Neural Computing and Applications (2020) 32:5285–5302
https://doi.org/10.1007/s00521-019-04030-1

ORIGINAL ARTICLE

DKD–DAD: a novel framework with discriminative kinematic descriptor and deep attention-pooled descriptor for action recognition

Ming Tong[1] · Mingyang Li[1] · He Bai[1] · Lei Ma[1] · Mengao Zhao[1]

Received: 8 October 2018 / Accepted: 11 January 2019 / Published online: 8 February 2019

Abstract
In order to improve action recognition accuracy, the discriminative kinematic descriptor and deep attention-pooled descriptor are proposed. Firstly, the optical flow field is transformed into a set of kinematic fields with more discriminativeness. Subsequently, two kinematic features are constructed, which more accurately depict the dynamic characteristics of action subject from the multi-order divergence and curl fields. Secondly, by introducing both of the tight-loose constraint and anti-confusion constraint, a discriminative fusion method is proposed, which guarantees better within-class compactness and between-class separability, meanwhile reduces the confusion caused by outliers. Furthermore, a discriminative kinematic descriptor is constructed. Thirdly, a prediction-attentional pooling method is proposed, which accurately focuses its attention on the discriminative local regions. On this basis, a deep attention-pooled descriptor (DKD–DAD) is constructed. Finally, a novel framework with discriminative kinematic descriptor and deep attention-pooled descriptor is presented, which comprehensively obtains the discriminative dynamic and static information in a video. Consequently, accuracies are improved. Experiments on two challenging datasets verify the effectiveness of our methods.

Keywords Action recognition · Deep learning · Kinematic feature · Attention mechanism

1 Introduction

Human action recognition in videos possesses a crucial academic value and an extensive market application prospect, which make it quickly become a focus and difficulty in computer vision and artificial intelligence. Consequently, it attracts great interests of researchers and institutions. However, action recognition is still a challenging problem while concentrating on some real-world data obtained from web videos [1, 2], movies [3], etc. Therefore, extracting effective features is undoubtedly very significant for action recognition.

1.1 Background and motivation

Extracting dynamic features is one of the important research directions for action recognition. Early works, including spatiotemporal interest point (STIP) [4], cuboids [5] and so on, usually adopt interest point detectors to capture pixels with salient change of intensity or gradients in a spatiotemporal video volume, then describe these interest points or small regions using statistics acquired from neighboring pixels, so as to obtain the motion information of action subject. Subsequently, some methods [6, 7] extend 2D image features to 3D features in videos to acquire spatiotemporal features for action recognition. In addition, quite a few research results show that the motion information of trajectories can obtain impressive performance, such as dense trajectories [8, 9] obtained by tracking densely sampled points using optical flow fields. In fact, the above methods based on interest points have

✉ Ming Tong
mtong@xidian.edu.cn

Mingyang Li
mingyangli@stu.xidian.edu.cn

He Bai
he_bai@stu.xidian.edu.cn

Lei Ma
leima_cn@stu.xidian.edu.cn

Mengao Zhao
mazhao_1@stu.xidian.edu.cn

[1] School of Electronic Engineering, Xidian University, Xi'an 710071, China

been turned out to be successful in the field of action recognition. However, they are highly dependent on localized statistics within a small spatiotemporal neighborhood [5, 9], and cannot describe the global characteristics of motion as a whole. Moreover, there are also some scholars [10, 11] who have built deep networks, such as two-stream convolutional networks [10] and 3D convolutional networks [11], so as to acquire spatiotemporal features for action recognition. However, these networks are not only difficult to train, but also unable to achieve the performance equivalent to hand-crafted features.

As an important tool for describing dynamic properties of videos, optical flow has been widely applied in the field of action recognition. However, many of the features based on local optical flow may simply summarize the flow according to histograms of its orientations [12, 13], thus arguably ignore other potentially discriminative properties. Actually, the optical flow may be regarded as a flow field, so some of its characteristics can be extracted using the fluid dynamic theory. By exploring the dynamic characteristics of optical flow field, optical flow can be described in a richer way to obtain the physical characteristics of flow pattern. However, they are less involved in existing features for action recognition.

With the increasing number of action classes, adopting motion features alone is not discriminative enough for dependable action recognition. In fact, the appearance information of action scene and discriminative object in a video also plays a quite significant role. Recently, due to the favorable learning and abstract abilities of deep learning, it has occupied an absolute dominant position in image processing field and has been widely used in various application fields [14–16]. For this reason, some scholars, by constructing a deep network, have attempted to extract the important static features from images for action recognition. Wang et al. [17] firstly captured the spatial relationship and the high-order correlations between parts. Then, they constructed a hierarchical spatial sum-product network (HS-SPN) to extract static deep features for action recognition. Kwak et al. [18] introduced the triplet-based rank constraints into a deep convolutional network, so as to automatically capture the pose embedding information from still image for action recognition. Subsequently, Qi et al. [19], by defining a joint loss function, integrated the pose hints into the convolutional neural networks (CNN) framework. Thus, the static deep features containing pose information are obtained for action recognition. However, these methods directly input the whole image into deep network for feature extraction without focusing on the discriminative object in background.

In order to overcome the above problem, some scholars attempted to extract features from the discriminative regions of video frames, so as to improve recognition performance. Peng et al. [20] firstly divided the whole human body region [21] into multiple regions. Then, a deep CNN network is used to extract features from individual discriminative regions for action recognition. Ni et al. [22] proposed a network composed by two connected deep convolutional neural networks (DCNNs). The first DCNN adopts video frames as inputs and creates response maps indicating locations for body parts. Then, these maps are fed into the second DCNN for learning discriminative and semantic-aligned action representations of each body part for action recognition. However, the above methods usually need to construct additional networks to obtain the discriminative regions, which is generally difficult for training networks and results in higher computational consumption. Moreover, these methods assume that the discriminative information always exists in the regions around human body, and therefore often focus on human or its parts. In fact, some actions may be easier to be distinguished using the appearance information of action scenes such as the ocean wave in "surfing" action; while others might need to pay close attention to the discriminative object that interacts with the human, such as the bicycle in "bike riding" action.

1.2 Overview of DKD–DAD

Motivated by the above methods, a framework with discriminative kinematic descriptor and deep attention-pooled descriptor (DKD–DAD) for action recognition is proposed, as shown in Fig. 1. Firstly, the optical flow field is transformed into a set of kinematic fields with more discriminativeness to construct two kinematic features; subsequently, a discriminative fusion method is proposed, by which a discriminative kinematic descriptor is obtained to depict the dynamic characteristics of action subject. Secondly, a prediction-attentional pooling method is proposed to automatically acquire the discriminative local regions in a video frame. Furthermore, a deep attention-pooled descriptor is presented to capture the discriminative static information in action scene. Finally, a DKD–DAD framework is constructed, which combines discriminative kinematic descriptor and deep attention-pooled descriptor together for action recognition. Consequently, accuracy is improved.

1.3 Working flow of DKD–DAD

In this section, the whole working flow of DKD–DAD is illustrated in Fig. 2, which includes two branches. Specifically, given an input action video, the left branch aims to obtain the proposed discriminative kinematic descriptor, which depicts the dynamic information of action video; the

