# Global Contrast Based Salient Region Boundary Sampling for Action Recognition

Zengmin Xu[1,3], Ruimin Hu[1,2(✉)], Jun Chen[1,2], Huafeng Chen[1], and Hongyang Li[1]

[1] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan, China
hrm@whu.edu.cn
[2] Collaborative Innovation Center of Geospatial Technology, Wuhan, China
[3] School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin, China

**Abstract.** Although the excellent representation ability of improved Dense Trajectory (iDT) based features for action video had been proved on several action datasets, the performance of action recognition still suffers from large camera motion of videos. In this paper, we improve the iDT method by advancing a novel salient region boundary based dense sampling strategy, which reduces the number of trajectories while preserves the discriminative power. We first implement the iDT sampling based on motion boundary image, then introduce a global contrast based salient object segmentation method in interest points sampling step of action recognition. To overcome the flaws of global color contrast-based salient region sampling, we apply morphological gradient to generate a more robust mask for sampling dense points, as motion boundaries are much clearer. To evaluate the proposed method, we conduct extensive experiments on two benchmarks including HMDB51 and UCF50. The results show that our sampling strategy can improve the performance of action recognition with minor computational cost of mask production. In particular, on the HMDB51 dataset, the improvement over the original iDT result is 3 %. Meanwhile, any other dense features of action recognition can achieve more competitive performance by utilizing our sampling strategy and Fisher vector encoding method simply.

**Keywords:** Salient region boundary · Sampling strategy · Action representation · Improved dense trajectories

## 1 Introduction

Human action recognition, as an important biometric technology, has become an active research topic for decades due to their wide applications in video surveillance, video understanding, *etc.* Researchers used to focus on simple datasets [1,2] collected from controlled experimental settings. As the increasing


Q. Tian et al. (Eds.): MMM 2016, Part I, LNCS 9516, pp. 187–198, 2016.
DOI: 10.1007/978-3-319-27671-7_16

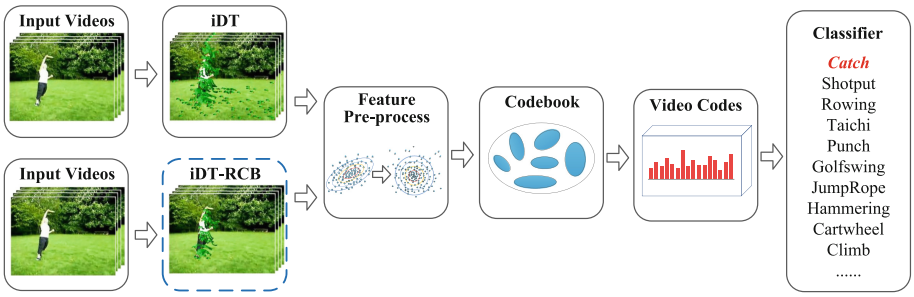


**Fig. 1.** Comparison of proposed approach (iDT-RCB) to traditional approach (iDT) for action recognition. Points sampled by iDT-RCB are more effective than iDT. Green trajectories indicate the sampled points have been tracked for fixed length of frames (Color figure online).

demand for understanding the content of real world video, action recognition still remains a challenging task on realistic data sets, which were collected from movies [3], web videos [4–6], *etc.* The diversity of realistic video data has resulted in significant challenges due to camera motion, viewpoint changes, occlusion, intra-class variations, complex background, *etc.*

How to represent human action in these realistic videos has been a fundamental problem in action recognition. By far, local space-time features [7–13] were shown to be successful on these datasets. Laptev [7] extracted space-time interest points (STIP) by extending the Harris detector from image to video. Dollar *et al.* [8] used 2D spatial Gaussian and 1D temporal Gabor filters to develop salient interest points in video. Willems *et al.* [9] applied the scale-space theory and Hessian matrix to detect interest points. Kläser *et al.* [10] proposed a HOG3D descriptor based on 3D-gradients. Sun *et al.* [11] modeled hierarchical spatio-temporal context via SIFT-based Trajectory. Wang *et al.* [12] sampled interest points on dense grid in each frame, and tracked them based on dense optical flow.

