3D $\mathcal{R}$ Transform on Spatio-Temporal Interest Points for Action Recognition

Chunfeng Yuan[1], Xi Li[2], Weiming Hu[1], Haibin Ling[3], and Stephen Maybank[4]

[1]National Laboratory of Pattern Recognition, Institute of Automation, {cfyuan, wmhu}@nlpr.ia.ac.cn
[2]School of Computer Science, University of Adelaide, lixichinanlpr@gmail.com
[3]Department of Computer and Information Sciences, Temple University, hbling@temple.edu
[4]School of Computer Science and Information Systems, Birkbeck College, sjmaybank@dcs.bbk.ac.uk

Abstract

Spatio-temporal interest points serve as an elementary building block in many modern action recognition algorithms, and most of them exploit the local spatio-temporal volume features using a Bag of Visual Words (BOVW) representation. Such representation, however, ignores potentially valuable information about the global spatio-temporal distribution of interest points. In this paper, we propose a new global feature to capture the detailed geometrical distribution of interest points. It is calculated by using the $\mathcal{R}$ transform which is defined as an extended 3D discrete Radon transform, followed by applying a two-directional two-dimensional principal component analysis. Such $\mathcal{R}$ feature captures the geometrical information of the interest points and keeps invariant to geometry transformation and robust to noise. In addition, we propose a new fusion strategy to combine the $\mathcal{R}$ feature with the BOVW representation for further improving recognition accuracy. We utilize a context-aware fusion method to capture both the pairwise similarities and higher-order contextual interactions of the videos. Experimental results on several publicly available datasets demonstrate the effectiveness of the proposed approach for action recognition.

1. Introduction

Many modern action recognition approaches are based on the combination of the spatio-temporal interest point feature and the Bag of Visual Words (BOVW) model [1, 2] due to its simplicity and efficiency. When utilizing the local texture or motion information (e.g., HOG, HOF [8]) encoded in spatio-temporal volumes (cuboid), however, most BOVW based representations ignores the location information of the interest points. In this paper, we propose a novel method to extract a global feature from the location information of interest points extracted from a video. We focus on the geometrical distribution of points in the 3D space and characterize the interest points from the perspective of geometry. We deduce the form and properties of the $\mathcal{R}$ transform, based on the 3D discrete Radon transform, and apply the 3D $\mathcal{R}$ transform to the problem of action recognition for spatio-temporal interest points representation. The 3D $\mathcal{R}$ transform has several unique advantages: (1) It captures the global distribution of spatio-temporal interest points; (2) It is invariant to geometric transformations and robust against noise, both are desirable properties of effective action representations; (3) It is easy to compute since it avoids the nontrivial steps such as foreground object segmentation and the tasks involved in the BOVW method such as selecting the optimal spatio-temporal descriptor, clustering for constructing a codebook, and determining the codebook size. Given an input video, the 3D $\mathcal{R}$ transform produces a result in the form of a 2D feature matrix. We then apply $(2D)^2$PCA [6] to reduce its dimensionality to create the final presentation, named as the $\mathcal{R}$ feature.

The $\mathcal{R}$ feature captures the global geometrical distribution information, while the BOVW representation encodes the discriminability of local features. The two features naturally complement each other. To take this benefit, we propose a new context-aware fusion strategy to combine the two features. Specifically, we first use one feature to decide the context for a video. Then, using the other feature, we design a context-aware kernel to measure the context-aware similarity between two videos. This new context-aware kernel is more robust to noise and outliers than the traditional context-free kernels, which consider only the pairwise relationships between samples. To summarize, the proposed fusion method combines two different features and captures the underlying contextual information from videos.

The remainder of the paper is organized as follows. Section 2 gives a review of the improved BOVW approaches and fusion approaches for action recognition. Section 3 introduces the proposed $\mathcal{R}$ feature. Section 4 discusses the video representations based on spatio-temporal interest points and describes the proposed fusion method. Section 5 reports experimental results on two human action datasets. Section 6 concludes the paper.

2. Related Work

Recently, several algorithms have been proposed to integrate geometrical information into BOVW. A common way is to use multi-scale pyramids [7] or spatio-temporal grids [8, 9] to produce a coarse description of the feature layout. These algorithms uniformly divide the 3D space into a spatio-temporal grid and then compute the histogram of local features in each sub-volume. The grid structure captures some simple location information, but richer geometrical distribution information is yet discarded. Bregonzio *et al.* [10] propose to treat the interest points inside a spatio-temporal window as a point "cloud". They perform the foreground object detection and segmentation. For each frame, ten features are extracted from the point cloud and detected object area, including the height and width ratio, speed, and relationship between the cloud and object area.

