Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition

Xiaojiang Peng[1,2]
xiaojiangp@gmail.com
Yu Qiao[†,2,3]
yu.qiao@siat.ac.cn
Qiang Peng[1]
qpeng@home.swjtu.edu.cn
Xianbiao Qi[2]
qixiaobiao@gmail.com

[1] Southwest Jiaotong University, Chengdu, P.R. China
[2] Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] The Chinese University of Hong Kong

Abstract

Feature representation is important for human action recognition. Recently, Wang et al. [25] proposed dense trajectory (DT) based features for action video representation and achieved state-of-the-art performance on several action datasets. In this paper, we improve the DT method in two folds. Firstly, we introduce a motion boundary based dense sampling strategy, which greatly reduces the number of valid trajectories while preserves the discriminative power. Secondly, we develop a set of new descriptors which describe the spatial-temporal context of motion trajectories. To evaluate the performance of the proposed methods, we conduct extensive experiments on three benchmarks including K-TH, YouTube and HMDB51. The results show that our sampling strategy significantly reduces the computational cost of point tracking without degrading performance. Meanwhile, we achieve superior performance than the state-of-the-art methods by utilizing our spatial-temporal context descriptors.

1 Introduction

Automatic recognition of human activity in videos has been an active research area in recent years due to its wide range of potential applications, such as smart video surveillance, video indexing and human-computer interface. Though various approaches have been proposed and many progresses have been achieved, it still remains a challenging task due to its large intra-class variations, clutter, occlusion and other fundamental difficulties [19].

The most important problem in action recognition is how to represent an action video. The approaches for action video representation can be roughly divided into four categories: (1) human pose based approaches which utilize human structure information [1]; (2) global action template based approaches which capture appearance and motion information on the whole motion body [4, 21]; (3) local feature based approaches which mainly extract valid

space-time cuboids [6, 12, 24]; (4) unsupervised feature learning based methods which learn the representation by hierarchical networks [8, 14].

Among most of the state-of-the-art methods, local spatial-temporal feature with bag-of-features (BoF) framework is perhaps the most popular and successful representation for action recognition [26]. Laptev [13] developed space-time interest points (STIP) by extending the Harris detector to 3D domain. Dollar et al. [6] detected space-time salient points by applying 2D spatial Gaussian and 1D temporal Gabor filters. Wang et al. [23] densely sampled cuboids at regular positions and scales. For descriptors, well-known approaches include HOG/HOF [12], Cuboids [6], HOG3D [9] and so on.

Recently, Wang et al. [24] proposed dense trajectory (DT) and a novel descriptor named motion boundary histogram (MBH) for action recognition. The motion boundary represents the gradient of optical flow which is initially introduced in the context of human detection [5]. A large number of experiments on nine popular human action datasets demonstrated the excellent performance of this approach [25]. The difference between the results of dense cuboids and DTs interests researchers. The superiority of DT compared to dense cuboids can be partly explained by the mechanism of human visual fixation system, in which moving objects remain in visual focus for a sustained amount of time through smooth pursuit of eye movements [15, 17]. We will give a visualized analysis in Section 2. Though its great power, the DT method is expensive in memory storage and computation due to the large number of dense sampled points.

In this paper, we first develop a motion boundary based sampling strategy named DT-MB to refine dense trajectory method. We start from densely sampled patches in a frame. Meanwhile, a binary motion boundary image is yielded from optical flow. Then we delete those regions that without motion boundary foregrounds. Central points of the rest patches are refined by the average location of foregrounds. Our DT-MB is partly motivated by the fact that those trajectories on motion boundary are the most meaningful ones, which is also implied by the superior performance of MBH [25]. Using our sampling method, the number of DTs can be sharply reduced without hurting the performance.

