Neurocomputing 236 (2017) 82–92

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Action recognition by saliency-based dense sampling

Zengmin Xu[a,b,c], Ruimin Hu[a,b,*], Jun Chen[b,d], Chen Chen[e], Huafeng Chen[b], Hongyang Li[b], Qingquan Sun[f]

[a] *State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China*
[b] *National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China*
[c] *School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin 541004, China*
[d] *Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China*
[e] *Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816, USA*
[f] *School of Computer Science and Engineering, California State University San Bernardino, San Bernardino, CA 92407, USA*

ARTICLE INFO

Keywords:
Region-based Contrast Boundary
Warped flow evaluation
Robust salient mask
Trajectory-pooled Deep-convolutional Descriptor(TDD)

ABSTRACT

Action recognition, aiming to automatically classify actions from a series of observations, has attracted more attention in the computer vision community. The state-of-the-art action recognition methods utilize dense sampled trajectories to build feature representations. However, their performances are limited due to action region clutters and camera motions in real world applications. No matter how the scenario changes in different backgrounds, the salient cues of actions are highly dependent on their appearances and motions. Based on this discovery, in this paper we propose a novel saliency-based dense sampling strategy named improved dense trajectories (iDT) on salient region-based contrast boundary (iDT-RCB). Without any external human detector, a robust mask is generated to overcome the limitations of global contrast based saliency in action sequences. Warped optical flow is exploited to adjust the interest points sampling to remove subtle motions. We show that an appropriate pruning of feature points can achieve a good balance between saliency and density of the sampled points. Experiments conducted on three benchmark datasets have demonstrated the effectiveness of the proposed method. More specifically, the fusion of deep-learned features and our hand-crafted features can even improve the recognition performance over baseline dense sampling methods. In particular, the fusion scheme achieves the state-of-the-art accuracy at 73.8% and 94.8% on Hollywood2 and UCF50, respectively.

1. Introduction

Human action recognition, one of the key technologies in computer vision domain, has been widely applied in human surveillance, scene understanding, human–computer interaction, etc. While reliable human action recognition has been achieved in simple scenes (KTH [1] and Weizmann [2]), the recognition task remains challenging in complex scenes. The diversity of realistic videos, such as movies [3] and web videos [4–6], has shown significant challenges with foreground clutter, background variations, camera motion, view changes and partial occlusions.

Human action modeling is a fundamental problem for action recognition. Modeling an action in video sequence starts with feature representation. Previous research efforts for action representation were mainly focused on the following four aspects:

- *Local features*: For each given detected interest point, a feature descriptor is computed for a 3D video patch descriptor. As local space–time features allow to build efficient action representation without object detection or motion segmentation, they have been successfully utilized in action recognition and thus, leads to a trend of generalizing descriptors such as STIP [7], Cuboids [8], 3D-SIFT [9], HOG3D [10], HOG/HOF [3], Hierarchical SIFT Trajectory [11], LTP [12], MoSIFT [13], MPEG Flow [14], and CGME [15], iMoSIFT [16].
- *Dense sampling*: Among the local space–time features, dense sampling methods have drawn more attention and provided promising results. The main idea is to densely sample feature points in each frame, and track them in video sequences based on optical flow. Multiple descriptors are computed along the trajectories of feature points to capture motion information, e.g., MBH [17], extended SURF [18], Dense [19], V-FAST [20], Stacked ISA [21], Saliency [22], OVDS [23], DT [24], LPM [25], DCS [26], MBI [27], Motionlets [28], iDT [29], DTD [30], Concept Relevance [31,32],

[*] Corresponding author at: State Key Laboratory of Software Engineering, Wuhan University, Wuhan 430072, China.
E-mail address: hrm@whu.edu.cn (R. Hu).

http://dx.doi.org/10.1016/j.neucom.2016.09.106
Received 21 March 2016; Received in revised form 18 August 2016; Accepted 21 September 2016
Available online 17 November 2016

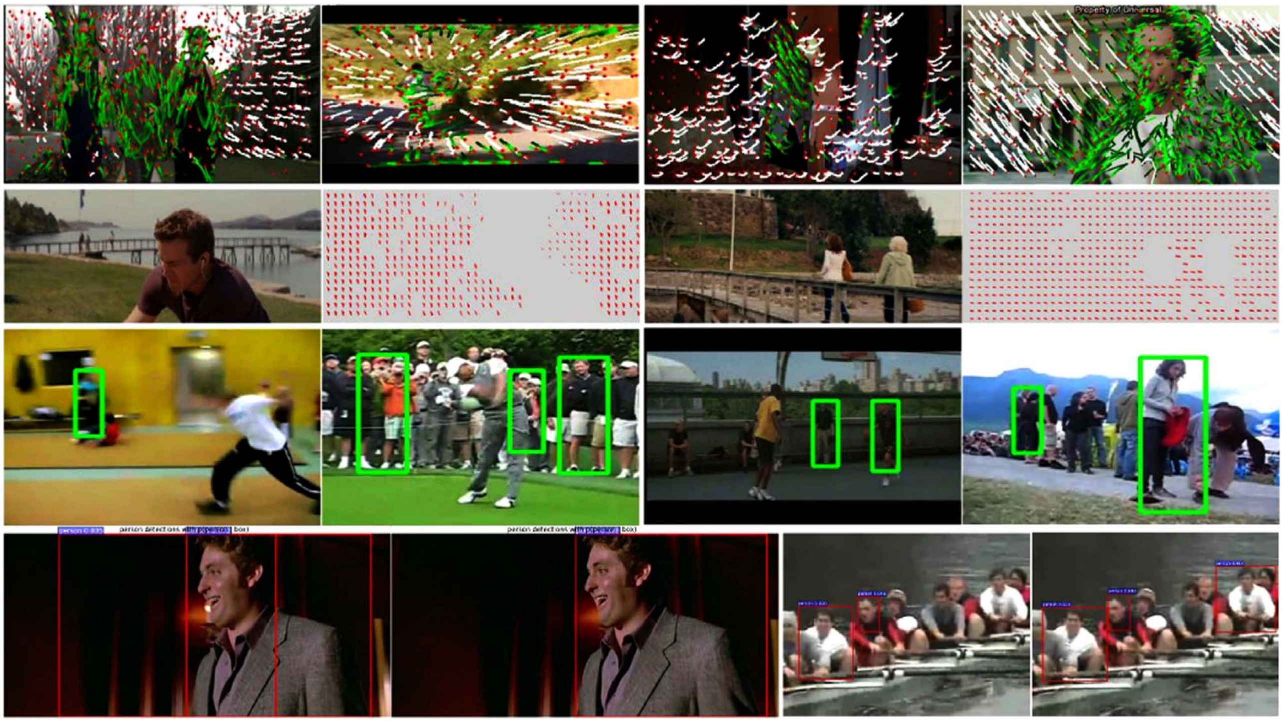

Fig. 1. Example illustrating the characteristics in action videos, e.g., background motions and foreground variations. The 1st row shows white removed trajectories under various camera motions. The 2nd row illustrates camera motion types via red underlying trajectories. The 3rd row demonstrates the failure cases of human detector due to complex human pose variations. The last row describes the failure cases of faster RCNN owe to illumination variations and partial occlusions. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

and iDT-RCB [33].