Usually, one kind of feature on its own is insufficient to fully describe a video. Therefore, a number of approaches, which propose feature fusion for improving action recognition in video sequences, have recently appeared in the literature [11, 12, 13, 14]. In [11, 12], the feature-level fusion is employed. All the feature vectors produced by different approaches are concatenated to form a larger feature vector. Liu *et al.* [13], Ye *et al.* [14] and Bregonzio *et al.* [15] employ the kernel-level fusion approach and utilize a multi-kernel classifier for combining different features. The above fusion approaches rely only on the pairwise similarities of videos without considering the high-order correlations among videos. This may cause sensitivity to noise and outliers of the data. Context-aware kernel methods [16, 17] have been proposed since they take advantage of the higher-order contextual information from samples. They are proven to achieve higher performances than the context-free kernels for image annotation [16] and object tracking [17]. Nevertheless, the context-aware kernel has not been explored for action recognition.

3. 3D $\mathcal{R}$ Transform on Spatio-Temporal Interest Points

The 2D $\mathcal{R}$ transform [4], as an improved representation of the 2D Radon transform, has shown to be an effective feature representation of human shape and silhouette in an image [18]. The 3D discrete Radon transform [3, 5] has been successfully applied to classifying objects in 3D models. However, there is little work on 3D $\mathcal{R}$ transform based on 3D Radon transform. In the following, we first deduce the new form and properties of $\mathcal{R}$ transform defined on the 3D discrete Radon transform and then use it to describe the distribution of the spatio-temporal interest points.

3.1. 3D $\mathcal{R}$ Transform

The definition of the 3D $\mathcal{R}$ transform is based on the 3D Radon transform. In other words, the 3D $\mathcal{R}$ transform is an extended representation of 3D Radon transform. Therefore, we start with a brief overview of 3D Radon transform.

Let $\mathbf{M}$ be a 3D model and $f(\mathbf{x})$ be the binary function defined on 3D space, where $\mathbf{x} = (x, y, t)$ denotes the position of a point in the 3D space. The binary function $f(\mathbf{x})$ is 1 when $\mathbf{x}$ lies within $\mathbf{M}$, and otherwise 0. The 3D discrete Radon transform is defined by summing the interpolated samples of a discrete 3D array lying on planes which satisfy certain constraints [3]. Given $\{\mathbf{x}_j\}_{j=1}^{J}$ be all the points in the model $\mathbf{M}$, the 3D discrete Radon transform of the 3D model $f(\mathbf{x})$ is defined by [5]:

$$T_f(\boldsymbol{\eta}, \rho) = \sum_{j=1}^{J} f(\mathbf{x}_j)\delta(\mathbf{x}_j^T \boldsymbol{\eta} - \rho) \qquad (1)$$

where $\boldsymbol{\eta}$ is a unit vector in 3D space, ρ is a real number, and $\delta(\cdot)$ is the Dirac delta function. The unit vector $\boldsymbol{\eta}$ can be written in spherical coordinates as: $\boldsymbol{\eta} = [\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta]$. Thus, Eq. (1) is rewritten as:

$$\begin{aligned} T_f(\rho, \theta, \phi) = \sum_{j=1}^{J} & f(x_j, y_j, t_j) \cdot \\ & \delta(x_j \cos\phi\sin\theta + y_j \sin\phi\sin\theta + t_j\cos\theta - \rho). \qquad (2) \end{aligned}$$

It can be easily calculated, but it is not invariant to translation, scaling or rotation. To overcome this problem, we define the $\mathcal{R}$ Transform of 3D Radon transform, inspired the 2D counterpart in [4]. The 2D $\mathcal{R}$ transform is defined as the integral of the square of the 2D Radon transform over the parameter ρ. Therefore, we define the 3D $\mathcal{R}$ transform as follows:

$$\mathcal{R}_f(\theta, \phi) = \int_{-\infty}^{\infty} T_f^2(\rho, \theta, \phi) d\rho \,. \qquad (3)$$

Next, we derive the following properties of the proposed new $\mathcal{R}$ Transform.