In addition, to enhance the discriminative power of DT based representation, a group of spatial-temporal context descriptors for trajectories are proposed which are partly inspired by [20, 27]. In [27], a descriptor based on co-occurrence HOG (CoHOG) is presented for human detection. Our motivation is that motion and appearance context in videos can provide important cues for action recognition. The novel context descriptors are comprised of spatial and temporal context descriptors. Spatial context descriptors contain spatial Co-HOG (S-CoHOG) which depicts complex structure of raw image, spatial co-occurrence HOF (S-CoHOF) which conveys complex motion structure, and spatial co-occurrence MBH (S-CoMBH) who captures the complex gradient structure of optical flow. Temporal context descriptors include temporal CoHOG (T-CoHOG) which expresses appearance changes along with time, temporal CoHOF (T-CoHOF) which conveys motion direction changes and temporal CoMBH (T-CoMBH) which extracts the changes of gradient orientations of flow.

It's worth noting that our spatial-temporal context descriptors are very different from spatial-temporal context information [28] and pairwise co-occurrence local spatial-temporal features [3]. Their methods try to capture spatial geometric relations among the features or temporal evolution of human poses by the temporal context information, while our descriptors aim to convey the local appearance and motion changes in the spatial-temporal context of a pixel. The spatial-temporal context information we used is more low-level than others.

The main contributions of this paper can be summarized into two folds. First, we develop a motion boundary based sampling strategy to reduce the number of dense trajectories

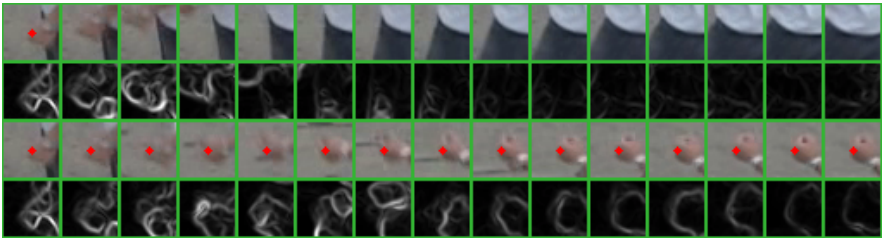


Figure 1: The 3rd and 1st rows: raw patches in cuboids with and without temporal slowness. The 2nd and 4th rows: motion boundary images corresponded to each patch.

which can save memory and computation without degrading performance. Second, a group of spatial-temporal context descriptors are presented for action video representation. Experiments show that our method outperforms several recent methods on three popular datasets.

