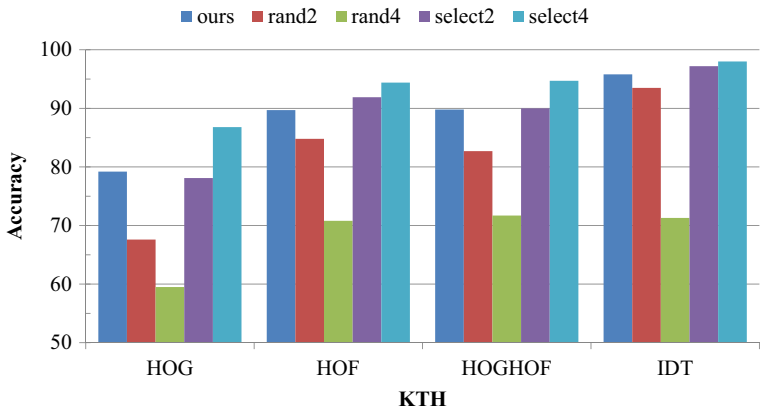

Fig. 7 Accuracy of our method, random select and human select from predicted classes on KTH dataset

method is much better than the method of random select. However, there is a large difference between our method and human selection. If we can estimate which part provides more useful information, further improvement of recognition performance is to be expected.

From Tables 1 and 2, we can observe another important result that our method gets higher improvements of the action recognition rate for Hollywood dataset than the KTH dataset. Now we want to discuss the reason for such difference. The theoretical basis of our approach is that the segmented parts can provide more useful information which can be directly obtained from the recognition results of each part. Take the IDT descriptor as an example, the experimental results show that the accuracies of "select4" on KTH and Hollywood dataset are 2.9% and 19% higher than the original method respectively. In other words, the additional information provided by the segmented parts is very limited on the simple KTH dataset, which also lead to a relatively moderate improvement of the action recognition accuracy. However on the complex Hollywood dataset, there exists lots of misclassified videos in which the segmented parts can be classified correctly. Because of this, our method achieves a significant performance improvement.

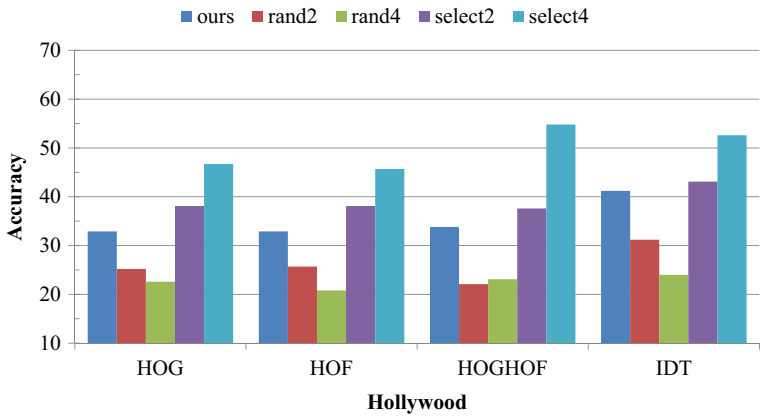

Fig. 8 Accuracy of our method, random select and human select from predicted classes on Hollywood dataset

5 Conclusions

In this paper, we propose a video partition based method for action recognition. The method is very simple and can be easily applied to lots of existing solutions. We segment every video into two child parts and four grandchild parts. Based on the operation of video partition, we construct a new video representation from the obtained decision value matrix which contains useful decision information of segmented parts in different levels. The widely used STIP and improved dense trajectories are adopted for evaluation. As demonstrated on two public human action datasets, our method achieves promising performance. Especially in the case of complex actions such as Hollywood dataset, the proposed DVM representation over a hierarchy of temporal granularities can substantially improve the accuracy of action recognition.

While the research of video understanding has long been inspired by human vision, we believe that the importance measurement of different parts in a video will help better recognize human action and find new paths forward. In the future, we will take the importance weight of segmented parts into account. Besides, we will also focus on making the new representation be able to generalize well on other large and complex human action datasets.

Acknowledgments The work was supported in part by the National Science Foundation of China (No. 61472103) and Australian Research Council (ARC) grant (DP150104645). We especially would like to thank the China Scholarship Council (CSC) for funding the first author to conduct the partially of this project at Australian National University.