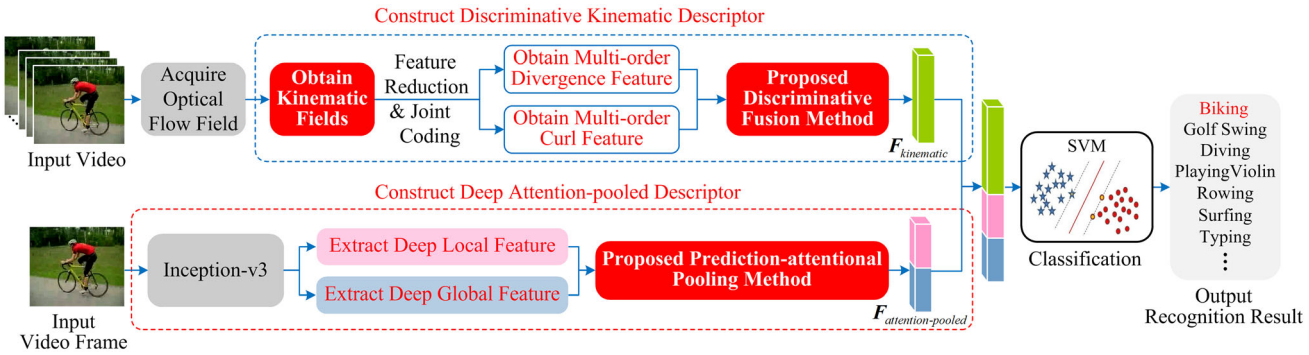

Fig. 1 Overview of the proposed DKD–DAD framework for action recognition

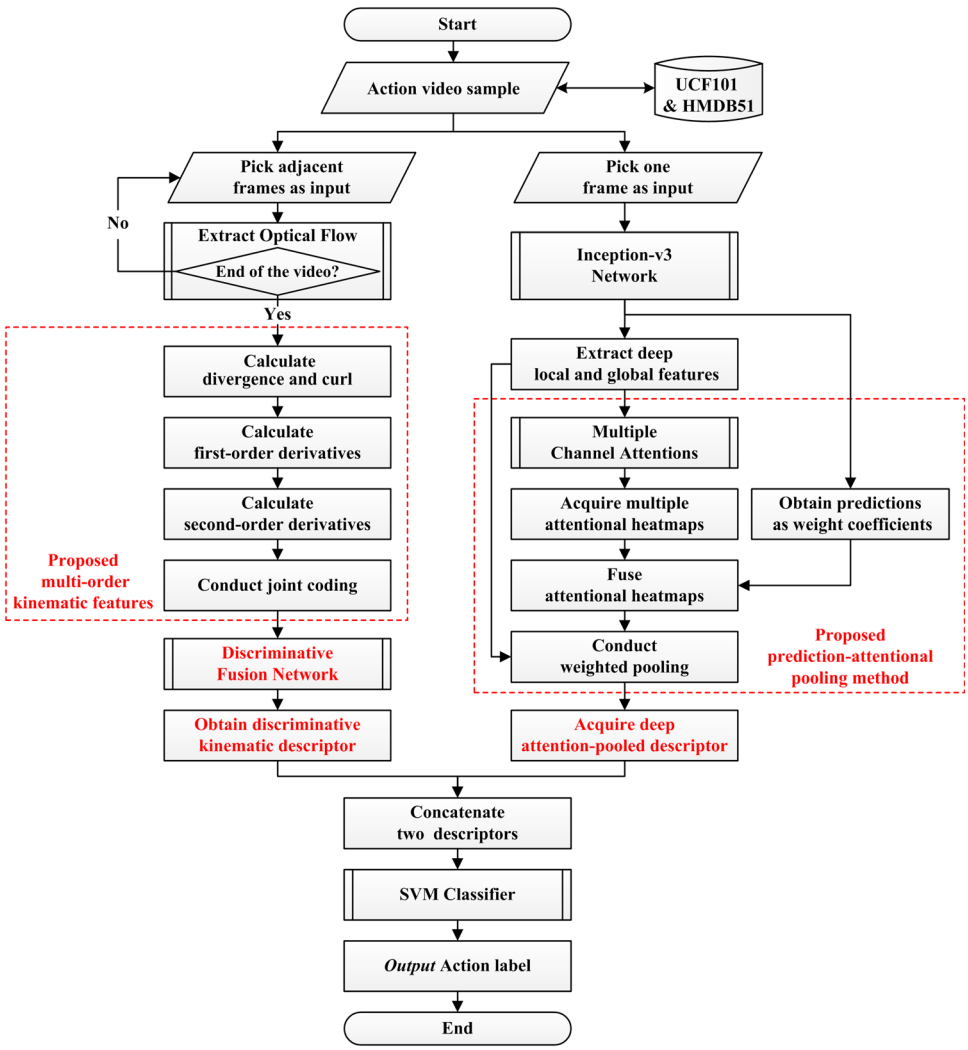

Fig. 2 Working flow of the proposed DKD–DAD framework for action recognition

right branch intends to acquire the proposed deep attention-pooled descriptor, which describes the static information.

The left algorithm branch includes the following steps: Firstly, the optical flow field is extracted. Then, the divergence and curl fields are calculated, respectively. Subsequently, the first-order divergence and curl fields are acquired. Similarly, the second-order divergence and curl fields are obtained. Furthermore, the multi-order divergence and curl fields are, respectively, jointly encoded to acquire the multi-order divergence and curl features. The obtained features are discriminatively fused, consequently the proposed discriminative kinematic descriptor is

obtained. Regarding the right branch, it shows the main procedures as below. Firstly, one frame is randomly picked as the key-frame, which is inputted into the Inception-v3 [23] network to acquire the static deep features. Then, the deep local and global features are, respectively, extracted from the Inception-v3. Subsequently, the extracted features are inputted into the multiple channel attentions to obtain the local and global attentional heatmaps, which are, respectively, fused by taking the predictions of deep network as weight coefficients. Furthermore, the fused attentional heatmaps are used to conduct weighted pooling on deep local and global features, respectively. Consequently the proposed deep attention-pooled descriptor is acquired. Finally, the discriminative kinematic descriptor and deep attention-pooled descriptor are concatenated as the feature representation of action video, then inputted into the support vector machine (SVM) classifier, and thus the action label is obtained.

In summary, the major contributions of this paper are as follows: (1) Two kinematic features of multi-order divergence and multi-order curl are constructed, which more accurately depict the dynamic characteristics of action subject. (2) A novel fusion method is proposed, which ensures the discriminativeness of two kinematic features. Furthermore, a discriminative kinematic descriptor is constructed. (3) A prediction-attentional pooling method is presented. Consequently, a deep attention-pooled descriptor is constructed. (4) A DKD–DAD framework for action recognition is proposed, which finally improves recognition accuracy. Experimental results demonstrate that, the proposed methods can provide promising performance compared to several state-of-the-art methods on two challenging datasets.

The rest of this paper is structured as follows. Section 2 constructs two kinematic features, and meanwhile presents the discriminative fusion method as well as discriminative kinematic descriptor. Section 3 proposes the prediction-attentional pooling method and deep attention-pooled descriptor. The related experiments and analysis of the proposed methods are shown in Sect. 4, followed by the conclusions with future work in Sect. 5.

2 Discriminative kinematic descriptor

Human action videos contain rich motion information that can characterize the intrinsic patterns of different actions. Most of the recent works usually used optical flow field to describe motion information. However, optical flow only records displacement vectors of pixels between two successive frames. While by calculating the kinematics of optical flow field, the physical properties of flow pattern can be captured, which describe motion in a richer way, and contain more details of motion as well as precise variations, such as local expansion, local spin, velocity, acceleration, etc. Therefore, researching field is equivalent to exploring motion itself. In order to better obtain the dynamic characteristics of action subject, this section firstly transforms the optical flow field into a set of kinematic fields with more discriminativeness, and then constructs two kinematic features. Finally, the discriminative fusion method is presented to obtain the proposed discriminative kinematic descriptor, as shown in Fig. 3.

2.1 Construction of kinematic features

In this section, two kinematic features, namely multi-order divergence feature and multi-order curl feature, are constructed. In order to do this, the optical flow is computed firstly [24]. Specifically, given a video, any a point q at time t is denoted as q_t, then the optical flow vector of q_t is denoted as $w(q_t) = (u(q_t), v(q_t))$, where $u(q_t)$ and $v(q_t)$ are the horizontal and vertical components of $w(q_t)$, respectively. Subsequently, by calculating optical flow vector from adjacent frames at every pixel position, the optical flow field is acquired. In the following, the construction process of kinematic features is detailedly given.

1. Extraction of divergence and curl

The divergence and curl are both the local first-order differential scalar quantities of optical flow field, which describe the physical pattern of flow, represent different characteristics of optical flow field, respectively, and meanwhile can well depict the different characteristics of motion in videos from distinct perspectives. In this paper, the divergence and curl of q_t are computed as follows:

$$div(q_t) = \frac{\partial u(q_t)}{\partial x} + \frac{\partial v(q_t)}{\partial y} \tag{1}$$

$$curl(q_t) = \frac{\partial v(q_t)}{\partial x} - \frac{\partial u(q_t)}{\partial y} \tag{2}$$

By, respectively, calculating $div(\cdot)$ and $curl(\cdot)$ for each point in a video frame, the divergence field $Field^{div}$ and curl field $Field^{curl}$ corresponding to the frame are obtained. The physical meaning of $Field^{div}$ derives from the fact that it acquires the amount of local expansion occurring in optical flow field, and can depict the regions of local expansion caused by action subject. The physical significance of $Field^{curl}$ is that it can delineate the local spin around the axis that is perpendicular to the plane of optical flow field, and can describe the dynamic characteristics resulting from human body motion in optical flow field.

2. Acquisition of the first-order derivatives of divergence and curl

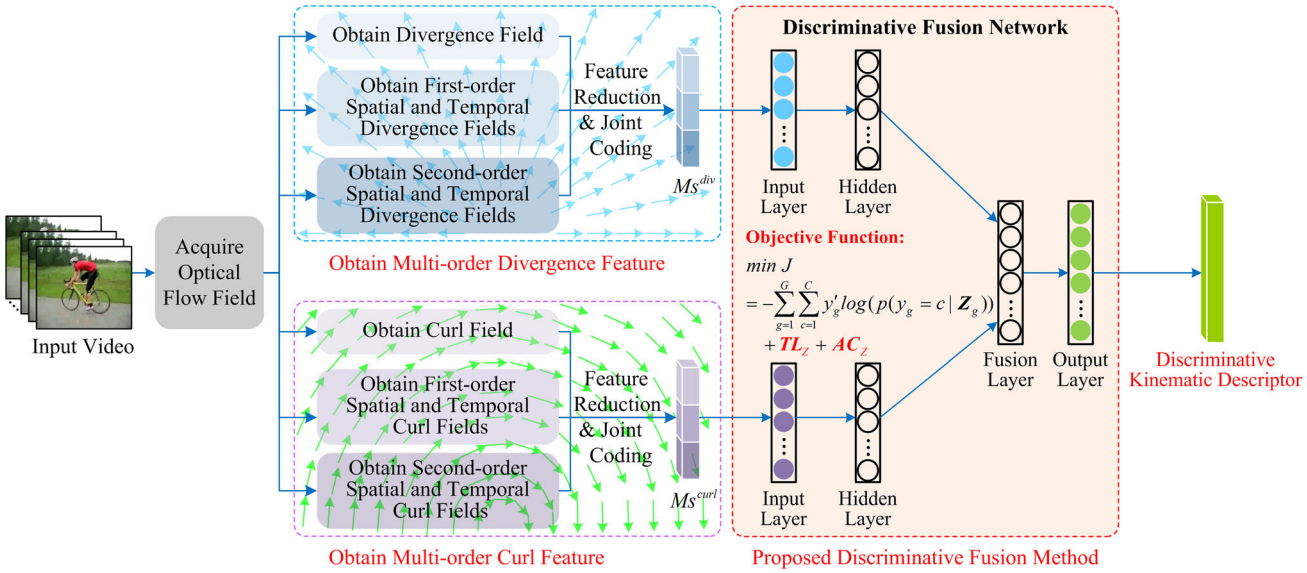

Fig. 3 Schematic of the proposed discriminative kinematic descriptor

In fact, it is not sufficient to describe the optical flow field generated by action subject only using divergence and curl. Therefore, the first-order derivatives of divergence and curl are, respectively, calculated to capture the precise variations of local expansion and local spin caused by the motion of action subject. Given a spatiotemporal point q_t, the first-order derivatives of divergence and curl for it along x, y and t directions are, respectively, computed by following Eqs. (3) and (4).

$$\left(div_x(q_t), div_y(q_t), div_t(q_t)\right)^T = \nabla \cdot div(q_t) \tag{3}$$

$$\left(curl_x(q_t), curl_y(q_t), curl_t(q_t)\right)^T = \nabla \cdot curl(q_t) \tag{4}$$

where $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial t)^T$ represents the gradient operator.

