Signal Processing 140 (2017) 198–206

Contents lists available at ScienceDirect

Signal Processing

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

Multi-stream deep networks for human action classification with sequential tensor decomposition

Huiwen Guo[a,b,c], Xinyu Wu[a,b,c,*], Wei Feng[a,b]

[a] Guangdong Provincial Key Lab of Robotics and Intelligent Systems, Chinese Academy of Sciences, Shenzhen Institutes of Advanced Technology, PR China
[b] Key Laboratory of Human-Machine-Intelligence Synergic Systems, Chinese Academy of Sciences, Shenzhen Institutes of Advanced Technology, PR China
[c] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, PR China

ARTICLE INFO

Article history:
Received 28 February 2017
Revised 18 May 2017
Accepted 22 May 2017
Available online 23 May 2017

Keywords:
Action classification
Global motion
Tensor decomposition
Gated Recurrent Unit
Recurrent Neural network

ABSTRACT

Effective spatial-temporal representation of motion information is crucial to human action classification. In spite of the attempt of most existing methods capturing spatial-temporal structure and learning motion representations with deep neural networks, such representations are failing to model action at their full temporal extent. To address this problem, this paper proposes a global motion representation by using sequential low-rank tensor decomposition. Specifically, we model an action sequence as a third-order tensor with spatiotemporal structure. Then, by using low-rank tensor decomposition, partial motion of objects in global context were preserved which will be feeding into deep architecture to automatically learning global-term motion features. To simultaneously exploit static spatial features, short-term motion and global-term motion in the video, we describe a multi-stream framework with recurrent convolutional architectures which is end-to-end trainable. Gated Recurrent Unit (GRU) is used as our recurrent unit which have fewer parameters than Long Short-Term Memory (LSTM). Extensive experiments were conducted on two challenging dataset: HMDB51 and UCF101. Experimental results show that our method achieves state-of-the-art performance on the HMDB51 dataset, and is comparable to the state-of-the-art methods on the UCF101 dataset.

1. Introduction

Video human action classification plays an important role in many applications, such as security surveillance, human-computer interaction [1] and annotation [2]. Human action classification means predict or classify the action in the video [3,4]. General speaking, a human action can be seen as spatial-temporal objects, and such a view finds support both in psychology [5] and in computer vision approaches [6]. Naturally, motion information is highly discriminative to recognize actions from a video. Thus, almost successful methods for action classification, indeed, efficiently learning the spatial-temporal representation of motion information.

Consistently with this fact, some traditional approaches [7–9] made their efforts on representing an action with motion-based video descriptors. Some approaches extract holistic motion representations to fully exploit the long-term motion information, such as Liu et al. [10] exploit local geometry for human action recognition by using Hessian regularized analysis [11,12]. Others finding holistic representation in video, such as Improved Dense Trajectory (IDT) [9]. However, Dollar et al. [13] claim that these are too rigid to capture possible variations of actions. To solve this problem, on the one hand, tensor analysis [14–16], with a global perspective, recently introduced to solve this problem and resulted in some successes. On the other hand, Some popular local features, such as Histogram of Gradients in 3D cuboid (HoG3D) [17], were extracted from optical flow fields, gradient field and pixel field. However, most of these methods typically could not deal with long-term action, e.g., people slow walking.

The recent rise of convolutional neural networks (CNNs) convincingly demonstrates the power of learning visual representations [18]. It has been proven empirically that the features automatically learnt from CNNs are much better than the handcrafted features. Equipped with large-scale training datasets, CNNs have quickly take over the majority of visual recognition tasks such as object and scene [19,20]. The attempt of extend CNNs architectures to video action classification often learn motion representations from short video intervals which ranging from 1 to 16 frames [21–23]. Despite the ability to obtain short-term motion features with good performance, these deep architectures were still ignore the long-term motion features of action.

* Corresponding author.
E-mail address: xy.wu@siat.ac.cn (X. Wu).

http://dx.doi.org/10.1016/j.sigpro.2017.05.022
0165-1684/

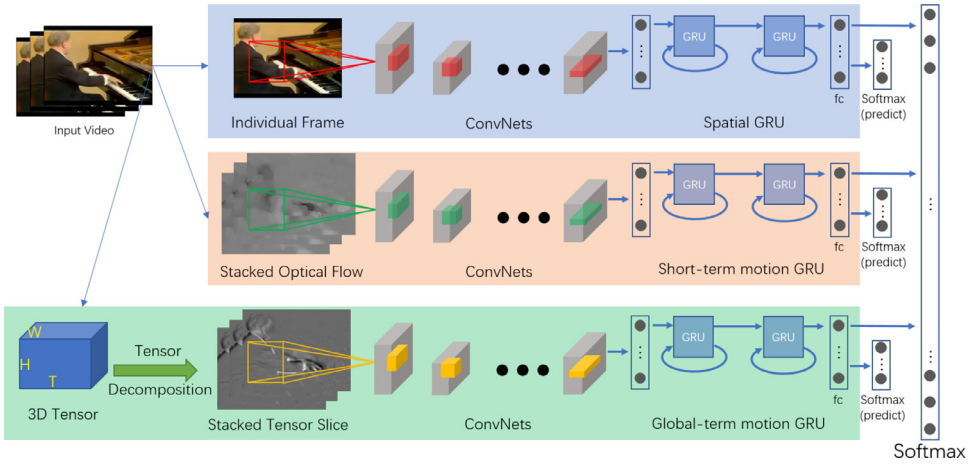

Fig. 1. Illustration of the proposed multi-stream framework.

More recently, Recurrent Neural Networks (RNNs) was introduced to model long-term dependencies of action. By combining convolutional layers and long-range temporal recursion, an architecture called CNN-RNN are built with deep over spatial as well as time dimension. RNNs encodes history information in memory units regulated with non-linear gates to discover temporal dependencies. Nevertheless, even with some complementary strategies such as reversing the video in RNNs [24] and feeding the video twice [25,26] has proved that only the temporal dependencies is not efficient enough to model global motion information in video action.