Among the state-of-the-art local space-time features, improved dense trajectories [13] have been shown to perform best on various datasets. The main idea is to remove camera motion from optical flow by homographic matrix, and explore the Fisher vector as a feature encoding approach. A large set of evaluations was presented to demonstrate the excellent performance of this feature. However, the iDT based representation is expensive in memory storage and computation due to the large number of densely sampled points.

In this paper, we develop a salient Region-based Contrast Boundary sampling strategy named iDT-RCB to refine improved dense trajectory approach. We start from densely sampled points on grid in each frame, and separate a large-scale region from its surroundings by a global contrast based method. Then we perform Morphological Gradient to construct a spatial saliency map with region boundary, and remove those sampled points which have no overlaps with foreground in the mask. The iDT-RCB is motivated by the fact that the trajectories on motion

boundary are the most meaningful ones. This is also inspired by DT-MB based sampling strategy [14] and the high performance of the MBH descriptor [23]. Under the control of our sampling method, the action representation can be focused on the most effective IDTs.

## 2 Methodology

In this section, we first briefly review the iDT features [13] including three steps: dense sampling, point tracking and trajectory estimating. We also illustrate the principle of motion boundary sampling method, then implement the improved dense trajectories sampling based on motion boundary.

### 2.1 Improved Dense Trajectories

Dense trajectories approach [12] densely sample feature points on a grid in each frame spaced by $W$ pixels. If the eigenvalues of the auto-correlation matrix are very small, it is impossible to track any point in homogeneous image areas. Hence the DT approach sets a threshold $T$ on the eigenvalues for each frame $I$ as:

$$T = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2), \tag{1}$$

where $(\lambda_i^1, \lambda_i^2)$ are the eigenvalues of point $i$ in the image $I$. Experiments showed that a value of 0.001 represents a good compromise between saliency and density of the sampled points. The sampled points are tracked through the video for $L$ =15 frames. Then they are removed and replaced by new interest points.

For each frame $I_t$, its dense optical flow field $\omega_t = (u_t, v_t)$ is computed w.r.t. the next frame $I_{t+1}$, where $u_t$ and $v_t$ are the horizontal and vertical components of the optical flow. Given a point $P_t = (x_t, y_t)$ in frame $I_t$, its tracked position in frame $I_{t+1}$ is smoothed by applying a median filter on $\omega_t$:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(x_t, y_t)}, \tag{2}$$

For each point, there are no feature points matching between every two frames, as the trajectory is only predicted by the points position in consecutive frames and the computed optical flow.

The improved dense trajectories approach samples and tracks feature points the same way as dense trajectories, but improve the dense trajectory by explicit camera motion estimation. The iDT approach utilizes human detector as a mask to remove feature matches on humans, the rest matches extracted from consecutive frames are applied to estimate the homography. Then iDT warps the second frame with the estimated homography and re-computes dense optical flow. HOF and MBH descriptors are computed on the warped optical flow. The homography and warped optical flow are estimated for every two frames.

For each trajectory, its shape is described by a sequence $(\Delta P_t, ..., \Delta P_{t+L-1})$ of displacement vectors $\Delta P_t = P_{t+1} - P_t = (x_{t+1} - x_t, y_{t+1} - y_t)$. If the maximal displacement of the trajectory vectors is less than a threshold, the trajectory is

considered to camera motion and will be removed. Due to the spatial neighborhood information and temporal motion properties of dense sampled points, the iDT method matches the visual fixation of video representation very well. Hence, iDT can always outperform DT, STIP and dense cuboids. The video frame and iDT sampling trajectories are illustrated in the 1st column of Fig. 2 respectively.