For a scale factor α, we have

$$\begin{aligned} \frac{1}{\alpha^2}\int_{-\infty}^{\infty} T_f^2(\alpha\rho, \theta, \phi) d\rho &= \frac{1}{\alpha^3}\int_{-\infty}^{\infty} T_f^2(\nu, \theta, \phi) d\nu \\ &= \frac{1}{\alpha^3}\mathcal{R}_f(\theta, \phi) \,. \qquad (4) \end{aligned}$$

For a spatio-temporal translation by (x_0, y_0, t_0), we have

$$\begin{aligned} \int_{-\infty}^{\infty} & T_f^2(\rho - x_0\cos\phi\sin\theta - y_0\sin\phi\sin\theta - t_0\cos\theta, \\ & \theta, \phi) d\rho = \int_{-\infty}^{\infty} T_f^2(\nu, \theta, \phi) d\nu = \mathcal{R}_f(\theta, \phi). \qquad (5) \end{aligned}$$

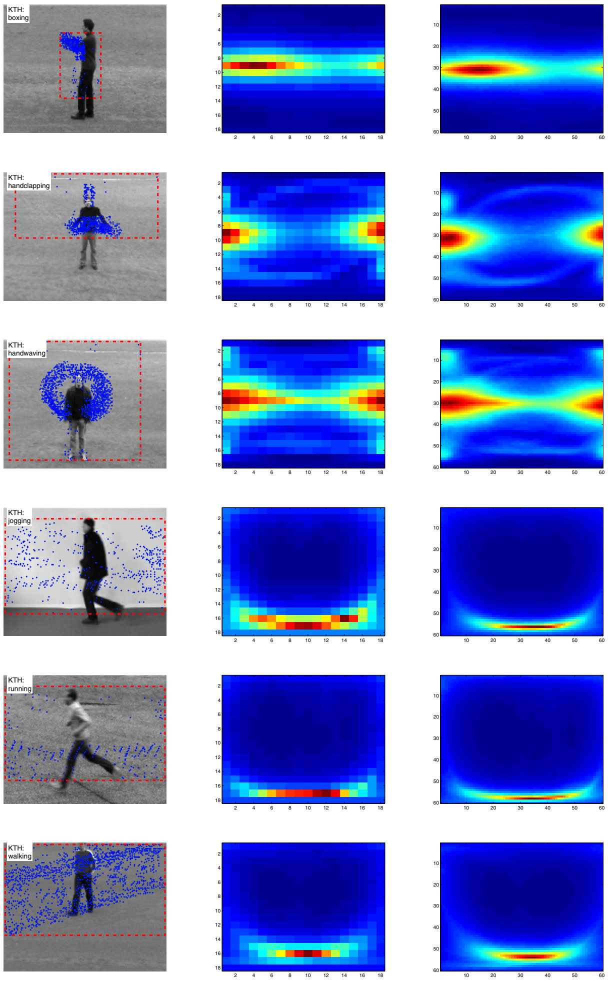

Figure 1. The 3D $\mathcal{R}$ transform of six videos belonging to different action classes in the KTH dataset. From left to right: interest points, $\mathcal{R}$ transform with $(\theta, \phi) = [1 : 10 : 180]$, and $\mathcal{R}$ transform with $(\theta, \phi) = [1 : 2 : 180]$.

For the rotation with angles (θ_0, ϕ_0), we have

$$\int_{-\infty}^{\infty} T_f^2(\rho, \theta + \theta_0, \phi + \phi_0) d\rho = \mathcal{R}_f(\theta + \theta_0, \phi + \phi_0). \quad (6)$$

From Equations (4)-(6), we can see: first, $\mathcal{R}$ transform is invariant to translation; second, scaling leads to amplitude scaling; and third, rotation results in phase shift. To achieve the robustness to rotation, we normalize the $\mathcal{R}$ transform to get the scaling invariance by the following equation:

$$\mathcal{R}'_f(\theta, \phi) = \frac{\mathcal{R}_f(\theta, \phi)}{\max_{\theta,\phi}\{\mathcal{R}_f(\theta, \phi)\}}. \quad (7)$$

These properties make the $\mathcal{R}$ transform useful for representing the distribution of the interest points for action recognition.

3.2. 3D $\mathcal{R}$ Transform on Spatio-Temporal Interest Points

We propose to apply $\mathcal{R}$ transform to 3D video sequence to describe the distribution structure of the spatio-temporal interest points extracted from a video. The minimal spatio-temporal window containing all the interest points extracted from a video is regarded as a 3D model. The binary function $f(\mathbf{x})$ on the 3D model is defined as:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is an interest point} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $\mathbf{x} = (x, y, t)$ denotes the position of each point in the 3D model. Denote $\{(x_j, y_j, t_j)\}_{j=1}^{J}$ be the positions of spatio-temporal interest points detected in a video, where J is the number of interest points. By Eq. (2) and (3), 3D $\mathcal{R}$ transform of the video is given by:

$$\mathcal{R}_f(\theta, \phi) = \int_\rho T_f^2(\rho, \theta, \phi) d\rho = \int_\rho \Big[\sum_{j=1}^{J} f(x_j, y_j, t_j) \cdot \delta(x_j \cos\phi \sin\theta + y_j \sin\phi \sin\theta + t_j \cos\theta - \rho)\Big]^2 d\rho. \quad (9)$$