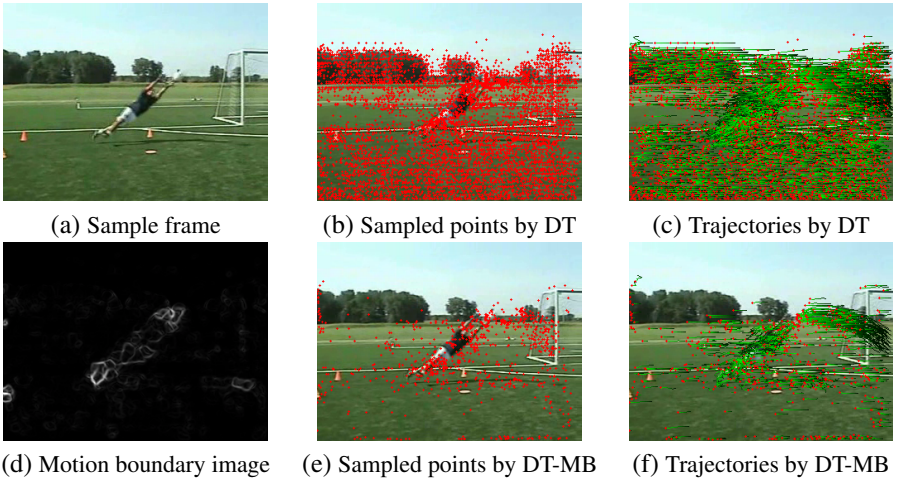
2 Dense Trajectories on Motion Boundary

Here, we give a brief review of dense trajectory method [25]. It mainly contains three steps: dense sampling, point tracking and trajectory checking. First, points are sampled in current frame on a grid by a step size w at S spatial scales. In order to track successfully, points in homogeneous areas are filtered out by checking the eigenvalues of their Harris matrices. Second, the valid points are tracked by median-filtered optical flow. Tracked points of consecutive frames at scale s are concatenated to form trajectories: $(P_t^s, P_{t+1}^s, ...)$, where $P_t^s = (x_t^s, y_t^s)$ represents spatial position. Some trajectories are shown in Fig. 2. Finally, once a trajectory's length reaches L, its average position drift and variation will be checked. Those trajectories with tiny or large mean drift and variation will be pruned, since they usually correspond to static or erroneous trajectories. Descriptors are extracted within a $N \times N \times L$ cuboid aligned by trajectories. To represent the video clip, a global histogram is yielded via the BoF framework. There exist two advantages for DT. One is that dense sampling can yield richer description of the action space than sparse interesting points (e.g. KLT, STIP). The other is that it's consistent with human visual fixation system to extract feature along with trajectories.

2.1 Visual Fixation Property of DT

Visual fixation is the maintaining of the visual gaze on a single location, also known as smooth pursuit or temporal slowness [15]. A number of species, including humans, other primates, cats and rabbits can perform this mechanism by three categories of eye movements: micro-saccade, ocular drift, and ocular micro-tremor. Dense trajectory based approaches match the fixation mechanism very well as illustrated in the 3rd row of Fig. 1.

There are two properties for temporal slowness in a view of video representation at least. On the one hand, it makes the feature **robust to velocity change**. Obviously, the appearance features in a cuboid with smooth pursuit will be very similar despite the difference of velocity. Motion features also keep similar after normalization. On the other hand, as shown in the last two rows of Fig. 1, **more meaningful appearance and motion information** are captured with temporal slowness. For these reasons, DT can always outperform dense cuboids.



(a) Sample frame (b) Sampled points by DT (c) Trajectories by DT

(d) Motion boundary image (e) Sampled points by DT-MB (f) Trajectories by DT-MB

Figure 2: Comparison of original DT and our Dense Trajectories on Motion Boundary.

2.2 Dense Trajectories on Motion Boundary

A limitation of DT is that too many points need to be tracked in the original dense sampling criterion [25]. However, only a few of them may lead to valid trajectories. We hold that those points on the motion boundary are the most discriminative ones. This is indeed partly implied by MBH descriptor [25] and motion boundary contour system (BCS) in neural dynamics of motion perception [7]. Our DT-MB sampling strategy constrains the sampled points on large magnitude regions of motion boundary image in the sampling step.

The implementation of dense trajectories on motion boundary (DT-MB) is straightforward. It needs two successive frames to sample points which is different from original DT. A comparative example is illustrated in Fig. 2. Fig. 2(d) shows a motion boundary image which indeed is the gradient magnitude of optical flow. After a thresholding operation on it, a mask is generalized to refine the original DT sampled points. Particularly, we create the mask by Otsu's algorithm [18] empirically. Those regions without motion boundary foregrounds will be deleted. Central points of the remaining patches will be refined by the average location of foregrounds. It's worth noting that the motion boundary image is a middle result of DT, so we never need to add complexity. As shown in the 2nd column of Fig. 2, our approach removes a large number of points which are not on the motion foreground. The 3rd column of Fig. 2 exhibits the trajectories from historical points by DT and DT-MB. The red marks are the end points of trajectories. Note that we do not force all the points of trajectories on the motion boundaries in case of inaccurate tracking. The detailed comparisons of complexity and performance are given in Section 4.

3 Spatio-temporal Context Descriptor

Pixels in videos are not lonely. It's beneficial to jointly encode the spatial-temporal context of a pixel. In this section, we present our spatial-temporal context descriptors which consist of spatial and temporal CoHOG, CoHOF and CoMBH.