- *Global representation*: Despite encouraging results have been obtained using the methods above on several datasets, low-level features limit the semantics of actions. Therefore, representing actions by pose-based methods [34–39] or global templates have been explored, for instance, MHI [40], Shapes [2], Action Bank [41], NTraj [42] and DMMs [43]. However, it is difficult to estimate poses or high quality templates due to the diversity of real world videos, except in special cases (e.g., puppet [42], simple scenes [2,34,37,40], accelerometers [39], and depth camera and inertial sensor [44–46]).
- *Deep learning*: As hand-crafted descriptors mentioned above may lack discriminative capacity for action representation, deep learning methods aim to automatically learn the semantic representation from raw video by using a deep neural network. Typical methods include 3D ConvNets [47], Deep ConvNets [48], Two-Stream ConvNet [49], TDD [50], Latent Concept Descriptors [51], H-FCN [52], FCLN [53] and Conv Two-Stream [54].

1.1. Motivation and contributions

As shown in the last two rows of Fig. 1, complex human body poses, partial occlusions and motion blurs often appear in action videos, the human detector [29] and faster RCNN [55] do not always work perfectly. These components may lead to incorrect region detection problem when estimating the homography with feature matching. To automatically detect action regions without expensive training data and any human detector, motivated by saliency detection research [22,29,56,57], a global contrast based segmentation algorithm was introduced to produce region-based contrast maps (RC-map).

Although RC-maps can constrain feature points on salient regions, the global contrast based segmentation algorithm, which uses image contrast under the assumption that a salient object exists in an image, was not suitable for action videos. Because it may result in unstable masks with respect to consecutive frames. Therefore, partly inspired by motion boundary researches [17,27], we applied morphological gradient to optimize RC-maps to generate more robust masks. It is named region-based boundary maps (RCB-map). The RCB-maps could capture discriminative appearance information on salient region boundaries.

However, action recognition becomes a challenging problem due to the motions of camera and the variations in pose, appearance, background, etc. It should be noted that action recognition cannot be achieved by merely employing an object detection or segmentation algorithm. As Fig. 1 shows, there are numerous irrelevance trajectories in real world videos due to camera motion. Hence, the action representation is prone to be inaccurate. Meanwhile, merely using improved dense sampling on RCB-maps (iDT-rawRCB) also cannot promote the recognition, as RCB-maps may not be able to capture the relative displacements while resisting background motions in a subset of consecutive frames.

To address the above problem, traditional motion estimation approaches model the global camera motions by using motion vector decomposition [58,59], or warp optical flow with a robustly estimated homography [5,29]. In this paper we assume that the true flow can be established by a normalized warped optical flow at each point of consecutive frames, and then a normalized magnitude of warped flow was defined to capture salient relative displacements, while the tiny ones lower than a threshold are regarded as meaningless. Since recognition task always benefit from dense features but sparse [60], we should replace feature points with those points sampled by original iDT after excluding minimum warped flow, when the sampled ratio becomes relative small.

The proposed method including iDT on RCB-map and warped flow pruning with dense feature supplementary scheme is named iDT-RCB. We extensively evaluate our method on Hollywood2 dataset. Inspired by the deep learning approaches [49,50,54], we also evaluate the fusion methods combining Convolutional Neural Network (CNN) architectures with our iDT-RCB. The fusion of them achieves the state-of-the-art performance on Hollywood2 and UCF50 datasets. The contributions of this paper are summarized as follows:

- We propose a salient region-based dense sampling method to conquer the region detection and motion evaluation problem. Different from the previous dense sampling methods [22,27,29] on pruning features, our method mines the saliency of distant regions from consecutive frames without automatic human detection.
- We evaluate the effectiveness of removing tiny motions from warped optical flow. When pruning tiny motions by a suitable magnitude threshold, the remainder of warped flow are regarded as salient motions between frames. In other words, the action recognition can benefit from salient region masks and salient motion displacements.
- We exploit information not only for the hand-crafted features but also for the fusion of deep-learned features. We separately present the results of iDT-RCB and our best results obtained with early fusion of TDD and iDT-RCB.

A preliminary version of this work appeared in [33]. This paper extends the earlier work [33] as follows. Firstly, we reveal the motion cues between salient displacements and warped optical flow through analysis and experiments. It demonstrates that our method can benefit from salient region boundaries (i.e., RCB-maps) and salient warped flow. Secondly, we propose a dense supplementary scheme to overcome the problem of extremely sparse features, when RCB-maps fail to capture enough features in some cases. Experimental results show that a suitable pruning of feature points represents good compromise between saliency and density of the sampled points. Thirdly, we comprehensively compare our method with state-of-the-art approaches. Extensive experiments on Hollywood2 dataset are also conducted. By combining CNN features with our method, the recognition results are further improved.

The remainder of this paper is organized as follows. In Section 2, a brief review of related work on dense sampling for action recognition is given. In Section 4, we describe the reason of region-based contrast boundary mapping. Then, we evaluate the influence of removing tiny motions according to warped optical flow, and provide a dense feature supplementary scheme in Section 5. Section 6 shows experimental results and finally Section 7 concludes this paper.

2. Related work

In this section we give a brief review of the related work on dense sampling. Current dense sampling researches can be generally categorized into two classes: saliency based and dense based approaches.

The saliency based approaches pay attention to seeking out a proper salient mask, which is a crucial aspect in feature points selection procedure. Dalal et al. [17] studied a descriptor using motion boundary based coding to capture shape, appearance and motion information. Willems et al. [18] proposed dense and scale-invariant spatio-temporal interest point, which is a spatiotemporal extension of the Hessian saliency measure. Vig et al. [22] applied saliency-mapping algorithms to prune background features. This results in a more compact video representation, and improves action recognition accuracy. Jain et al. [26] decomposed visual motion into dominant and residual motions, and designed a new descriptor to capture additional information on the local motion patterns. Ballas et al. [56] identified prominent regions in videos content through motion, illumination and cornerness saliencies, and introduced a new space–time invariant pooling scheme. Peng et al. [27] constrained the sampled points on large magnitude regions of motion boundary image in the sampling step. Mathe et al. [61] pruned background features based on visual saliency. Li et al. [62] applied multiple instance learning on top of dense trajectory features in order to learn mid-level action to better represent human actions. However, it is obviously inappropriate to merely employ a uniform algorithm for action representation, which has been discussed in Section 1.

The dense based approaches aim to capture more tiny body motion that can easily separate different actions in various videos. Wang et al. [24] tried to sample feature points on dense grid in each frame, and tracked them based on dense optical flow. Jiang et al. [63] clustered dense trajectories, and utilized the cluster centers as reference points so that the relationship between them can be modeled. Wang et al. [23] densely sampled video patches with the optimizing position and scale parameters to guarantee that the features are shift and scale invariant. Shi et al. [25] explored sampling over high density with local spatio-temporal features extracted from a Local Part Model. Peng et al. [64] stacked two FV encoding layers via a hierarchical structure, and described a max-margin dimensionality reduction algorithm to compress densely sampled subvolumes. Simonyan et al. [49] built a Two-Stream ConvNet architecture which incorporates spatial and temporal networks. Wang et al. [50] presented a trajectory-pooled deep convolutional descriptor, which shares the merits of both hand-crafted features and deep-learned features. Nevertheless, constructing deep-learning features leads to high computational complexity problem. On the other hand, the hand-crafted feature can also be complementary to CNN. For example, by combining the simple local features (e.g, iDT [29]) and deep CNN features (e.g., Two-Stream ConvNet [49]), Wang et al. [50] could promote the recognition result on HMDB51 from 63.2% to 65.9%. So in this paper, we focus the hand-crafted feature in the traditional methods.