Observed from Eq. (9), each interest point is first projected into all planes with parameters (ρ, θ, ϕ) and then the $\mathcal{R}$ feature is obtained by the integral of the square of projections over ρ. Therefore, the 3D $\mathcal{R}$ transform efficiently describes the geometrical distribution of interest points. Afterwards, $\mathcal{R}_f(\theta, \phi)$ is normalized by Eq. (7). For convenience, hereafter we use $\mathcal{R}_f(\theta, \phi)$ to represent the normalized $\mathcal{R}$ transform.

The $\mathcal{R}$ transform uses a two dimensional variable $\mathcal{R}_f(\theta, \phi)$ to represent the distribution of the interest points. By sampling two parameters θ and ϕ, $\mathcal{R}_f(\theta, \phi)$ turns out to be a 2D matrix. Figure 1 shows the 3D $\mathcal{R}$ transform of six videos belonging to different action classes in the KTH dataset. In the first column, all the interest points detected in a video are superposed on a single frame. It can be seen that the geometrical distribution of interest points varies according to the different action classes and is very helpful for improving the action recognition accuracy. The second column and third column respectively exhibit the 3D $\mathcal{R}$ transform with $(\theta, \phi) = [1 : 10 : 180]$ and $(\theta, \phi) = [1 : 2 : 180]$. The more samples of θ and ϕ, the more detailed the characterization of the interest points' distribution, but the larger the matrix.

In order to reduce the dimension and improve the robustness of the $\mathcal{R}$ feature, we apply the 2-Directional 2DPCA, i.e. $(2D)^2$PCA, to the matrix obtained from the $\mathcal{R}$ transform. The $(2D)^2$PCA, introduced by [6], simultaneously calculates 2DPCA in the row and column directions and obtains higher recognition accuracy than PCA and two-dimensional PCA (2DPCA) [19]. Finally, by applying the $(2D)^2$PCA on the obtained matrix $\mathcal{R}_f(\theta, \phi)$, we obtain the corresponding low-dimensional matrix as the final feature.

4. Context-aware Feature Fusion for Action Recognition

4.1. Video Representation based on Spatio-temporal Interest Points

We represent each video sequence by two types of features of spatio-temporal interest points: the global $\mathcal{R}$ feature and the BOVW representation of the local cuboid features.

We first perform the spatio-temporal interest point detection for a given video using the Harris3D detector [1]. Afterward, we employ the HOG/HOF feature [2] to describe the cuboid extracted at each interest point. So, a video V is denoted as $(\mathbf{x}_i, \boldsymbol{\alpha}_i)$, $1 \le i \le N$, where $\mathbf{x}_i$ is the spatio-temporal position vector of the i^{th} detected interest point, $\boldsymbol{\alpha}_i$ is the HOG/HOF feature, and N is the total number of interest points detected in the video.

Subsequently, we extract two different types of features to characterize each video. The BOVW based representation only utilizes the HOG/HOF feature $\boldsymbol{\alpha}_i$ of each interest point, while the global $\mathcal{R}$ feature utilizes spatio-temporal position feature $\mathbf{x}_i$.

For the BOVW based representation, several local HOG/HOF features $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_m\}$ from a training set are quantized to form a codebook (i.e. BOVW) by using the k-means clustering method. According to the obtained BOVW, each HOG/HOF feature is mapped into a visual word. Then, each video is represented as a histogram with regard to all visual words, formulated as:

$$H = (n_1, n_2, \cdots, n_K) \tag{10}$$

where $n_i = N_i/N$ denotes the occurrence frequency of the i^{th} visual word in this video, and K is the number of visual words.

For the detected interest points, we use the new $\mathcal{R}$ transform to characterize their spatio-temporal distribution and then refine it by $(2D)^2$PCA, as described in Section 3. Obviously, these two features complement each other. The BOVW based representations rely on the discriminative power of individual local cuboid features, whilst the $\mathcal{R}$ features exploit the global spatio-temporal distribution of the interest points. In addition, the two features are versatile and easy to compute.