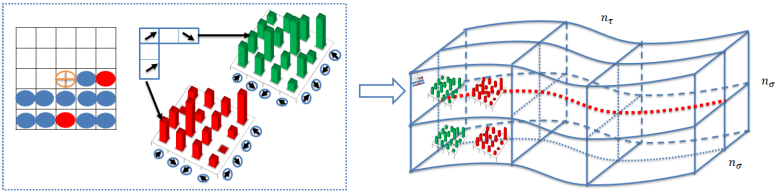


Figure 3: An example of spatial co-occurrence features with grid of size $n_\sigma \times n_\sigma \times n_\tau$.

3.1 Spatial Context Descriptors

Our spatial context descriptors aim to capture complex spatial structures of appearance and motion. We employ three kinds of co-occurrence descriptors for this motivation.

S-CoHOG.The spatial CoHOG is initially introduced in the context of pedestrian detection [27]. Specially, it uses pairs of gradient orientations as units and employs the co-occurrence matrix for image representation. The co-occurrence matrix expresses the joint distribution of gradient orientations between anchor points and given offset points over a patch as illustrated in Fig. 3. Considering two offsets, shown as the red solid circles in Fig. 3, each pair of which will vote for its own co-occurrence matrix.

S-CoHOF and S-CoMBH. The implementations of these are very similar with S-CoHOG except for the inputs. S-CoHOF applies spatial pairs of optical flow orientations as units and S-CoMBH utilizes spatial pairs of the gradient orientations in the horizontal and vertical flow components, separately. So there will be two S-CoMBH components, namely S-CoMBHx and S-CoMBHy.

In our case, we use two offsets (i.e., (2,0) and (0,2)) and process the trajectory cuboids pixel-wise with grid of size $n_\sigma \times n_\sigma \times n_\tau$. Specially, a co-occurrence matrix C over a $m \times n$ patch I, parameterized by an offset (x,y), is defined as:

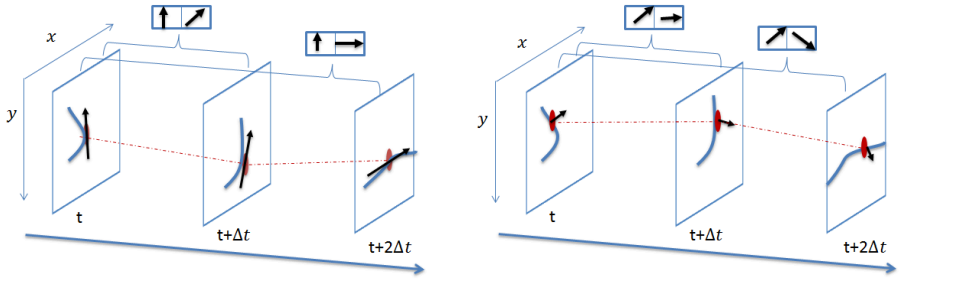
$$C_{x,y}(p,q) = \sum_{i=1}^{m}\sum_{j=1}^{n} \begin{cases} \frac{G(i,j)+G(i+x,j+y)}{2}, & \text{if } O(i,j)=p \text{ and } O(i+x,j+y)=q; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where p and q are the quantization bins, $G(i,j)$ is the gradient magnitude and $O(i,j)$ is the gradient orientation at position (i,j). As shown in Eq. (1), the average gradient magnitude is used to weight co-occurrence matrices. To reduce the boundary effects, we also apply linear interpolation. Finally, a $2 \times n_{bins} \times n_{bins} \times n_\sigma \times n_\sigma \times n_\tau$ dimensional vector is created for a cuboid by accumulating all the flatten matrices.

3.2 Temporal Context Descriptors

We expect the novel temporal context descriptors to depict clear motion and appearance changes from successive patches. Three temporal co-occurrence descriptors are designed for this purpose, namely T-CoHOG, T-CoHOF and T-CoMBH. Fig. 4(a) shows the basic units of T-CoHOG and T-CoMBH, and Fig. 4(b) is the case of T-CoHOF.

