Motion Boundary Based Sampling and 3D Co-occurrence Descriptors for Action Recognition

Xiaojiang Peng[a,b], Yu Qiao[b,c], Qiang Peng[a]

[a] *Southwest Jiaotong University, Chengdu, P.R. China*
[b] *Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, P.R. China*
[c] *The Chinese University of Hong Kong, Hong Kong, P.R. China*

Abstract

Recent studies witness the success of Bag-of-Features (BoF) frameworks for video based human action recognition. The detection and description of local interest regions are two fundamental problems in BoF framework. In this paper, we propose a motion boundary based sampling strategy and spatial-temporal (3D) co-occurrence descriptors for action video representation and recognition. Our sampling strategy is partly inspired by the recent success of dense trajectory (DT) based features [1] for action recognition. Compared with DT, we densely sample spatial-temporal cuboids along motion boundary which can greatly reduce the number of valid trajectories while preserve the discriminative power. Moreover, we develop a set of 3D co-occurrence descriptors which take account of the spatial-temporal context within local cuboids and deliver rich information for recognition. Furthermore, we decompose each 3D co-occurrence descriptor at pixel level and bin level and integrate the decomposed components with a multi-channel framework, which can improve the performance significantly. To evaluate the proposed methods, we conduct extensive experiments on three benchmarks including KTH, YouTube and HMDB51. The results show that our sampling strategy significantly reduces the computational cost of point tracking without degrading performance. Meanwhile, we achieve superior performance than the state-of-the-art methods. We report 95.6% on KTH, 87.6% on YouTube and 51.8% on HMDB51.

Keywords: Dense Trajectory, Action Recognition, 3D Co-occurrence Descriptors, Motion Boundary, Bag of Features

1. Introduction

Automatic recognition of human action in videos has been an active research area in recent years due to its wide range of potential applications, such as smart video surveillance, video indexing, human-computer interface, etc. Though various approaches have been proposed and significant progresses have been made, action recognition still remains a challenging task due to the high dimension and complexity of video data, the large intra-class variations, clutter, occlusion and other fundamental difficulties [2].

A fundamental problem in action recognition is how to represent an action video. The approaches for action video representation can be roughly divided into five categories: (1) dynamic model based approaches which apply statistical sequential models such as HMM and Bayesian network to describe the temporal states of actions [3, 4]; (2) human pose based approaches which utilize pose structure information [5, 6]; (3) global action template based approaches which construct global templates to capture appearance and motion information of the whole motion body [7, 8, 9]; (4) local feature based approaches which mainly extract spatial-temporal cuboids [10, 11, 12, 13, 14, 15, 16, 17] or motion parts [18, 19]; (5) unsupervised feature learning based methods which learn the representation by hierarchical networks or other models [20, 21, 22, 23].

Among the state-of-the-art methods, the representation of local spatial-temporal feature with Bag-of-Features (BoF) framework [24] is perhaps the most popular and successful one for action recognition. Laptev [25] developed space-time interest points (STIP) by extending the Harris detector to 3D domain. Dollar et al. [10] detected space-time salient points by applying 2D spatial Gaussian and 1D temporal Gabor filters. Willems et al. [26] utilized Hessian matrix to extract scale-invariant spatial-temporal interest points in videos. Wang et al. [14] densely sampled cuboids at regular positions and scales. For descriptors, well-known approaches include HOG/HOF [11], Cuboids [10], HOG3D [13], 3D-SIFT [27], and so on.

Recently, Wang et al. [15] proposed dense trajectory for sampling spatial-temporal interest points and introduced a novel descriptor named motion boundary histogram (MBH) for action recognition. The motion boundary is defined by the gradient magnitude of optical flow which is initially introduced in the context of human detection [28]. Extensive experiments on nine popular human action datasets demonstrated the excellent performance of this approach [1]. Though its great power, the DT based representation is

expensive in memory storage and computation due to the large number of densely sampled points.

In this paper, we first develop a motion boundary based sampling strategy named DT-MB to reduce the computation and storage consumption of previous DT based method. We start from densely sampled patches with grids in a frame. Meanwhile, motion boundary (Figure 1) is derived from optical flow and a binary mask is estimated from motion boundary. Then we remove those sampled regions that have very few overlaps with foreground in the mask. Central points of the rest patches are refined by averaging the location of occupied foregrounds within the patches. Our DT-MB is partly motivated by the fact that those trajectories on motion boundary are the most meaningful ones. This is also implied by the superior performance of the MBH descriptor [1]. Using our sampling method, the number of DTs can be sharply reduced without hurting the performance.

In addition, to further enhance the discriminative power of DT based representation, we propose a set of spatial-temporal (3D) co-occurrence descriptors to describe the local appearance and motion features along trajectories. This is partly inspired by the success of co-occurrence feature in image domain [29, 30, 31]. In [30], a descriptor based on co-occurrence HOG (CoHOG) is presented for human detection. In [29], gray-level co-occurrence matrix (GLCM) is introduced to extract textural features for image classification. Our motivation is that the spatial-temporal co-occurrence features, which depict the local tiny context of motion and appearance in videos, can provide important cues for action recognition. The novel descriptors are composed of 3D-CoHOG, 3D-CoHOF and 3D-CoMBH. We find that (1) 3D-CoHOG depicts more complex structure of spatial patch and the appearance changes along with time; (2) 3D-CoHOF conveys complex motion structure and motion direction changes; (3) 3D-CoMBH captures the complex gradient structure of optical flow and the changes of gradient orientations of flow. Furthermore, we thoroughly exploit two types of multi-channel pipelines for these descriptors, namely the pixel level pipeline and the bin level pipeline. Considering 3D-CoHOG in a given cuboid aligned by trajectory, we set several offsets at horizontal, vertical and temporal axes for each point, and the co-occurrence matrices in all the offsets are vectorized and concatenated to form 3D co-occurrence descriptors. For pixel level multi-channels of 3D-CoHOG, we vectorize the co-occurrence matrices for each offset and model them by the BoF pipeline individually, and then combine all the BoF pipelines by a multi-channel kernel SVM. For the channels in bin level, we split the co-