### 2.2 Motion Boundary Based Sampling

Although the iDT is benefited from the camera motion compensation, the performance of action recognition still suffers from the large camera movements. The truth is that most of challenging action datasets contains lots of camera motion, for example, HMDB51 has 59.9 % videos including camera motion [5]. Hence, we should study how to improve dense trajectories approaches.

Among the approaches improving dense trajectories, Vig *et al.* [15] uses saliency-mapping algorithms to prune background features. Wang *et al.* [16] extracts video patches only from human body regions instead of the whole videos, it only works well on those action videos including simple background. Shi *et al.* [17] explores sampling over high density with local spatio-temporal features extracted from a *Local Part Model*, but its sampling process cost a lot of time. Jain *et al.* [18] decomposes visual motion into dominant and residual motions, and designs a new descriptor to capture additional information on the local motion patterns. Jiang *et al.* [19] clusters dense trajectories, and use the cluster centers as reference points so that the relationship between them can be modeled. Ballas *et al.* [20] does not use saliency information to sample features but to pool them. Simonyan *et al.* [21] propose a two-stream ConvNet architecture which incorporates spatial and temporal networks.

All approaches mentioned above cannot solve the problem of reducing irrelevant trajectories caused by large camera movements. Therefore, we focus on discovering points need to be tracked. Unlike the Dense Trajectories based on Motion Boundary (DT-MB) sampling method [14], we implement improved Dense Trajectories sampling based on Motion Boundary (iDT-MB) to save meaningful points. We follow [14] to create the mask named Motion Boundary Image (MBI) by Otsus algorithm, and retain the regions with motion boundary foregrounds. The magnitude of each position in MBI is calculated as

$$MBI(i,j) = Otsu(\max(\sqrt[2]{\mathcal{I}_u^u * \mathcal{I}_u^u + \mathcal{I}_v^u * \mathcal{I}_v^u}, \sqrt[2]{\mathcal{I}_u^v * \mathcal{I}_u^v + \mathcal{I}_v^v * \mathcal{I}_v^v})), \quad (3)$$

where $\mathcal{I}^u, \mathcal{I}^v$ denote images containing the $u$ (horizontal) and $v$ (vertical) components of optical flow, $\mathcal{I}^\omega = (\mathcal{I}^u, \mathcal{I}^v)$ denote the 2D flow image ($\omega = (u, v)$), e.g., $\mathcal{I}_v^u = \frac{d}{dv}\mathcal{I}^u$ is the $v$-derivative of the $u$ component of optical flow.

The MBI is a middle result of iDT, so we do not need to add complexity. Note that iDT-MB can save fewer trajectories than DT-MB because of the difference between iDT and DT. The 4th column of Fig. 2 exhibits an MBI and the trajectories from historical points by iDT-MB. The detailed comparisons of complexity and performance are given in Sect. 4.