T-CoHOG. We build T-CoHOG via the temporal pairs of gradient orientations in raw frames. The anchor points are first tracked by middle-filtered optical flow with a given temporal offset, and then the corresponding points in two frames will vote for the co-occurrence matrices. We also use the mean gradient magnitude and linear interpolation to weight the



(a) T-CoHOG and T-CoMBH (b) T-CoHOF

Figure 4: Temporal co-occurrence descriptors. (a): temporal pairs of gradient orientations in T-CoHOG or T-CoMBH. (b): temporal pairs of optical flow orientations in T-CoHOF.

desired matrices. As illustrated in Fig. 4(a), the change of appearance within a certain time (i.e., Δt and $2\Delta t$) can be captured clearly.

T-CoHOF. T-CoHOF is built by using the temporal pairs of optical flow orientations. One can imagine that it needs three frames at least to obtain the basic units of T-CoHOF. The implementation of T-CoHOF is similar with T-CoHOG. As can be seen from Fig. 4(b), the alteration of motion orientations is reflected in the co-occurrence of basic units.

T-CoMBH. The temporal pairs of gradient orientations in the horizontal and vertical optical flow components are used for T-CoMBH. We can also get T-CoMBHx and T-CoMBHy, separately. The computational processes of them are similar with T-CoHOG as shown in Fig. 4(a). T-CoMBH expresses the local changes of motion boundary intensity.

Tracking is a necessary step in dense trajectory based approach, so our descriptors can benefit from the computational process of DT. It is worth noting that our temporal co-occurrence descriptors may share some common features with temporal grids. Both of them can capture temporal structure, but ours are embodied in more low-level than temporal grids, and they can be used with temporal grids at the same time. In this paper, we also use the grid of size $n_\sigma \times n_\sigma \times n_\tau$ to yield all the temporal co-occurrence descriptors and we only use the temporal offset $2\Delta t$ with Δt equals to 1.

4 Experiments

We conduct experiments on three popular human action datasets, namely KTH, YouTube and HMDB51. In this section, we first give a brief introduction for these datasets, and then compare the performance and complexity of DT and DT-MB. Finally, we give a comprehensive comparison among our novel descriptors and other individual descriptors of trajectory.

4.1 Datasets and Setup

These datasets we used are collected from controlled experimental setting and web videos. Some sample frames are illustrated in Fig. 5. We totally evaluate more than 10,000 video clips for our experiments.

The **KTH** dataset [22] is one of the most popular datasets in action recognition, which consists of 2,391 video clips acted by 25 subjects. It contains 6 action classes: *walking*,

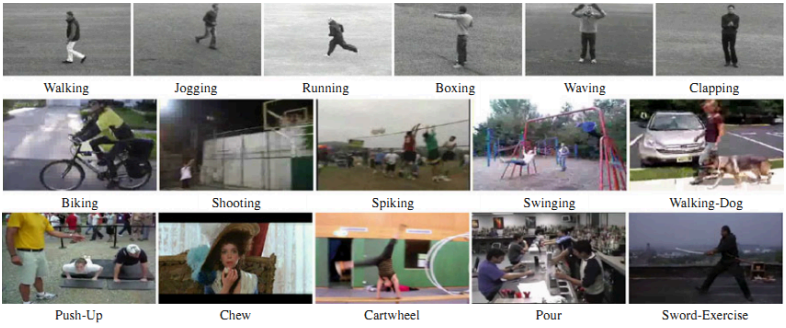


Figure 5: Sample frames from KTH, YouTube and HMDB51.

jogging, *running*, *boxing*, *hand-waving*, and *hand-clapping*. Actions are recorded at 4 environment settings: outdoors, outdoors with camera motion, outdoors with clothing change, and indoors. We follow the experimental settings in [22] where clips are divided into the training set (16 subjects) and the testing set (9 subjects).