Realizing the above limitations, in this paper, we propose a sequential low-rank tensor decomposition approach to effectively characterize global motion in video, consequently facilitating human action classification. Firstly, we represent an action sequence as a third-order tensor with spatiotemporal structure. Then, low-rank tensor decomposition approach is introduced to obtain the sparse component which has removed irrelevant background information and preserved global motion information with a global view. To exploit all cues uniformity, a multi-stream framework of deep neural networks is proposed for human action classification. Fig. 1 shows the structure of our method. Our framework is composed of spatial stream, short-term temporal stream and global-term motion stream, which are designed to capture spatial feature, short-term motion feature and global-term motion feature, respectively. The short-term stream is computed on stacked optical flows over a short temporal windows and thus can capture short-term motion. The global-term stream is fed with the slices of sparse component and automatically learning global-term features. In addition, to model the long-term dependencies, we employ a RNN model in our framework. With the empirically study, we choose a recurrent unit called the Gated Recurrent Unit (GRU) on all features extracted by their streams. Compared with the popular Long Short Term Memory (LSTM), GRU has fewer parameters and scarcely decline of performance. The contributions of our work are summarized as follows:

- We introduce a low-rank tensor decomposition approach to effectively extracting global motion information in long term action. We demonstrate the importance of the global-term motion information for high performance of human action classification.
- We propose a multi-stream framework that integrates spatial, short-term motion and global-term motion clues in videos. To complete exploit these information, Recurrent Neural Networks with GRU unit which model the temporal dependencies is applied in our framework. We demonstrate that the multi-stream networks are able digest complementary information to receive significantly improved performance.

The effectiveness of the proposed method is evaluated on two challenging datasets: UCF101 and HMDB51. The experimental results show that our method achieves the state-of the-art performance on the HMDB51 datasets, and out performances most methods on the UCF101 dataset.

The rest of this paper is organized as below: in Section 2, we introduce the related work of the proposed method. In Section 3, we describe our method in details. Then the experimental results are shown in Section 4, and finally the conclusion is given in Section 5.

2. Related work

In the past decade, many researchers have been focusing on human action classification, and comprehensive surveys of this problem can be found in several review papers [27,28]. As aforementioned, the existing approaches of human action classification were committed to designing or learning the most effective feature description.

Typical pipelines resemble earlier methods for object recognition, the use of handcraft local motion features and, in particular, Space-Time Interest Points (STIP) [29] has been found important for action classification. With a global motion view, many approaches were trying to describe global dynamics with all frame information. Fan et al. [30] extract high-level pose features from video to code the pose energy change. Liu et al. [31] using p-Laplacian regularized sparse coding to preserve the local motion geometry.

As a kind of dimension reduction algorithm, tensor decomposition [15,32] is widely applied to action classification with holistic representation of video data [33–35]. By stacked the frames as a 3D tensor, Krausz et al. [36] propose an nonnegative tensor factorization approach to represent the global motion as combination of partial motion (vector basis). Zhang et al. [37] design a tensor descriptor using Histograms of Oriented Gradients. To deal with the unequal length problem in video, Su et al. [38] propose a spatial-temporal iterative tensor decomposition technique. Similar with [38,39] using sparse canonical temporal alignment approach to solve it, and deep tensor decomposition technique is then applied to find a effective representation in tensor subspace [40,41]. Low-rank or sparse decomposition of tensor is one of the methods to extracting more discriminable global motion features. Several algorithms have been proposed to cope with low-rank and sparse decomposition problem in computer vision [42–44]. For example, Candes et al. [45] designed robust PCA(RPCA) method to

decompose an observation matrix into low-rank and sparse components. Goldfarb et al. [46] developed some high-order RPCA method for robust tensor recovery. Despite the successful application in action classification, the handcraft features based on tensor decomposition can not express the variability of the video on the one hand, on the other hand, only the single clue is not effective enough to action recognition. Followed by a deep architecture, our tensor decomposition based characteristics were automatically learned which describe the global motion of video action.

Motivated by the promising results of deep networks on image analysis tasks [47–49], several works have exploited deep architectures for action classification. Ji et al. [50] extended the traditional CNN to 3D-CNN, which gets input from multiple channels and performs 3D convolution. It achieved lower performance compared with the hand-crafted representation [9]. To fed more temporal information into the convolutional networks, Ng et al. [51] explored temporal pooling and concluded that max pooling in the temporal domain is preferable. Empirically, Trans et al. [23] show that a network with 3×3×3 homogeneous filters performs better than varying the temporal depth on filters. A generic descriptor named C3D, is proposed by averaging the outputs of the first fully connected layer of the C3D network. Varol et al. [19] explore 3D convolutions over longer temporal durations at the input layer. Improvements are observed by extending the temporal depth [27]. Recently, a class of multi-stream deep neural network architecture were proposed according to the fact that the motion of an object and its location is handled separately through the *Dosaral Stream* [52]. Simonyan et al. [21] proposed a two-stream deep convolutional networks where input of one stream is static images and the other one is stacked optical flow. The structure of two-stream in [53] were C3D and 3D-CNN. Feichtenhofer et al. [54] and Karpathy et al. [22] shows that a fusion at an intermediate layer improves the performance. Extensions of the two stream network include the work of Wu et al. [55] where a third stream using audio signal is added to the network and Shi et al. [26] where Deep Trajectory Descriptor (sDTD) as the third network. In addition, multi-modal representation of information is widely applied [56,57]. From a variety of perspectives, these algorithms generally extract a variety of features [58].

To exploit the temporal dependencies information, some studies resort to the use of recurrent structures. Baccouche et al. [59] and Donahue et al. [60] tackle the problem of action classification through a cascade of convolutional networks and a class of Recurrent Neural Networks (RNN) [61] known as Long-Short Term Memory (LSTM) [62] networks. In [63] and [64], an widely empirical exploration of various recurrent network architectures were down, and prove that The Gated Recurrent Unit (GRU) [65] outperformed the LSTM on all tasks with the exception of language modeling.

Benefit from above previous works, we design a multi-stream deep networks to simultaneously model spatial, short-term motion and global-term motion clues. GRU networks are then adopted to explore long-term temporal dependencies.

3. Methodology

In this section, we first describe the low-rank tensor decomposition for global motion extraction and analyze the effectiveness of the approach for obtaining global motion information, then introduce the proposed multi-stream architecture, followed by implementation details.