References

1. Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: a large-scale video classification benchmark. arXiv:1609.08675
2. Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv 43(3):16
3. Benmokhtar R (2014) Robust human action recognition scheme based on high-level feature fusion. Multimed Tools Appl 69(2):253–275
4. Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S (2016) Dynamic image networks for action recognition. In: IEEE conference on computer vision and pattern recognition, pp 3034–3042
5. Borges PVK, Conci N, Cavallaro A (2013) Video-based human behavior understanding: a survey. IEEE Trans Circuits Syst Video Technol 23(11):1993–2008
6. Boureau YL, Bach F, LeCun Y, Ponce J (2010) Learning mid-level features for recognition. In: IEEE conference on computer vision and pattern recognition, pp 2559–2566
7. Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (2015) Activitynet: a large-scale video benchmark for human activity understanding. In: IEEE conference on computer vision and pattern recognition, pp 961–970
8. Cao Y, Barrett D, Barbu A, Narayanaswamy S, Yu H, Michaux A, Lin Y, Dickinson S, Siskind JM, Wang S (2013) Recognize human activities from partially observed videos. In: IEEE conference on computer vision and pattern recognition, pp 2658–2665
9. Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27
10. Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, pp 65–72
11. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. arXiv:1604.06573
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
13. Jain A, Gupta A, Rodriguez M, Davis LS (2013) Representing videos using mid-level discriminative patches. In: IEEE conference on computer vision and pattern recognition, pp 2571–2578

14. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
15. Kantorov V, Laptev I (2014) Efficient feature extraction, encoding, and classification for action recognition. In: IEEE conference on computer vision and pattern recognition, pp 2593–2600
16. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: IEEE conference on computer vision and pattern recognition, pp 1725–1732
17. Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: British machine vision conference, vol 275, pp 1–10
18. Kong Y, Kit D, Fu Y (2014) A discriminative model with multiple temporal scales for action prediction. In: European conference on computer vision, pp 596–611
19. Kovashka A, Grauman K (2010) Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: IEEE conference on computer vision and pattern recognition, pp 2046–2053
20. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: a large video database for human motion recognition. In: IEEE international conference on computer vision, pp 2556–2563
21. Lan T, Zhu Y, Roshan Zamir A, Savarese S (2015) Action recognition by hierarchical mid-level action elements. In: IEEE international conference on computer vision, pp 4552–4560
22. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
23. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE conference on computer vision and pattern recognition, pp 1–8
24. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE conference on computer vision and pattern recognition, pp 3361–3368
25. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild". In: IEEE conference on computer vision and pattern recognition, pp 1996–2003
26. Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: IEEE conference on computer vision and pattern recognition, pp 3337–3344
27. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput Vis Image Underst 150:109–125
28. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6):976–990
29. Raptis M, Kokkinos I, Soatto S (2012) Discovering discriminative action parts from mid-level video representations. In: IEEE conference on computer vision and pattern recognition, pp 1242–1249
30. Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981
31. Ryoo M (2011) Human activity prediction: early recognition of ongoing activities from streaming videos. In: IEEE international conference on computer vision, pp 1036–1043
32. Ryoo MS, Aggarwal JK (2010) UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)
33. Sadanand S, Corso JJ (2012) Action bank: a high-level representation of activity in video. In: IEEE conference on computer vision and pattern recognition, pp 1234–1241
34. Schüldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: IEEE international conference on pattern recognition, vol 3, pp 32–36
35. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: ACM international conference on multimedia, pp 357–360
36. Shen H, Yan Y, Xu S, Ballas N, Chen W (2015) Evaluation of semi-supervised learning method on action recognition. Multimed Tools Appl 74(2):523–542
37. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp 568–576
38. Soomro K, Zamir AR, Shah M (2012) Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
39. Sun L, Jia K, Yeung DY, Shi BE (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: IEEE international conference on computer vision, pp 4597–4605
40. Tamrakar A, Ali S, Yu Q, Liu J, Javed O, Divakaran A, Cheng H, Sawhney H (2012) Evaluation of low-level features and their combinations for complex event detection in open source videos. In: IEEE conference on computer vision and pattern recognition, pp 3681–3688
41. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: IEEE international conference on computer vision, pp 4489–4497
42. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2:28
43. Wang H, Kläser A, Schmid C, Liu CL (2011) Action recognition by dense trajectories. In: IEEE conference on computer vision and pattern recognition, pp 3169–3176

44. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: IEEE international conference on computer vision, pp 3551–3558
45. Wang H, Ullah MM, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: British machine vision conference, vol 124, pp 1–11
46. Wang L, Qiao Y, Tang X (2013) Mining motion atoms and phrases for complex action recognition. In: IEEE international conference on computer vision, pp 2680–2687
47. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In: IEEE conference on computer vision and pattern recognition, pp 4305–4314
48. Wang L, Ouyang W, Wang X, Lu H (2015) Visual tracking with fully convolutional networks. In: IEEE international conference on computer vision, pp 3119–3127
49. Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. Comput Vis Image Underst 115(2):224–241
50. Xu H, Tian Q, Wang Z, Wu J (2016) A survey on aggregating methods for action recognition with dense trajectories. Multimed Tools Appl 75(10):5701–5717
51. Xu Z, Qing L, Miao J (2015) Activity auto-completion: predicting human activities from partial videos. In: IEEE international conference on computer vision, pp 3191–3199
52. Xu Z, Yang Y, Hauptmann AG (2015) A discriminative cnn video representation for event detection. In: IEEE conference on computer vision and pattern recognition, pp 1798–1807
53. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: IEEE conference on computer vision and pattern recognition, pp 4694–4702
54. Zhang S, Yao H, Sun X, Wang K, Zhang J, Lu X, Zhang Y (2014) Action recognition based on overcomplete independent components analysis. Inf Sci 281:635–647
55. Zhang S, Zhou H, Yao H, Zhang Y, Wang K, Zhang J (2015) Adaptive normalhedge for robust visual tracking. Signal Process 110:132–142
56. Zhang S, Lan X, Yao H, Zhou H, Tao D, Li X (2016) A biologically inspired appearance model for robust visual tracking. In: IEEE transactions on neural networks and learning systems
57. Zhang W, Zhu M, Derpanis KG (2013) From actemes to action: a strongly-supervised representation for detailed action understanding. In: IEEE international conference on computer vision, pp 2248–2255
58. Zhou Y, Ni B, Hong R, Wang M, Tian Q (2015) Interaction part mining: a mid-level approach for fine-grained action recognition. In: IEEE conference on computer vision and pattern recognition, pp 3323–3331
59. Zhu J, Wang B, Yang X, Zhang W, Tu Z (2013) Action recognition with actons. In: IEEE international conference on computer vision, pp. 3559–3566

Ying Zheng is currently a Ph.D. candidate at Harbin Institute of Technology, Harbin, China. He received the B.S. degree in computer science from Jiangxi University of Finance and Economics, Jiangxi, China, in 2012. He received the M.S. degrees in computer science from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2014. Her research interests include computer vision and machine learning, especially focusing on analysis of human action in video understanding.

Hongxun Yao received the B.S. and M.S. degrees in computer science from the Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and in 1990, respectively, and received Ph.D. degree in computer science from Harbin Institute of Technology in 2003. Currently, she is a professor with the School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include computer vision, pattern recognition, multimedia computing, human-computer interaction technology. She has 6 books and over 200 scientific papers published, and won both the honor title of "the new century excellent talent" in China and "enjoy special government allowances expert" in Heilongjiang Province, China.

Xiaoshuai Sun received the B.S. degree in computer science from Harbin Engineering University, Harbin, China, in 2007. He received the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2009 and 2015 respectively. He was a Research Intern with Microsoft Research Asia (2012-2013) and also a winner of Microsoft Research Asia Fellowship in 2011. He owns 2 authorized patents and was the author or co-author of over 50 journal and conference papers in the field of multimedia and computer vision.

Xuesong Jiang is currently a Ph.D. candidate at Harbin Institute of Technology, Harbin, China. He received the B.S. and M.S. degrees in computer science from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2011 and 2013 respectively. His research interests include deep learning and image dehazing.

Fatih Porikli received the Ph.D. degree from the New York University, NY. He is currently a Professor in the Research School of Engineering, Australian National University (ANU). Until 2013, he was a Distinguished Research Scientist with Mitsubishi Electric Research Labs (MERL), Cambridge, USA. His research interests include computer vision, pattern recognition, manifold learning, sparse optimization, online learning, and image enhancement with commercial applications in video surveillance, intelligent transportation, satellite, and medical systems. He authored more than 140 publications and invented 66 patents.