The **YouTube** dataset [16] is collected from YouTube videos. It contains 11 action categories: basketball *shooting*, volleyball *spiking*, trampoline *jumping*, soccer *juggling*, horse back *riding*, *cycling*, *diving*, *swinging*, *golf-swinging*, *tennis-swinging*, and *walking* (with a dog). A total of 1,168 video clips are available. Following [16], we use Leave-One-Group-Out cross-validation and report the average class accuracy.

The **HMDB51** dataset is a large action video database with 51 action categories. Totally, there are 6,766 manually annotated clips which are extracted from a variety of sources ranging from digitized movies to YouTube [11]. It contains facial actions, general body movements and human interactions. It is a very challenging benchmark due to its high intra-class variation and other fundamental difficulties. We follow the experimental settings in [11] where three train/test splits are available.

We employ the standard BoF framework to represent videos. In particular, we set the parameters $(w,S,N,L,n_\sigma,n_\tau)$ mentioned in previous section to be $(5,8,32,15,2,3)$. Given all the local features, codebooks with size 4k are constructed separately using k-means from subsets of 100k randomly selected features. Vector quantization is used for each descriptor and the statistical histograms of visual word occurrences are used as video representations.

4.1.1 Setting of Classification

For classification we use the RBF-SVM with a multi-channel χ^2 kernel which is slightly different from [12]. The multi-channel Gaussian kernel is defined by:

$$K(H_i,H_j) = \exp(-\alpha \sum_{c\in\mathcal{C}} \frac{1}{A_c} D_c(H_i^c,H_j^c)) \tag{2}$$

where $D_c(H_i^c,H_j^c)$ is the χ^2 distance between two video representations H_i and H_j in the cth feature channel. A_c is the average value of all the distances in training set for the cth channel. α is a scaling factor ranging from 0 to 1 which is obtained by cross-validation, and we set it to be 1 when dealing with single channel feature. For multi-class classification, we use the *one-against-rest* approach and select the class with the highest score.

4.2 Experimental Results and Analysis

4.2.1 DT versus DT-MB

Our first purpose is to investigate the effect caused by constraining dense trajectories on the motion boundary. We compare the recognition accuracy, frame rate in the whole process of feature extraction including the time of I/O (fps), the average tracking time of dense points per frame (T_{track}) and the average number of trajectories per video clip. Specially, we quantize orientations into eight bins (an additional zero bin is added for HOF) with full orientation and weight with magnitudes. For the accuracy comparison, we only report the performance of all the raw DT descriptors combination by using (2). We evaluate the fps and T_{track} within 10 videos randomly selected from each dataset and the run-time is obtained on an Acer laptop with a 2.5 GHz Intel Core i5 CPU and 4 GB RAM.

Table 1: Comparison of DT and DT-MB with all the raw DT descriptors.

Datasets		$T_{track}(ms)$	Trajectories/clip	fps	Accuracy (%)
HMDB51	DT	46.33	16,133	3.43	46.60
	DT-MB	**12.82**	4,512	4.63	46.03
YouTube	DT	39.01	37,542	4.71	84.25
	DT-MB	**6.60**	10,878	5.85	85.10
KTH	DT	11.72	2,185	12.85	94.81
	DT-MB	**4.00**	1,178	16.05	94.79