Till then, by, respectively, calculating $div_x(\cdot)$, $div_y(\cdot)$, $curl_x(\cdot)$ and $curl_y(\cdot)$ for each point q_t in a video frame, a set of first-order spatial kinematic fields is obtained, including the first-order spatial divergence fields $Field^{div_x}$ and $Field^{div_y}$ as well as the first-order spatial curl fields $Field^{curl_x}$ and $Field^{curl_y}$, which acquire the relative motion between pixels along x and y directions, and meanwhile remove the camera motion. Similarly, by calculating $div_t(\cdot)$ and $curl_t(\cdot)$ for each point q_t, a set of first-order temporal kinematic fields is obtained as well, including the first-order temporal divergence field $Field^{div_t}$ and the first-order temporal curl field $Field^{curl_t}$, which obtain the velocities of divergence and curl, and meanwhile directly remove the slowly changing background in a video through the subtraction of two consecutive kinematic fields.

3. Acquisition of the second-order derivatives of divergence and curl

In order to more detailedly describe the kinematic characteristics of optical flow field, the second-order derivatives of divergence and curl for q_t along x, y and t directions are further computed, respectively, as shown in Eqs. (5) and (6).

$$\begin{aligned}&\left(div_{xx}(q_t), div_{yy}(q_t), div_{tt}(q_t)\right)^T \\ &= \nabla \odot \left(div_x(q_t), div_y(q_t), div_t(q_t)\right)^T\end{aligned} \tag{5}$$

$$\begin{aligned}&\left(curl_{xx}(q_t), curl_{yy}(q_t), curl_{tt}(q_t)\right)^T \\ &= \nabla \odot \left(curl_x(q_t), curl_y(q_t), curl_t(q_t)\right)^T\end{aligned} \tag{6}$$

where $\odot$ denotes the element-wise multiplication.

In the above formulas, the second-order derivatives $div_{xx}(q_t)$, $div_{yy}(q_t)$, $curl_{xx}(q_t)$ and $curl_{yy}(q_t)$ can, respectively, describe the change rates of the first-order derivatives of divergence and curl along x and y directions. And they construct second-order spatial kinematic fields, including the second-order spatial divergence fields $Field^{div_{xx}}$ and $Field^{div_{yy}}$, as well as the second-order spatial curl fields $Field^{curl_{xx}}$ and $Field^{curl_{yy}}$, which depict the more detailed motion information. Whereas the second-order derivatives $div_{tt}(q_t)$ and $curl_{tt}(q_t)$, respectively, correspond to the change rates of the first-order derivatives of divergence and curl along t direction. Thus, the second-order temporal kinematic fields, including the second-order temporal divergence field $Field^{div_{tt}}$ and curl field $Field^{curl_{tt}}$, acquire the accelerations of divergence and curl.

4. Joint coding

A set of kinematic fields obtained above usually possesses high dimensions and strong correlation, which results in great challenges for the subsequent joint feature coding.

Therefore, above kinematic features are firstly reduced in dimension, respectively, then they are jointly encoded to obtain the proposed multi-order divergence feature and multi-order curl feature. The specific process is as follow. Here, the divergence fields are taken as examples.

(a) Feature dimension reduction. For the j-th frame in the i-th video, its fields $Field_{i,j}^{div}$, $Field_{i,j}^{div_x}$, $Field_{i,j}^{div_y}$, $Field_{i,j}^{div_t}$, $Field_{i,j}^{div_{xx}}$, $Field_{i,j}^{div_{yy}}$ and $Field_{i,j}^{div_{tt}}$ are reduced in dimension, respectively, by two-dimension principle component analysis (2DPCA) [25], so as to obtain the corresponding low-dimensional representation $FIELD_{i,j}^{div} = \left[\hat{F}ield_{i,j}^{div}, \hat{F}ield_{i,j}^{div_x}, \hat{F}ield_{i,j}^{div_y}, \hat{F}ield_{i,j}^{div_t}, \hat{F}ield_{i,j}^{div_{xx}}, \hat{F}ield_{i,j}^{div_{yy}}, \hat{F}ield_{i,j}^{div_{tt}}\right] \in R^d$, where $i = 1, 2, \ldots, G$, G represents video number; $j = 1, 2, \ldots, Q_i$ and Q_i denotes the frame number in the i-th video; d is the dimension of $FIELD_{i,j}^{div}$.

(b) Feature coding. Fisher vector [26] is used to jointly code for above low-dimensional representation. A Gaussian mixture model (GMM) of K components is utilized to create the Fisher vectors. Then, L2 normalization is applied to the Fisher vectors to obtain the multi-order divergence feature set $Ms^{div} = \left\{Ms_1^{div}, Ms_2^{div}, \ldots, Ms_G^{div}\right\} \in R^{G\times O}$ for all videos, where $O = 2dK$. By the same way, the multi-order curl feature set $Ms^{curl} = \left\{Ms_1^{curl}, Ms_2^{curl}, \ldots, Ms_G^{curl}\right\} \in R^{G\times O}$ for all videos is also obtained.

2.2 Construction of discriminative kinematic descriptor

The proposed Ms^{div} and Ms^{curl}, respectively, depict the dynamic characteristics of action subject from multiple levels and different perspectives, between which there exists a certain complementarity information. Therefore, fusing them will necessarily acquire a more complete feature representation to delineate action subject in complex environment more precisely. This section aims to propose a discriminative neural network fusion method to achieve the fusion of Ms^{div} and Ms^{curl}. Consequently, the proposed discriminative kinematic descriptor is obtained, as shown in Fig. 3. The specific process is presented as follow.

1. Introduction of a single tight-loose constraint term

Given a feature set $\boldsymbol{Z} = f\left(Ms^{div}, Ms^{curl}; \boldsymbol{\Theta}\right)$, where $f(\cdot)$ denotes the feature projection function, and $\boldsymbol{\Theta} = \{\Theta_1, \Theta_2, \ldots, \Theta_\omega\}$ represents model parameter set, ω is the number of parameters. A single tight-loose constraint $\boldsymbol{TL}_Z$ is firstly introduced, as indicated in Eq. (7).

$$\boldsymbol{TL}_Z = \frac{1}{G}\sum_{\xi=1}^{C}\sum_{\eta=1}^{C^\xi}\sum_{\gamma=1,\gamma\neq\xi}^{C}\frac{\left\|\boldsymbol{Z}_\eta^\xi - \bar{\boldsymbol{Z}}^\xi\right\|}{\left\|\boldsymbol{Z}_\eta^\xi - \bar{\boldsymbol{Z}}^\gamma\right\|} \tag{7}$$

where C is the number of action classes; C^ξ represents the feature number of the ξ-th class; $\boldsymbol{Z}_\eta^\xi \in R^{1\times 2O}$ denotes the η-th feature of the ξ-th class in $\boldsymbol{Z}$; $\bar{\boldsymbol{Z}}^\gamma \in R^{1\times 2O}$ and $\bar{\boldsymbol{Z}}^\xi \in R^{1\times 2O}$, respectively, represent the feature centers of the γ-th class and the ξ-th class, namely mean values of features.

2. Introduction of an anti-confusion constraint term

It is known that there are usually a large number of outliers in feature space. The distances of these outliers to the feature centers of their own classes are usually larger than the distances to the feature centers of other classes, which seriously affects the discriminativeness of features. In order to reduce the confusion caused by outliers, an anti-confusion constraint $\boldsymbol{AC}_Z$ is introduced as a penalty term to measure the degree of confusion between different classes of features, as shown in Eq. (8).

$$\boldsymbol{AC}_Z = \frac{1}{G}\sum_{\xi=1}^{C}\sum_{\eta=1}^{C^\xi}\sum_{\gamma=1,\gamma\neq\xi}^{C} relu\left(\left\|\boldsymbol{Z}_\eta^\xi - \bar{\boldsymbol{Z}}^\xi\right\| - \left\|\boldsymbol{Z}_\eta^\xi - \bar{\boldsymbol{Z}}^\gamma\right\|\right) \tag{8}$$

where $relu(\cdot)$ represents the rectified linear unit (ReLU) [27].

3. Construction of the objective function for the proposed fusion method

By introducing both of the constraint terms $\boldsymbol{TL}_Z$ and $\boldsymbol{AC}_Z$ into the cross-entropy loss function, the objective function of the proposed fusion method is obtained, as shown in Eq. (9).

$$\begin{aligned} min\, J &= -\sum_{g=1}^{G}\sum_{c=1}^{C} y_g' \, log\left(p\left(y_g = c|\boldsymbol{Z}_g\right)\right) + \boldsymbol{TL}_Z + \boldsymbol{AC}_Z \\ &= -\sum_{g=1}^{G}\sum_{c=1}^{C} y_g' \, log\left(p\left(y_g = c|\boldsymbol{Z}_g\right)\right) \\ &\quad + \frac{1}{G}\sum_{\xi=1}^{C}\sum_{\eta=1}^{C^\xi}\sum_{\gamma=1,\gamma\neq\xi}^{C}\frac{\left\|\boldsymbol{Z}_\eta^\xi - \boldsymbol{Z}^\xi\right\|}{\left\|\boldsymbol{Z}_\eta^\xi - \boldsymbol{Z}^\gamma\right\|} \\ &\quad + \frac{1}{G}\sum_{\xi=1}^{C}\sum_{\eta=1}^{C^\xi}\sum_{\gamma=1,\gamma\neq\xi}^{C} relu\left(\left\|\boldsymbol{Z}_\eta^\xi - \boldsymbol{Z}^\xi\right\| - \left\|\boldsymbol{Z}_\eta^\xi - \boldsymbol{Z}^\gamma\right\|\right) \end{aligned} \tag{9}$$

where y_g and y_g' are, respectively, the predicted label and true label of the g-th sample; $\boldsymbol{Z}_g$ represents the g-th feature in $\boldsymbol{Z}$.

It can be seen from Eq. (9) that, during the optimization solution process, the proposed $\boldsymbol{TL}_Z$, by calculating the relative distances between each feature and its feature center as well as the feature centers of other classes, respectively, makes each feature point be closer to its own feature center, and meanwhile be farther from the feature centers of other classes. That is, the within-class compactness is enhanced, and the between-class separability is increased simultaneously. Consequently, the discriminativeness of features is improved. Further, it can be seen that the proposed $\boldsymbol{AC}_Z$, by gathering the statistics for the sum of error distances in feature space, reduces the between-class confusion caused by outliers.