There are still few methods trying to refine the multi-class problem by active learning or semi-supervised learning. Yang et al. [65] proposed a semi-supervised batch mode multi-class active learning algorithm for visual concept recognition. Recently, semi-supervised learning [66–68] also have been proposed. However, the effectiveness of these methods is highly depending on the batch sizes or fine-tuned parameters with respect to a different dataset. Therefore, this kind of methods is beyond the scope of this paper.

3. Improved dense trajectories revisited

As shown in Fig. 2, our proposed method (iDT-RCB) is based on

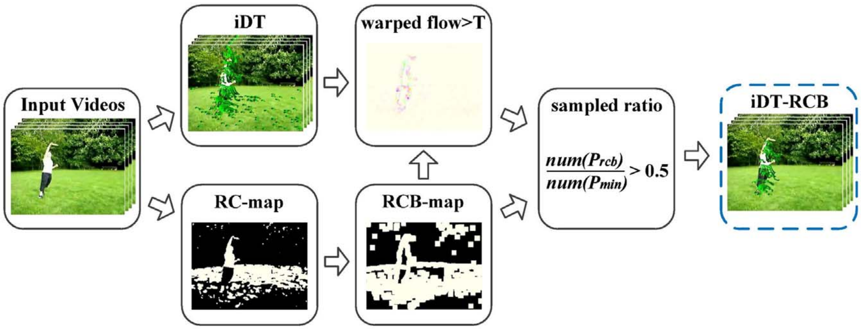

Fig. 2. Comparison of proposed approach (iDT-RCB) with traditional approach (iDT) for action recognition. Points sampled by iDT-RCB are more effective than iDT, because action regions have been detected by saliency-mapping of region-based contrast boundary and salient warped optical flow. Green trajectories indicate that the sampled points have been tracked for fixed length of frames. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

low level trajectory extraction and we choose improved dense trajectories [29]. In this section, we briefly review the extraction process of improved dense trajectories. We use improved dense trajectories due to its good performance.

Improved dense trajectories are extended from dense trajectories [24]. To compute dense trajectories, the first step is to densely sample a set of points on 8 spatial scales on a grid with step size of 5 pixels. Points in homogeneous areas are eliminated by setting a threshold for the smaller eigenvalue of their autocorrelation matrices. Then these sampled points are tracked by median filtering of dense flow field

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

where a point $P_t = (x_t, y_t)$ is given in frame I_t, its tracked position in frame I_{t+1} is smoothed by applying a median filter kernel M on ω_t. For each frame I_t, its dense optical flow field ω_t is computed with respect to the next frame I_{t+1}.

To avoid the drifting problem of tracking, the sampled points are tracked for 15 frames. Then they are removed and replaced by new feature points. Those static trajectories and others with suddenly large displacement are both ignored, since they are incorrect due to inaccurate optical flow.

The iDT approach boosts the recognition performance of dense trajectory by explicit camera motion estimation. The iDT approach first finds the correspondence between two consecutive frames. According to the SURF feature matching and optical flow based matching, they use the RANSAC algorithm to estimate the homography matrix. Then, they warp the second frame with the estimated homography and recompute dense optical flow, called warped flow. Warped flow brings advantages to the descriptors calculated from optical flows, in particular for HOF and MBH.

We adopt many steps of the iDT sampling method and make a modification as described in Algorithm 1. Different from the iDT, we constrain feature points with RCB-map and salient warped flow instead of human detector. We observe that tracking those points is effective for action representation. In summary, given a video V, we obtain a set of points $P_{rcb}(V)$ after applying RCB-map and salient warped flow, while a set of points $P_{min}(V)$ is also obtained merely employing salient warped flow. When the sampled ratio is lower than a threshold, it means that the set of points $P_{rcb}(V)$ may have too few features to represent an action, thus we will automatically replace $P_{rcb}(V)$ with $P_{min}(V)$. Tracking these points for consecutive frames resulting in a set of trajectories $Tr(V)$.

4. Region-based contrast boundary sampling

In this section, we introduce a global contrast based salient region detection algorithm [57] in feature points sampling step. We also explain why this algorithm does not perform well in action videos, then we present our new sampling strategy on salient region boundary in detail.

4.1. Global contrast based salient region sampling

Although the iDT can benefit from the camera motion compensation, the performance of action recognition still suffers from the inevitable movements of cameras. The truth is that most of the challenging action datasets contain lots of camera motions. For example, HMDB51 has 59.9% videos including camera motion [5]. Hence, we should study how to capture salient appearance in consecutive frames.

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 [29]. 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 4th column of Fig. 3(a). Hence, in order to find out the attractive salient regions, we follow [27] to create the mask named Motion Boundary Image (MBI). But improved dense trajectories on motion boundary images (iDT-MB) are not stable, since the motion boundaries are significantly influenced by the threshold value on gradient variation of optical flow. See the 2nd column of Fig. 3(d), it shows the effective sampling example, but fails to capture the meaningful ones in Fig. 3(b) due to the unstable performance of MBI threading. A worse result is given in the 2nd column of Fig. 3(c), almost nothing is left in some cases.

Motivated by saliency detection research [22,29,56,57], we introduce a contrast based segmentation algorithm to produce region-based contrast maps. This improved dense trajectory on region-based contrast maps (iDT-RC) is partly inspired by Global Contrast based Salient Region Detection [57]. The main idea of iDT-RC is to automatically estimate salient object regions across every frame and enhance iDT sampling method without any prior knowledge of the video content. The iDT-RC sampling includes three steps:

(1) We first use a graph-based image segmentation method [57] 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:

$$S(r_k) = \sum_{r_k \neq r_i} w(r_i) D_r(r_k, r_i), \tag{2}$$

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.

(2) We further incorporate spatial information by introducing a spatial weighting term in Eq. (2) 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

$$S(r_k) = \sum_{r_k \neq r_i} \exp\left(-\frac{D_s(r_k, r_i)}{\sigma_s^2}\right) w(r_i) D_r(r_k, r_i), \tag{3}$$

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

(3) To save the useful feature points in every frame, we follow the RC-map [57] approach to obtain a segmentation mask, and apply the estimated salient mask to iDT sampling method. Those feature points sampled by the iDT-RC but not in global contrast based salient regions will be deleted.

4.2. 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, aims 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 obtain some unexpected masks due to its global color contrast, see the 3rd column of Fig. 3(c). Secondly, sometimes the salient region generated by RC-map is too limited to track enough feature points for representing an action, the discriminative ones may be not saved, see the 3rd column of Fig. 3(a). Last but not the least, not all the appearance regions are valid to represent actions of people.