3.1. Low-rank tensor decomposition for global motion learning

For uniformly description, scalars, vectors, matrices and tensors are denoted by lowercase letters, lowercase boldface letters, uppercase boldface and calligraphic letters, respectively. Only real-valued data are considered in this paper. A tensor can be considered as a multidimensional or N-way array. An Nth-order is denoted as: $\mathcal{A} \in R^{I_1 \times I_2 \times \ldots \times I_N}$. Tensor can also be expressed in the form of matrix, which is called tensor matricization. By unfolding a tensor along a mode, a tensors unfolding matrix corresponding to this mode is obtained. This operation is also known as mode-n matricization. For a Nth-order tensor $\mathcal{A}$, its unfolding matrices are denoted by $\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(N)}$. Accordingly, its inverse operator fold can be defined as $fold(\mathcal{A}_{(k)}, k) := \mathcal{A}$. There are two types of higher-order tensor decompositions, the PARAFAC decomposition and the Tucker decomposition. The Tucker decomposition naturally generalizes the orthonormal subspaces corresponding to the left/right singular matrix computed by the matrix SVD. The n-mode tensor $\mathcal{A} \in R^{I_1 \times I_2 \times \ldots \times I_N}$ can be decomposed as:

$$\mathcal{A} = \mathcal{Z} \times_1 U_1 \times_2 U_2, \ldots, \times_n U_n \tag{1}$$

where $U_i \in R^{I_i \times R_i}$ are n orthogonal matrix. U_i spans the R_i dimensional subspace of the original R^{I_i} space, with its orthonormal columns as the basis. U_i accounts for the implicit factor of the ith-mode dimension of tensor $\mathcal{A}$. $\mathcal{Z}$ is the core tensor associating each of the n subspace.

Consider a real $n-$model tensor $\mathcal{A} \in R^{I_1 \times I_2 \times \ldots \times I_n}$, the best rank-$(R_1, R_2, \ldots, R_N)$ approximation is to find a tensor $\widetilde{\mathcal{A}} \in R^{I_1 \times I_2 \times \ldots \times I_n}$ with pre-specified $rank_k(\widetilde{\mathcal{A}}) = R_k$, that minimizes the least-square cost function:

$$\min_{\widetilde{\mathcal{A}}} \|\mathcal{A} - \widetilde{\mathcal{A}}\|_F^2 \quad s.t. \quad rank_i(\widetilde{\mathcal{A}}) = R_i \quad \forall i \tag{2}$$

The $n-$rank conditions imply that $\widetilde{\mathcal{A}}$ should have the Tucker decomposition as Eq. (1): $\widetilde{\mathcal{A}} = \widetilde{\mathcal{Z}} \times_1 \widetilde{U_1} \times_2 \widetilde{U_2}, \ldots, \times_n \widetilde{U_n}$. Problem arises that how to design a method that could automatically find the optical $n-$rank condition of the given tensor A. To simplify the problem, consider a ideal model that corruption is produced by additive irregular patterns ϵ:

$$\mathcal{A} = \mathcal{X} + \mathcal{S} \tag{3}$$

where $\mathcal{A}$, $\mathcal{X}$ and $\mathcal{S}$ are $n-$mode tensors with identical size in each mode. The underlining assumption of Eq. (3) is that the tensor data $\mathcal{A}$ is generated by a highly structured tensor $\mathcal{X}$, and then corrupted by an additive irregular patterns $\mathcal{S}$. One straightforward assumption may be that the $n-$rank of $\mathcal{X}$ should be small and the corruption $\mathcal{S}$ is bounded. To impose these constraints on the Eq. (3), suggesting that the corruption of the irregular patterns $\mathcal{S}$ is bounded. The constraint could be the case in certain situations. However, the irregular patterns in real world visual data is unknown and unbounded in general. A reasonable observation is that the irregular patterns $\mathcal{S}$ usually occupy only a small portion of the data. Therefore, l_0 norm penalization is imposed on $\mathcal{S}$. However, l_0 norm is highly nonconvex optimization. Given the fact that $\|\mathcal{S}\|_1$ is the tightest convex approximation of $\|\mathcal{S}_0\|$, one can relax $\|\mathcal{S}_0\|$ by $\|\mathcal{S}_1\|$. Then, form the Eq. (3) as follows:

$$\min_{\mathcal{X},\mathcal{S}} \frac{1}{2} \sum_{i=1}^{N} \|\mathcal{A}_i - \mathcal{X}_i - \mathcal{S}_i\|_F^2 + \lambda_1 \|\mathcal{X}_i\|_* + \lambda_2 \|\mathcal{S}\|_1 \tag{4}$$

where $\|\mathcal{X}_i\|_*$ and $\|\mathcal{S}_i\|_1$ denote the nuclear and l_1 norm of each mode-i unfolding matrices of $\mathcal{X}$ and $\mathcal{S}$, respectively. The constant λ_1 and λ_2 balance between the low-dimensional structure and sparse irregularity. When the optimal $\mathcal{X}$ is achieved, similar to the Tucker decomposition, the core tensor $\mathcal{Z}$ can be computed by [66]:

$$\mathcal{Z} = \mathcal{X} \times_1 U_1^T \times_2 U_2^T \ldots \times_n U_n^T \tag{5}$$

where U_i is the left singular matrix of $\mathcal{X}_i$. Accordingly, we can get the rank-$(R_1, R_2, \ldots, R_N)$ decomposition of $\mathcal{X} = \mathcal{Z} \times_1 U_1 \times_2 U_2, \ldots, \times_n U_n$. We call the correspondent decomposition in Eq. (1) to be the optimal rank-$(R_1, R_2, \ldots, R_N)$ decomposition of tensor $\mathcal{A}$ under the sense of l_1 norm.

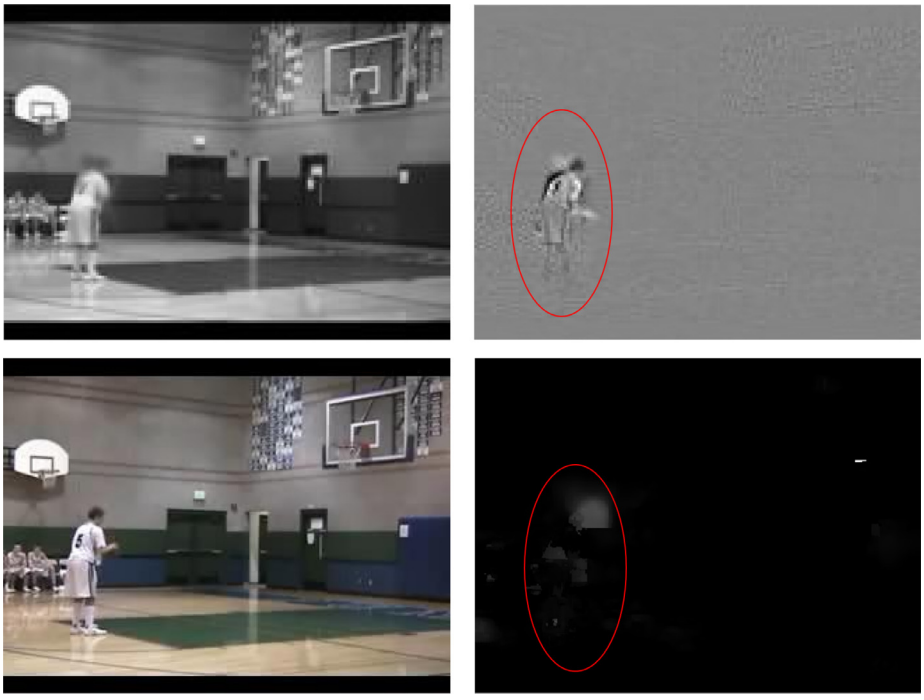

Fig. 2. Example result of tensor decomposition on basketball action. The upper column are the low-rank and the sparse component of original frame, the below column are the original image and optical flow magnitude image. Red ellipses mark their biggest difference between optical flow and tensor decomposition.