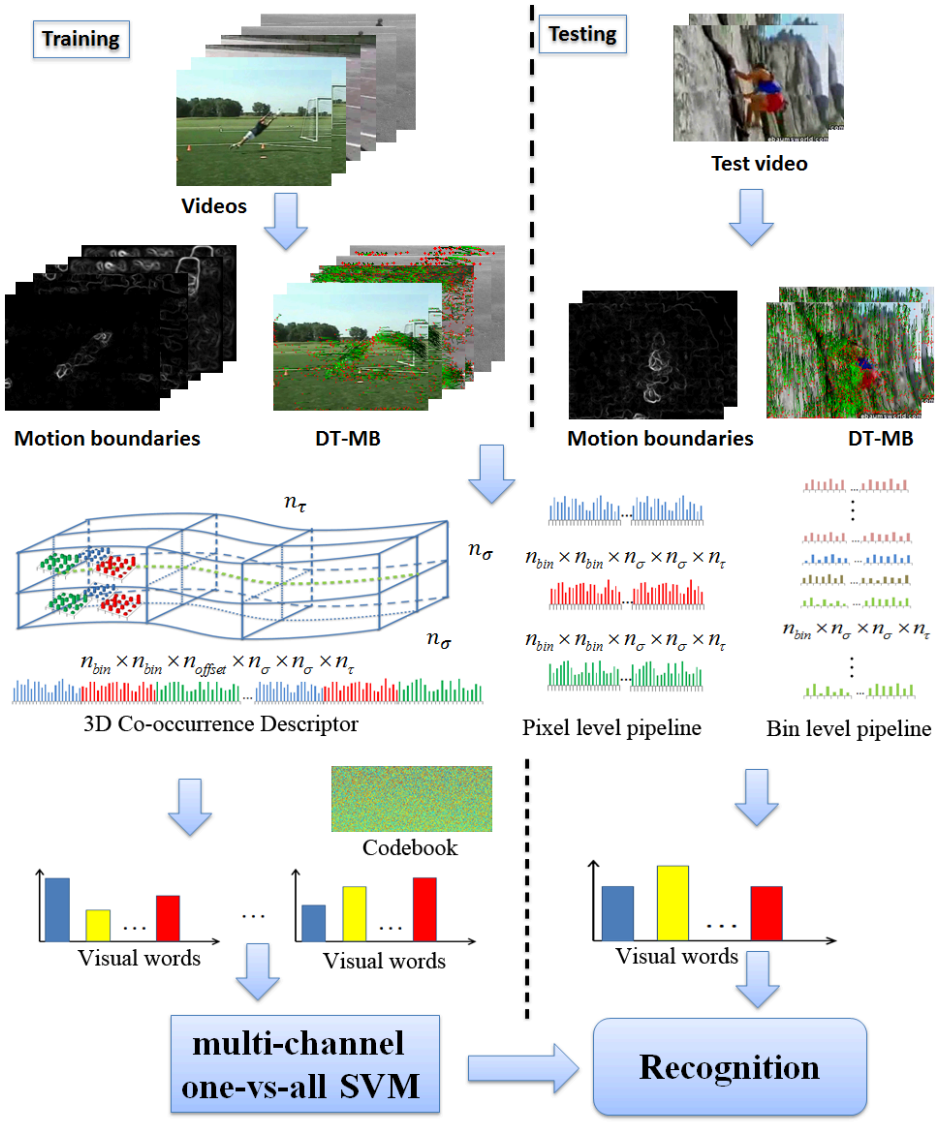


Figure 1: The framework of our approach. To represent action videos, we first dense sampling trajectories based on motion boundary in videos and describe them by a set of 3D co-occurrence descriptors. Then they are clustered into set of visual words. Each video is represented as a histogram of words. Finally, we train a multi-channel SVM classifier for testing videos.

occurrence matrices into several channels by their co-occurrence bins for each offset. The idea of using multi-channels for 3D co-occurrence is partly inspired by the fact that MBHx and MBHy perform differently [1] and the complementarities can be better investigated in multi-channel way as shown in [32].

To evaluate our sampling strategy and the proposed descriptors, we perform action classification with a standard BoF framework and a kernel SVM classifier [11] on three widely-used datasets, namely KTH [33], YouTube [34] and HMDB51 [35]. Our framework is illustrated in Figure 1. We investigate our DT-MB sampling strategy over original dense trajectory [1] in the view of computation and memory cost. We evaluate the improvement of our new descriptors over original HOG, HOF and MBH [1]. Furthermore, we provide a theoretical analysis on the advantages of using co-occurrence feature.

The main contributions of this paper are summarized as follows:

1) we develop a motion boundary based sampling strategy to reduce the number of dense trajectories which can save memory and computation without degrading performance;

2) we propose a set of 3D co-occurrence descriptors, namely 3D-CoHOG, 3D-CoHOF and 3D-CoMBH, which can depict the spatial-temporal contextual information within local cuboids;

3) we present two decomposition strategies for 3D co-occurrence descriptors (pixel level and bin level) and integrate the decomposed components with a multi-channel framework, which can further improve the performance;

4) we achieve state-of-the-art results on several widely-used human action datasets.

It's worth noting that our new descriptors are independent of the spatial-temporal cuboid detectors (e.g., DT [1], STIP [25], dense cuboids [14]). Though we mainly discuss our novel descriptors with dense trajectory, one can easily extend them with other detectors as well. The analysis and results presented here extend our preliminary work in BMVC 2013 [36]. Here, we develop more general spatial-temporal co-occurrence descriptors and further improve the performance by exploiting their multi-channel versions. We also provide an information theory analysis to validate the advantages of using co-occurrence descriptors.

The rest of this paper is organized as follows. In Section 2, we give a brief review of dense trajectory based method and present our DT-MB method in detail. In Section 3, we present our 3D co-occurrence descriptors. The two levels of decomposition for 3D co-occurrence descriptors are presented in Section 4. Section 5 shows the experimental results and gives a comprehensive comparison for each individual descriptor. We conclude our work in Section 6.

2. Dense Trajectories on Motion Boundary

In this section, we first give a brief review of dense trajectory method [1] and explain the advantage of DT from a view of human visual fixation system. Then, we present our new sampling strategy based on motion boundary in details.

2.1. Preliminaries on Dense Trajectory

Dense sampling strategy has been widely used in extracting local image features, and achieved great success in image classification. This fact inspires researchers to develop dense sampling approaches for video based action recognition, which can yield richer description of action than spare interesting points. Two successful examples are dense cuboid [14] and dense trajectory [1]. Dense trajectory based method mainly consists of the follow steps.

Dense sampling and filtering. Feature points are sampled in the current frame on a grid by a step size w at S spatial scales. To track successfully, points in homogeneous image areas are filtered out by examining the eigenvalues of their auto-correlation matrix.

Trajectories. Dense points are tracked by median-filtered optical flow on each spatial scale separately. Tracked points in successive frames at scale s are concatenated to form trajectories: $(P_t^s, P_{t+1}^s, ...)$, where $P_t^s = (x_t^s, y_t^s)$ represents the spatial position. To prevent the trajectories from drifting, the length is limited to L frames. Once a trajectory's length reaches L, its mean position drift and variation will be checked. Trajectories with tiny or large mean drift and variation will be pruned since they usually correspond to static or erroneous trajectories.

Descriptors for DT. There are four types of descriptors for each cuboid aligned by trajectories [1]. The trajectory shape is described by a sequence

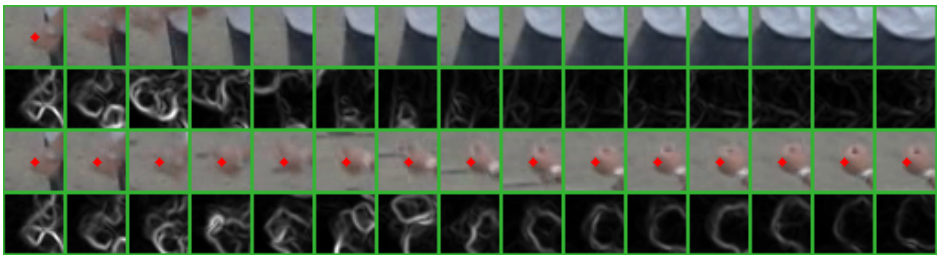
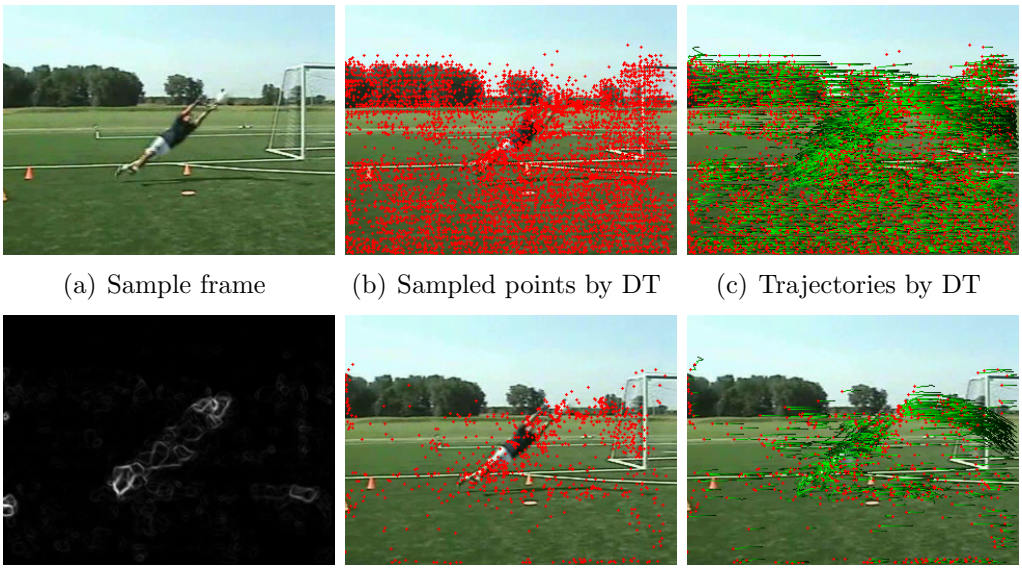


Figure 2: The 3rd and 1st rows: raw patches in cuboids with and without pursuit. The 2nd and 4th rows: motion boundaries of patches. Dense trajectory based cuboid (the 3rd row) fixs the visual field on the moving fist, which is consistent with *visual fixation*.