Therefore, in order to handle the above issues, we try to sample improved dense trajectories on raw region-based contrast boundary (iDT-rawRCB). Unlike [22,23,25,26], our iDT-rawRCB sampling strategy constrains the sampled points on salient region boundaries in the sampling step. We use three iterations of the morphological gradient on RC-map to generate a robust RCB-map. The morphological gradient can be expressed as

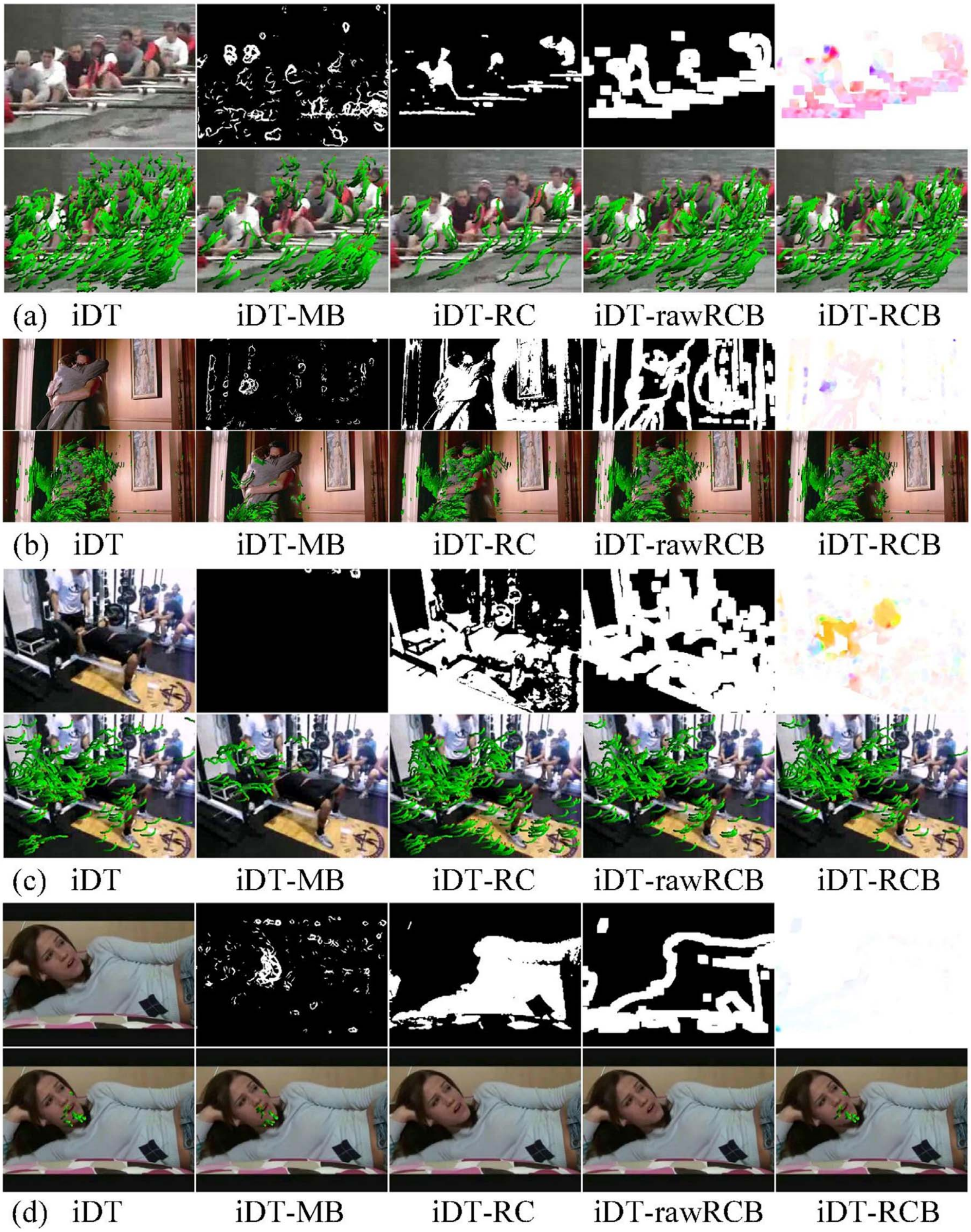

Fig. 3. Visualization of iDT, iDT-MB, iDT-RC, iDT-rawRCB and iDT-RCB sampling strategies for 4 actions. Compared to iDT, iDT-MB can reduce irrelevant motions, but it is not stable. iDT-RC can handle salient regions, but it cannot capture the salient boundaries accurately. Although iDT-rawRCB merely employing RCB-maps is robust to salient regions, particular at shot boundaries, the iDT-RCB can benefit from salient appearance cues (i.e., RCB-maps) and salient motion cues (i.e., warped flow), as shown in (a)–(c). The last two rows demonstrate the failure case of RCB-map due to global contrast variations, resulting in no facial expressions on shot boundaries by iDT-rawRCB. To conquer the above issue, a dense supplementary scheme is applied to iDT-RCB. The RCB-maps are replaced by warped flow for saliency-mapping, when feature points are too few to capture informative motions, as described in (d).

$$RCBmap = morph_{grad}(RCmap) = dilate(RCmap) - erode(RCmap), \tag{4}$$

5. Warped optical flow evaluation

This section gives a brief evaluation of removing tiny motions from warped optical flow. A dense supplementary scheme is provided for better feature space distribution. We also detail our new sampling strategy based on salient region boundary and salient warped flow.

5.1. Salient motions evaluation

As trajectories generated by camera motion can be removed by warped flow [29], it is reasonable to assume that the true flow can be modeled by a normalized warped optical flow at each point of

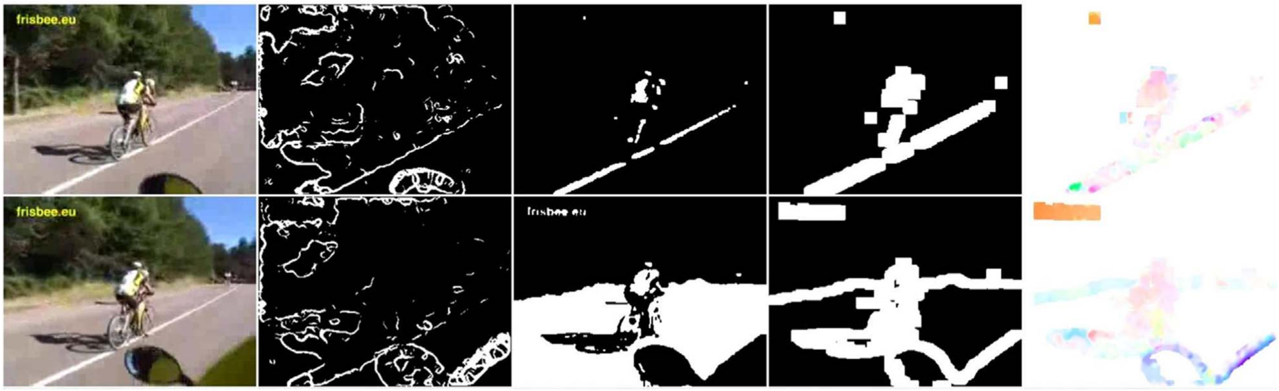

Fig. 4. RCB-map using morphological gradient is more robust than RC-map and MBI for salient region segmentation in action videos, see the 4th column. Note that iDT-rawRCB merely apply the RCB-map, whereas iDT-RCB combine RCB-map with salient warped optical flow, and dense supplementary scheme.

consecutive frames. For performance measures of optical flow, Baker et al. [69] extended the sum-of-squared-difference (SSD) to compute frame interpolation. Inspired by his work, we define a normalized magnitude of warped optical flow

$$\mathcal{I}(i) = \left[\frac{(\mathcal{I}_i^u)^2 + (\mathcal{I}_i^v)^2}{\max_{\forall i \in \mathcal{I}}(\| \mathcal{I}_i^{\omega}(u, v)\|^2) + \varepsilon}\right]^{\frac{1}{2}}, \tag{5}$$

where $\mathcal{I}_i^u$ and $\mathcal{I}_i^v$ stands for each point i of image $\mathcal{I}$ containing the u (horizontal) and v (vertical) components of optical flow, respectively, $\mathcal{I}_i^{\omega} = (\mathcal{I}_i^u, \mathcal{I}_i^v)$ stands for the 2D flow image $\mathcal{I}^{\omega}(u, v)$ at point i. In our experiments $\varepsilon = 1.0$ (grey-levels per pixel squared).