Then, we will describe how to learn the global motion information from the output of low-rank tensor decomposition. For an action sequence, $\mathcal{A} \in R^{w\times h\times t}$ was created by stacking all frames into a three dimension tensor, where w and h are the width and height of the frame and t is the length of the video . By using a tensor decomposition method (such as the Rank Sparsity Tensor Decomposition (RSTD) [67]), $\mathcal{X} \in R^{w\times h\times t}$ and $\mathcal{S} \in R^{w\times h\times t}$ were obtained. According to the characteristics of the tensor decomposition, $\mathcal{X}$ and $\mathcal{S}$ were represent the static redundant irrelevant information and the motion information in global context in video, respectively. Then, we feed each time slice of $\mathcal{S}$, i.e., $\mathcal{S}_i \in R^{w\times h}$, where $i = 1, \ldots, t$, into CNNs which belongs to the first half of the global-motion stream.

We empirically analyze the effectiveness of the extracting of the global motion information. One tensor decomposition example of basketball action is shown in Fig. 2. Firstly, it can be seen that irrelevant background information is eliminated, which has been proved that will affect the modeling of human action in the unconstrained video [68]. Secondly, and the most importantly, each moving partial of human body were preserved in video. Although both optical flow and tensor decomposition record the motion of the spatial space, they represent motion information in different manners. The optical flow record the instantaneous motion between two consecutive frames, while the tensor decomposition record all local motion parts throughout all frames. For example, the optical flow image and the tensor decomposition image are shown in Fig. 2. In current frame the player's legs are not moving, while the legs are moved in the previous frames. Thus, there are no information reserved in the position of the leg in the optical flow image. Our tensor decomposition image did not treat the leg information as the background information and preserved the leg information by taking all local time motion into account, which are benefit to the holistic action classification.

3.2. Multi-stream architecture

In this subsection, we will describe our multi-stream architecture for action classification. The core of our framework is shown in Fig. 1. Basically, a good action classification system should contain both spatial and temporal subsystems. In our model, we further consider the temporal subsystem as two modules: short-term temporal subsystem and global-term temporal subsystem. As a result, our multi-stream framework includes spatial stream, short-term stream and global-term stream. Then, we introduce GRU networks to model the temporal dependencies.

The spatial stream is designed to capture static appearance features, by training on single frame images (224 × 224). The temporal stream takes dense optical flow fields as inputs and aims to describe the short-term motion. Like the two-stream networks in [21], whose temporal stream input is volumes of stacking optical flow fields (224 × 224 × 2F, where F is the number of stacking flows and is set to 10), our temporal stream input is also stacked optical flow. An optical flow field is computed from two consecutive frames. As mentioned in the previous subsection, the global-term stream mainly focuses on the global motion. The input of the global-term stream is stacked frames of sparse component slices of tensor decomposition with the length of 5(224 × 224 × 5).

As all clues (spatial, short-term motion, global-term motion) have been captured, we further employ GRU [65] to model temporal dependencies. GRU is a popular RNN model that incorporates memory cells with several gates to learn long-term dependencies without suffering from vanishing and exploding gradients as the traditional RNNs [69]. It is able to exploit temporal information of a data sequence with arbitrary length through recursively mapping the input sequence to output labels with hidden GRU units [55]. Fig. 3 illustrates the typical structure of a hidden GRU unit. Denote x_t as the feature representation from the output of the CNNs. Generally, an GRU maps an input sequence $(x_1, x_2, \ldots, x_T)$ to output labels $(y_1, y_2, \ldots, y_T)$ through computing activations of the units in the network recursively from $t = 1$ to $t = T$. At time t, the activation vectors of hidden state h_t is computed as:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\ \widetilde{h}_t &= tanh(W \cdot [r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t \end{aligned} \tag{6}$$

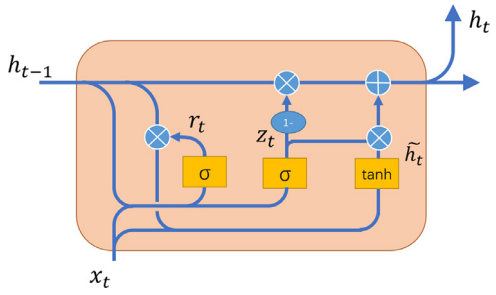

Fig. 3. The structure of a GRU unit.

where W_z, W_r and W are the weight matrices connecting two different units. σ is the sigmoid function, and $[\cdot]$ is an element-wise product operator.

As a neural network, the GRU model can be easily deepened by stacking the hidden states from a layer as inputs of the next layer. Two layers of GRU model is adopted in our framework. Finally, to get the final prediction, we apply late fusion to the three streams.

3.3. Implementation details

In this work, like the [55], we adopt two Convolutional Networks architectures, the VGG19 architecture [47] for the spatial stream and the CNN-M [21] model for capturing the short-term motion and the global-term motion. The CNN-M contain five convolutional layers followed by three fully connected layers. As the frame length of each unit video slip is 20, the output of spatial, short-term and global-term stream for each unit video slip are 20×4096 (time length$\times$ feature dimension), 11×4096 (each stacked optical flow need 10 frames), 16×4096 (each stacked sparse component need 5 frames), respectively. Our implementation is based on the Keras with tensorflow backend. The training process is divided into three steps, the training of convolutional networks(ConvNet) in each stream, the training of each complete stream(contain GRU model) and the jointly training of all streams.