$(\Delta P_t^s, ..., \Delta P_{t+L}^s)$ of displacement vectors $\Delta P_t^s = (x_{t+1}^s - x_t^s, y_{t+1}^s - y_t)$. Usually, this vector is normalized by the ℓ_1-norm. Therefore, we obtain a $2L$ dimensional shape descriptor. To catch the motion and structure information, HOG, HOF and MBH are extracted within a space-time cuboid whose size is $N \times N \times L$ aligned with the trajectory. HOG and HOF are among popular descriptors which yield excellent results on many datasets [11]. The MBH is derived from the gradients of optical flow which is originally introduced for human detection [28]. To embed more structure information, we usually subdivide the cuboid into a spatial-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$. Assume n_{bin} is the number of quantized bins for HOG and MBH, and $n_{bin}+1$ (1 for static) for HOF. Then we can obtain a $n_\sigma \times n_\sigma \times n_\tau \times n_{bin}$ feature vector for HOG, $n_\sigma \times n_\sigma \times n_\tau \times (n_{bin}+1)$ for HOF and $n_\sigma \times n_\sigma \times n_\tau \times n_{bin}$ for MBHx and MBHy, respectively.

Dense trajectory based approaches whose features are extracted along with trajectories are consistent with human visual fixation system as illustrated in the 3rd row of Figure 2. *Visual fixation* refers to the maintaining of the visual gaze on a single location, also known as *smooth pursuit* or temporal slowness [37]. A number of species, including humans, other primates, cats and rabbits can perform this mechanism by three categories of eye movements: micro-saccade, ocular drift, and ocular micro-tremor. There are two basic properties for smooth pursuit from a view of video representation. On the one hand, it makes the feature *robust to velocity change*. Obviously, the appearance features in a cuboid with smooth pursuit can be very similar despite the difference of motion velocity. Motion features also remain similar after normalization. On the other hand, as shown in the 4th row of Figure 2, *more meaningful appearance and motion information* are captured with tem-



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

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

Figure 3: Comparison of original DT and our Dense Trajectories on Motion Boundary.

poral slowness. For these reasons, DT can always outperform dense cuboids with identical parameters [1] in theory.

2.2. Dense Trajectories on Motion Boundary

A limitation of DT is that too many points need to be tracked in the original dense sampling criterion [1]. However, only a few of them may lead to valid trajectories. We notice that those points on the motion boundary are the most discriminative ones. This is indeed partly implied by MBH descriptor [1] and motion boundary contour system (BCS) in neural dynamics of motion perception [38]. In this paper, we introduce motion boundary based sampling strategy which constrains the sampled points to the sharp regions of motion boundary.

The implementation of DT-MB is straightforward. Different from original DT, it needs two successive frames to sample points. A comparative example is illustrated in Figure 3. Figure 3(d) shows an example of motion boundary image. We calculate the gradient magnitudes for both the horizontal and the vertical components of optical flow, and set the maximum of them as motion boundary image. After a thresholding operation on the motion boundary

image, a mask is generalized to refine the original DT sampled points. Particularly, we estimate the mask by Otsu's algorithm [39] empirically. Those regions outside the foreground of mask will be removed. Central points of the remaining patches will be refined by the average location of foregrounds. It's worth noting that the motion boundary image is a middle result of DT, so we donot need to add complexity. As shown in the 2nd column of Figure 3, our approach removes a large number of points which are not on the motion foreground. The 3rd column of Figure 3 exhibits the trajectories from historical points by DT and DT-MB. The red marks are the end points of trajectories. Note that we do not force all the points of trajectories on the motion boundaries in case of inaccurate tracking. Our DT-MB can be viewed as *an effective strategy to reduce the influence of camera motion in the early stage*. The detailed analysis of complexity and performance are given in Section 5.

3. 3D Co-occurrence Descriptors

Generally, there always exist strong correlations among spatial-temporal neighborhoods of pixels. Traditional HOG, HOF and MBH descriptors are statistical histograms counted pixel-wise which ignore the correlation of pairwise pixels. To jointly encode the spatial-temporal correlations of pixels, we present 3D co-occurrence descriptors which consist of 3D-CoHOG, 3D-CoHOF and 3D-CoMBH.

3D-CoHOG. The spatial CoHOG (2D-CoHOG) is initially introduced in the context of pedestrian detection [30]. Specially, it uses pairs of gradient orientations as units, and employs the co-occurrence matrix for image representation. As for 3D-CoHOG in video domain, offsets in time domain are taken into account, which is an extension from 2D. The co-occurrence matrix expresses the joint distribution of gradient orientations between anchor points and offset points over a cuboid as illustrated in Figure 4, and it can jointly depict more complex structure of spatial patch and the appearance changes along with time. A straightforward co-occurrence strategy is to obtain the statistics of all the possibilities of joint occurrences. This results in a $n_{bin}^{n_{offset}+1}$ co-occurrence matrix where n_{offset} is the number of offset points and n_{bin} is the quantized bins for HOG. This strategy leads an excessively redundant matrix whose high dimensionality not only increases the computation and storage cost but also makes the classification expensive. To overcome this problem, we use pairwise co-occurrence representation. Con-

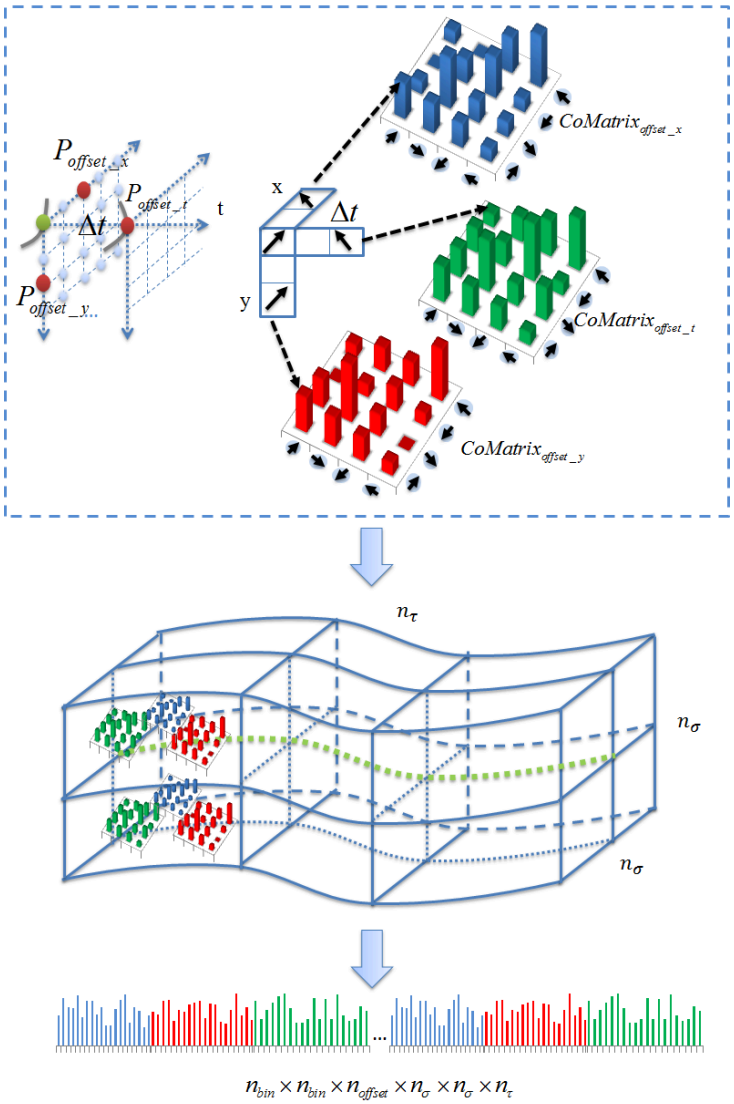


Figure 4: An example of 3D co-occurrence descriptor with grid of size $n_\sigma \times n_\sigma \times n_\tau$.

sidering three offsets shown as the red points in Figure 4, each pair of which will vote for one co-occurrence matrix. After voting, we vectorize all co-occurrence matrices and concatenate them into a vector for each cell of grid, and then concatenate these vectors cell-wise to yield the final descriptor for a cuboid. Given a cuboid with grid size $n_\sigma \times n_\sigma \times n_\tau$, the final 3D-CoHOG descriptor is a $n_{bin} \times n_{bin} \times n_{offset} \times n_\sigma \times n_\sigma \times n_\tau$ dimensional vector.