We illustrate the difference between the original and corrected optical flow in the middle two columns of Fig. 5. The original optical flow field contain lots of inaccurate motion vectors due to camera motions. We follow [29] to warp optical flow resulting in stabilized motion vectors. We not only compute the maximal magnitude of the motion vectors during its length of each trajectory as demonstrated in [29], but also compute the maximal normalized magnitude using Eq. (5). If the normalized magnitude of $\mathcal{I}(i)$ is lower than a threshold, the displacement at point i is considered to be consistent with camera motion, and thus removed.

Although most of the motions in human body areas have been removed when a threshold of 0.4 is set, salient motions still benefit from a suitable threshold like 0.001, see the 3rd column of Fig. 5. Hence, we can sample improve dense trajectories merely excluding minimum warped flow (iDT-min). This gives us similar effects as sampling features based on visual saliency maps [22,27,61].

5.2. Dense feature supplementary scheme

In this subsection, we validate the proposed approach and describe the importance of denseness for the contribution of salient warped flow.

Influence of T: To reduce the impact of tiny motions, we set a threshold T in this stage. When $T \to 0$, it is equal to only apply the iDT method; inversely, fewer points are sampled to generate result for $T \to 1$. We change T from 1E−6 to 0.4, the corresponding results are shown in Fig. 6. It is obvious that 0.01 is a good choice for on the HMDB51 dataset.

However, recognition task always benefit from dense features but sparse [60]. Since iDT-min has pruned tiny motions, merely applying RCB masks on iDT-min sampling will lead to much fewer trajectories, and then result in performance degradation. Hence, we propose a dense feature supplementary scheme. When the sampled ratio is less than a threshold, we will automatically replace points P_{rcb} sampled by iDT-RCB with points P_{min} sampled by iDT-min. The sampled ratio is defined as $num(P_{rcb})/num(P_{min})$, under the current scale of each frame. In this paper we set the threshold of sampled ratio as 0.5 empirically.

We conduct experiments for evaluating the effectiveness of iDT-RCB with dense supplementary scheme. Fortunately, iDT-RCB has derived benefit from salient motions and dense features, as was expected in Fig. 6.

5.3. Summary of iDT-RCB sampling

In our approach, we follow [29] to initialize the sampling parameters, such as trajectory length, spatial scale, spatial cells, temporal cells, eigenvalues threshold, etc. These parameters are set the same as iDT [29]. This initialization will not affect the sampling results (i.e., the number of sampled points). But the iDT-RCB sampling results will be affected by RCB-maps. This is because the SaliencyCut [51] using graphcut and GMM mode leads to iterative process, there may be a slight difference in generalized results of RC-map. Although the RCB-map produced by RC-map is not very stable, the discriminative power from salient region boundary still remained. All recognition results on three datasets are better than iDT with Human Detector, as we can see in Table 1. Nevertheless, the final recognition results may be affected by a trained codebook (e.g., GMM) from randomly selected trajectories.

The proposed iDT-RCB sampling is described below in detail. Note that there are three differences between iDT-RCB and iDT sampling [29]. Firstly, we add step 3 for generating RCB-map. Secondly, we add steps 10 and 11 for evaluating warped optical flow. Thirdly, we modify steps 4 and 6 for replacing Human Detector with RCB-map.

Algorithm 1. iDT-RCB sampling procedure.

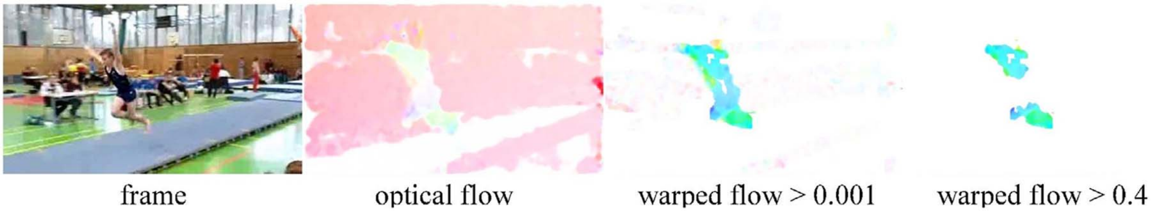

Fig. 5. Examples of removing tiny motions via threshold magnitude of warped optical flow. While setting threshold value to 0.001 can preserve plenty of tiny motions, valid human body motions may be pruned by a value of 0.4.

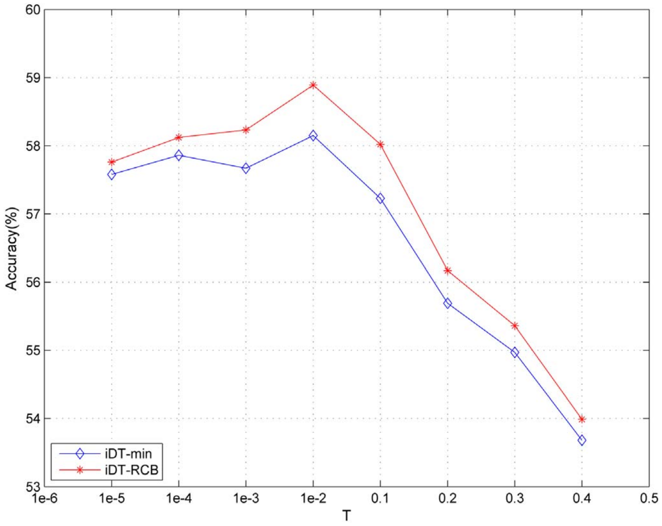

Fig. 6. Comparative results of different T on HMDB51 datasets. Note that iDT-min and iDT-RCB stands for different T on iDT and iDT-rawRCB, respectively.