4. Acquisition of the proposed discriminative kinematic descriptor

Here, a three-layer neural network called the discriminative fusion network is constructed to finally achieve the fusion of Ms^{div} and Ms^{curl}. This network takes the training samples from Ms^{div} and Ms^{curl} as inputs, and Eq. (9) is used as the objective function for optimization. By minimizing Eq. (9) using the stochastic gradient descent (SGD) algorithm, the optimal model parameter set $\boldsymbol{\Theta}^*$ is acquired. Consequently, the discriminative fusion of Ms^{div} and Ms^{curl} is achieved. That is, the proposed discriminative kinematic descriptor $\boldsymbol{F}_{kinematic} = f\left(Ms^{div}, Ms^{curl}; \boldsymbol{\Theta}^*\right)$ is obtained.

Overall, the multi-order divergence feature Ms^{div} and multi-order curl feature Ms^{curl} are firstly obtained from a set of kinematic fields, which possess better discriminativity, and meanwhile remove the camera motion and slowly changing background. Then, in order to acquire a more complete feature representation, the discriminative fusion method is proposed to achieve the fusion of Ms^{div} and Ms^{curl}. Consequently, the discriminative kinematic descriptor is obtained, which possesses better within-class compactness and between-class separability, and meanwhile it is robust to outliers. Moreover, the additional detection of interest points is not needed in this paper, thus the computational consumption is significantly reduced, and the negative effects caused by inaccurate interest point detection on action recognition are effectively avoided. All of these are very useful for action recognition.

3 Deep attention-pooled descriptor

When performing action recognition, both dynamic information and static information are very significant clues. In fact, when recognizing the action classes that are closely related to specific objects or action scenes, static features play a crucial role. This section aims to obtain the discriminative static information in background for action recognition. For this purpose, a deep attention-pooled descriptor is constructed.

Firstly, the architecture of Inception-v3 network is briefly introduced. Then, the prediction-attentional pooling method is proposed. Subsequently, it is applied to both lower layer and higher layer of Inception-v3 for acquiring the proposed deep local attentional feature and deep global attentional feature. Finally, by concatenating the two attentional features, the proposed deep attention-pooled descriptor is constructed, as shown in Fig. 4.

3.1 Introduction of architecture of Inception-v3

Inception-v3 deep neural network was developed by Google, which is a 42 layer deep convolutional neural network with 130 layers, and consists of multiple Inception modules. There exist 4 convoluting modules in each Inception module, and the receptive fields of convoluting modules for each Inception module are allowed to freely select from 5×5, 3×3 and 1×1, which can synthesize the different scale information. Compared with Inception-v2 network [28], Inception-v3 adopts a combination of $1 \times n$ and $n \times 1$ convolutional kernel sizes instead of the original $n \times n$ size, which significantly reduces parameter number. In addition, Inception-v3 adopts the global average pooling (GAP), rather than the traditional fully connected layer, to obtain the feature vector at the end of network.

3.2 Extraction of deep local and global features

In fact, a deep network can learn different features at each layer of layer hierarchy. To be specific, the activations in lower layers possess smaller receptive fields, meanwhile, they are much more sensitive to edge-like patterns and corners; while activations in higher layers possess larger receptive fields, which can learn the more global and high-level feature representation and obtain more complex invariances. However, Inception-v3 only adopts the top layer of network, which is not enough for describing the fine-grained detail.

In order to obtain a more complete feature representation, the local feature and global feature are, respectively, extracted from the lower layer and higher layer of Inception-v3, which lay the foundation for further obtaining the proposed deep local attentional feature and deep global attentional feature. Specifically, (1) the Mixed_5c layer with size $35 \times 35 \times 288$ of Inception-v3 is selected and served as the deep local feature $\boldsymbol{X}^L$, where 35×35 denotes the number of regions in a video frame and 288 is the dimension of feature vector for each region. The reason for selecting the 35×35 region is that, the classical hand-

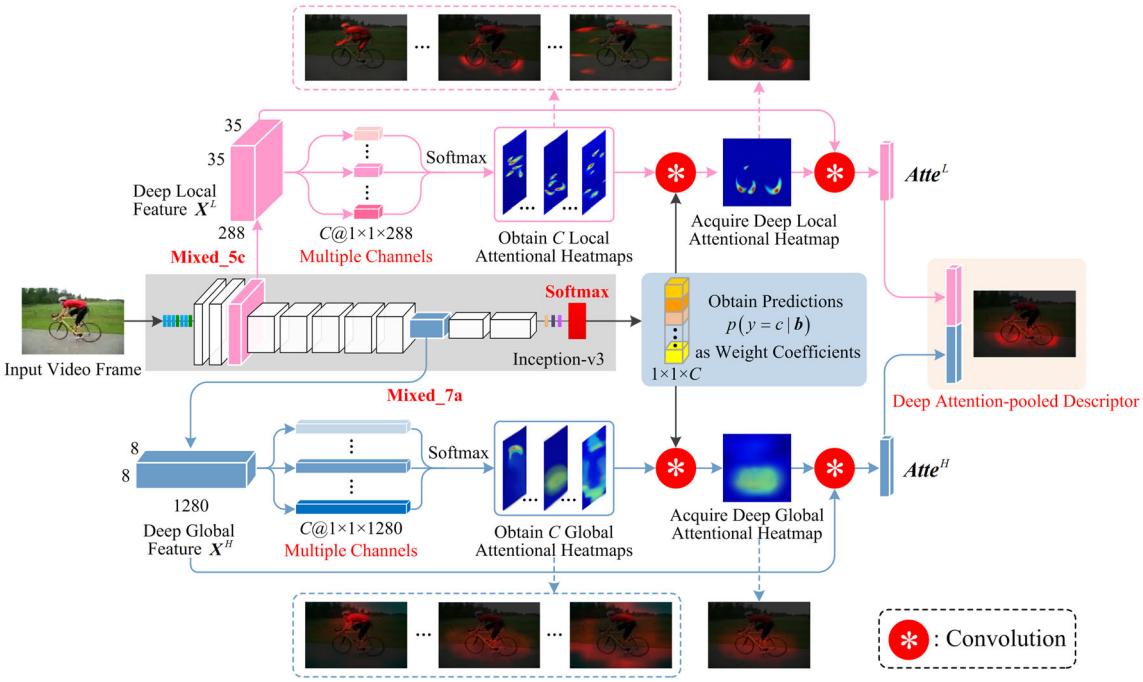

Fig. 4 Schematic of the proposed deep attention-pooled descriptor

crafted feature usually adopts an 8×8 region for local feature extraction, and when it is mapped to Inception-v3, the most similar window scale 35×35 is obtained. (2) the Mixed_7a layer with size $8 \times 8 \times 1280$ of Inception-v3 is selected and taken as the deep global feature $\boldsymbol{X}^H$, where 8×8 denotes the number of regions in a video frame and 1280 is the dimension of feature vector for each region. The reason for choosing Mixed_7a instead of the last Mixed_7c is that, the feature maps of Mixed_7c have very large receptive fields, which means that each pixel point in feature maps corresponds to all regions of input image [29], thus different locations cannot be assigned different weights. That is to say, it is impossible to highlight the discriminative regions, which is disadvantageous for further acquiring the proposed deep global attentional feature.

3.3 Proposed prediction-attentional pooling method

It is well known that the current deep network usually adopts GAP, rather than fully connected layer, to compress the feature map at end of network for obtaining global features. However, GAP considers all the regions inside feature maps equally important, which may reduce the discriminativeness of features [30]. Therefore, some methods [31, 32] use the attention mechanism to highlight the discriminative regions. Furthermore, other ones extend the single-channel attention mechanism to multiple channels for enhancing the discriminativeness of features. Yan et al. [33] proposed a multi-branch attention networks, which obtains the attention maps from scene-level context and region-level context perspectives, respectively. Then, the two context branches are further integrated to acquire the final attentional regions. Girdhar et al. [29] utilized the low-rank second-order pooling to obtain multiple attention maps from bottom-up and top-down perspectives, respectively. Then, these attention maps are combined to acquire the final attentional regions. However, the above methods adopted simple ways to fuse different attention maps without highlighting the more discriminative attention maps, which makes the final acquired attention regions not accurate enough. Depending on the problem to solve, a novel prediction-attentional pooling method is proposed, which aims to more accurately focus on the significant discriminative regions, and meanwhile suppress irrelevant background interference. Details are as follows.

Given an extracted deep feature map $\boldsymbol{X}=\left[\boldsymbol{X}_1^T,\ldots,\boldsymbol{X}_i^T,\ldots,\boldsymbol{X}_N^T\right]^T \in R^{N\times D}$, and $\boldsymbol{X}_i \in R^{1\times D}$ maps to distinct overlapping regions in input space, where N denotes the number of regions in a video frame and D is the dimension of feature vector for each region. Thereupon, the proposed prediction-attentional pooling method is briefly summarized as following. Firstly, the attentions with C channels are constructed, where the number of channels equals to the number of classes, and a single weight is learned for each channel, aiming to pay attention to distinct aspects of deep feature. Then, the attentional heatmap for $\boldsymbol{X}$ in each channel is, respectively, calculated to obtain the

attentional heatmap set $\{\boldsymbol{M}^1,\ldots,\boldsymbol{M}^c,\ldots,\boldsymbol{M}^C\}$, where $\boldsymbol{M}^c \in R^{1\times N}$ is the c-th attentional heatmap. Secondly, the predictions of deep network are taken as weights to fuse $\{\boldsymbol{M}^1,\ldots,\boldsymbol{M}^c,\ldots,\boldsymbol{M}^C\}$, so as to obtain the weighted fusion attentional heatmap $\boldsymbol{M}^{fuse} \in R^{1\times N}$. Thirdly, $\boldsymbol{M}^{fuse}$ is utilized as the weight of $\boldsymbol{X}$ to further enhance the effect of important local regions. Consequently, the more accurate and discriminative deep feature is obtained. The specific calculation process is shown below.