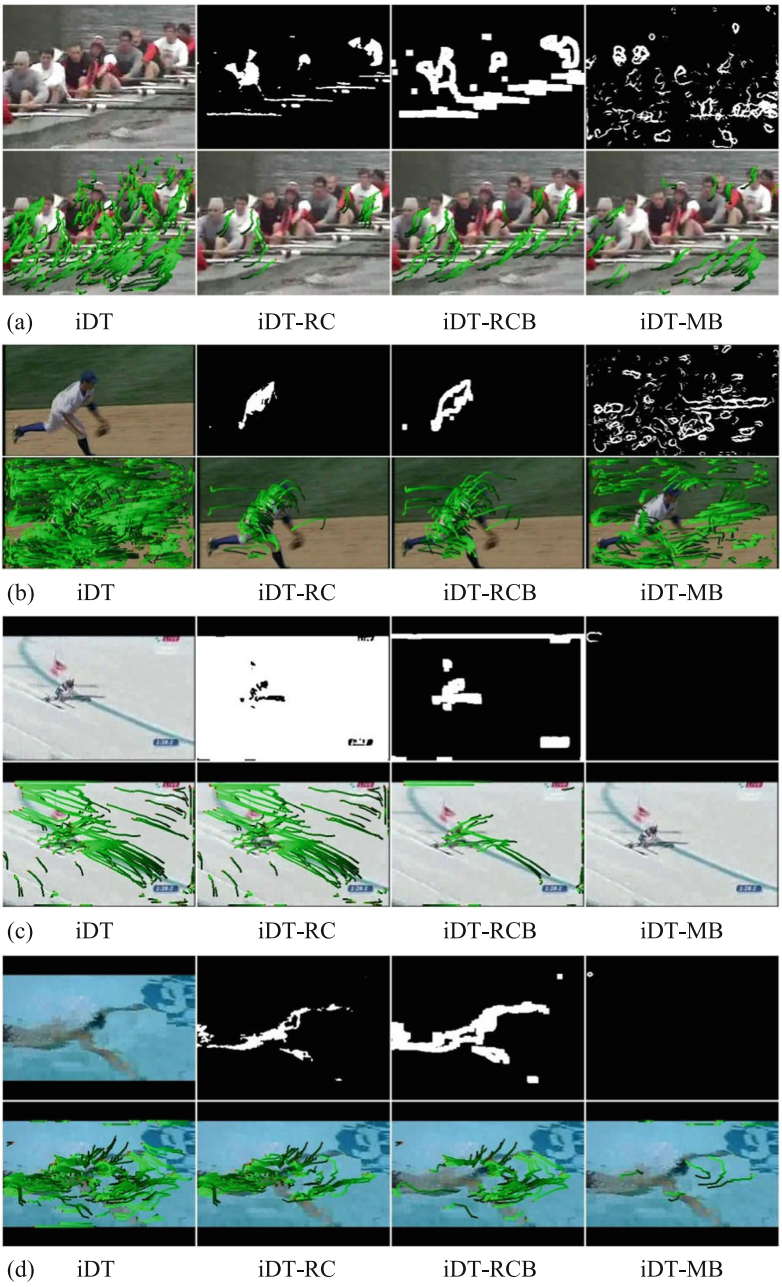


**Fig. 2.** Visualization of iDT, iDT-RC, iDT-RCB and iDT-MB sampling strategies for 4 actions. Compared to iDT, iDT-RCB is more robust to salient regions, in particular at shot boundaries (see 3rd column). iDT-RC can also handle salient regions, but it cannot capture the salient boundaries accurately. iDT-MB can reduce irregular motions, but it is not stable.

## 3 Our Approach

In this section, we introduce a Global Contrast based Salient Region Detection Algorithm in interest points sampling step of action recognition. We also explain why this method does not perform well in points sampling, then we present our new sampling strategy based on Salient Region Boundary in details.

### 3.1 Global Contrast Based Salient Region Sampling

A limitation of iDT-MB is that many trajectories are not in the foreground area, as iDT approach only considers salient points on dense grid of each frame, not the whole salient region of image. Meanwhile, the iDT-MB method is not stable since the motion boundaries are significantly influenced by the threading on gradient variation of optical flow. See the 4th column of Fig. 2(a), it shows the effective sampling example, but fail to capture the meaningful ones in Fig. 2(b) and Fig. 2(d) due to the unstable performance of MBI threading. Another worse result is given in the 4th column of Fig. 2(c), nothing is left in some cases.

To highlight the salient regions for action representation, we take into account the human detection algorithm. Unfortunately, even the state-of-the-art human detector cannot work well on action video datasets [13]. Furthermore, the salient region may be not in human body area but other objects, like the oars are more attractive in rowing action, see the 2nd column of Fig. 2(a). Hence, in order to find out the attractive salient regions, we propose a sampling method for applying the state-of-the-art salient region detection algorithm on improved dense trajectories. It is implemented by the iDT sampling based on salient Region-based Contrast (iDT-RC). This is partly inspired by Global Contrast based Salient Region Detection [22].