4.2. Context-aware Feature Fusion

Given the two feature representations, i.e., the $\mathcal{R}$ feature and the BOVW one, our aim is to obtain the labels of the test videos from the labeled training videos according to the between-video similarity. A crucial step is to compute the similarity between videos, and, since kernel-based classifier is used in our study, to build a kernel matrix from the similarity measure. Traditionally, a kernel matrix is computed based on the pairwise comparison between videos, but such a kernel matrix can be sensitive to noise, outliers, etc. Addressing this issue, we propose a context-aware feature fusion method which not only combines the two feature representations but also captures the underlying contextual information from videos.

The proposed context-aware feature fusion method includes two steps: context selection and context-aware kernel construction. In order to efficiently combine the obtained two features of a video, we use one feature to compute the context of each video and use the other feature to calculate the context-aware kernel for action recognition.

The context of each video is computed using the k nearest neighbor method. We obtain the r nearest neighboring videos of each video in one feature space as its context. Let V_i^1, V_j^1, V_i^2 and V_j^2 denote two feature representations of two video sequences i and j. Let $\mathcal{N}_j$ denote the r nearest neighboring videos of video j. Then $\mathcal{N}_j$ is computed by the following equation:

$$\mathcal{N}_j = \{m | Sort\{d(V_j^1, V_m^1)\} \le r\}, \tag{11}$$

where the distance d is computed by the first feature, and $Sort\{d(V_j^1, V_m^1)\} \le r$ denotes that the video m belongs to the r neighborhood of the video j in the first feature space.

Secondly, the context-aware kernel is composed of three parts, the similarity between the two videos, the similarity between the first video and the context of the second video and vice versa. The context-aware kernel is defined as:

$$\mathbb{K}(i,j) = k(V_i^2, V_j^2) + k_N(V_i^2, V_{\mathcal{N}_j}^2) + k_N(V_j^2, V_{\mathcal{N}_i}^2), \tag{12}$$

where the similarities k and k_N are based on the second feature. The similarity between one video and the context of the other video is defined as:

$$k_N(V_i^2, V_{\mathcal{N}_j}^2) = \frac{1}{r} \sum_{m \in \mathcal{N}_j} k(V_i^2, V_m^2). \tag{13}$$

Finally, we compute the context-aware similarity of two videos by substituting Eq. (13) into Eq. (12). Moreover, when computing the basic similarity of two single videos in any feature space, we employ the intersection kernel [7]:

$$k(V_i, V_j) = \sum_n \min(V_i(n), V_j(n)), \tag{14}$$

where $V_i(n)$ and $V_j(n)$ are the n^{th} feature elements of videos i and j and $k(V_i, V_j)$ measures the "overlap" between two feature bins.

Subsequently, we incorporate the context-aware kernel computed by Eq. (12) into SVM classifier directly. The kernel k from Eq. (14) is a positive definite kernel [7], and therefore Eq. (12) summing several values computed by Eq. (14) is a positive definite kernel. Thereby, Eq. (12) satisfies the Mercer's condition and is directly incorporated into the kernel function of the SVM classifier.

5. Experiments

We test our approach on three human action datasets: KTH [20], UCF sports dataset [21], and UCF films [21]. We emphasize that our approach requires no preprocessing steps such as object segmentation and tracking.

5.1. Parameter Evaluation of the $\mathcal{R}$ Feature

There are two parameters θ and ϕ in $\mathcal{R}$ transform during the computation of the proposed $\mathcal{R}$ feature. Therefore, we evaluate these two parameters in our approach on the KTH dataset. Moreover, we test if the performance is improved by applying $(2D)^2$PCA to refine the feature obtained from the $\mathcal{R}$ transform. We perform leave-one-person-out cross-validation to make the performance evaluation on the KTH video database.

Parameters θ and ϕ are sampled in the range of $[0, 180]$. Figure 2 shows the performances of seven different numbers of samples for θ and ϕ, namely $(\theta, \phi) = [1 : 30 : 180]$, $[1 : 25 : 180]$, $[1 : 20 : 180]$, $[1 : 15 : 180]$, $[1 : 10 : 180]$, $[1 : 5 : 180]$ and $[1 : 3 : 180]$. The blue curve is the obtained recognition accuracy using the $\mathcal{R}$ transform feature without $(2D)^2$PCA, and the red one is recognition accuracy using the $(2D)^2$PCA to refine the $\mathcal{R}$ transform feature. From Figure 2, the following points are observed: i) the sampling frequency has little influence on the final result, and the best accuracy of 91.67% is obtained under $(\theta, \phi) = [1 : 10 : 180]$; ii) the features obtained by $(2D)^2$PCA gain a higher recognition accuracy in most cases than the $\mathcal{R}$ transform features on their own. The former achieves 90.31% average recognition accuracy, and the latter achieves 84.9%. It demonstrates that $\mathcal{R}$ transform feature is an effective descriptor and the $(2D)^2$PCA further improves the discriminative of the $\mathcal{R}$ transform feature. In all other experiments on both datasets, we set $(\theta, \phi) = [1 : 10 : 180]$ and employ $(2D)^2$PCA on $\mathcal{R}$ transform as the final $\mathcal{R}$ feature. The final $\mathcal{R}$ feature is an 18×18 matrix.