1. Acquisition of attentional heatmaps for C channels. Firstly, in order to obtain $\{\boldsymbol{M}^1,\ldots,\boldsymbol{M}^c,\ldots,\boldsymbol{M}^C\}$, a convolutional kernel $\boldsymbol{a}^c \in R^{1\times D}$ is applied on each channel aiming to acquire attentional heatmaps from different perspectives. Specifically, a softmax function for generating the attention distribution on the regions of the image is adopted for each channel to, respectively, obtain M_i^c, as shown in Eq. (10).

$$M_i^c = \frac{exp\left(\boldsymbol{a}^c \boldsymbol{X}_i^T\right)}{\sum_{j=1}^{N} exp\left(\boldsymbol{a}^c \boldsymbol{X}_j^T\right)} \tag{10}$$

   where M_i^c represents the i-th element in $\boldsymbol{M}^c$, namely the attentional weight of the i-th vector $\boldsymbol{X}_i$ in the c-th channel. The larger the M_i^c is, the higher the importance degree of $\boldsymbol{X}_i$ in the c-th channel is. Equation (10) is adopted for each channel, then the attentional heatmap set $\{\boldsymbol{M}^1,\ldots,\boldsymbol{M}^c,\ldots,\boldsymbol{M}^C\}$ is obtained.
2. Acquisition of the weighted fusion attentional heatmap $\boldsymbol{M}^{fuse}$ for C channels. The motivation is that different actions activate different attentional heatmap sets. In fact, different channels in attentional heatmaps capture different regions related to action subject, discriminative objects and background. In certain circumstance, some channels are more important than the others. Therefore, the higher weights should be assigned to these discriminative channels that play more significant roles in action recognition.

For the sake of highlighting the contributions of discriminative channels related to $\boldsymbol{X}$, the prior probability of $\boldsymbol{X}$ belonging to the c-th class is adopted as the weight of M_i^c to conduct weighted fusion on $\{\boldsymbol{M}^1,\ldots,\boldsymbol{M}^c,\ldots,\boldsymbol{M}^C\}$, so as to obtain the weighted fusion attentional heatmap $\boldsymbol{M}^{fuse}$. For the i-th element M_i^{fuse} in $\boldsymbol{M}^{fuse}$, the calculation is shown in Eq. (11).

$$M_i^{fuse} = \sum_{c=1}^{C} p(y=c|\boldsymbol{X}) M_i^c \tag{11}$$

where y is class label; $p(y=c|\boldsymbol{X})$ represents the prior probability of $\boldsymbol{X}$ belonging to the c-th class. As can be seen from Eq. (11), the larger the $p(y=c|\boldsymbol{X})$ is, the larger the weight of M_i^c is, then the larger the contribution of M_i^c to M_i^{fuse} is, that is to say, Eq. (11) assigns larger weights to the more discriminative channels.

Furthermore, as for the calculation of $p(y=c|\boldsymbol{X})$, according to the structure of deep network, it is known that in the process of forward propagation, the feature map of each layer is obtained from the former layer feature map through basic matrix operations. That is, the conditional probability $p(\boldsymbol{b}|\boldsymbol{X})=1$ holds, in which $\boldsymbol{b}$ is the bottleneck vector of deep network. Thereby, the following derivation holds:

$$\begin{aligned} p(y=c|\boldsymbol{X}) &= p(y=c,\boldsymbol{X})/p(\boldsymbol{X}) \\ &= p(y=c,\boldsymbol{b},\boldsymbol{X})/p(\boldsymbol{b},\boldsymbol{X}) \\ &= p(y=c|\boldsymbol{b},\boldsymbol{X}) \\ &= p(y=c|\boldsymbol{b}) \end{aligned} \tag{12}$$

where $p(y=c|\boldsymbol{b})$ represents the probability of $\boldsymbol{b}$ belonging to the c-th class, namely the prediction of deep network.

It can be seen from Eq. (12) that the probability of $\boldsymbol{X}$ belonging to the c-th class is equal to the prediction of deep network, where the prediction can be obtained by fine-tuning the network on video dataset. Thus, Eq. (11) is transformed as follow:

$$M_i^{fuse} = \sum_{c=1}^{C} p(y=c|\boldsymbol{b}) M_i^c \tag{13}$$

Till then, the weighted fusion attentional heatmap $\boldsymbol{M}^{fuse}$ is acquired.

3. Acquisition of more accurate and discriminative deep feature ***Atte***. $\boldsymbol{M}^{fuse}$ is used to conduct weighted pooling on $\boldsymbol{X}$ for obtaining the attentional feature ***Atte***, as shown in Eq. (14).

$$\boldsymbol{Atte} = \boldsymbol{M}^{fuse}\boldsymbol{X} = \sum_{i=1}^{N}\sum_{c=1}^{C} p(y=c|\boldsymbol{b}) M_i^c \boldsymbol{X}_i \tag{14}$$

In order to obtain ***Atte*** automatically, the SGD algorithm is utilized to minimize the objective function of network, as shown in Eq. (15).

$$\begin{aligned} min\, J = &-\sum_{g=1}^{G}\sum_{c=1}^{C} y'_g \, log\left(p\left(y_g=c|\boldsymbol{Atte}_g\right)\right) \\ &+ \zeta_1 \sum_{i=1}^{N}\left(M_i^{fuse}\right)^2 + \zeta_2 \sum_{c=1}^{C} \|\boldsymbol{a}^c\|_2 \end{aligned} \tag{15}$$

where $\boldsymbol{Atte}_g$ is the deep attentional feature of the g-th sample; $p\left(y_g=c|\boldsymbol{Atte}_g\right)$ denotes the possibility of the g-th sample belonging to the c-th class; ζ_1 and ζ_2, respectively, denote the attentional regularization coefficient and weight decay coefficient; $\|\cdot\|_2$ is l_2-norm.

Conclusively, the proposed prediction-attentional pooling method adopts predictions as weights to conduct weighted fusion on the attentional heatmaps of multiple channels, so as to obtain the weighted fusion attentional heatmap $\boldsymbol{M}^{fuse}$, which highlights the contributions of the discriminative channels and meanwhile suppresses irrelevant background interference. Furthermore, $\boldsymbol{M}^{fuse}$ is utilized as the weight for deep feature map $\boldsymbol{X}$ to enhance the effect of important local regions that are significant for action recognition. Consequently, the more accurate and discriminative deep feature $\boldsymbol{Atte}$ is obtained.

3.4 Construction of deep attention-pooled descriptor

In this section, the proposed deep local attentional feature $\boldsymbol{Atte}^L$ and deep global attentional feature $\boldsymbol{Atte}^H$ are firstly obtained. Then, the proposed deep attention-pooled descriptor is constructed.

Specifically, the proposed prediction-attentional pooling method is applied to both deep local feature $\boldsymbol{X}^L$ and deep global feature $\boldsymbol{X}^H$, thus $\boldsymbol{Atte}^L$ and $\boldsymbol{Atte}^H$ are obtained. It is obviously that, $\boldsymbol{Atte}^L$ mainly focuses on detail information like texture and edge orientation, while $\boldsymbol{Atte}^H$ usually contains global body information and possesses a whole abstract description for action. Therefore, in order to comprehensively depict the discriminative information of action scene, $\boldsymbol{Atte}^L$ and $\boldsymbol{Atte}^H$ are further concatenated to finally construct the proposed deep attention-pooled descriptor $\boldsymbol{F}_{attention-pooled} = \left[\boldsymbol{Atte}^L, \boldsymbol{Atte}^H\right]$.

In summary, by combining $\boldsymbol{Atte}^L$ and $\boldsymbol{Atte}^H$, the proposed deep attention-pooled descriptor can more comprehensively and accurately depict the static visual appearance information of action scene and discriminative object in a video, which improves the discriminativeness of features, and is very useful for action recognition.

4 Experiments and analysis

In this section, the comparisons and analysis on experimental results of the proposed methods for action recognition are reported on two challenging video datasets, namely UCF101 and HMDB51. The illustrations of their representative frames are provided in Fig. 5.

4.1 Datasets and experimental settings

4.1.1 Introduction of datasets

UCF101 [1] dataset includes 13,320 videos with 101 action classes. This dataset gives the largest diversity in terms of actions and large variations in camera motion, viewpoint, object appearance and pose, illumination conditions, cluttered background, object scale, etc. Videos of each action class are divided into 25 groups, where videos from the same group may share similar background and viewpoint. The standard protocol of three train-test splits [34] is used in our experiments, and average accuracy is adopted as the eventual performance measure.

HMDB51 [2] dataset is collected from various sources and represents a fine multifariousness of light conditions, surroundings and situations in which action happens. The camera motion consists of traveling shots, camera shaking, zooming, etc. In total, the dataset contains 6766 video clips divided into 51 action classes, each including at least 101 video clips. The original evaluation scheme of three train-test splits [2] is adopted in our experiments. Each split includes 30 videos for testing and 70 videos for training in each class. The average result over three splits is utilized to evaluate the final performance.

4.1.2 Experimental settings

(1) Parameter setting for the proposed discriminative kinematic descriptor. The number of Gaussians K in the Fisher vector is set to 128. (2) Parameter settings for Inception-v3. Inception-v3 is utilized in this paper, and is trained on the ILSVRC2012 dataset [35]. TensorFlow open source software library [36] provided by Google is utilized to build the CNN framework, and the parameters of Inception-v3 are fine-tuned on UCF101 and HMDB51 datasets using 4 NVIDIA Titan X GPUs. The SGD algorithm is adopted for training the network. The batch size is set as 50; the momentum is set to 0.9; the learning rate is set as 0.0001; the weight decay is set as 0.0005, and the dropout ratio is selected as 0.9. (3) Parameter settings for the proposed deep attention-pooled descriptor. Similarly, SGD is adopted to train the proposed deep network. Specifically, the batch size is set as 200 and learning rate is set to 0.001. (4) Classifier settings. For the proposed discriminative kinematic descriptor and DKD–DAD framework, the linear SVM is used as a classifier. As for the proposed deep attention-pooled descriptor, the output of softmax layer in network is directly used for action recognition.

4.2 Experiment on parameter selection

In this section, UCF101 dataset is taken as an example, and the regularization coefficient ζ_1 and weight decay ζ_2 are optimized to show their significance for action recognition by using the proposed deep attention-pooled descriptor. Specifically, ζ_1 and ζ_2 are, respectively, set by searching the grids $\{0.0005, 0.005, 0.05, 0.5\}$ and

(a) UCF101 **(b)** HMDB51

Fig. 5 Representative frames from videos in UCF101 and HMDB51 datasets

$\{0.00005, 0.0005, 0.005, 0.05\}$. The following Fig. 6 shows the recognition result under different parameters.