The main idea of iDT-RC is to automatic estimate salient object regions across every frame, enhances iDT sampling method without any prior knowledge of the video content. The iDT-RC sampling includes three steps:

(a) We first use a graph-based image segmentation method [22] to cut every frame into regions, and build the color histogram for each region. For a region $r_k$, we assign its saliency value by measuring its color contrast to other regions:

$$S(r_k) = \sum\nolimits_{r_k \neq r_i} w(r_i) D_r(r_k, r_i), \tag{4}$$

where $w(r_i)$ is the weight of region defined by the number of pixels in $r_i$, and $D_r(r_k, r_i)$ is the color distance metric between regions $r_k$ and $r_i$.

(b) We further incorporate spatial information by introducing a spatial weighting term in Eq. (4) to increase the effects of closer regions and decrease the effects of farther regions. Specifically, for any region $r_k$, the spatially weighted region contrast based saliency is

$$S(r_k) = \sum\nolimits_{r_k \neq r_i} \exp(-\frac{D_s(r_k, r_i)}{\sigma_s^2}) w(r_i) D_r(r_k, r_i), \tag{5}$$

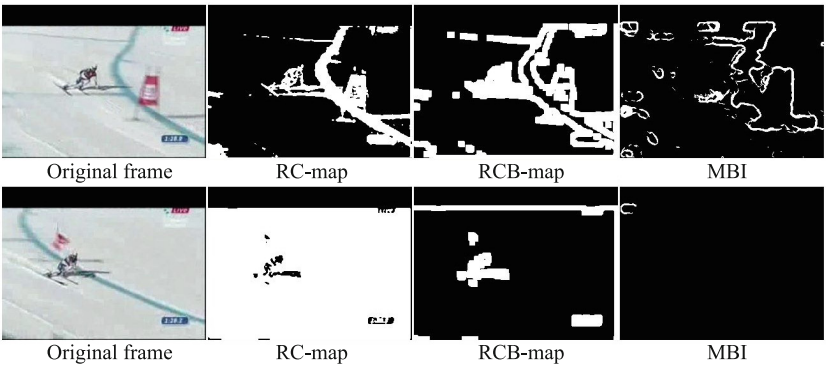


**Fig. 3.** RCB-map using Morphological Gradient is more robust than RC-map and MBI for salient region segmentation in action videos.

where $D_s(r_k, r_i)$ is the spatial distance between the two regions, $\sigma_s$ controls the strength of spatial distance weighting.

(c) To save the useful interest points in every frame, we follow the RC-map [22] approach to get a segmentation mask, and apply the estimated salient mask to iDT sampling method. Those interest points sampled by the iDT-RC but not in global contrast based salient regions will be deleted.

### 3.2 Optimization with Salient Region Boundary

However, the iDT-RC combined iDT with salient regions straightly does not perform well in points sampling. Several reasons may account for this issue: Firstly, the Global Contrast based Salient Region Detection, which uses image contrast under the assumption that a salient object exists in an image, aim to model saliency for image pixels using color statistics of the input image. Hence, the RC-map approach does not always work perfectly, it will get some unexpected masks due to its global color contrast, see the 2nd column of Fig. 2(c). Secondly, sometimes the salient region generated by RC-map is too limited to track enough interest points for representing an action, the discriminative ones may be not saved, see the 2nd column of Fig. 2(a). Last but not the least, not all trajectories from salient regions may lead to valid trajectories, the performance of the codebook is influenced by the noise trajectory samples. Therefore, in order to handle the issue mentioned above, we propose another iDT sampling method based on salient Region Boundary named iDT-RCB. Unlike [15–18], our iDT-RCB sampling strategy constrains the sampled points on salient region boundaries in the sampling step. We use two iterations of the Morphological Gradient on RC-map to generate a robust RCB-map. The Morphological Gradient can be expressed as

$$RCBmap = morph_{grad}\,(RCmap) = dilate\,(RCmap) - erode\,(RCmap), \quad (6)$$

The proposed iDT-RCB sampling process is described below in detail.