The spatial stream are first pre-trained using the ImageNet [70] training set and fine-tuned using the training video data. The input video frames is uniformly fixed to the size of 224×224. To fine-tune process, we gradually decrease the learning rate from 10^{-2} to 10^{-3} after 1K iterations, to 10^{-4} after 10K iterations and to 10^{-5} after 20K iterations. The dropout is applied to the fully connected layers with a ratio of 0.5 to avoid over-fitting. To training the short-term stream, the optical flow is computed by using the GPU implementation of [71] and stack the optical flow in each 10-frame window to obtain a 20-channel optical flow image as the input. We train the short-term stream from scratch by adopting 0.6 dropout ratio and setting the learning rate gradually decreasing with the increase of training iteration. Initially, it set as 10^{-1}, which is reduced to 10^{-2}, 10^{-3}, 10^{-4} after 50K, 100K, 200K iterations, respectively. The training process of global-term stream is similar to the short-term one, with the input of 5-channel tensor decomposition frames. By consider the sparse component slices as static images, we also tried to use the spatial stream to train it, but observed worse results than motion stream. Note that we augment our data by using crops and mirroring.

Two-layer GRU model [55] is adopt for temporal dependencies modeling. Each GRU has 1024 hidden units in the first layer and 512 hidden units in the second layer. Recent work performing joint training of the GRU with a convolutional networks improves the results on the UCF-101 benchmark with 0.6% . Thus, we jointly training each complete stream. We setting a mini-batch size of 100 to train the network weights, where the learning rate is set as 10^{-3} with 200K iterations.

Table 1
Exploration of the performance of different models on the UCF101 and HMDB51 datasets.

Model	UCF101	HMDB51
Global-term Stream(TD,L=25)	78.9%	51.4%
Global-term Stream(TD,L=50)	78.6%	49.3%
Global-term Stream(TD,L=*max*)	80.4%	47.6%
Spatial ConvNet	81.1%	52.1%
Short-term ConvNet	77.5%	50.3%
Global-term ConvNet	80.4%	51.4%
Spatial ConvNet + LSTM	83.9%	53.9%
Short-term ConvNet + LSTM	81.1%	51.6%
Global-term ConvNet + LSTM	81.6%	51.9%
Spatial ConvNet + GRU	84.5%	54.5%
Short-term ConvNet + GRU	82.7%	52.8%
Global-term ConvNet + GRU	81.9%	52.1%
Spatial + Short-term Stream	90.1%	62.3%
Spatial + Global-term Stream	91.5%	64.7%
Short-term + Global-term Stream	91.9%	64.9%
Multi-Stream	**93.3%**	**67.8%**

The proposed multi-stream framework fusing spatial, short-term motion and global-motion clues. Each stream is trained separately. Finally, by using Softmax layer at the end of all streams, we jointly training our whole deep networks. Despite the complexity and the time consumption of the jointly training process, it improve nearly 1.0% accuracy in application.

4. Experiment

In this section, we will first introduce the detail of datasets. Then, experiments are designed to study the effectiveness of each individual stream. Finally, we report the experimental results compare with the state-of-the-art methods.

4.1. Datasets

UCF-101 [72] is a widely adopted dataset for human action classification, which containing 13,320 video clips annotated into 101 action classes. All the video clips have a fixed frame rate of 25 fps with a spatial resolution of 320×240. This dataset is challenging because most videos were captured under uncontrolled environments with camera motion, cluttered backgrounds and large intra-class variations. We train our networks with unit video clip of 20 frames. To produce a single label prediction for an entire video clip (longer than 20 frames), we average the label probabilities-the outputs of the network's softmax layer-across all frames and choose the most probable label. At test time, we extract 20 frame clips with a stride of 10 frames from each video and average across all clips from a single video.

The HMDB51 dataset [73] is a large collection of realistic videos from various sources, including movies and web videos. It is composed of 6,766 video clips from 51 action categories, with each category containing at least 100 clips. We follow the original evaluation scheme. And the single label prediction is conduct similar to UCF-101 setting.

4.2. Exploration experiments

We report the performance with different setting on two datasets. Firstly, the tensor decomposition algorithm is sensitive to the moving of camera. Shortening the length of the constructed tensor with time dimension may relax the influence while reduces the ability of capturing the global motion information. We set the length L as 25, 50 and *max length* to evaluate the influence for tensor decomposition, the results are shown in Table 1. Comparing the top three cells of results, it match the intuitive expectations that the best result appear in the case of $L = 25$ on HMDB51

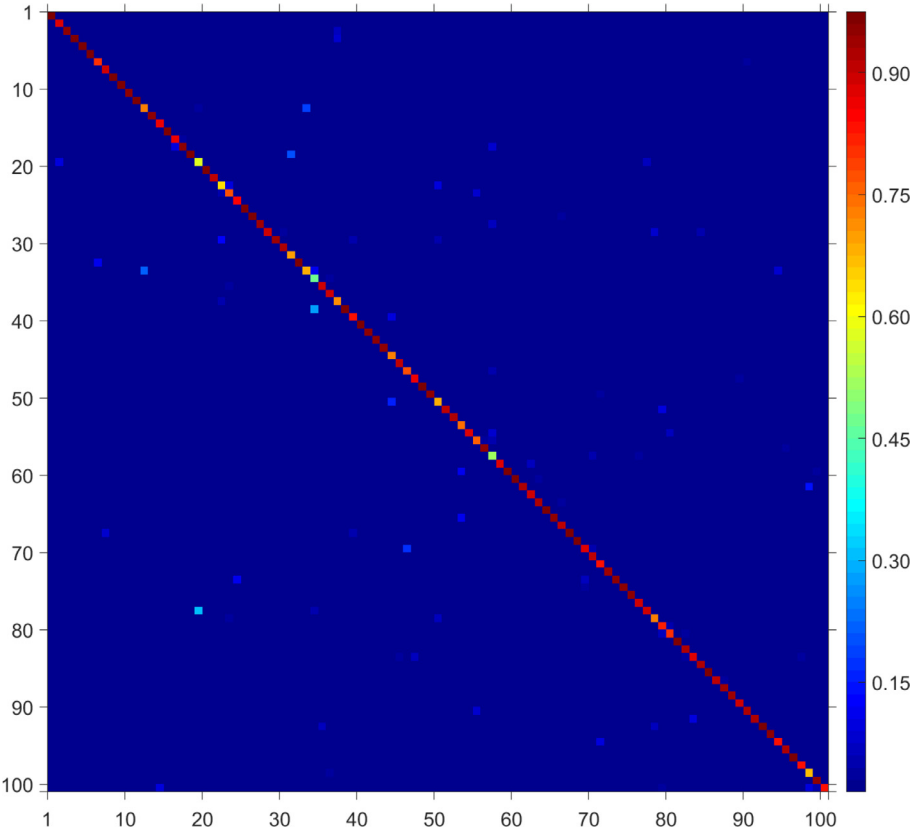

Fig. 4. Performance confusion matrix for our method on the UCF101 dataset.

and $L = maxlength$ on UCF101. This is largely due to the fact that HMDB51 contain more unconstrained camera moving video.