3D-CoHOF and 3D-CoMBH. We can also apply the above 3D co-occurrence strategy to HOF and MBH descriptors. The implementations of these are very similar with 3D-CoHOG except for the inputs. 3D-CoHOF applies spatial and temporal pairs of optical flow orientations as units, and 3D-CoMBH utilizes spatial and temporal pairs of the gradient orientations in the horizontal and vertical flow components, separately. So there will be two 3D-CoMBH components, namely 3D-CoMBHx and 3D-CoMBHy.

In our case, considering the computation and discriminative ability, we use three offsets (i.e., (2,0) and (0,2) for spatial offsets, and $\Delta t = 2$ for temporal offset) and process the trajectory-aligned cuboids pixel-wise with a grid of size $n_\sigma \times n_\sigma \times n_\tau$. Specially, a co-occurrence matrix $\mathcal{C}$ over a $M \times N \times T$ cell I, parameterized by an offset (x, y), is defined as:

$$\mathcal{C}_{x,y}(p,q) = \sum_{t=1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{N} \begin{cases} \frac{G_t(i,j)+G_t(i+x,j+y)}{2}, & \text{if } O_t(i,j)=p, O_t(i+x,j+y)=q; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where p and q are the quantization bins, $G_t(i,j)$ is the gradient magnitude and $O_t(i,j)$ is the assigned bin (e.g., gradient orientation, flow orientation) at position (i,j) of the t-th frame. As shown in Equation (1), the average gradient magnitude is used to weight co-occurrence matrices. To reduce the boundary effects, we also apply linear interpolation for voting the co-occurrence matrix. For example, given a pre-quantized gradient bin 1.4, we would vote this for both bin 1 and bin 2 with weights 0.6 and 0.4, respectively.

One can imagine that it needs $\Delta t + 1$ frames at least to calculate the co-occurrence matrices for 3D-CoHOG and $\Delta t + 2$ for 3D-HOF and 3D-MBH. The offsets in time domain we used are aligned by trajectories. It is worth noting that tracking is a necessary step in dense trajectory based approach, so our descriptors can benefit from the computational process of DT.

3.1. Information Theory Analysis for Co-occurrence Descriptors

Here, we give an information theory analysis for co-occurrence descriptors, and explain why co-occurrence can yield extra information for classification and how to select the offsets.

Suppose we have K categories denoted by set $C = \{c_1, c_2, ..., c_K\}$. The prior probabilities of C can be denoted as $p(C) = \{p(c_1), p(c_2), ..., p(c_K)\}$. We utilize the mutual information to analyze the different discrimination

between the distributions yielded by individual and co-occurrence (pairwise) units or pixels.

Suppose we have N bins for individual unit distribution $H = \{h_1, ..., h_N\}$ (e.g., HOG) and the probability of the ith element h_i is denoted by $p(h_i)$. For histogram features, each bin is assumed to be independent. Thus, for an individual unit H, its contribution to the classification can be defined as the mutual information:

$$I(C; H) = I(H; C) = \sum_{n=1}^{N} \sum_{k=1}^{K} p(h_n) p(c_k|h_n) \log \frac{p(c_k|h_n)}{p(c_k)}. \tag{2}$$

The joint distribution between H and another individual unit distribution $H' = \{h'_1, h'_2, ..., h'_M\}$ can be denoted as $p(\mathcal{F}) = \{p(f_{1,1}), p(f_{1,2}), ..., p(f_{N,M})\}$ where $p(f_{n,m}) = p(h_n, h'_m)$. Then, the information gain of co-occurrence feature $\mathcal{F}$ with respect to H is given by,

$$\begin{aligned} I(\mathcal{F}; C) - I(H; C) = & \sum_{n,m} \sum_{k} p(f_{n,m}) p(c_k|f_{n,m}) \log \frac{p(c_k|f_{n,m})}{p(c_k)} \\ & - \sum_{n} \sum_{k} p(h_n) p(c_k|h_n) \log \frac{p(c_k|h_n)}{p(c_k)}. \end{aligned} \tag{3}$$

Recall,

$$p(h_n) = \sum_{m} p(f_{n,m}) = \sum_{m} p(h_n, h'_m), \tag{4}$$

$$p(c_k|h_n) = \frac{1}{p(h_n)} \sum_{m} p(f_{n,m}) p(c_k|f_{n,m}). \tag{5}$$

Then, we can rewrite Equation (3) to,

$$\begin{aligned} I(\mathcal{F}; C) - I(H; C) &= \sum_{k} \sum_{n,m} p(f_{n,m}) p(c_k|f_{n,m}) (\log \frac{p(c_k|f_{n,m})}{p(c_k|h_n)}) \\ &= \sum_{n,m} p(f_{n,m}) KL(p(c_k|f_{n,m}), p(c_k|h_n)). \end{aligned} \tag{6}$$

where KL represents the Kullback-Leibler divergence [40] between two distributions. We conclude two important properties from Equation (6):

1) Since $KL(;) \geq 0$, we have $I(\mathcal{F}; C) - I(H; C) \geq 0$. This suggests the co-occurrence feature can provide extra information for classification. In

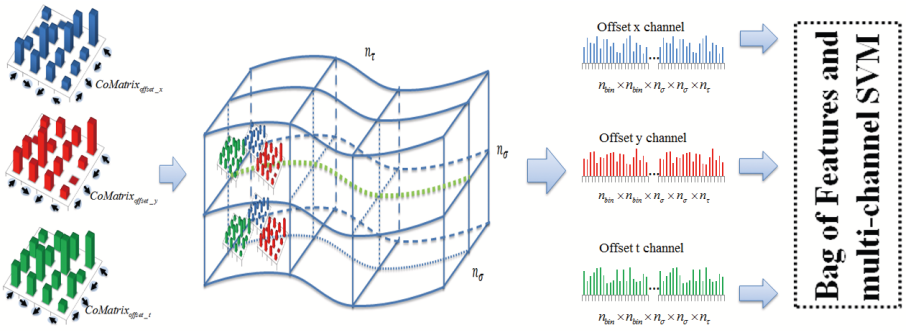


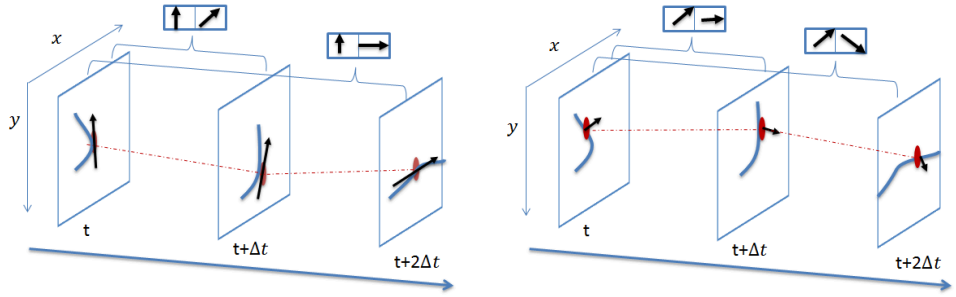
Figure 5: Pixel level of multi-channel 3D Co-occurrence descriptors with grid of size $n_\sigma \times n_\sigma \times n_\tau$.

fact, this property can be interpreted by the so-called Data Processing Inequality (DPI) [41] as well. DPI states that post-processing cannot increase information [41]. In particular, if $Z = f(Y)$, then $I(X;Y) \geq I(X;f(Y))$, where X, Y, Z are random variables and f is a (probabilistic) function. Equation (4) shows that H can be seen as a function of $\mathcal{F}$, then we have $I(\mathcal{F};C) \geq I(H;C)$ from the view of DPI.