**Algorithm 1**. iDT-RCB Sampling Procedure

**Input:**

$Video\ Frames = \{I_1, I_2, ..., I_N\}$;

**Output:**

$Valid\ Trajectories = Tr_1, Tr_2, ..., Tr_M$;

Initialize the sampling parameters
**for** $i = 1$ to $N$ **do**
  generate the $RCB\text{-}map$ by using two iterations of Eq. (6)
  $P_j^{(1)} \Leftarrow denseSample(greyI_i, RCB\text{-}map)$ for each scale. $Tr_j^{(1)} \Leftarrow P_j^{(1)}$
  $\omega_i \Leftarrow$ compute dense optical flow by Farnebäck algorithm
  $matches_i \Leftarrow matchFromSurfandFlow(greyI_i, \omega_i, RCB\text{-}map)$
  $H_i \Leftarrow findHomography(matches_{i-1}, matches_i, RANSAC)$
  warp the second frame with $H_i$
  $\omega_i^{'} \Leftarrow$ re-compute dense optical flow by warped second frame
  predict the motion of $P_j^{(t+1)}$ by using $\omega_i^{'}$ and Eq. (2)
  $Tr_j \Leftarrow \{P_j^{(1)}, P_j^{(2)}, ..., P_j^{(t)}, P_j^{(t+1)}, ..., P_j^{(L)}\}$
  **if** $Tr_j$ is valid &&$Tr_j$ is not camera motion **then**
    $Valid\ Trajectories \Leftarrow Tr_j$
  **end if**
**end for**

where $P_j^{(1)}$ denote the first position of the $j$-th sampled point. Points from $P_j^{(1)}$ to $P_j^{(L)}$ of subsequent $L$ frames are concatenated to form the $j$-th trajectory $Tr_j$.

We hold that those points on the salient region boundary are the most discriminative ones. This is indeed partly implied by MBH descriptor [23], Dmask including narrow strip surrounding the persons contour [24], and motion boundary contour system in neural dynamics of motion perception [25].

Although many action recognition approaches have been developed and inspiring progresses can achieve advanced levels, our iDT-RCB sampling method is more effective for large camera motion. It is very suitable for feature extraction in action videos, see the 3rd column of Fig. 3.

## 4 Experiments

In this section, we describe the details of extensive experiments to evaluate the usefulness of the proposed method in action recognition.

### 4.1 Datasets

We conduct experiments on two action datasets, namely HMDB51 [5] and UCF50 [6]. Some example frames are illustrated in Fig. 4. We summarize them and the experimental protocols as follows.


**Fig. 4.** Sample frames from HMDB51 (top) and UCF50 (bottom) datasets.

The HMDB51 dataset is collected from a variety of sources ranging from digitized movies to YouTube videos. There are 51 action categories and 6,766 video sequences in HMDB51. We follow the original protocol using three train-test splits and perform experiments on the original videos not the stabilized ones. We report average accuracy over the three splits as performance measure.

The UCF50 dataset has 50 action categories, consisting of real-world videos taken from YouTube. The actions range from general sports to daily life exercises. For all 50 categories, the videos are split into 25 groups. For each group, there are at least 4 action clips. In total, there are 6,618 video clips in UCF50. We apply the Leave-One-Group-Out Cross Validation for UCF50 dataset and report average accuracy over the twenty five splits.

### 4.2 Experimental Setup

In all the following experiments, we densely extract improved trajectories based on the code from Wang [13]. The iDT-MB is implemented by using the code from Peng [14]. The iDT-RC and iDT-RCB is partly implemented by using the code from Wang [13] and Cheng [22].