Next, we evaluate the performance of each individual stream on both datasets. The short-term ConvNet gets the worst performance because no pre-trained model is available. Although global-term ConvNet also without pre-trained, it can capture more temporal information than the short-term ConvNet. We can also find that all streams outperform corresponding ConvNet architecture. The short-term stream is 3.6% better than short-term ConvNet. The lowest improvement is obtained in global-term stream. The reason may be that global-term clue contain long-term temporal dependencies inherently. All the remarkable improvements indicate that CNN-RNN is a better structure than the pure CNN. The performance of the spatial, short-term and global-term with LSTM are also provided. The results proved that GRU is better than LSTM, even with less parameters.

We evaluate the combinations of multiple networks to study whether fusion can compensate the limitations of a single stream in describing complex video data. The simple average fusion is adopted. Results are summarized in the bottom three groups of Table 1. We first assess the gain from integrating the spatial and the short-term motion information. On UCF101, significant improvements (about 5.6% for spatial stream and 7.4% for short-term stream) are observed over the best single stream results. The gain on HMDB51 is consistent and as significant as that on UCF101. The most gain obtained by the combination with the short-term stream and the global-term stream both on UCF101 and HMDB51, indicating that the motion information is more critical for human action analysis.

Finally, we training the whole multi-stream networks jointly and the accuracy achieves 93.3% and 67.8% on UCF101 and HMDB51. Thus, we can conclude that the spatial, short-term and global-term streams are complementary to each other. And the result proves that their complementary properties can be utilized to improve the overall recognition performance.

Table 2
Comparison with state-of-the-art results.

Method	UCF101	Method	HMDB51
Donahue et al. [60]	82.9%	Husain et al. [53]	53.9%
Trans et al. [75]	86.7%	Wang et al. [9]	57.2%
Husain et al. [53]	86.7%	Wang et al. [76]	59.4%
Simonyan et al. [21]	88.0%	Simonyan et al. [21]	59.4%
Ng et al. [51]	88.6%	Shi et al. [26]	65.2%
Wu et al. [55]	92.6%	Lev. et al. [74]	67.7%
Lev. et al. [74]	**94.0%**		
Ours	**93.3%**	Ours	**67.8%**

The confusion matrixes for our multi-stream approach on UCF101 and HMDB51 datasets are as shown in Figs. 4 and 5. On the UCF101 dataset, our method performs perfectly on many categories such as Pizza Tossing. However, the confusion matrix on the HMDB51 dataset shows that some categories are easily misclassified, despite our method still performs well on most categories.

4.3. Comparison with state of the arts

We compare our approach with the state of the arts on both datasets. Results are listed in Table 2. Our proposed multi-stream approach achieves the competitive performance on both datasets, some examples are as shown in Fig. 6. On UCF-101, many works with competitive results are based on the hand-engineered dense trajectory features, while our approach fully relies on the deep networks. Compared with the original result of the two-stream approach [21], our approach captures a more comprehensive set of useful clues with a more effective long-term dependencies strat-

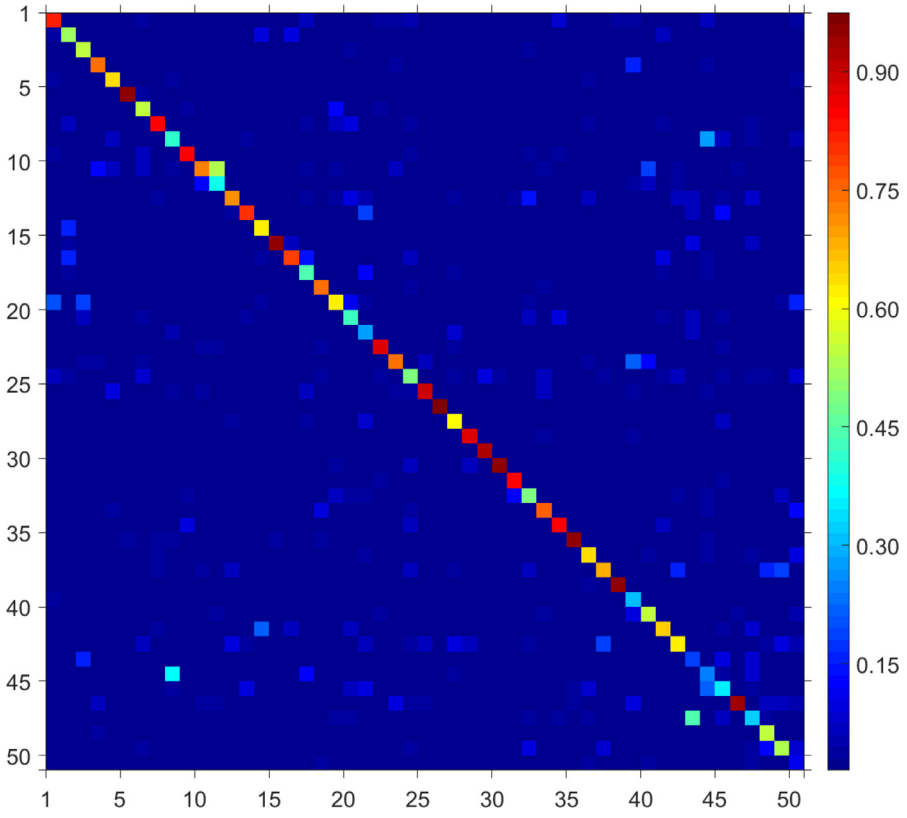

Fig. 5. Performance confusion matrix for our method on the HMDB51 dataset.