5.2. Parameter Evaluation of the Context-aware Feature Fusion

We evaluate the parameter r, which is the number of neighbors for context construction in fusion method, on the KTH dataset. In the BOVW model, the size of codebook is set to 500. Our proposed context-aware fusion methods include two manners. The $\mathcal{R}$ feature can be used for context calculation by Eq. (11) and the BOVW feature for kernel calculation according to Eq. (12), referred to as '$\mathcal{R}$+BOVW'. Alternatively, the BOVW feature can be used for context calculation and the $\mathcal{R}$ feature for kernel calculation, referred to as 'BOVW+$\mathcal{R}$'. Besides, we test single feature based and context-aware kernel methods. Namely, we employ the same type of feature for both context calculation and kernel calculation, referred to as 'BOVW+BOVW' or '$\mathcal{R}$+$\mathcal{R}$' respectively.

The above four experiments are shown in Figure 3. In each experiment, we use the first feature of the legend to calculate the context by Eq. (11) and the second one to achieve the kernel for SVM according to Eq. (12). Namely, 'BOVW+$\mathcal{R}$' and '$\mathcal{R}$+BOVW' are our proposed fusion approaches; while 'BOVW+BOVW' and '$\mathcal{R}$+$\mathcal{R}$' are the one feature based and context aware approaches. From Figure 3, it can be seen that the specific value of r is not very sensitive. We set r to 5 in other experiments to reduce computational cost. Moreover, Figure 5 shows that the $\mathcal{R}$ feature boosts the recognition performance by 6.91% with respect to the 'BOVW+BOVW' kernel averagely.

5.3. Experiments on the KTH Database

To evaluate the two proposed fusion methods, we compared our fusion approaches with two other feature fusion approaches: the feature-level fusion approach [11][12] and the similarity kernel-level fusion approach. All of these fusion approaches combine the proposed two features. Specifically, the feature-level fusion approach concatenates the two normalized feature vectors to form a larger feature vector as input to the SVM classifier. For the similarity kernel-level fusion approach, we separately compute the similarity matrix by each kind of feature, and then utilize the weighted sum of the two obtained similarity matrices as the final kernel of the SVM, which is formulated as follows:

$$\begin{aligned} k_2(V_i, V_j) &= \alpha \sum_n \min(V_i^1(n), V_j^1(n)) + \\ &\quad (1-\alpha) \sum_m \min(V_i^2(m), V_j^2(m)). \end{aligned} \tag{15}$$

Table 1 lists the recognition accuracies of eight approaches on KTH dataset, including one feature based approaches (e.g., 'BOVW', '$\mathcal{R}$'), One feature based and context aware approaches (e.g., 'BOVW+BOVW', '$\mathcal{R}$+$\mathcal{R}$'), two other feature fusion approaches (e.g., 'feature-level fusion', 'kernel-level fusion'), and our proposed fusion approaches (e.g., 'BOVW+$\mathcal{R}$', '$\mathcal{R}$+BOVW'). It shows the following points:

- The $\mathcal{R}$ feature based approach achieves 91.67% accuracy, which is 6.95% higher than the BOVW-based approach. This proves the powerful discrimination of the proposed $\mathcal{R}$ feature.

- One feature based and context aware approaches (e.g., 'BOVW+ BOVW', '$\mathcal{R}$+$\mathcal{R}$') outperform the corresponding one feature based approaches (e.g., 'BOVW', '$\mathcal{R}$'). It illustrates that the extended similarity kernel considering context can improve the recognition performance. Namely, the proposed context aware kernel method obtains higher performance than the traditional context-free kernel method.