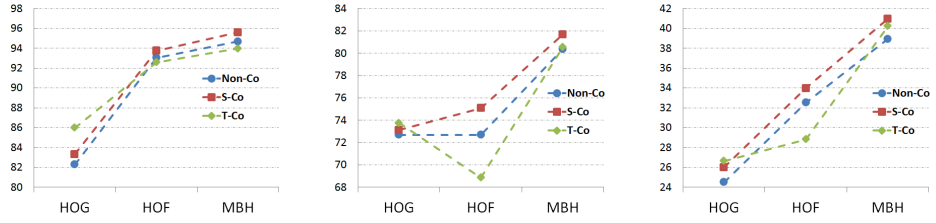
We show the comparison with original DT in Table 1. The computational cost decreases significantly for tracking points by using DT-MB. It is about 6 times less than DT on the YouTube. The numbers of trajectories also fall dramatically on all datasets. Specially, it is reduced by about 4 times on the HMDB51 dataset. The average class accuracies of DT and DT-MB on all the three datasets are very similar and it is even better than DT on the YouTube dataset. When comparing the confusion matrices of DT and DT-MB on the YouTube, we find out that the accuracies degrade only for those actions which are strongly related to the backgrounds like *tennis-swinging* and volleyball *spiking*. Overall, our DT-MB is able to save memory and time while keeping the accuracy.

4.2.2 Spatio-temporal Context Descriptors

We evaluate the raw DT, spatial and temporal context descriptors using our DT-MB sampling scheme. We just quantize orientations into four bins except an additional bin for HOF in this evaluation. One can certainly use eight bins which is not the key point in our evaluation. The performances of all descriptors (i.e., HOG, HOF, MBH and their spatial and temporal co-occurrence descriptors) are compared in Fig. 6. The results for MBH, S-CoMBH and T-CoMBH are obtained by combining x and y channels.

The HOG descriptor by adding spatial-temporal context (S-Co and T-Co) gives better results than original HOG. The performance degrades by using temporal context for HOF while aggrades by joining spatial context. S-CoMBH works always better than MBH while T-CoMBH sometimes does not. As can be seen from Fig. 6, a striking conclusion is that temporal context information for pure spatial feature (i.e., HOG) is more effective, and spatial context information for pure temporal features (i.e., HOF and MBH) is better.

Descriptor combination. Table 2 reports the results obtained by combining descriptors in the standard BOF framework. The baseline [25] is the combination of Trajectory, HOG,



(a) KTH (b) YouTube (c) HMDB51

Figure 6: Percentage accuracies of all the individual descriptors on three datasets. "Non-Co" corresponds to the original descriptors in [24].

HOF and MBH, we reimplement it in our evaluation. When adding our spatial-temporal context descriptors to the baseline there is an improvement of 0.47%, 2.05% and 3.32% on KTH, YouTube and HMDB51, separately. The combination of all spatial-temporal context descriptors can work better than that of original descriptors. The best combination is S-CoMBH+T-CoMBH for KTH, trajectory+S-CoHOG+T-Co for YouTube, and all the descriptors for HMDB51. The improvements demonstrate that spatial-temporal context contains useful complementary information.

Table 2: Different combinations of descriptors using standard BOF.

Combination	KTH	YouTube	HMDB51
Trajectory+HOG+HOF+MBH	93.63	84.25	45.90
HOG+HOF+MBH	93.98	83.48	45.88
Trajectory+S-Co + T-Co	94.79	85.70	48.98
S-Co + T-Co	94.21	85.33	48.89
All combined	94.10	86.30	49.22
Best combined	95.60	86.56	49.22

4.3 Comparison with State-of-the-Art Results

Table 3 presents the comparison to several recent results on each dataset. For fair comparison, we do not show the spatio-temporal pyramids (STP) post-processing results for Wang's approach which is also inferior to ours. Our method outperforms all these previously reported results. In particular, on the HMDB51 dataset, the improvement over the best reported result to date[1] is about 3%.

Table 3: Compare our results to the state-of-the-art results without STP.