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

Hollywood2		HMDB51		UCF50	
Jain et al. [26]	62.5%	Simonyan et al. [49]	59.4%	Reddy et al. [6]	76.9%
Ni et al. [74]	66.7%	Wang et al. [50]	65.9%	Shi et al. [25]	83.3%
Hoai et al. [73]	72.7%	Feichtenhofer et al. [54]	**69.2%**	Wang et al. [72]	85.7%
Lan et al. [75]	68.0%	Lan et al. [75]	65.4%	Lan et al. [75]	94.4%
iDT without HD [29]	63.0%	iDT without HD [29]	55.9%	iDT without HD [29]	90.5%
iDT with HD [29]	64.3%	iDT with HD [29]	57.2%	iDT with HD [29]	91.2%
iDT-MB	61.9%	iDT-MB	53.3%	iDT-MB	88.4%
iDT-RC	62.7%	iDT-RC	55.7%	iDT-RC	90.8%
iDT-rawRCB	64.3%	iDT-rawRCB	57.8%	iDT-rawRCB	91.3%
iDT-RCB	65.2%	iDT-RCB	58.9%	iDT-RCB	92.0%
TDD[50]+iDT-RCB	**73.8%**	TDD[50]+iDT-RCB	66.4%	TDD[50]+iDT-RCB	**94.8%**

```
Input:
    $VideoFrames = \{I_1, I_2, \dots, I_N\}$;
Output:
    $ValidTrajectories = Tr_1, Tr_2, \dots, Tr_M$;
Initialize the sampling parameters
for $i$=1 to $N$ do
    generate the $RCB-map$ by using three iterations of Eq. (4)
    $P_j^{(1)} \Leftarrow denseSample(greyI_i, RCB-map)$ for each scale. $Tr_j^{(1)} \Leftarrow P_j^{(1)}$
    $\omega_i \Leftarrow$ compute dense optical flow by Farnebäck algorithm
    $matches_i \Leftarrow matchFromSurfandFlow(greyI_i, \omega_i, RCB-map)$
    $H_i \Leftarrow findHomography(matches_{i-1}, matches_i, RANSAC)$
    warp the second frame with $H_i$
    $\omega_i' \Leftarrow$ re-compute dense optical flow by warped second frame
    remove $P_j^{(1)}$ when $\mathcal{I}(P_j^{(1)}) < T$ according to Eq. (5)
    replace $P_{rcb}$ with $P_{min}$ if sampled ratio <0.5
    predict the motion of $P_j^{(t+1)}$ by using $\omega_i'$
    $Tr_j \Leftarrow \{P_j^{(1)}, P_j^{(2)}, \dots, P_j^{(t)}, P_j^{(t+1)}, \dots, P_j^{(L)}\}$
    if $Tr_j$ is valid & & $Tr_j$ is not camera motion then
        $ValidTrajectories \Leftarrow Tr_j$
    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 into 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 [17], Dmask including narrow strip surrounding the persons contour [42], and motion boundary contour system in neural dynamics of motion perception [70].

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 5th column of Fig. 4.

6. Experiments

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

6.1. Datasets

We conduct experiments on three action datasets, namely Hollywood2 [3], HMDB51 [5] and UCF50 [6]. Some example frames are illustrated in Fig. 7. We summarize them and the experimental protocols as follows.

The Hollywood2 dataset has been collected from 69 different Hollywood movies and includes 12 action classes. It contains 1707 videos split into a training set (823 videos) and a test set (884 videos). Training and test videos come from different movies. The performance is measured by mean average precision (mAP) over all classes.

The HMDB51 dataset is collected from a variety of sources ranging from digitized movies to YouTube videos. There are 51 action categories and 6766 video sequences in HMDB51. We follow the

(a) HandShake (a) GetOutCar (b) Push-Up (b) Chew (c) Horse-Race (c) Playing-Guitar
(a) HugPerson (a) AnswerPhone (b) Cartwheel (b) Pour (c) Punch (c) Ski-Jet

Fig. 7. Example frames from (a) Hollywood2, (b) HMDB51 and (c) UCF50.

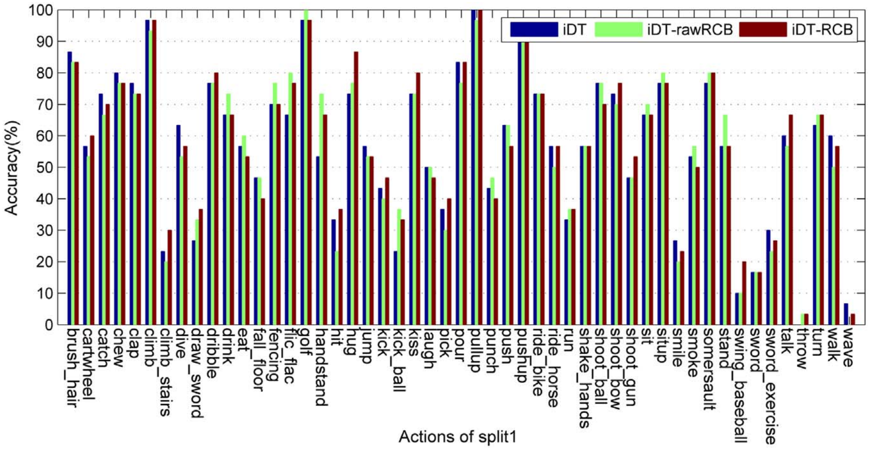

Fig. 8. The split1 results of all the action categories from HMDB51 dataset.

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 6618 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.

6.2. Experimental setup

In the following experiments, we densely extract improved trajectories based on the code [29]. The iDT-MB is implemented by the code [27]. The iDT-RC and iDT-RCB is partly implemented by the code [29,57].

To recognize actions, we evaluate the presented sampling methods at the same server cluster with multithreading, and follow [29,64] 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 [29]. 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 Whitening and L2-normalization. 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 [71]. For multi-class classification, we use the one-vs-rest approach and select the class with the highest score.

We compare our methods with recent works [6,25,26,29,50,49,54,72–75]. The mean run-time of sampling process and the mean number of sampled trajectories are compared with iDT. The processing speed is reported in frames per second (fps), run at a single-core Intel Xeon X3430 (2.4 GHz) without multithreading.

6.3. Combining with convolutional neural network

In this experiment we also evaluate the effect of combining CNN features which are encoded by trajectory-pooled deep-convolutional descriptor [50]. To reduce the influence of illumination, we use the combined representation obtained from spatiotemporal normalization and channel normalization for TDDs. To keep the dimensionality manageable, we fix the dimension D=64 reduced by PCA. Then, we train a GMM with K=256 Gaussians, and finally the video is represented with a $2DK$ dimensional vector. This is exactly the same setup used by Wang et al. [50]. However, combining ConvNets with high performance is not the final goal of this paper, and we aim to verify the effectiveness of iDT-RCB. Hence, we combine trajectory features like iDT-RCB with CNN features obtained by the code of TDD[50], using early fusion of Fisher vector representation.

6.4. Results and analysis

Since the SaliencyCut [57] 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 4th column of Fig. 3. Furthermore, the influence of removing tiny motions from warped optical flow is evaluated. We also provide a dense feature supplementary scheme to remain the denseness of feature distribution.

On all the datasets we used, the proposed method achieves comparable performance with respect to traditional iDT. To further investigate the effects of RCB masks on traditional iDT, we illustrate the split1 recognition results of all the action classes from HMDB51 dataset in Fig. 8.

We also compare our method with the recent results reported in Table 1. The iDT without HD combining with FV encoding [29] is taken as baseline, the accuracy of iDT-RCB on Hollywood2, HMDB51 and UCF50 is improved by 2.2%, 3% and 1.5%, respectively. Our iDT-RCB implementation achieves the best result on UCF50, while the result on HMDB51 is slightly decreased than [49], which have used the trained deep Convolutional Networks.

As indicated by Table 1, when combining with MBI, the iDT-MB approach gets worse results on these datasets. One probable reason for this degradation is that they use improper mask for sampling, whereas the iDT-RC also miss many discriminative trajectories due to unstable mask, as it is shown in Fig. 4.

For verifying the effectiveness of our methods, we combine CNN features (e.g., TDD) with iDT-RCB. As frame based features improve from CNN, the fusion of them can further boost the performance. This further improvement indicates that our iDT-RCB are complementary to those deep-learned features. The recognition results are shown in Table 1, an interesting comparison is against the Conv Two-Stream [54], which employs VGG-16 for both streams with fusion by 3D Conv and 3D Pooling. Although the convolutional layers of TDD are fewer than Conv Two-Stream [54], our results on three datasets still underline the importance of our proposed method. Meanwhile, our results on Hollywood2 and UCF50 obtain the state-of-the-art performance.

Evaluation results of sampling strategy are presented in Table 2. We report the average number of trajectories per video clip and the fps

Table 2
Comparison of sampled trajectories number and features extraction speed to iDT [29]. Note that we only randomly select 10 videos from each dataset.

Sampling strategy	Hollywood2		HMDB51		UCF50	
	Trajectories/clip	fps	Trajectories/clip	fps	Trajectories/clip	fps
iDT without HD [29]	151,714	1.81	12,223	3.59	24,679	3.90
iDT with HD [29]	153,946	1.85	12,300	3.60	24,672	3.90
iDT-MB	7682	1.63	4710	3.35	9542	3.68
iDT-RC	91,967	1.57	9041	2.93	18,267	3.25
iDT-rawRCB	96,184	1.53	9253	2.86	18,657	3.16
iDT-RCB	117,233	1.49	10,528	2.82	21,802	3.13

within 10 videos randomly selected from each dataset. The trajectories reduction does not reduce the accuracy. Indeed, a limited reduction and efficient selection tend to improve the accuracy with minor computational cost, as we can see from the last row of Table 2. Taking into account the subsequent recognition procedure, fewer trajectories also lead to faster video encoding process.