2) One key parameter to design our co-occurrence descriptor is spatial and temporal offsets. Considering Equation (6), $p(c_k|h_n)$ denotes the probability distribution by given a special bin. It is an approximate uniform distribution in our investigation, which means an individual bin has very limited discriminative power. Thus, the information gain in Equation (6) is mainly dependent on the distribution $p(c_k|f_{n,m})$. The gain is zero when $p(c_k|f_{n,m})$ tends to uniform. Intuitively, $p(c_k|f_{n,m})$ will tend to be uniform in two cases: very small offset and large offset. The extreme case for small offset is zero offset, which yields uniform distribution obviously. Large offset tends to be independent with the anchor point, then the joint distribution can be rewritten as $p(c_k|h_n)p(c_k|h'_m)$, which is near to an uniform distribution as well. We evaluate the offset experimentally in Subsection 5.3.3.

4. Multi-channels of 3D Co-occurrence Descriptors

In this section, we first present the multi-channel scheme at pixel level for 3D co-occurrence descriptors. Then, we revisit our previous spatial-temporal context descriptors. Finally, we give the details of bin level multi-channel scheme.



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

Figure 6: Temporal co-occurrence descriptors. (a): temporal pairs of gradient orientations in T-CoHOG or T-CoMBH. (b): temporal pairs of optical flow orientations in T-CoHOF.

4.1. Pixel Level of Multi-Channel Pipeline

Inspired by the fact that MBHx and MBHy perform differently and their combination leads to better results [1], we also split all the 3D co-occurrence descriptors according to the offsets and integrate them in a multi-channel scheme with BoF model. In our case, we get four types of descriptors to split, namely 3D-CoHOG, 3D-CoHOF, 3D-CoMBHx and 3D-CoMBHy. As shown in Figure 5, we split each 3D co-occurrence descriptor along with offset_x, offset_y and offset_t. After voting in the spatial-temporal grids, we apply BoF model for each vectoried co-occurrence matrix, and leverage multi-channel kernel SVM to combine channels for each 3D co-occurrence descriptor. The complementarity among each co-occurrence pair can be effectively exploited by multi-channel SVM as shown in [32].

4.2. Spatial and Temporal Context Descriptors

In a previous work [36], we have developed another type of spatial and temporal context descriptors which employ spatial co-occurrence and temporal co-occurrence, separately. They are composed of spatial co-occurrence HOG (S-CoHOG), S-CoHOF, S-CoMBH, temporal co-occurrence HOG (T-CoHOG), T-CoHOF and T-CoMBH. Each S-Co descriptor is yielded by concatenating the co-occurrence matrices from offset_x and offset_y. As the previous case in Figure 4, only the blue and red co-occurrence matrices will be vectorized and concatenated for S-Co descriptors. T-Co descriptors are designed to depict the appearance and motion changes from successive patches as illustrated in Figure 6. These spatial and temporal context descriptors

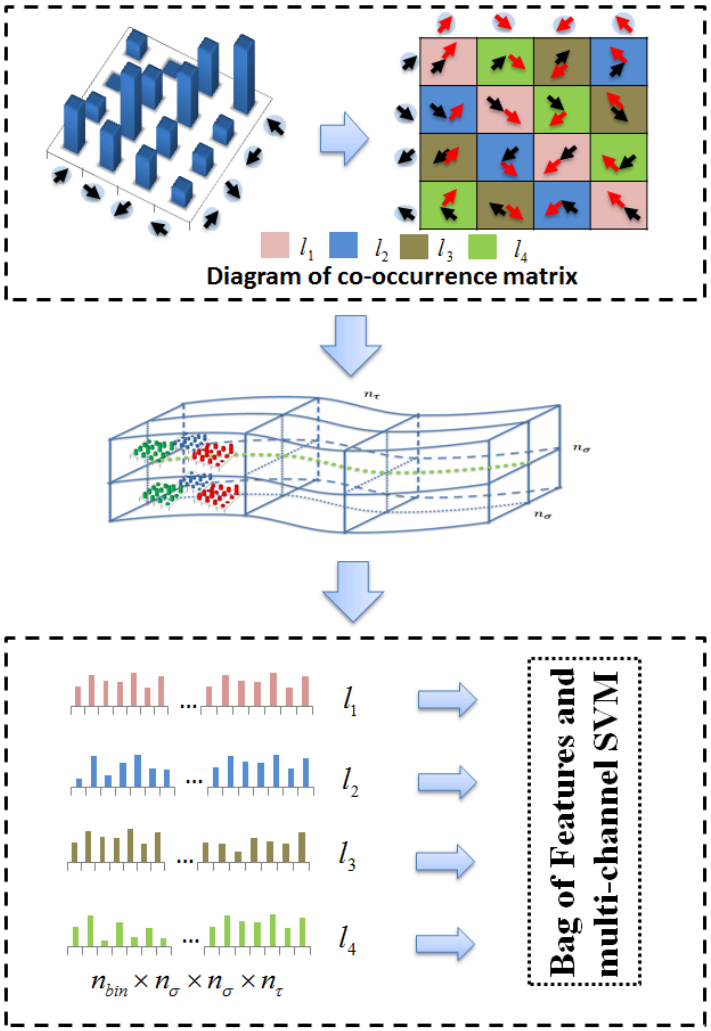


Figure 7: Bin level of multi-channel 3D Co-occurrence descriptors with grid $n_\sigma \times n_\sigma \times n_\tau$.

can be seen as spatial channel and temporal channel of the 3D co-occurrence descriptors.

4.3. Decomposing Co-occurrence Matrix: Bin Level Pipeline

The diagram of bin level decomposition is illustrated in Figure 7. The co-occurrence matrices can be seen as the interchanges of gradient orientations or flow orientations between anchor points and their offsets. Inspired by the multi-channels of Motion Interchange Pattern [42], we also decompose each

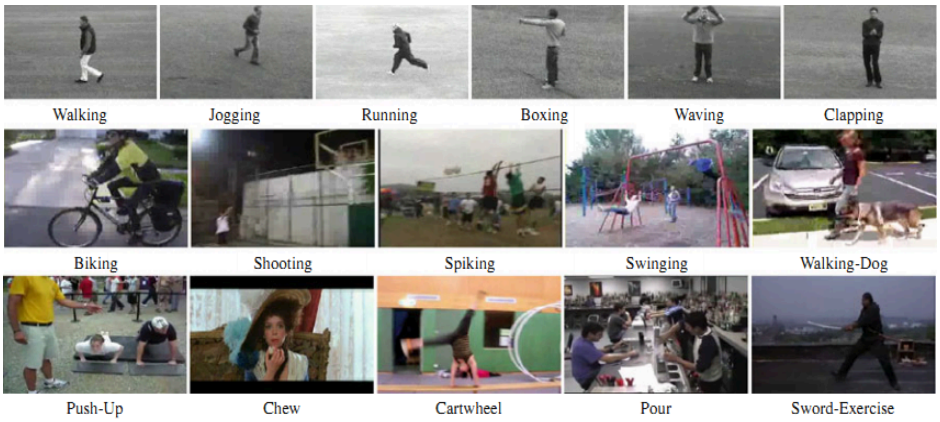


Figure 8: Sample frames from KTH, YouTube and HMDB51.

co-occurrence matrix by the relative angle of pairwise bins. As shown in Figure 7, the relative angles are 0°, 90°, 180° and 270° from channel ℓ_1 to channel ℓ_4, respectively. Although the use of angle channels in [42] is aimed at computing conveniently, we find this strategy significantly improve the performance in our case, especially by the channels from offset$_t$. We explain that different angle channels can reflect the difference of action categories. Take the offset t of 3D-CoHOF for example, high value can occur in the ℓ_3 channel for action *boxing*, while strong values happen in the ℓ_2 and ℓ_4 channels for action *waving*.

5. Experiments

We evaluate the performance of the proposed methods on three popular human action datasets, namely KTH [33], YouTube [34] and HMDB51 [35]. In this section, we first give a brief introduction for these datasets, and then compare the performance and complexity between DT and DT-MB. Finally, we give a comprehensive comparison between our descriptors and other descriptors.