KTH		**YouTube**		**HMDB51**	
Laptev *et al.* [12]	91.8%	Liu *et al.* [16]	71.2%	Kuehne *et al.* [11]	23%
Le *et al.* [14]	93.9%	Le *et al.* [14]	75.8%	Sadanand *et al.* [21]	26.9%
Ji *et al.* [8]	90.2%	B. *et al.* [2]	76.5%	Orit *et al.* [10]	29.2%
Wang *et al.*[25]	95%	Wang *et al.*[25]	84.1%	Wang *et al.*[25]	46.6%
Our Method	**95.6%**	Our Method	**86.56%**	Our Method	**49.22%**

[1] http://serre-lab.clps.brown.edu/resources/HMDB/eval/

5 Conclusion

This paper first introduced a new dense sampling approach for dense trajectories. It constrains sampled points on the motion boundary which significantly save memory and time cost without degrading performance. Another contribution is the new spatial-temporal context descriptors which make full use of spatial and temporal co-occurrence information. The comparisons of the individual descriptors demonstrate that temporal context information for pure spatial feature is more effective, and spatial context information for pure temporal features is beneficial. Finally, our method improves the performance of current action recognition systems on three challenging datasets.

Acknowledgments This work is partly supported by National Natural Science Foundation of China (61002042), Shenzhen Basic Research Program (JC201005270350A, JCYJ20120903092050890, JCYJ20120617114614438), 100 Talents Programme of Chinese Academy of Sciences, Guangdong Innovative Research Team Program (No.201001D0104648280), and the 2013 Doctoral Innovation Funds of Southwest Jiaotong University. This work is mainly conducted in SIAT and Yu Qiao is the corresponding author.

References

[1] Gabriele Fanelli Angela Yao, Juergen Gall and Luc Van Gool. Does human action recognition benefit from pose estimation? In *BMVC*, pages 67.1–67.11, 2011.

[2] Subhabrata Bhattacharya, Rahul Sukthankar, et al. A probabilistic representation for efficient large scale visual recognition tasks. In *CVPR*, pages 2593–2600, 2011.

[3] Piotr Bilinski and Francois Bremond. Statistics of pairwise co-occurring local spatio-temporal features for human action recognition. In *ECCV*, pages 311–320, 2012.

[4] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *TPAMI*, 23(3):257–267, 2001.

[5] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, pages 428–441, 2006.

[6] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, pages 65–72, 2005.

[7] Stephen Grossberg and Ennio Mingolla. Neural dynamics of motion perception: direction fields, apertures, and resonant grouping. *Perception & psychophysics*, 53(3): 243–278, 1993.

[8] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, pages 221–231, 2013.

[9] Alexander Klaser, Marcin Marszałek, Cordelia Schmid, et al. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[10] Orit Kliper-Gross, Yaron Gurovich, et al. Motion interchange patterns for action recognition in unconstrained videos. In *ECCV*, volume 7577, pages 256–269. 2012.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008.

[13] Ivan Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[14] Quoc V Le et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, 2011.

[15] Nuo Li and James J DiCarlo. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321(5895):1502–1507, 2008.

[16] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, pages 1996–2003, 2009.

[17] Susana Martinez-Conde et al. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3):229–240, 2004.

[18] N OTSU. A threshold selection method from gray-level histogram. *Trans. on Systems, Man, and Cybernetics*, 9:62–66, 1979.

[19] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[20] Xianbiao Qi, Rong Xiao, Jun Guo, and Lei Zhang. Pairwise rotation invariant co-occurrence local binary pattern. In *ECCV*, pages 158–171. 2012.

[21] Sreemanananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *CVPR*, pages 1234–1241, 2012.

[22] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, volume 3, pages 32–36, 2004.

[23] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, et al. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[24] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[25] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, pages 1–20, 2012.

[26] X. Wang, L. Wang, and Y. Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In *ACCV*, 2012.

[27] Tomoki Watanabe, Satoshi Ito, and Kentaro Yokoi. Co-occurrence histograms of oriented gradients for pedestrian detection. *Advances in Image and Video Technology*, pages 37–47, 2009.

[28] Xinxiao Wu, Dong Xu, Lixin Duan, and Jiebo Luo. Action recognition using context and appearance distribution features. In *CVPR*, pages 489–496, 2011.