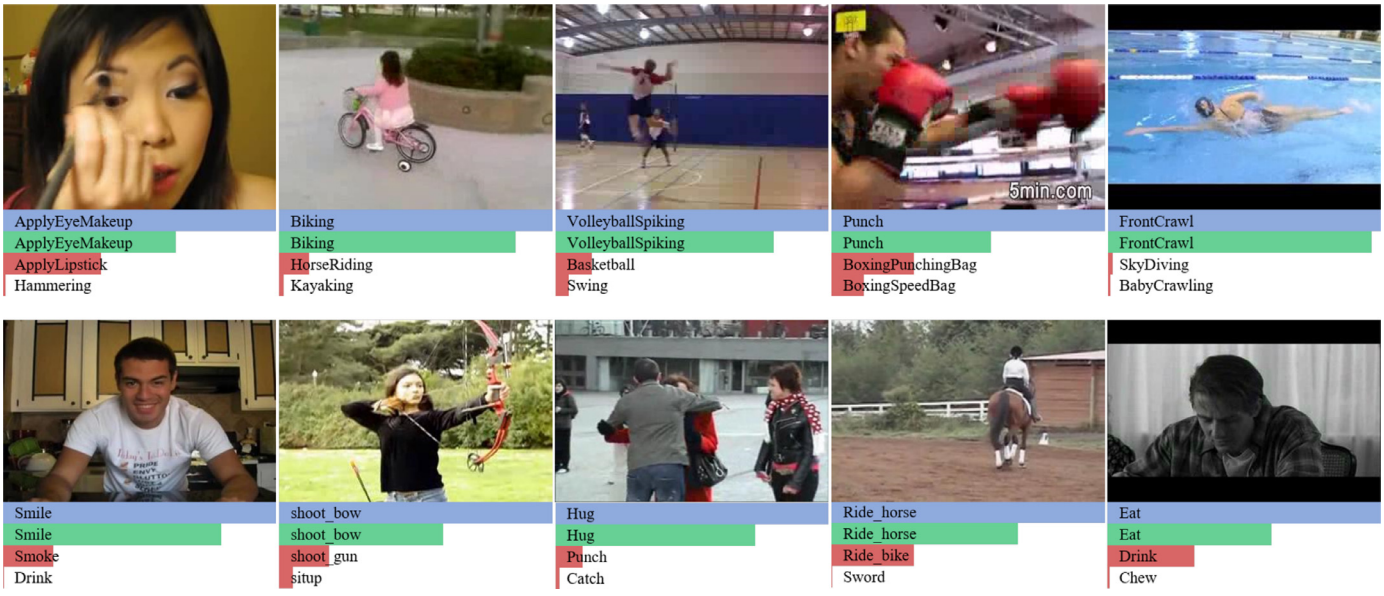

Fig. 6. Human action classification results on UCF101(upper row) and HMDB51(below row). Blue indicates ground truth label and the bars below show model classification results sorted in decreasing confidence. Green and red distinguish correct and incorrect results, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

egy. Note that a gain of even just 1% on the widely adopted UCF-101 dataset is generally considered as a significant progress. In addition, the recent works in [51,60] also adopted the LSTM to model the temporal clues for video classification and reported promising performance, but did not explore the global-term stream and employ GRU model. Zha et al. [74] using Fisher Vectors achieve the best results.

On the HMDB51 dataset, all the recent approaches were developed based on multiple features, either the hand-engineered descriptors or the ConvNet-based representations. Our approach produces better results than all of them.

5. Conclusion

This paper has proposed an effective descriptor for global-term motion of actions. We construct the action video as an three-order tensor, low-rank tensor decomposition is then applied to obtain the sparse component which preserved partial motion of target

at full temporal extent. A multi-stream framework is then employed to identify actions from a video sequence. Followed by uniformly considering spatial, short-term motion and global-term motion clues, GRU were introduced to model the long-term temporal dependencies for all clues. Our method achieves state-of-the-art performance on the HMDB51 dataset and outperforms most of existing methods on the UCF101 dataset. In the future, one promising direction is to pre-train the short-term motion and the global-term motion stream using large video datasets, which may improve the results significantly.

Acknowledgment

The work described in this paper is supported by National Natural Science Foundation of China (61473277).