5.1. Datasets and Setup

These datasets we used are collected from controlled experimental setting or web videos. Some sample frames are illustrated in Figure 8. We totally evaluate more than 10,000 video clips for our experiments.

The **KTH** dataset [33] is one of the most popular datasets in action recognition, which consists of 2,391 video clips acted by 25 subjects. It contains 6 action classes: *walking*, *jogging*, *running*, *boxing*, *hand-waving*, and *hand-clapping*. Actions are recorded at 4 environment settings: outdoors, outdoors with camera motion, outdoors with clothing change, and indoors. We follow the experimental settings in [33] where clips are divided into the training set (16 subjects) and the testing set (9 subjects).

The **YouTube** dataset [34] is collected from YouTube videos. It contains 11 action categories: basketball *shooting*, volleyball *spiking*, trampoline *jumping*, soccer *juggling*, horse back *riding*, *cycling*, *diving*, *swinging*, *golf-swinging*, *tennis-swinging*, and *walking* (with a dog). A total of 1,168 video clips are available. Following [34], we use Leave-One-Group-Out cross-validation and report the average accuracy over all classes.

The **HMDB51** dataset [35] is a large action video database with 51 action categories. Totally, there are 6,766 manually annotated clips which are extracted from a variety of sources ranging from digitized movies to YouTube. It contains facial actions, general body movements and human interactions. It is a very challenging benchmark due to its high intra-class variation and low video quality. We follow the experimental settings in [35] and report the mean average accuracy over all classes.

We employ the standard BoF framework [24] to represent videos. In particular, we set the parameters $(w, S, N, L, n_\sigma, n_\tau)$ mentioned in the previous sections to be $(5, 8, 32, 15, 2, 3)$ following [1]. After calculating the descriptors of videos in training set, we construct codebooks with size 4k for each channel separately using k-means on subsets of 100k randomly selected features. Then we quantize the local descriptors with the codebook and employ the statistical histograms of code words as video representations.

5.2. Setting of Classification

For classification we use the RBF-SVM with a multi-channel χ^2 kernel which is slightly different from [11]. The multi-channel Gaussian kernel is defined by:

$$K(H_i, H_j) = \exp(-\alpha \sum_{c \in \mathcal{C}} \frac{1}{A_c} D_c(H_i^c, H_j^c)) \tag{7}$$

where $D_c(H_i^c, H_j^c)$ is the χ^2 distance between two video representations, H_i and H_j in the cth feature channel. A_c is the average value of all the distances in training set for the cth channel. α is a scaling factor ranging from 0 to

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

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

1 which is obtained by cross-validation, and we set it to be 1 when dealing with single channel feature. For multi-class classification, we use the *one-against-rest* approach and select the class with the highest score.

5.3. Experimental Results and Discussion

5.3.1. DT versus DT-MB

Our first purpose is to investigate the effect caused by constraining dense trajectories on motion boundary. We compare the recognition accuracy, frame rate in the whole process of feature extraction including the time of I/O (fps), the average tracking time of dense points per frame (T_{track}) and the average number of trajectories per video clip between DT and DT-MB. Specially, we quantize orientations into eight bins (an additional zero bin is added for HOF) with full orientation and weight the bins with magnitudes. For the accuracy comparison, we only report the performance of all the raw DT descriptors combination by using (7). We evaluate the fps and T_{track} within 10 videos randomly selected from each dataset and the run-time is obtained on an Acer laptop with a 2.5 GHz Intel Core i5 CPU and 4 GB RAM.

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

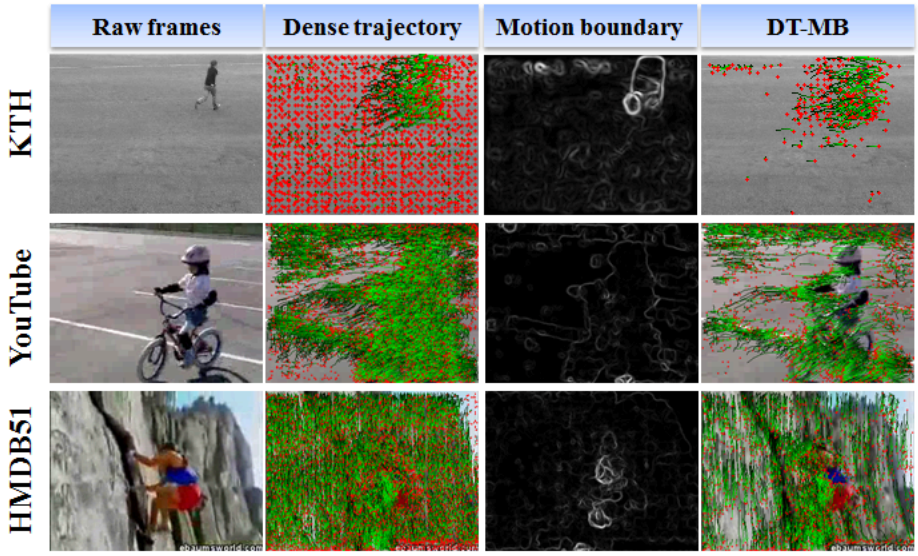


Figure 9: Samples of DT vs DT-MB from KTH, YouTube and HMDB51 datasets.

More examples of DT-MB are depicted in Figure 9. As it is noted that the motion boundaries in Figure 9 are less pure than the one shown in Figure 3 and salient regions can exist in background motion boundaries due to large irregular camera motion. Nevertheless, as illustrated in Figure 9 and Figure 3, our DT-MB is able to reduce most of the irrelevant trajectories, which can save memory and computation cost while preserve the accuracy.

5.3.2. Evaluation of Proposed Descriptors

We evaluate all the proposed descriptors using our DT-MB sampling scheme since it is a good alternative to original DT sampling strategy. We quantize orientations into four bins except an additional bin for HOF in this evaluation. One can certainly use eight bins which is not the key point in our evaluation. We fix the offsets for direction x, y, t to 2. The comparisons with original DT descriptors are shown in Table 2 and Figure 10. The results for MBH are obtained by combining x and y channels. The results for ST-Co are achieved by combining S-Co and T-Co channels.

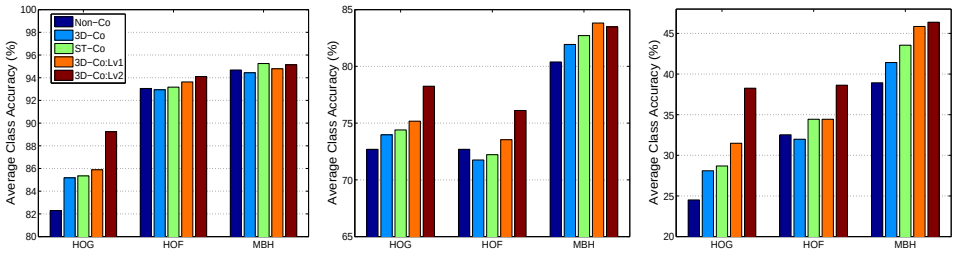
As for the HOG type of feature, all of the co-occurrence descriptors extended from original HOG achieve better results than original HOG on the three datasets, the improvements for "3D-CoHOG:Lv2" (bin level multi-channel pipeline of 3D-CoHOG) are 6.14%, 5.56% and 13.75%, respectively.

Table 2: Accuracies of all the types of individual descriptors on three datasets using standard BOF. The "Lv1" denotes the pixel level multi-channel pipeline and the "Lv2" represents the bin level pipeline.