To recognize actions, we run five feature sampling methods at the same server cluster with multithreading, follow [13,26] to train a GMM codebook with $K = 256$ Gaussians based on 256,000 randomly sampled trajectories. The default parameters of descriptor in the spatio-temporal grid, the size of the volume and the tracked frames length are the same as [13]. Each trajectory is described by concatenating HOG, HOF, and MBH descriptors, which is a 396-dimensional vector. We reduce the descriptors dimension to 200 by performing PCA and Whitening. Then, each video is represented by a $2DK$ dimensional Fisher vector for each descriptor type. Finally, we apply Power L2-normalization to the Fisher vector. To combine different descriptor types, we concatenate their normalized Fisher vectors. In our experiments, we choose linear SVM as our classifier with the implementation of LIBSVM [27]. For multi-class classification, we use the one-vs-rest approach and select the class with the highest score.

We compare our approach to recent methods [13–21]. The mean run-time of sampling process and mean number of sampled trajectories are compared to iDT. The processing speed is reported in frames per second (fps), run at a single-core Intel Xeon X3430 (2.4 GHz) without multithreading.

### 4.3 Results and Analysis

Since the SaliencyCut [22] is an iterative process of using graphcut and GMM appearance mode, there may be a slight difference in generalized results. However, its performance still improves, as salient region boundaries are much clearer, see the 3rd column of Fig. 2. Compared to the baseline (iDT without HD [13]), we have 3 % improvement on HMDB51 and 1.5 % improvement on UCF50.

**Table 1.** Comparison of our results (HOG+HOF+MBH) to the state of art. We present our results for FV encoding without automatic human detection (HD).

| HMDB51 | | UCF50 | |
|---|---|---|---|
| Peng *et al.* [14] | 49.2 % | Reddy *et al.* [6] | 76.9 % |
| Jain *et al.* [18] | 52.1 % | Shi *et al.* [17] | 83.3 % |
| Simonyan *et al.* [21] | **59.4 %** | Wang *et al.* [28] | 85.7 % |
| iDT without HD [13] | 55.9 % | iDT without HD [13] | 90.5 % |
| iDT with HD [13] | 57.2 % | iDT with HD [13] | 91.2 % |
| iDT-MB | 53.3 % | iDT-MB | 88.4 % |
| iDT-RC | 55.7 % | iDT-RC | 90.8 % |
| iDT-RCB | 58.9 % | iDT-RCB | **92.0 %** |

We use the bounding box provided from [13] in iDT sampling with HD. As the human detector does not always work perfectly, it will miss humans due to pose or viewpoint changes. Table 1 reports action recognition average accuracy compared to other dense trajectories approaches. Our iDT-RCB sampling method achieves the best result on UCF50. The result on HMDB51 is slightly decreased than [21], which have used the trained deep Convolutional Networks.

We evaluate the average number of trajectories per video clip and fps within 10 videos randomly selected from each dataset, and the run-time is obtained. Table 2 illustrates the minor computational cost of iDT-RCB. Fewer trajectories also can lead to faster speed in the feature encoding process.

**Table 2.** Comparison of sampled trajectories number and features extraction speed to iDT [13]. Note that we only randomly select 10 videos from each dataset.

| Sampling strategy | HMDB51 | | UCF50 | |
|---|---|---|---|---|
| | Trajectories/clip | fps | Trajectories/clip | fps |
| iDT without HD [13] | 489,865 | 3.59 | 2,383,147 | 3.90 |
| iDT with HD [13] | 492,456 | 3.60 | 2,452,230 | 3.97 |
| iDT-MB | 164,643 | 3.75 | 911,280 | 4.36 |
| iDT-RC | 263,670 | 2.81 | 1,329,635 | 3.25 |
| iDT-RCB | 380,511 | 2.75 | 1,737,075 | 3.16 |

## 5 Conclusion

This paper proposes a novel dense sampling approach named iDT-RC for improved dense trajectories. We first implement iDT sampling based on MBI, and applies a salient region contrast based segmentation method in interest points sampling step. To overcome the flaws of salient region contrast based method in action recognition, we apply morphological gradient to RC-map for generating more robust salient mask named RCB-map. The improved sampling method named iDT-RCB constrains sampled points on the salient region boundary which can improve the performance with minor computational cost. The comparisons of the sampling strategies demonstrate that salient region boundary information is more effective. Finally, our method improves the performance of current action recognition systems on two challenging datasets which represents a good compromise between speed and accuracy.