Figure 6 shows that the best performance is achieved when $\zeta_1 = 0.05$ and $\zeta_2 = 0.0005$. Ulteriorly, it can be observed that, when ζ_1 and ζ_2 become larger or smaller, the accuracy begins to decline. Thereby, the trade-off on ζ_1 and ζ_2 is very necessary. In fact, when ζ_1 and ζ_2 become larger or smaller, the deep attention-pooled descriptor cannot more accurately highlight the discriminative regions of key-frame, then the discriminability of static features is weakened, thereby the recognition performance becomes worse. Consequently, ζ_1 and ζ_2 are, respectively, set to 0.05 and 0.0005 on UCF101 dataset in subsequent experiments. Furthermore, similar results are demonstrated on HMDB51 dataset.

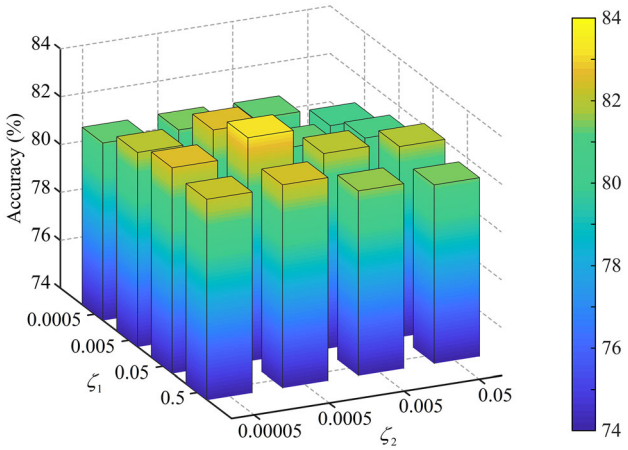

Fig. 6 Accuracy of the proposed deep attention-pooled descriptor versus regularization coefficient and weight decay parameters on UCF101 dataset

4.3 Action recognition with kinematic features

In this section, the proposed kinematic features are applied for action recognition to verify their effectiveness. Tables 1 and 2, respectively, show the recognition results of the proposed kinematic features, namely multi-order divergence feature and multi-order curl feature, as well as contrastive methods on UCF101 and HMDB51.

As can be seen from Tables 1 and 2, the proposed kinematic features achieve better accuracies than all contrastive methods. The reasons lie in that: both of the features, by transforming optical flow field into a set of kinematic fields with more discriminativeness, acquire the different dynamic characteristics of optical flow field from multiple levels and various perspectives. In fact, they capture the spatiotemporal characteristics of action subject, thus they can more accurately depict the detailed motion information of subject, and meanwhile remove the camera motion and slowly changing background. Consequently, the accuracies are improved.

4.4 Action recognition with discriminative kinematic descriptor

This section aims to demonstrate the effectiveness of the proposed discriminative kinematic descriptor, namely validity of the proposed discriminative fusion method. Taking UCF101 and HMDB51 datasets as examples, Tables 3 and 4, respectively, show the recognition results of the proposed discriminative kinematic descriptor obtained by the proposed discriminative fusion method, as well as the results of the concatenation and linear weighted fusion for the proposed multi-order divergence feature and multi-order curl feature. Meanwhile, the results of contrastive methods are also given. In addition, the weight coefficients of linear weighted fusion are obtained by use of the grid search algorithm.

It can be observed from Tables 3 and 4 that: (1) The proposed discriminative kinematic descriptor outperforms all contrastive methods. (2) The result of concatenating the multi-order divergence feature and multi-order curl feature

Table 1 Recognition result of the proposed kinematic features and contrastive methods on UCF101 dataset

Method	Feature	Accuracy (%)
Miao et al. [37]	Trajectory	71.00
Shi et al. [38]	HOG3D	68.50
Peng et al. [39]	HOG	74.79
	HOF	78.63
Nguyen and Mirza [40]	MBH	79.80
Kobayashi [41]	LMS	80.48
Proposed	Multi-order divergence feature	88.86
	Multi-order curl feature	87.92

Table 2 Recognition result of the proposed kinematic features and contrastive methods on HMDB51 dataset

Method	Feature	Accuracy (%)
Jain et al. [42]	DCS	52.10
Miao et al. [37]	Trajectory	39.00
Yu et al. [43]	LFF	46.90
Wang et al. [8]	HOG	44.40
	HOF	52.30
	MBH	56.90
Caetano et al. [44]	OFCM	56.91
Xu et al. [45]	IDT-RCB	58.90
Proposed	Multi-order divergence feature	61.81
	Multi-order curl feature	61.11

is superior to the results of using any of them alone, which indicates that there is really some complementary information between them. (3) The performance of linear weighted fusion is better than that of concatenation, which indicates that the proposed multi-order divergence feature and multi-order curl feature have different contribution degrees to action recognition. Based on the above observations, the best performance of the proposed discriminative kinematic descriptor is owing to the following contributions. The proposed discriminative fusion method, by introducing the tight-loose constraint term, reduces the within-class variations while also increasing the between-class differences. That is, the proposed discriminative kinematic descriptor possesses better within-class compactness and between-class separability simultaneously. In addition, by further introducing the anti-confusion constraint term, the confusion caused by outliers is reduced, which enhances the discriminativeness and robustness of the proposed kinematic descriptor. Consequently, the performance is improved effectively.

4.5 Action recognition with prediction-attentional pooling method

This section aims to verify the effectiveness of the proposed prediction-attentional pooling method. Taking the split 1 of UCF101 dataset for example, Fig. 7 demonstrates the recognition accuracies with applying the proposed pooling method, GAP, max pooling (MAX) and classical attention pooling method on the extracted deep local feature $\boldsymbol{X}^L$ and deep global feature $\boldsymbol{X}^H$.

It can be seen from Fig. 7 that: (1) Regardless of the deep local or global feature, the classical attention pooling method achieves better accuracies than GAP and MAX. The reason lies in that, the introduction of the attention mechanism highlights the contribution of discriminative local regions. (2) The proposed pooling method outperforms classical attention pooling method. The reason lies in that, the proposed pooling method adopts the predictions of network output as weights to weighted fuse the attentions of different channels, which further highlights the contributions of discriminative channels. Consequently, the discriminative regions are accurately obtained, and the accuracy is effectively improved.

4.6 Action recognition with deep attention-pooled descriptor

In order to verify the effectiveness of the proposed deep attention-pooled descriptor, Tables 5 and 6, respectively, show the recognition results of this descriptor and contrastive methods on UCF101 and HMDB51 datasets.

As is shown in Tables 5 and 6, the proposed deep attention-pooled descriptor performs better than all contrastive methods. The reason lies in that, the proposed descriptor, by combining the proposed deep local attentional feature and global attentional feature, further accurately depicts the static visual appearance information of action scene and discriminative object in a video, and enhances the discriminativeness of static deep features. Consequently, accuracies are improved effectively.

Table 3 Recognition result of the proposed discriminative kinematic descriptor and contrastive methods on UCF101 dataset

Method	Feature	Accuracy (%)
Miao et al. [46]	HOF + MBH + RBH + HOG	78.90
Kihl et al. [47]	HOG + MrP + GMrP	79.40
Feichtenhofer et al. [48]	IDT + LATE + S_T	87.70
Wang et al. [8]	HOG + HOF + MBH	86.00
Fernando et al. [49]	HOG + HOF + MBH	86.50
Peng et al. [39]	HSV + STP	87.20
Wang et al. [50]	MoFAP	88.30
Kobayashi [41]	IDT + LMS	86.38
Tu et al. [51]	ML-HDP (IDT)	83.40
Zheng et al. [52]	Action sketch	83.85
Proposed	Concatenation	89.86
	Linear weighted fusion	90.25
	Discriminative kinematic descriptor	92.01

Table 4 Recognition result of the proposed discriminative kinematic descriptor and contrastive methods on HMDB51 dataset

Method	Feature	Accuracy (%)
Wang et al. [50]	MoFAP	61.70
Jiang et al. [53]	IDT + FV + TrajMF	57.30
Bilinski and Bremond [54]	IDT + VCML	58.60
Shao et al. [55]	KMP	49.80
Caetano et al. [44]	OFCM	56.91
Fernando et al. [49]	HOG + HOF + MBH	60.00
Wang et al. [8]	HOG + HOF + MBH	60.10
Yang et al. [56]	HOG + HOF + MBHx + MBHy	60.84
Yao et al. [57]	Multiview dictionary learning	54.00
Kobayashi [41]	IDT + LMS	60.22
Proposed	Concatenation	62.88
	Linear weighted fusion	63.12
	Discriminative kinematic descriptor	64.51

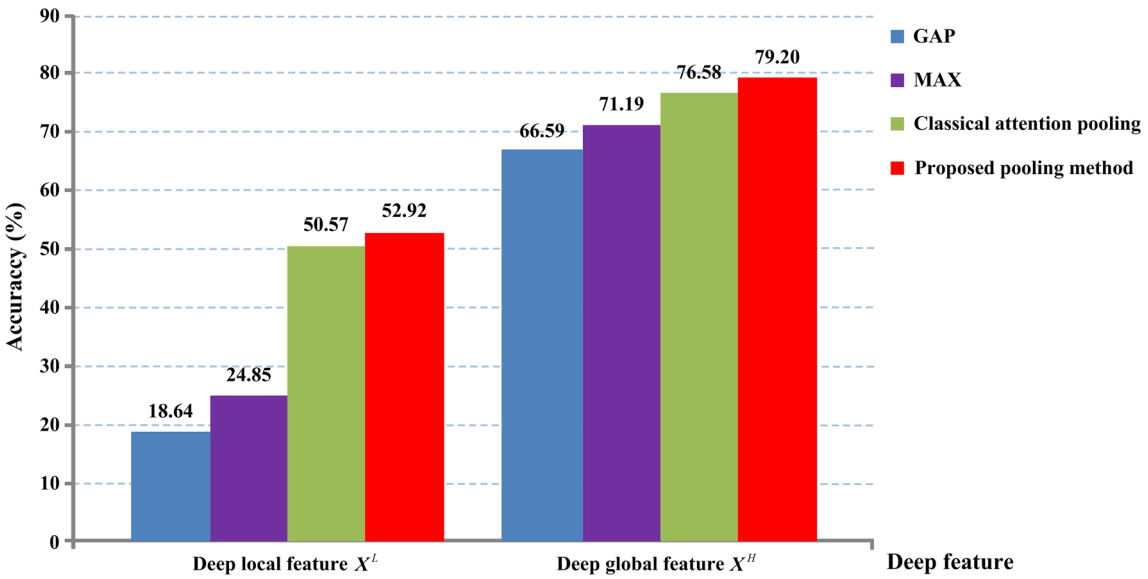

Fig. 7 Recognition accuracy of different pooling methods on UCF101 dataset (split 1)

4.7 Heatmap visualization of prediction-attentional pooling method

To intuitively demonstrate the validity of the proposed prediction-attentional pooling method, Fig. 8 illustrates the visualization examples of heatmaps obtained by the proposed pooling method. For comparative analysis, the visualization examples of heatmaps obtained by classical attention pooling method are given simultaneously.