	KTH (%)	YouTube (%)	HMDB51(%)
Trajectory	87.96	68.32	28.50
HOG	83.10	72.69	24.51
3D-CoHOG	85.18	73.97	28.10
ST-CoHOG	85.35	74.40	28.69
3D-CoHOG:Lv1	85.88	75.17	31.48
3D-CoHOG:Lv2	**89.24**	**78.25**	**38.26**
HOF	93.05	72.09	32.53
3D-CoHOF	92.94	71.75	31.98
ST-CoHOF	93.17	72.22	34.44
3D-CoHOF:Lv1	93.63	73.54	35.16
3D-CoHOF:Lv2	**94.10**	**76.11**	**38.63**
MBH	94.68	80.39	38.93
3D-CoMBH	94.44	81.93	41.42
ST-CoMBH	**95.25**	82.71	43.55
3D-CoMBH:Lv1	94.79	**83.82**	45.86
3D-CoMBH:Lv2	95.14	83.48	**46.38**

As illustrated in Figure 10, the improvements of 3D-CoHOG, ST-CoHOG and "3D-CoHOG:Lv1"(pixel level multi-channel pipeline) are very similar and present a rising trend. The significant difference between "3D-CoHOG:Lv2" and other pipelines indicates the complementarity of different co-occurrence bins, which can be better explored by multiple channel pipelines. One impressive result of "3D-CoHOG:Lv2" is achieved on HMDB51 dataset where the performance is similar with original MBH as shown in Table 2. This indicates that the 3D co-occurrence for HOG is very effective. This can be ascribed to the fact that 3D-CoHOG can depict the spatial and temporal structure changes which suggest motion information implicitly.

The multi-channel versions (i.e., level 1 and level 2) of 3D CoHOF perform consistently better than the original HOF on all the datasets. The direct 3D-CoHOF seems to slightly degrade the performance. We argue that the high dimension of 3D-CoHOF makes Euclidean distance measure unstable which counteracts the advantage of co-occurrence. That's why our reproduced result (i.e., 32.53% by 5 bins) of HOF is superior than that (i.e., 31.5% by 9



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

Figure 10: The corresponding graph of Table 2. "Non-Co" corresponds to the original descriptors in [15].

bins) in [1].

It is beneficial to incorporate 3D co-occurrence information to MBH descriptor, especially on YouTube dataset and HMDB51 dataset where the maximum improvements are 3.43% and 6.45%, respectively. As for the KTH dataset, all the versions of MBH perform similarly. Observing recent related publications [43, 44, 1], we find the approximate accuracy of 95% on the KTH dataset might be the upper bound by using low-level features since there are some confused videos even difficult for human to classify.

Overall, our co-occurrence schemes can boost performance on all the datasets, and the multi-channel versions of "3D-Co" always perform better than original "3D-Co". Compared with "3D-Co", the multi-channel versions of "3D-Co" own several advantages: first, splitting the long vectors into short ones and performing BoF separately have the similar effect as increasing the codebook size significantly for long vectors, which is beneficial to performance. In fact, when using the multi-channel version, the codebook of original "3D-Co" is defined as Cartesian product, which increases the codebook size exponentially [45]; second, the mutual influence among different channels can be reduced by splitting the "3D-Co" into multi-channels. Though multi-channel versions could lead to perform BoF more times, all the implementations can be easily conducted in parallel and the processing speed is faster since the dimension of feature is lower than the original one.

5.3.3. Evaluation of Parameter

To evaluate the key parameter (i.e., offset) for co-occurrence features, we conduct a verification experiment using the channel x and channel y of 3D-CoHOG on the KTH dataset. We show the information gains (IGs) and

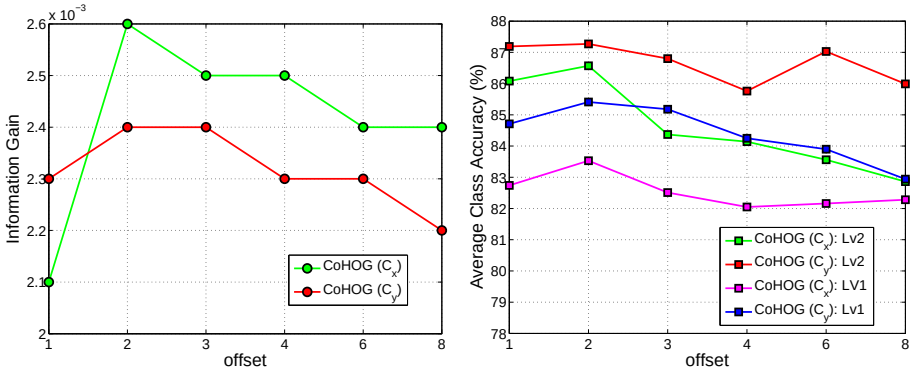


Figure 11: The information gain and performance with varying offsets on the KTH dataset.

performance with different offsets in Figure 11. For the the statistics of IGs, we randomly select 10 videos[1] for each action class and count the distributions mentioned in Subsection 3.1 pixel-wise within trajectories, and then compute IGs for different offsets using Equation 6. Specially, IGs are computed for each channel at pixel level separately because we use pairwise co-occurrence scheme. For the performance, we evaluate it on the whole dataset with the same parameters as Subsection 5.3.2 except for the offsets. As shown in Figure 11, co-occurrence descriptors can bring information gain on all the offsets we used. Increasing the offset decreases the information gains of both channels when offsets are above 2, which validates our qualitative analysis in Subsection 3.1. We achieve the highest IGs and performance at $offset = 2$. The trend of performance (right of Figure 11) is not as clear as that of IGs because there are some uncertain factors for the process of BoF like the sampling of descriptors and the generation of codebooks by K-means. In the rest of this paper, the offsets for x, y, t channels are fixed to 2 unless otherwise stated.

5.3.4. Further Analysis on Multi-channel 3D Co-occurrence Descriptors

We show the results of channel x, y and t (i.e., the spatial and temporal offsets) for all the 3D co-occurrence descriptors separately in Table 3. The channel x and y show similar performance for all kinds of descriptor in both levels of multi-channel pipeline. The bin level outperforms the pix-

[1]The filenames of selected videos is available at http://mmlab.siat.ac.cn/personal/pxj/. All the source code will be released as well.

Table 3: Results of offset x, y, t for different level multi-channel pipelines on YouTube dataset and HMDB51 dataset.

Dataset	YouTube(%)				HMDB51(%)			
Channel	C_x	C_y	C_t	C_{xyt}	C_x	C_y	C_t	C_{xyt}
3D-CoHOG:Lv1	72.3	72.4	72.8	75.2	25.2	26.5	28.3	31.5
3D-CoHOF:Lv1	72.1	71.9	70.8	73.5	31.9	32.2	30.4	34.5
3D-CoMBHx:Lv1	76.2	75.3	74.9	78.3	30.6	29.9	30.6	35.0
3D-CoMBHy:Lv1	76.1	75.0	76.4	79.5	36.5	36.0	34.5	40.5
3D-CoHOG:Lv2	75.7	74.9	77.4	78.3	31.0	31.8	35.8	38.3
3D-CoHOF:Lv2	75.7	75.3	73.8	76.1	36.6	36.5	35.2	38.6
3D-CoMBHx:Lv2	77.3	77.7	79.7	79.6	32.9	33.8	32.3	36.5
3D-CoMBHy:Lv2	77.8	77.1	78.5	79.1	39.6	39.0	38.7	42.8

el level pipeline significantly. Interestingly, the channel t for 3D-CoHOG is superior to the channel x and y in both pipeline but the results are inverse for 3D-CoHOF. A possible explanation is that temporal co-occurrence for spatial-aware descriptors contains motion information implicitly, and spatial co-occurrence for temporal-aware ones can capture more complementary information than their temporal co-occurrence.

To further investigate the effects of bin level pipelines, we show the results of different bin levels of 3D-CoHOG in Table 4. The four channels $\{\ell_1, ..., \ell_4\}$ denote the gradient orientation changes from 0° to 270° between the co-occurrence pairs as shown in Figure 7. There is no evident trend except that the ℓ_3 channel in temporal co-occurrence outperforms other channels. A possible explanation is that this channel reflects the tiny changes occurred on edges of patches aligned by trajectory, since inverse orientations usually exist beside edges and the changes on edges or boundaries are discriminative.

Table 4: Results of bin level channels for 3D-CoHOG in all offset points on three datasets.

	C_x(%)				C_y(%)				C_t(%)			
	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4
KTH	82.4	83.7	86.3	79.7	85.2	84.0	76.6	85.0	82.6	84.3	90.2	83.0
YouTube	71.2	67.5	69.8	68.9	71.8	69.3	63.9	70.0	71.3	68.5	71.7	68.4
HMDB51	25.5	22.3	22.9	22.7	26.9	24.8	23.8	23.6	26.1	23.2	29.8	22.6

Table 5: Results of different combinations on three datasets. The "3D-Co" includes 3D co-occurrence descriptors of HOG, HOF and MBH.