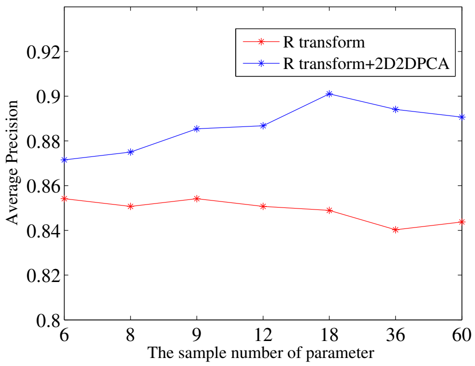

Figure 2. Recognition accuracies obtained by the $\mathcal{R}$ transform feature and the $(2D)^2$PCA feature with respect to seven different samplings of the two parameters θ and ϕ on KTH dataset.

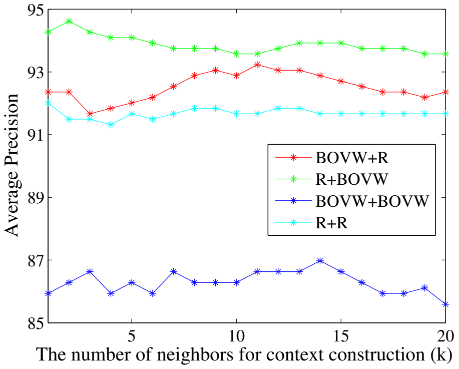

Figure 3. Recognition accuracies of four methods with respect to r, the number of neighbors for context construction on the KTH dataset.

Table 1. Comparison of eight methods on the KTH dataset. (%)

KTH	box	hand clap	hand wave	jog	run	walk	Average
BOVW	90.63	87.50	92.71	76.04	69.79	91.66	84.72
BOVW+BOVW	92.71	90.63	92.71	77.08	78.13	90.63	86.98
$\mathcal{R}$	94.79	95.83	92.71	82.29	89.58	94.79	91.67
$\mathcal{R}$+$\mathcal{R}$	95.83	95.83	93.75	81.25	88.54	96.88	92.01
feature-fusion	98.96	97.92	95.83	86.46	90.63	95.83	94.27
kernel-fusion	98.96	98.96	95.83	87.50	86.46	96.88	94.09
ours: BOVW+$\mathcal{R}$	97.92	95.83	94.79	83.33	89.58	97.92	93.23
ours: $\mathcal{R}$+BOVW	100	98.96	96.88	87.50	89.58	100	**95.49**

- The approaches of fusing two features all achieve higher accuracies than the one feature based approaches. It shows that the $\mathcal{R}$ feature and the BOVW-based feature are complementary and are feasible to be combined for action recognition.

- Our fusion approach (‘$\mathcal{R}$+BOVW’) obtains higher accuracy than the other two common fusion approaches, which demonstrates the effectiveness of our proposed fusion strategy. Moreover, our fusion approach with the $\mathcal{R}$ feature for context calculation and the BOVW feature for kernel calculation (e.g., ‘$\mathcal{R}$+BOVW’) achieves the best accuracies.

5.4. Experiments on the UCF Sports Dataset

The UCF sports database is tested in a leave-one-out manner, cycling each example in as a test video one at a time, following [21] [22] [23]. In the BOVW model, the size of the codebook is set to 800. On the UCF sports database, we perform experiments similar to those on the KTH dataset. The results are shown in Table 2. The similar results as in KTH are obtained in the Table 2, which demonstrate the effectiveness of our proposed cloud feature and fusion method on the realistic and complicated dataset. The overall average accuracy for the UCF dataset using our approach is 87.33%, which is reported 69.2% in [21].

Table 3. Evaluation results on the KTH dataset and the UCF dataset. (%)

Method	KTH	UCF
Our approach	**95.49**	87.33
Bregonzio *et al.* [10]	93.17	-
Sun *et al.* [12]	94.0	-
Yeffet *et al.* [23]	90.1	79.2
Wang *et al.* [2]	92.1	85.6
Kovashka *et al.* [22]	94.53	87.27
Le *et al.* [24]	93.9	86.5
Wang *et al.* [25]	94.2	**88.2**

Besides, Table 3 presents a comparison of our results with state-of-the-art results, which indicate that our approach achieves or even outperforms the listed approaches. With our approach, the overall average accuracies are 87.33% for the UCF dataset and 95.49% for the KTH dataset.