It can be seen from Fig. 8a that for "biking" action, compared with the visualization result of classical attention pooling method, the deep global attentional heatmap obtained by the proposed pooling method can highlight the discriminative region (bicycle) and meanwhile suppress other irrelevant regions. Furthermore, by superimposing the deep local attentional heatmap on deep global attentional heatmap, the more discriminative local regions (bicycle wheels) are ulteriorly highlighted. Similarly, for "riding horse" in Fig. 8b, "swinging on the pommel horse" in Fig. 8c and "playing violin" in Fig. 8d, the proposed pooling method can also more accurately focus on the discriminative objects "horse," "pommel horse" and "violin" in video frames. For "table tennis shot" in Fig. 8e and "surfing" in Fig. 8f, the same conclusions can be obtained.

4.8 Action recognition with DKD–DAD framework

This section aims to demonstrate the effectiveness of the proposed DKD–DAD framework. Tables 7 and 8, respectively, show the recognition results of the DKD–DAD and contrastive methods on UCF101 and HMDB51 datasets.

From the above experimental results, it can be seen that the proposed DKD–DAD achieves better accuracy than all contrastive methods. Through analysis, this is due to the following contributions. DKD–DAD combines the discriminative kinematic descriptor and deep attention-pooled descriptor together for action recognition, which shares the benefits of both hand-crafted feature and deep feature, and thus comprehensively acquires important dynamic characteristics and discriminative static information in a video. Consequently, accuracies are effectively improved.

4.9 Experiments on running time

Running time plays a significant role in performance assessment, thereby the time consumption of the proposed methods are simply presented. The UCF101 dataset is taken as an example. (1) For a video containing 55 frames, it approximately takes 51.50 s to extract the discriminative kinematic descriptor. Since the proposed kinematic descriptor does not require interest point detection and trajectory tracking, the time consumption is chiefly on calculating optical flow. (2) As for the proposed deep attention-pooled descriptor, it takes about 46.14 ms for each frame. (3) For the proposed DKD–DAD framework, the overall processing time of a 55-frame video is about 53.00 s. These experiments are run on a workstation with a 2.60 GHz CPU.

5 Conclusions

The following conclusions are drawn from this paper. Firstly, by transforming the optical flow field into a set of kinematic fields with more discriminativeness, the dynamic characteristics hidden within the optical flow field are captured. Subsequently, two kinematic features are constructed, which more accurately depict the dynamic characteristics of action subject from the multi-order divergence and curl fields, meanwhile remove the camera motion and slowly changing background. Secondly, a discriminative fusion method is proposed. By introducing a single tight-loose constraint, the better within-class compactness and between-class separability are guaranteed. At the same time, the introduction of the other anti-confusion constraint reduces the confusion caused by outliers. On this

Table 5 Recognition result of the proposed deep attention-pooled descriptor and contrastive methods on UCF101 dataset

Method	Feature	Accuracy (%)
Simonyan and Zisserman [10]	Spatial stream ConvNet	73.00
Zhu and Newsam [58]	STDN	59.10
Bilen et al. [59]	MDI-end-to-end + static-rgb	76.90
Fernando et al. [49]	HRP. (CNN)	78.80
Lan et al. [60]	LOG	80.00
Feichtenhofer et al. [61]	Appearance stream	82.29
Wang et al. [62]	Transformations	80.80
Wang et al. [63]	Appearance	69.60
Kar et al. [64]	AdaScan (Spatial network)	78.60
Proposed	Deep attention-pooled descriptor	83.01

Table 6 Recognition result of the proposed deep attention-pooled descriptor and contrastive methods on HMDB51 dataset

Method	Feature	Accuracy (%)
Zhu and Newsam [58]	STDN	38.30
Bilen et al. [59]	MDI-end-to-end + static-rgb	42.80
Feichtenhofer et al. [61]	Appearance stream	43.42
Fernando et al. [49]	HRP. (CNN)	47.50
Ye and Tian [65]	spatial-C3D-LSTM	51.20
Lan et al. [60]	LOG	52.40
Wang et al. [63]	Appearance	41.30
Kar et al. [64]	AdaScan (Spatial network)	41.40
Girdhar and Ramanan [29]	Pose regularized Attentional Pooling	52.20
Proposed	Deep attention-pooled descriptor	58.19

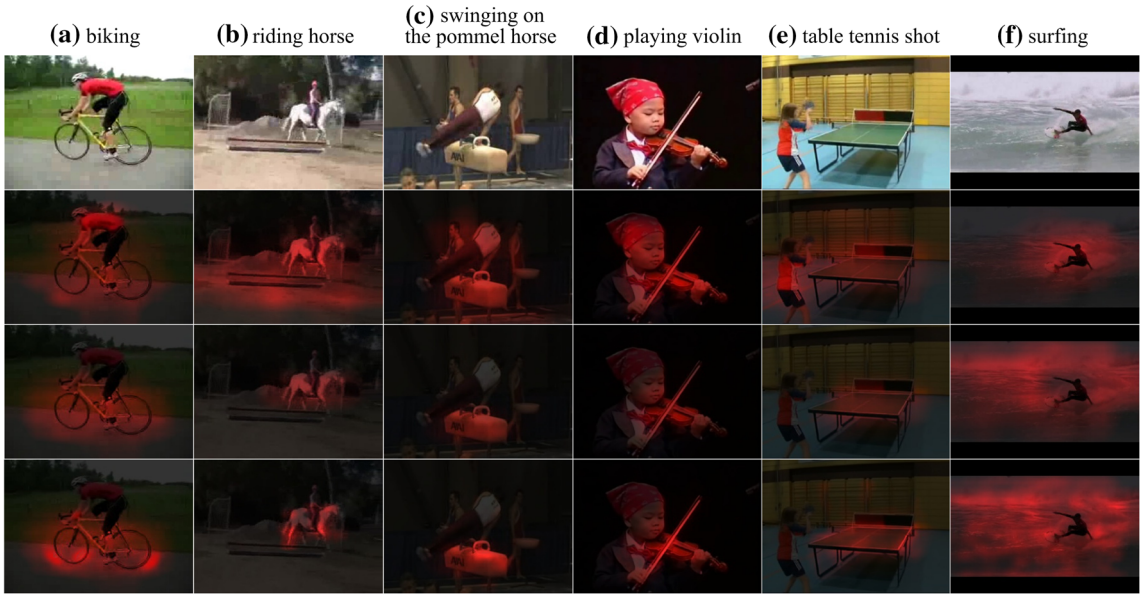

Fig. 8 Heatmap visualization of the proposed prediction-attentional pooling method and classical attention pooling method on UCF101 dataset. Row 1: original video frames; row 2: heatmaps obtained by classical attention pooling method; row 3: deep global attentional heatmaps obtained by the proposed pooling method; row 4: heatmaps obtained by superimposing the deep local attentional heatmaps on deep global attentional heatmaps

Table 7 Recognition result of the proposed DKD–DAD framework and contrastive methods on UCF101 dataset

Method	Feature	Accuracy (%)
Zhang et al. [66]	EMV + RGB-CNN	86.40
Bilen et al. [59]	MDI-end-to-end + static-rgb + trj	89.10
Lan et al. [60]	Hybird	90.60
Fernando et al. [49]	HRP	91.40
Wang et al. [63]	Temporal pyramid CNNs	89.10
Ma et al. [67]	VGG16 + Images + IDT-FV	91.10
Yang et al. [68]	L-SCNN-16	92.00
Cherian et al. [69]	GRP + IDT-FV	92.30
Zhang et al. [70]	DTMV + RGB-CNN	87.50
Varol et al. [71]	$LTC_{Flow+RGB}$ + IDT	92.70
Proposed	DKD–DAD	93.16

Table 8 Recognition result of the proposed DKD–DAD framework and contrastive methods on HMDB51 dataset

Method	Feature	Accuracy (%)
Bilen et al. [59]	MDI-end-to-end + static-rgb + trj	65.20
Fernando et al. [49]	HRP	66.90
Lan et al. [60]	Hybird	67.20
Wang et al. [63]	Temporal pyramid CNNs	63.10
Yang et al. [68]	L-SCNN-16	64.50
Kar et al. [64]	AdaScan + IDT + C3D	66.90
Cherian et al. [69]	GRP + IDT-FV	67.00
Zhang et al. [70]	DTMV + RGB-CNN	55.30
Varol et al. [71]	$LTC_{Flow+RGB}$ + IDT	67.20
Proposed	DKD–DAD	68.06

basis, the discriminative kinematic descriptor is constructed, which possesses better discriminativeness and robustness. Thirdly, a prediction-attentional pooling method is proposed, which adopts the predictions of deep network as weights to weighted fuse different channel information of attentions, and thus highlights the contributions of discriminative channels. Consequently, its attention is more accurately focused on discriminative regions while suppressing irrelevant background interference. Furthermore, the deep attention-pooled descriptor is constructed, which obtains the significant static visual appearance information of action scene and discriminative object in a video. Finally, a DKD–DAD framework is constructed by combining the proposed discriminative kinematic descriptor and deep attention-pooled descriptor, which comprehensively obtains the dynamic characteristics and static information, and further improves the accuracies of action recognition. The proposed methods are extensively evaluated on two challenging datasets of UCF101 and HMDB51, where the superior performance is achieved in comparison with a number of state-of-the-art methods. The future work will focus on researching and designing deeper network as well as more effective pooling method, so as to handle complex video concepts.