	KTH (%)	YouTube (%)	HMDB51 (%)
Traj.+HOG+HOF+MBH	93.63	84.25	45.90
HOG+HOF+MBH	93.98	83.48	45.88
Traj.+ 3D-Co	93.98	84.93	48.54
3D-Co	94.56	84.93	47.67
Traj.+S-Co + T-Co	94.79	85.70	48.98
S-Co + T-Co	94.21	85.33	48.89
Traj.+ 3D-Co:Lv1	94.68	87.33	50.54
3D-Co:Lv1	94.68	86.90	50.50
Traj.+ 3D-Co:Lv2	95.14	**87.59**	51.51
3D-Co:Lv2	95.14	87.16	**51.76**
Best Combination	**95.60**	87.59	51.76

5.3.5. Descriptor combination

We also conduct experiments to examine the performance of the combination of different descriptors. Table 5 reports the results of several combinations for previous descriptors and our proposed co-occurrence ones using Equation 7. The baseline [1] is the combination of Trajectory, HOG, HOF and MBH, we reimplement it in our evaluation. The combinations of our new descriptors consistently outperform that of original descriptors. Without the trajectory descriptor, the improvements of our bin level pipelines compared to original one are 1.16%, 2.91% and **5.88**% on KTH, YouTube and HMDB51, respectively. The best combination is "S-CoMBH+T-CoMBH" for KTH, the "trajectory+3D-Co:Lv2" for YouTube, and the "3D-Co:Lv2" for HMDB51. An additional finding is that the trajectory descriptor is not important for all the combinations on these three datasets. We explain that the trajectory descriptor is actually the optical flow information of a successive point, which is implicitly included in the HOF (HOF contains neighbor flows as well).

5.4. Comparison with State-of-the-Art Results

Table 6 presents the comparison between our best results and several recent results on all datasets. For fair comparison, we do not show the spatio-temporal pyramids (STP) post-processing results for Wang's approach which is also inferior to ours. Our method outperforms all these previously reported

Table 6: Comparison with the state-of-the-art results.

KTH(%)		**YouTube**(%)		**HMDB51**(%)	
Laptev *et al.* [11]	91.8	Liu *et al.* [34]	71.2	Sadanand *et al.* [9]	26.9
Le *et al.* [22]	93.9	Le *et al.* [22]	75.8	Orit *et al.* [42]	29.2
Ji *et al.* [23]	90.2	B. *et al.* [46]	76.5	Wang *et al.* [1]	46.6
Wang *et al.*[1]	95	Wang *et al.*[1]	84.1	Wang *et al.* [18]	42.1
Our Method	**95.6**	Our Method	**87.6**	Our Method	**51.8**

results. In particular, the improvement over the best reported result to date[2] is 5.2% on the HMDB51 dataset, and it is 3.5% on the YouTube dataset.

6. Conclusion

This paper first introduced a new dense sampling strategy (i.e., DT-MB) for dense trajectories. This scheme constrains sampled points on the motion boundary which can significantly save memory and time cost without degrading performance. Another important contribution is that we propose a set of 3D co-occurrence descriptors, namely 3D-CoHOG, 3D-CoHOF and 3D-CoMBH, which can depict the spatial-temporal contextual information within local cuboids. We also exploit these 3D-Co descriptors by using the decomposition at pixel level and bin level, respectively. The comparisons of the individual descriptors demonstrate that our new features are beneficial. Finally, our method improves the performance of the state-of-the-art action recognition methods on several challenging datasets.

References

[1] H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Dense trajectories and motion boundary descriptors for action recognition, IJCV (2013) 1–20.

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

[3] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden markov model, in: CVPR, 1992, pp. 379–385.

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

[4] T. Starner, A. Pentland, Real-time american sign language recognition from video using hidden markov models, in: Motion-Based Recognition, 1997, pp. 227–243.

[5] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and viterbi path searching, in: CVPR, 2007, pp. 1–8.

[6] G. F. Angela Yao, Juergen Gall, L. V. Gool, Does human action recognition benefit from pose estimation?, in: BMVC, 2011, pp. 67.1–67.11.

[7] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, TPAMI 23 (3) (2001) 257–267.

[8] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: ICCV, Vol. 2, 2005, pp. 1395–1402.

[9] S. Sadanand, J. J. Corso, Action bank: A high-level representation of activity in video, in: CVPR, 2012, pp. 1234–1241.

[10] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: PETS, 2005, pp. 65–72.

[11] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, 2008, pp. 1–8.

[12] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, IJCV 79 (3) (2008) 299–318.

[13] A. Klaser, M. Marszałek, C. Schmid, et al., A spatio-temporal descriptor based on 3d-gradients, in: BMVC, 2008.

[14] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, et al., Evaluation of local spatio-temporal features for action recognition, in: BMVC, 2009.

[15] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: CVPR, 2011, pp. 3169–3176.

[16] X. Wang, L. Wang, Y. Qiao, A comparative study of encoding, pooling and normalization methods for action recognition, in: ACCV, 2012.

[17] S. Feng, P. Emil, L. Robert, Sampling strategies for real-time action recognition, in: CVPR, 2013.

[18] W. LiMin, Q. Yu, T. Xiaoou, Motionlets: Mid-level 3d parts for human motion recognition, in: CVPR, 2013.

[19] J. Arpit, G. Abhinav, R. Mikel, S. D. Larry, Representing videos using mid-level discriminative patches, in: CVPR, 2013.

[20] G. W. Taylor, R. Fergus, Y. LeCun, C. Bregler, Convolutional learning of spatio-temporal features, in: ECCV, 2010, pp. 140–153.

[21] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: Human Behavior Understanding, 2011, pp. 29–39.

[22] Q. V. Le, et al., Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: CVPR, 2011, pp. 3361–3368.

[23] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, TPAMI (2013) 221–231.

[24] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV, 2003, pp. 1470–1477.

[25] I. Laptev, On space-time interest points, IJCV 64 (2) (2005) 107–123.

[26] G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: ECCV, 2008, pp. 650–663.

[27] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM Multimedia, 2007, pp. 357–360.

[28] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, ECCV (2006) 428–441.

[29] R. M. Haralick, K. Shanmugam, I. H. Dinstein, Textural features for image classification, Systems, Man and Cybernetics, IEEE Transactions on (6) (1973) 610–621.

[30] T. Watanabe, S. Ito, K. Yokoi, Co-occurrence histograms of oriented gradients for pedestrian detection, Advances in Image and Video Technology (2009) 37–47.

[31] X. Qi, R. Xiao, J. Guo, L. Zhang, Pairwise rotation invariant co-occurrence local binary pattern, in: ECCV, 2012, pp. 158–171.

[32] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, IJCV 73 (2) (2007) 213–238.

[33] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local svm approach, in: ICPR, Vol. 3, 2004, pp. 32–36.

[34] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos "in the wild", in: CVPR, 2009, pp. 1996–2003.

[35] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: A large video database for human motion recognition, in: ICCV, 2011, pp. 2556–2563.

[36] X. Peng, Y. Qiao, Q. Peng, X. Qi, Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition, in: BMVC, 2013, pp. 1–11.

[37] N. Li, J. J. DiCarlo, Unsupervised natural experience rapidly alters invariant object representation in visual cortex, Science 321 (5895) (2008) 1502–1507.

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

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

[40] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79–86.

[41] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[42] O. Kliper-Gross, Y. Gurovich, et al., Motion interchange patterns for action recognition in unconstrained videos, in: ECCV, Vol. 7577, 2012, pp. 256–269.

[43] H. Liu, R. Feris, M.-T. Sun, Benchmarking datasets for human activity recognition, in: Visual Analysis of Humans, Springer London, 2011, pp. 411–427.

[44] H. Tal, A critical review of action recognition benchmarks, in: CVPR, 2013.

[45] H. Jegou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, PAMI 33 (1) (2011) 117–128.

[46] S. Bhattacharya, R. Sukthankar, et al., A probabilistic representation for efficient large scale visual recognition tasks, in: CVPR, 2011, pp. 2593–2600.