References

[1] Y. Guo, L. Li, W. Liu, J. Cheng, D. Tao, Multiview cauchy estimator feature embedding for depth and inertial sensor-based human action recognition, IEEE Trans. Syst. Man Cybern. PP (99) (2016) 1–11.
[2] W. Liu, D. Tao, Multiview Hessian regularization for image annotation, IEEE Trans. Image Process. 22 (7) (2013) 2676–2687.
[3] L. Chen, P. Huang, J. Cai, Z. Meng, Z. Liu, A non-cooperative target grasping position prediction model for tethered space robot, Aerosp. Sci. Technol. 58 (2016) 571–581.
[4] J. Cai, P. Huang, B. Zhang, D. Wang, A TSR visual servoing system based on a novel dynamic template matching method, Sensors 15 (12) (2015) 32152–32167.
[5] A. Meltzoff, WolfgangPrinz, The Imitative Mind, Cambridge University Press, 2002.
[6] W. Bian, D. Tao, Y. Rui, Cross-domain human action recognition., IEEE Trans. Syst. Man Cybern. 42 (2) (2012) 298–307.
[7] Z. Zhang, D. Tao, Slow feature analysis for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 436–450.
[8] J. Miao, X. Xu, S. Qiu, C. Qing, D. Tao, Temporal variance analysis for action recognition, IEEE Trans. Image Process. 24 (12) (2015) 5904–5915.
[9] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision(ICCV), 2013, pp. 3551–3558.
[10] W. Liu, H. Liu, D. Tao, Y. Wang, K. Lu, Multiview Hessian regularized logistic regression for action recognition, Signal Process. 110 (2014) 101–107.
[11] D. Tao, L. Jin, W. Liu, X. Li, Hessian regularized support vector machines for mobile image annotation on the cloud, IEEE Trans. Multimedia 15 (4) (2013) 833–844.
[12] C. Hong, J. Yu, J. You, X. Chen, Hypergraph regularized autoencoder for 3D human pose recovery, Signal Process. 124 (2015) 132–140.
[13] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
[14] Y. Luo, D. Tao, Y. Wen, R. Kotagiri, C. Xu, Tensor canonical correlation analysis for multi-view dimension reduction, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3111–3124.
[15] D. Tao, X. Li, X. Wu, W. Hu, S.J. Maybank, Supervised tensor learning, Knowl. Inf. Syst. 13 (1) (2007a) 1–42.
[16] D. Tao, X. Li, X. Wu, S.J. Maybank, General tensor discriminant analysis and gabor features for gait recognition., IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007b) 1700–1715.
[17] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR), 2005, pp. 886–893.
[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[19] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, Eprint Arxiv (2016).
[20] S.Z. Su, Z.H. Liu, S.P. Xu, S.Z. Li, R. Ji, Sparse auto-encoder based feature learning for human body detection in depth image, Signal Process. 112 (C) (2015) 43–52.
[21] D. Annane, J.C. Chevrolet, S. Chevret, J.C. Rapha?l, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 1 (4) (2014) 568–576.
[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2014, pp. 1725–1732.
[23] T. Du, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: IEEE Conference on Computer Vision(ICCV), 2015, pp. 4489–4497.
[24] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inf. Process. Syst. 4 (2014) 3104–3112.
[25] W. Zaremba, I. Sutskever, Learning to execute, Eprint Arxiv (2014).
[26] Y. Shi, Y. Tian, Y. Wang, T. Huang, Sequential deep trajectory descriptor for action recognition with three-stream CNN, IEEE Trans. Multimedia PP (99) (2017) 1–11.
[27] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: a survey, Image Vis. Comput. 60 (2017) 4–21.
[28] R. Poppe, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990.
[29] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2) (2005) 107–123.
[30] J. Fan, Z. Zha, X. Tian, Action recognition with novel high-level pose features, in: IEEE International Conference on Multimedia and Expo Workshops, 2016, pp. 1–6.
[31] W. Liu, Z. Wang, D. Tao, J. Yu, Hessian Regularized Sparse Coding for Human Action Recognition, Springer International Publishing, 2015.
[32] X. Yang, W. Liu, D. Tao, J. Cheng, Canonical correlation analysis networks for two-view image recognition, Inf. Sci. 385 (2017) 338–352.
[33] D. Tao, X. Li, X. Wu, S.J. Maybank, General tensor discriminant analysis and gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1700.
[34] P. Huang, Y. Xu, Svm-based learning control of space robots in capturing operation, Int. J. Neural Syst. 17 (6) (2007) 467–477.
[35] S. Cai, S. Wu, G. Bao, Cylinder position servo control based on fuzzy PID, J. Appl. Math. 2013 (2013) 1–10.
[36] B. Krausz, C. Bauckhage, Action recognition in videos using nonnegative tensor factorization, in: International Conference on Pattern Recognition(ICPR), 2010, pp. 1763–1766.
[37] J. Zhang, Y. Han, J. Jiang, Tucker decomposition-based tensor learning for human action recognition, Multimedia Syst. 22 (3) (2016) 343–353.
[38] Y. Su, H. Wang, P. Jing, C. Xu, A spatial-temporal iterative tensor decomposition technique for action and gesture recognition, Multimedia Tools Appl. (2015) 1–18.
[39] C. Jia, M. Shao, Y. Fu, Sparse canonical temporal alignment with deep tensor decomposition for action recognition, IEEE Trans. Image Process. PP (99) (2016) 1.
[40] C. Jia, Y. Fu, Low-rank tensor subspace learning for RGB-D action recognition, IEEE Trans. Image Process. 25 (10) (2016) 4641–4652.
[41] D. Tao, X. Li, X. Wu, S.J. Maybank, Geometric mean for subspace selection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2008) 260–274.
[42] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (5) (2014) 2019–2032.
[43] W. Liu, Z.J. Zha, Y. Wang, K. Lu, D. Tao, *p*-Laplacian regularized sparse coding for human activity recognition, IEEE Trans. Ind. Electron. 63 (8) (2016) 5120–5129.
[44] L. Chen, B. Zhang, P. Huang, Z. Liu, Z. Meng, Autonomous rendezvous and docking with nonfull field of view for tethered space robot, Int. J. Aerosp. Eng. 2017 (2017) 1–11.
[45] Cand, E.J. S, X. Li, Y. Ma, J. Wright, Robust principal component analysis, J. ACM 58 (3) (2009) 11–20.
[46] D. Goldfarb, Z. Qin, Robust low-rank tensor recovery: models and algorithms, SIAM J. Matrix Anal. Appl. 35 (1) (2013) 225–253.
[47] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations (ICLR) (2015) 1–13.
[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp. 1–9.
[49] J. Yu, B. Zhang, Z. Kuang, D. Lin, J. Fan, Iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, IEEE Trans. Inf. Forensics Secur. PP (99) (2017) 1.
[50] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.
[51] Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015, pp. 4694–4702.
[52] M.A. Goodale, A.D. Milner, Separate visual pathways for perception and action, Trends Neurosci. 15 (1) (1992) 20.
[53] F. Husain, B. Dellen, C. Torras, Action recognition based on efficient deep feature learning in the spatio-temporal domain, IEEE Rob. Autom. Lett. 1 (2) (2016) 984–991.
[54] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016, pp. 1933–1941.
[55] Z. Wu, Y.G. Jiang, X. Wang, H. Ye, X. Xue, Multi-stream multi-class fusion of deep networks for video classification, ACM on Multimedia Conference (2016) 791–800.
[56] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659.
[57] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking., IEEE Trans. Cybern. PP (99) (2016) 1–11.
[58] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Ind. Electron. 62 (6) (2015) 3742–3751.
[59] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in: International Conference on Human Behavior Unterstanding, 2011, pp. 29–39.
[60] J. Donahue, L.A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual

recognition and description, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2016) 2625–2634.
[61] A.J. Robinson, F. Failside, Static and dynamic error propagation networks with application to speech coding., in: Neural Information Processing Systems(NIPS), 1987, pp. 632–641.
[62] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[63] R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in: International Conference on Machine Learning(ICML), 2015, pp. 2342–2350.
[64] K. Greff, R.K. Srivastava, J. Koutnik, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space odyssey, IEEE Transactions on Neural Networks and Learning Systems PP (99) (2016) 1–18.
[65] K. Cho, B.V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, Eprint Arxiv (2014).
[66] L.D. Lathauwer, B.D. Moor, J. Vandewalle, On the best rank-1 and rank-(r1, r2,...,rn) approximation of higher-order tensor, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1324–1342.
[67] Y. Li, J. Yan, Y. Zhou, J. Yang, Optimum subspace learning and error correction for tensors, in: European Conference on Computer Vision(ECCV), 2010, pp. 790–803.
[68] I.R. Years, Human action recognition based on fusion features extraction of adaptive background subtraction and optical flow model, Math. Probl. Eng. 2015 (4) (2015).
[69] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166.
[70] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2009, pp. 248–255.
[71] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: European Conference on Computer Vision(ECCV), 2004, pp. 25–36.
[72] K. Soomro, A.R. Zamir, M. Shah, UCF101: a dataset of 101 human actions classes from videos in the wild, Eprint Arxiv (2012).
[73] H. Kuehne, H. Jhuang, R. Stiefelhagen, T. Serre, HMDB51: a large video database for human motion recognition, in: High Performance Computing in Science and Engineering, 2012, pp. 571–582.
[74] G. Lev, G. Sadeh, B. Klein, L. Wolf, RNN fisher vectors for action recognition and image annotation, in: European Conference on Computer Vision(ECCV), 2016, pp. 833–850.
[75] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, IEEE International Conference on Computer Vision (ICCV) (2015) 4489–4497.
[76] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Val Gool, Temporal segment networks: towards good practices for deep action recognition, in: European Conference on Computer Vision(ECCV), 2016.