**Acknowledgement.** The research was supported by the National Nature Science Foundation of China (61231015, 61170023, 61367002), the National High Technology Research and Development Program of China (863 Program) (2015AA016306, 2013AA014602), the Internet of Things Development Funding Project of Ministry of industry in 2013(25), the Technology Research Program of Ministry of Public Security (2014JSYJA016), the Major Science and Technology Innovation Plan of Hubei Province (2013AAA020), the Nature Science Foundation of Hubei Province (2014CFB712).

## References

1. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR 2004, vol. 3, pp. 32–36 (2004)
2. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Pattern Anal. Mach. Intell. **29**(12), 2247–2253 (2007)
3. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR 2008, pp. 1–8 (2008)
4. Liu, J.G., Luo, J.B., Shah, M.: Recognizing realistic actions from videos in the wild. In: CVPR 2009, pp. 1996–2003 (2009)
5. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video data-base for human motion recognition. In: ICCV 2011, pp. 2556–2563 (2011)
6. Reddy, K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
7. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2), 107–203 (2005)
8. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: PETS 2005, pp. 65–72 (2005)
9. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
10. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC 2008 (2008)

11. Sun, J., Wu, X., Yan, S.C., Cheong, L.F., Chua, T.S., Li., J.T.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR 2009, pp. 2004–2011 (2009)
12. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
13. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV 2013, pp. 3551–3558 (2013)
14. Peng, X.J., Qiao, Y., Peng, Q.: Motion boundary based sampling and 3D co-occurrence descriptors for action recognition. Image Vis. Comput. **32**(9), 616–628 (2014)
15. Vig, E., Dorr, M., Cox, D.: Space-variant descriptor sampling for action recognition based on saliency and eye movements. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 84–97. Springer, Heidelberg (2012)
16. Wang, B., Liu, Y., Xiao, W.H., Xiong, Z.H., Wang, W., Zhang, M.J.: Human action recognition with optimized video densely sampling. In: ICME 2013, pp. 1–6 (2013)
17. Shi, F., Petriu, E., Laganiere, R.: Sampling strategies for real-time action recognition. In: CVPR 2013, pp. 2595–2602 (2013)
18. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: CVPR 2013, pp. 2555–2562 (2013)
19. Jiang, Y.-G., Dai, Q., Xue, X., Liu, W., Ngo, C.-W.: Trajectory-based modeling of human actions with motion reference points. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 425–438. Springer, Heidelberg (2012)
20. Ballas, N., Yang, Y., Lan, Z.Z., Delezoide, B., Preteux, F., Hauptmann, A.: Space-time robust video representation for action recognition. In: ICCV 2013, pp. 2704–2711 (2013)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS 2014 (2014)
22. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
23. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
24. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV 2013, pp. 3192–3199 (2013)
25. Grossberg, S., Mingolla, E.: Neural dynamics of motion perception: direction fields, apertures, and resonant grouping. Percept. Psychophysics **53**(3), 243–278 (1993)
26. Peng, X., Zou, C., Qiao, Y., Peng, Q.: Action recognition with stacked fisher vectors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 581–595. Springer, Heidelberg (2014)
27. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Tech. **2**(3), 27 (2011)
28. Wang, L.M., Qiao, Y., Tang, X.O.: Mining motion atoms and phrases for complex action recognition. In: ICCV 2013, pp. 2680–2687 (2013)