7. Conclusion

This paper proposes a novel dense sampling approach without human detection. We introduce a salient region contrast based segmentation method in feature points sampling step. To overcome the flaws of salient region contrast based method in action videos, we apply morphological gradient to RC-map for generating more robust salient mask. This improved sampling method 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. We also evaluate the salient motions by setting minimum threshold of warped optical flow. Experimental results describe that a suitable threshold (depending on model) represents a good compromise between salient motions and feature denseness. Finally, our method improves the recognition on three benchmark datasets, the fusion of CNN features and our iDT-RCB can achieve the state-of-the-art performance on Hollywood2 and UCF50.

Acknowledgment

The research was supported by the National High Technology Research and Development Program of China (863 Program) (2015AA016306), the National Nature Science Foundation of China (61231015, 61367002, 61502152, 61501413, 11561015), 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 Nature Science Foundation of Hubei Province (2014CFB712).

References

[1] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004, ICPR 2004, vol. 3, IEEE, Cambridge, UK, 2004, pp. 32–36.
[2] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.
[3] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008, IEEE, Anchorage, AK, USA, 2008, pp. 1–8.
[4] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE, Miami, Florida, USA, 2009, pp. 1996–2003.
[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, Barcelona, Spain, 2011, pp. 2556–2563.
[6] K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos, Mach. Vis. Appl. 24 (5) (2013) 971–981.
[7] I. Laptev, On space–time interest points, Int. J. Comput. Vis. 64 (2–3) (2005) 107–123.
[8] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, IEEE, Beijing, China, 2005, pp. 65–72.
[9] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th International Conference on Multimedia, ACM, Augsburg, Germany, 2007, pp. 357–360.
[10] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in: BMVC 2008—19th British Machine Vision Conference, British Machine Vision Association, Leeds, UK, 2008, pp. 275–1.
[11] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE, Miami, Florida, USA, 2009, pp. 2004–2011.
[12] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, Japan, 2009, pp. 492–497.
[13] M.-y. Chen, A. Hauptmann, Mosift: Recognizing Human Actions in Surveillance Videos, Carnegie Mellon University Technical Report, 2009.
[14] V. Kantorov, I. Laptev, Efficient feature extraction, encoding and classification for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2593–2600.
[15] H. Chen, J. Chen, H. Li, Z. Xu, R. Hu, Compressed-domain based camera motion estimation for realtime action recognition, in: Advances in Multimedia Information Processing—PCM 2015, Springer, Gwangju, South Korea, 2015, pp. 85–94.
[16] Chen Huafeng, Chen Ruimin, Structural imosift for human action recognition, Wuhan Univ. J. Nat. Sci. 21 (3) (2016) 262–266.
[17] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, in: Computer Vision—ECCV 2006, Springer, Graz, Austria, 2006, pp. 428–441.
[18] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Computer Vision—ECCV 2008, Springer, Marseille, France, 2008, pp. 650–663.
[19] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: BMVC 2009—British Machine Vision Conference, BMVA Press, London, UK, 2009, pp. 124–1.
[20] T.-H. Yu, T.-K. Kim, R. Cipolla, Real-time action recognition by spatiotemporal semantic and structural forests., in: BMVC, Aberystwyth, Wales, UK, vol. 2, 2010, p. 6.
[21] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Colorado Springs, CO, USA, 2011, pp. 3361–3368.
[22] E. Vig, M. Dorr, D. Cox, Space-variant descriptor sampling for action recognition based on saliency and eye movements, in: Computer Vision—ECCV 2012, Springer, Florence, Italy, 2012, pp. 84–97.
[23] B. Wang, Y. Liu, W. Xiao, Z. Xiong, W. Wang, M. Zhang, Human action recognition with optimized video densely sampling, in: 2013 IEEE International Conference on Multimedia and Expo (ICME), IEEE, London, U.K., 2013, pp. 1–6.
[24] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, Int. J. Comput. Vis. 103 (1) (2013) 60–79.
[25] F. Shi, E. Petriu, R. Laganiere, Sampling strategies for real-time action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2595–2602.
[26] M. Jain, H. Jégou, P. Bouthemy, Better exploiting motion for better action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2555–2562.
[27] X. Peng, Y. Qiao, Q. Peng, Motion boundary based sampling and 3D co-occurrence descriptors for action recognition, Image Vis. Comput. 32 (9) (2014) 616–628.
[28] L. Wang, Y. Qiao, X. Tang, Motionlets: Mid-level 3D parts for human motion recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2674–2681.
[29] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.
[30] Y. Shi, W. Zeng, T. Huang, Y. Wang, Learning deep trajectory descriptor for action recognition in videos using deep neural networks, in: 2015 IEEE International Conference on Multimedia and Expo (ICME), IEEE, Chengdu, China, 2015, pp. 1–6.
[31] X. Chang, Y. Yang, E.P. Xing, Y.-L. Yu, Complex event detection using semantic saliency and nearly-isotonic svm, in: International Conference on Machine

Learning—ICML 2015, vol. 37, Lille, France, 2015.
[32] X. Chang, Y.-L. Yu, Y. Yang, E.P. Xing, They are not equally reliable: Semantic event search using differentiated concept classifiers, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[33] Z. Xu, R. Hu, J. Chen, H. Chen, H. Li, Global contrast based salient region boundary sampling for action recognition, in: MultiMedia Modeling, Springer, Miami, USA, 2016, pp. 187–198.
[34] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and Viterbi path searching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, CVPR'07, IEEE, Minneapolis, Minnesota, USA, 2007, pp. 1–8.
[35] A. Yao, J. Gall, G. Fanelli, L.J. Van Gool, Does human action recognition benefit from pose estimation?, in: BMVC, vol. 3, 2011, p. 6.
[36] M. Raptis, L. Sigal, Poselet key-framing: a model for human activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2650–2657.
[37] C. Wang, Y. Wang, A. Yuille, An approach to pose-based action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 915–922.
[38] A. Cherian, J. Mairal, K. Alahari, C. Schmid, Mixing body-part sequences for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2353–2360.
[39] F. Zhou, F. De la Torre, Spatio-temporal matching for human detection in video, in: Computer Vision—ECCV 2014, Springer, Zürich, Switzerland,2014, pp. 62–77.
[40] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.
[41] S. Sadanand, J.J. Corso, Action bank: a high-level representation of activity in video, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, Rhode Island, USA, 2012, pp. 1234–1241.
[42] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. Black, Towards understanding action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3192–3199.
[43] C. Chen, B. Zhang, Z. Hou, J. Jiang, M. Liu, Y. Yang, Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features, Multimed. Tools Appl. (2016) 1–19.
[44] C. Chen, R. Jafari, N. Kehtarnavaz, Improving human action recognition using fusion of depth camera and inertial sensors, IEEE Trans. Hum.-Mach. Syst. 45 (1) (2014) 51–61.
[45] C. Chen, R. Jafari, N. Kehtarnavaz, A survey of depth and inertial sensor fusion for human action recognition, Multimed. Tools Appl. (2015) 1–21.
[46] C. Chen, R. Jafari, N. Kehtarnavaz, A real-time human action recognition system using depth and inertial sensor fusion, IEEE Sens. J. 16 (3) (2015) 773–781.
[47] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[48] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[49] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.
[50] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.
[51] Z. Xu, Y. Yang, A.G. Hauptmann, A discriminative cnn video representation for event detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1798–1807.