Table 2. Evaluation results on the UCF Sports dataset. (%)

Method	dive	golf	lift	kick	ride	run	skate	swing1	swing2	walk	Average
BOVW	100	66.67	100	80	58.33	53.85	58.33	95	76.92	77.27	76.67
BOVW+BOVW	100	72.22	100	80	58.33	53.85	58.33	95	76.92	77.28	77.33
$\mathcal{R}$	85.71	66.67	83.33	95	66.67	69.23	91.67	100	76.92	63.64	80.00
$\mathcal{R}$+$\mathcal{R}$	92.86	55.56	100	95	66.67	69.23	91.67	100	76.92	68.18	82.00
feature-fusion	100	77.78	100	100	58.33	61.54	83.33	100	76.92	86.36	85.33
kernel-fusion	100	83.33	100	85	58.33	61.54	75	100	84.62	86.36	84.00
BOVW+$\mathcal{R}$	92.86	72.22	100	95	83.33	76.92	100	100	76.92	72.73	86.00
$\mathcal{R}$+BOVW	100	88.89	100	95	58.33	69.23	83.33	100	76.92	90.91	**87.33**

6. Conclusion

In this paper we presented a new action recognition framework based on spatio-temporal interest points. We first proposed a new holistic video representation, the 3D $\mathcal{R}$ transform on spatio-temporal interest points, to capture the information of the global geometrical distribution. We then proposed a new fusion strategy to combine the local cuboid feature and the global $\mathcal{R}$ feature for action recognition. A context-aware kernel between two video sequences has been designed in order to overcome the disadvantage of the traditional pairwise context-free kernels, which is sensitive to noise and outliers in the data. Experimental results on several datasets have demonstrated the effectiveness of our proposed $\mathcal{R}$ feature and fusion method.

Acknowledgement

This work is partly supported by NSFC (Grant No. 60935002, 61100099), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081). Ling is supported in part by NSF (Grant No. IIS-1218156).

References

[1] I. Laptev. On space-time interest points. *IJCV*, Vol.64, No.2, pp.107-123, 2005.

[2] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[3] Y. Shkolnisky, and A. Averbuch. 3D Fourier Based Discrete Radon Transform. *Applied and Computational Harmonic Analysis*, Vol.15, No.1, pp.33-69(37), 2003.

[4] S. Tabbone, L. Wendling, J. Salmon. A new shape descriptor defined on the Radon transform. *CVIU*, pp.42-51, 2006.

[5] P. Daras, D. Zarpalas, D. Tzovaras, and M. G. Strintzis. Shape Matching using the 3D Radon Tranform. In *3DPVT*, 2004.

[6] D. Zhang, and Z. Zhou. $(2D)^2$PCA: 2-Directional 2-Dimensional PCA for Efficient Face Representation and Recognition. *Neurocomputing*, Vol.69, No.1-3, pp.224-231, 2005.

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pp.2169-2178, 2006.

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

[9] J. Choi, W. J. Jeon, and S. C. Lee. Spatio-Temporal Pyramid Matching for Sports Videos. In *ACM MIR*, 2008.

[10] M. Bregonzio, S. Gong and T. Xiang. Recognising Action as Clouds of Space-Time Interest Points. In *CVPR*, 2009.

[11] L. Wang, H. Zhou, S.-C. Low, and C. Leckie. Action Recognition via Multi-Feature Fusion and Gaussian Process Classification. In *WACV*, 2009.

[12] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *CVPR*, 2009.

[13] J. Liu, J. Luo, and M. Shah. Action Recognition in Unconstrained Amateur Videos. In *ICASSP*, pp.3549-3552, 2009.

[14] Y. Ye, L. Qin, Z. Cheng, Q. Huang. Recognizing Realistic Action Using Contextual Feature Group. In *PCM*, 2011.

[15] M. Bregonzio, T. Xiang, S. Gong. Fusing appearance and distribution information of interest points for action recognition. *Pattern Recognition*, Vol.45, No.3, pp.1220-1234, 2012.

[16] H. Sahbi, and X. Li. Context Dependent SVMs for Interconnected Image Network Annotation. In *ACM MM*, 2010.

[17] X. Li , A. Dick, H. Wang, C. Shen, and A. van den Hengel. Graph mode-based contextual kernels for robust SVM tracking. In *ICCV*, 2011.

[18] Y. Wang, K. Huang, and T. Tan. Human Activity Recognition Based on $\mathcal{R}$ Transform. In *CVPR*, 2007.

[19] J. Yang, D. Zhang, A.F. Frangi, and J. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *PAMI*, VOL. 26, NO. 1, pp.131-137, 2004.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing Human Actions: A Local SVM Approach. In *ICPR*, pp.32-36, 2004.

[21] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: a spatiotemporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[22] A. Kovashka, and K. Grauman. Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition. In *CVPR*, 2010.

[23] L. Yeffet, and L.Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.

[24] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pp. 3361-3368, 2011.

[25] H. Wang, A. Kläser, I. Laptev, C. Schmid, C. L. Liu. Action Recognition by Dense Trajectories. In *CVPR*, pp.3169-3176, 2011.