Acknowledgement This work was supported partially by Shaanxi Province key project of Research and Development Plan research Project S2018-YF-ZDGY-0187 and International Cooperation Project of Shaanxi Province research project S2018-YF-GHMS-0061.

Compliance with ethical standards

Conflict of interest All the authors of the manuscript declared that there are no potential conflicts of interest.

Human and animal rights All the authors of the manuscript declared that there is no research involving human participants and/or animal.

Informed consent All the authors of the manuscript declared that there is no material that required informed consent.

References

1. Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
2. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: Proceedings of IEEE international conference on computer vision (ICCV), pp 2556–2563
3. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 1–8
4. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
5. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: Proceedings of the 14th international conference on computer communications and networks (ICCCN), pp 65–72
6. Yuan C, Li X, Hu W, Ling H, Maybank SJ (2014) Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition. IEEE Trans Image Process 23(2):658–672
7. Zhang J, Shum HP, Han J, Shao L (2018) Action recognition from arbitrary views using transferable dictionary learning. IEEE Trans Image Process 27(10):4709–4723
8. Wang H, Oneata D, Verbeek J, Schmid C (2016) A robust and efficient video representation for action recognition. Int J Comput Vis 119(3):219–238
9. Wang H, Kläser A, Schmid C, Liu CL (2013) Dense trajectories and motion boundary descriptors for action recognition. Int J Comput Vis 103(1):60–79
10. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems (NIPS), pp 568–576
11. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of IEEE international conference on computer vision (ICCV), pp 4489–4497
12. Yi Y, Lin M (2016) Human action recognition with graph-based multiple-instance learning. Pattern Recognit 53:148–162
13. Singh S, Arora C, Jawahar CV (2017) Trajectory aligned features for first person action recognition. Pattern Recognit 62:45–55
14. Zhang H, Sun Y, Liu L, Wang X, Li L, Liu W (2018) ClothingOut: a category-supervised GAN model for clothing segmentation and retrieval. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3691-y

15. Ji Y, Zhang H, Wu QMJ (2018) Saliency detection via conditional adversarial image-to-image network. Neurocomputing 316:357–368
16. Zhang H, Ji Y, Huang W, Liu L (2018) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3579-x
17. Wang J, Wang G (2018) Hierarchical spatial sum-product networks for action recognition in still images. IEEE Trans Circuits Syst Video Technol 28(1):90–100
18. Kwak S, Cho M, Laptev I (2016) Thin-slicing for pose: Learning to understand pose without explicit pose estimation. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 4938–4947
19. Qi T, Xu Y, Quan Y, Ling L (2017) Image-based action recognition using hint-enhanced deep neural networks. Neurocomputing 267:475–488
20. Peng X, Schmid C (2016) Multi-region two-stream R-CNN for action detection. In: Proceedings of European conference on computer vision (ECCV), pp 744–759
21. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems (NIPS), pp 91–99
22. Ni B, Li T, Yang X (2018) Learning semantic-aligned action representation. IEEE Trans Neural Netw Learn Syst 29(8):3715–3725
23. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 2818–2826
24. Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Proceedings of the Scandinavian conference on image analysis (SCIA), pp 363–370
25. Yang J, Zhang D, Frangi AF, Yang J (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Int 26(1):131–137
26. Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vis 105(3):222–245
27. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NIPS), pp 1097–1105
28. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
29. Girdhar R, Ramanan D (2017) Attentional pooling for action recognition. In: Advances in neural information processing systems (NIPS), pp 34–45
30. Zhou Q, Fan H, Su H, Yang H, Zheng S, Ling H (2018) Weighted bilinear coding over salient body parts for person re-identification. arXiv:1803.08580
31. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. arXiv:1511.04119
32. Zhuang B, Liu L, Shen C, Reid I (2017) Towards context–aware interaction recognition for visual relationship detection. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 589–598
33. Yan S, Smith JS, Lu W, Zhang B (2017) Multi-branch attention networks for action recognition in still images. IEEE Trans Cognit Develop Syst. https://doi.org/10.1109/TCDS.2017.2783944
34. Jiang YG, Liu J, Zamir AR, Toderici G, Laptev I, Shah M, Sukthankar R (2013) THUMOS challenge: action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14. Accessed 30 June 2018
35. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
36. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. http://tensorflow.org/. Accessed 30 June 2018. Software available from tensorflow.org
37. Miao J, Xu X, Qiu S, Qinf C, Tao D (2015) Temporal variance analysis for action recognition. IEEE Trans Image Process 24(12):5904–5915
38. Shi F, Laganière R, Petriu E (2016) Local part model for action recognition. Image Vis Comput 46(11):18–28
39. Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Comput Vis Image Underst 150:109–125
40. Nguyen TV, Mirza B (2017) Dual-layer kernel extreme learning machine for action recognition. Neurocomputing 260:123–130
41. Kobayashi T (2017) Flip-invariant motion representation. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 5628–5637
42. Jain M, Jegou H, Bouthemy P (2013) Better exploiting motion for better action recognition. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 2555–2562
43. Yu M, Liu L, Shao L (2016) Structure-preserving binary representations for RGB-D action recognition. IEEE Trans Pattern Anal Mach Int 38(8):1651–1664
44. Caetano C, dos Santos JA, Schwartz WR (2016) Optical flow co-occurrence matrices: a novel spatiotemporal feature descriptor. In: Proceedings of international conference pattern recognition (ICPR), pp 1947–1952
45. Xu Z, Hu R, Chen J, Chen C, Chen H, Li H, Sun Q (2017) Action recognition by saliency-based dense sampling. Neurocomputing 236:82–92
46. Miao J, Xu X, Mathew R, Huang H (2015) Residue boundary histograms for action recognition in the compressed domain. In: Proceedings of IEEE international conference on image processing (ICIP), pp 2825–2829
47. Kihl O, Picard D, Gosselin PH (2015) A unified framework for local visual descriptors evaluation. Pattern Recognit 48(4):1174–1184
48. Feichtenhofer C, Pinz A, Wildes RP (2015) Dynamically encoded actions based on spacetime saliency. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 2755–2764
49. Fernando B, Anderson P, Hutter M, Gould S (2016) Discriminative hierarchical rank pooling for activity recognition. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR), pp 1924–1932
50. Wang L, Qiao Y, Tang X (2016) MoFAP: a multi-level representation for action recognition. Int J Comput Vis 119(3):254–271
51. Tu NA, Huynh-The T, Khan KU, Lee YK (2018) ML-HDP: a hierarchical bayesian nonparametric model for recognizing human actions in video. IEEE Trans Circuits Syst Video Technol. https://doi.org/10.1109/TCSVT.2018.2816960

52. Zheng Y, Yao H, Sun X, Zhao S, Porikli F (2018) Distinctive action sketch for human action recognition. Signal Process 144:323–332
53. Jiang YG, Dai Q, Liu W, Xue X, Ngo CH (2015) Human action recognition in unconstrained videos by explicit motion modeling. IEEE Trans Image Process 24(11):3781–3795
54. Bilinski PT, Bremond F (2015) Video covariance matrix logarithm for human action recognition in videos. In: Proceedings of the 24th international conference on artificial intelligence (IJCAI), pp 2140–2147
55. Shao L, Liu L, Yu M (2016) Kernelized multiview projection for robust action recognition. Int J Comput Vis 118(2):115–129
56. Yang Y, Liu R, Deng C, Gao X (2016) Multi-task human action recognition via exploring super-category. Signal Process 124:36–44
57. Yao T, Wang Z, Xie Z, Gao J, Feng DD (2017) Learning universal multiview dictionary for human action recognition. Pattern Recognit 64:236–244
58. Zhu Y, Newsam S (2016) Depth2action: exploring embedded depth for large-scale action recognition. In: Proceedings of European conference on computer vision (ECCV), pp 668–684
59. Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S (2016) Dynamic image networks for action recognition. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 3034–3042
60. Lan Z, Yu SI, Yao D, Lin M, Raj B, Hauptmann A (2016) The best of both worlds: Combining data-independent and data-driven approaches for action recognition. In: Proceedings of IEEE international conference on computer vision and pattern recognition workshops (CVPR Workshops), pp 123–132
61. Feichtenhofer C, Pinz A, Wildes R (2016) Spatiotemporal residual networks for video action recognition. In: Advances in neural information processing systems (NIPS), pp 3468–3476
62. Wang X, Farhadi A, Gupta A (2016) Actions $\sim$ transformations. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 2658–2667
63. Wang P, Cao Y, Shen C, Liu L, Shen HT (2017) Temporal pyramid pooling-based convolutional neural network for action recognition. IEEE Trans Circuits Syst Video Technol 27(12):2613–2622
64. Kar A, Rai N, Sikka K, Sharma G (2017) Adascan: adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 3376–3385
65. Ye Y, Tian Y (2016) Embedding sequential information into spatiotemporal features for action recognition. In: Proceedings of the IEEE international conference on computer vision workshops (ICCV Workshops), pp 37–45
66. Zhang B, Wang L, Wang Z, Qiao Y, Wang H (2016) Real-time action recognition with enhanced motion vector CNNs. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 2718–2726
67. Ma S, Bargal SA, Zhang J, Sigal L, Sclaroff S (2017) Do less and achieve more: training CNNs for action recognition utilizing action images from the web. Pattern Recognit 68:334–345
68. Yang H, Yuan C, Xing J, Hu W (2017) SCNN: Sequential convolutional neural network for human action recognition in videos. In: Proceedings of the IEEE international conference on image processing (ICIP), pp 355–359
69. Cherian A, Fernando B, Harandi M, Gould S (2017) Generalized rank pooling for activity recognition. In: Proceedings of IEEE international conference on computer vision and pattern recognition (CVPR), pp 1581–1590
70. Zhang B, Wang L, Wang Z, Qiao Y, Wang H (2018) Real-time action recognition with deeply transferred motion vector CNNs. IEEE Trans Image Process 27(5):2326–2339
71. Varol G, Laptev I, Schmid C (2018) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Int 40(6):1510–1517