[52] L. Wang, Y. Qiao, X. Tang, L. Van Gool, Actionness estimation using hybrid fully convolutional networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[53] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[54] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[55] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. (2016) 1.
[56] N. Ballas, Y. Yang, Z.-Z. Lan, B. Delezoide, F. Preteux, A. Hauptmann, Space–time robust representation for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2704–2711.
[57] M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[58] S. Wu, O. Oreifej, M. Shah, Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories, in: 2011 International Conference on Computer Vision, IEEE, Barcelona, Spain, 2011, pp. 1419–1426.
[59] M.A. Hasan, M. Xu, X. He, C. Xu, Camhid: camera motion histogram descriptor and its application to cinematographic shot classification, IEEE Trans. Circuits Syst. Video Technol. 24 (10) (2014) 1682–1695.
[60] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: Computer Vision—ECCV 2006, Springer, Graz, Austria, 2006, pp. 490–503.
[61] S. Mathe, C. Sminchisescu, Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition, IEEE Trans. Pattern Anal. Mach.Intell. 37 (7) (2015) 1408–1424.
[62] H. Li, J. Chen, Z. Xu, H. Chen, R. Hu, Multiple instance discriminative dictionary learning for action recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2016.
[63] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, C.-W. Ngo, Trajectory-based modeling of human actions with motion reference points, in: Computer Vision—ECCV 2012, Springer, Florence, Italy, 2012, pp. 425–438.
[64] X. Peng, C. Zou, Y. Qiao, Q. Peng, Action recognition with stacked fisher vectors, in: Computer Vision—ECCV 2014, Springer, Zürich, Switzerland, 2014, pp. 581–595.
[65] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. 113 (2) (2015) 113–127.
[66] Y. Yang, F. Wu, F. Nie, H.T. Shen, Y. Zhuang, A.G. Hauptmann, Web and personal image annotation by mining label correlation with relaxed visual graph embedding, IEEE Trans. Image Process. 21 (3) (2012) 1339–1351.
[67] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, A.G. Hauptmann, Action recognition by exploring data distribution and feature correlation, in: Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, Rhode Island, USA, 2012, pp. 1370–1377.
[68] J. Jiang, J. Ma, C. Chen, X. Jiang, Z. Wang, Noise robust face imagesuper-resolution through smooth sparse representation, IEEE Trans. Cybernetics. 99, 2016, pp.1-12
[69] S. Baker, D. Scharstein, J. Lewis, S. Roth, M.J. Black, R. Szeliski, A database and evaluation methodology for optical flow, Int. J. Comput. Vis. 92 (1) (2011) 1–31.
[70] S. Grossberg, E. Mingolla, Neural dynamics of motion perception: direction fields, apertures, and resonant grouping, Percept. Psychophys. 53 (3) (1993) 243–278.
[71] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[72] L. Wang, Y. Qiao, X. Tang, Mining motion atoms and phrases for complex action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2680–2687.
[73] M. Hoai, A. Zisserman, Improving human action recognition using score distribution and ranking, in: Asian Conference on Computer Vision, Springer, Singapore, 2014, pp. 3–20.
[74] B. Ni, P. Moulin, X. Yang, S. Yan, Motion part regularization: Improving action recognition via trajectory selection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3698–3706.
[75] Z. Lan, M. Lin, X. Li, A.G. Hauptmann, B. Raj, Beyond gaussian pyramid: Multi-skip feature stacking for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 204–212.

Zengmin Xu received the B.S. and M.S. degrees from Changsha University of Science and Technology, Changsha, China, in 2003 and 2006, respectively, and is currently working toward the Ph.D. degree at the National Engineering Research Center for Multimedia Software (NERCMS), School of Computer, Wuhan University. His research interests include multimedia content analysis, computer vision, and pattern recognition.

Ruimin Hu received the B.S. and M.S. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and 1990, respectively, and the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 1994. He is the Dean of School of Computer, Wuhan University, Wuhan, China, and the Director of the National Engineering Research Center for Multimedia Software and the Key Laboratory of Multimedia Network Communication Engineering with Wuhan University. He is also the Executive Chairman of the Audio Video coding Standard (AVS) workgroup of China in Audio Section. He has authored two books and over 100 scientific papers. His research interests include audio/video coding and decoding, video surveillance, and multimedia data processing.

Jun Chen received the M.S. degree in Instrumentation from Huazhong University of Science and Technology, Wuhan, China, in 1997, and the Ph.D. degree in Photogrammetry and Remote Sensing from Wuhan University, Wuhan, China, in 2008. He is the Deputy Director of the National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University. His research interests include multimedia communications and security emergency information processing.

Chen Chen received the B.E. degree in automation from Beijing Forestry University, Beijing, China, in 2009, and the M.S. degree in Electrical Engineering from Mississippi State University, Starkville, in 2012 and the Ph.D. degree in the Department of Electrical Engineering at the University of Texas at Dallas, Richardson, TX, in 2016. He is currently a Post-Doc in the Center for Research in Computer Vision at University of Central Florida (UCF). His research interests include compressed sensing, signal and image processing, pattern recognition and computer vision. He has published over 40 papers in refereed journals and conferences in these areas.

Dr. Chen is the recipient of the David Daniel Fellowship Award (Best Doctoral Dissertation Award) from the University of Texas at Dallas in 2016. He received the top 10% paper award in the IEEE International Conference on Image Processing (ICIP), in 2015. He served as TPC Member and the Reviewer for various conferences. He is also an Active Reviewer for the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Transactions on Image Processing, the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, Image and Vision Computing, and Neurocomputing. He is an editor of KSII Transactions on Internet and Information Systems.

Huafeng Chen received the M.S. degree from Kunming University of Science and Technology, Kunming, China, in 2007, and is currently working toward the Ph.D. degree at the National Engineering Research Center for Multimedia Software (NERCMS), School of Computer, Wuhan University. His research interests include multimedia content analysis, computer vision, and pattern recognition.

Hongyang Li received the M.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2005, and is currently working toward the Ph.D. degree at the National Engineering Research Center for Multimedia Software (NERCMS), School of Computer, Wuhan University. His research interests include computer vision, and pattern recognition.

Qingquan Sun is currently an Assistant Professor in the School of Computer Science and Engineering at California State University San Bernardino, San Bernardino, USA. He received the Ph.D. degree in the field of Electrical and Computer Engineering at the University of Alabama, Tuscaloosa, AL, USA, in 2013. Dr. Sun has published around 20 journal/conference papers and book chapters. He also has led 3 NSF projects. Dr. Sun's research has been supported by U.S. National Science Foundation, U.S. Air Force Research Laboratory, and other sources. His research interests include intelligent sensing, distributed computing, signal processing, and machine learning in cyber physical systems.

Dr. Sun is internationally recognized in intelligent sensing and machine learning fields. He has served as an Associate Editor for three international journals, a Chair for an international conference, a Technical Program Committee (TPC) Member for 7 international conferences, and a Reviewer for around 20 international journals including IEEE Transactions on Human Machine Systems, IEEE Transactions on Systems, Man, and Cybernetics: System; IEEE Wireless Communication Magazine; IEEE Transactions on Industry Informatics; IEEE Transactions on Emerging Topics in Computing, etc.