Action Recognition using Spatial-Optical Data Organization and Sequential Learning Framework

Yuan Yuan[a], Yang Zhao[a,b], Qi Wang[c,*]

[a] *Center for OPTical IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics of CAS, Xi'an 710119, China*
[b] *University of Chinese Academy of Sciences, Beijing 100049, China*
[c] *School of Computer Science, Center for OPTical IMagery Analysis and Learning (OPTIMAL), and Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China*

Abstract

Recognizing human actions in videos is a challenging problem owning to complex motion appearance, various backgrounds and semantic gap between low-level features and high-level semantics. Existing methods have scored some achievements and many new thoughts have been proposed for action recognition. They focus on designing a robust feature description and training an elaborate learning model, and many of them can benefit from a two-stream network with a stack of RGB frames and optical flow frames. However, these features for human action representation are struggling with the limited feature representation as RGB videos are confused by static appearance redundancy and optical flow videos cannot represent the detailed appearance. To solve these problems, we propose an efficient algorithm based on the spatial-optical data organization and the sequential learning framework. There are two contributions of our method: a novel data organization based on hierarchical weighting segmentation and optical flow for video representation, and a lightweight deep learning model based on the Convolutional 3D (C3D) network and the Recurrent Neural Network (RNN) for complicated action recognition. The new data organization aggregates the merits of motion appearance, movement trajectories and optical flow in a creative way to highlight the meaningful information. And the proposed lightweight model has an insight

[*] Corresponding author

Email addresses: `y.yuan1.ieee@gmail.com` (Yuan Yuan), `zhaoyang.opt@gmail.com` (Yang Zhao), `crabwq@gmail.com` (Qi Wang)

into patterns and semantics of sequential data by low-level spatiotemporal feature extraction and high-level information mining. The proposed method is evaluated on the state-of-the-art dataset and the results demonstrate that our method have a good performance for complex human action recognition.

Keywords: action recognition, spatiotemporal feature, deep learning, sequential learning framework

1. Introduction

Recognizing human action and interaction [1][2] in videos is a hot topic in computer vision as it has a wide range of applications in both academia and industry such as video surveillance [3][4], robot vision [5] and intelligent transportation systems [6][7][8]. Many algorithms [9][10][11][12][13] have been carried out for action recognition but it is still a tough problem as videos have more spatiotemporal information and sequential relationships than still images. What's more, various human appearances, complicated backgrounds and a wide diversity of intra-class variations [14][15][16] make it more difficult to generate a general learning framework for various human actions.

Motivated by image processing, many researches expect to extract a discriminative feature containing both action appearance and motion information. And existing features can be divided into two types, hand-crafted features [17][18][19] and deep-learned features [20][21]. The former features have various feature extraction frameworks and most of them are extended from image processing techniques by statistics and geometry. Some traditional features such as Scale-Invariant Feature Transform (SIFT) [22], Speeded Up Robust Feature (SURF) [23], Histogram of Oriented Gradient (HOG) [24], Histograms of Oriented Optical Flow (HOF) [25] of this type have been applied to various fields in both research and industry. However, these features have an inherent limitation which are not capable of extracting a discriminative spatiotemporal feature in long term videos [18]. Consequently, some other work try to find a new thought to represent the spatiotemporal features in terms of trajectory tracking and motion appearance. Improved Dense Trajectory (IDT) proposed by Wang *et* *al.* [26] stands out among many others as they skilfully take the advantages of motion trajectories and local hand-crafted features. And it outperforms the other work in two popular datasets, UCF101 [16] and HMDB51 [15].

IDT has achieves a good performance in human action recognition, but the hand-crafted features still have many drawbacks such as a limited discriminative capacity and high parameter sensibilities [27]. For tackling these problems, recent researches turn their attention to deep-learned features since they have a strong ability to automatically discovery the high-level semantics in videos with a hierarchical learning framework [28][29][30][31][32]. And extensive experiments demonstrate that the deep-learned features outperform the hand-crafted features for complex and enormous data. What's more, the deep-learned features can simplify the procedure involved in designing hand-crafted features and leave most work to the framework itself. On the other hand, sequential information is significant for action recognition as complex actions are commonly combined by several related acts, and it's extremely difficult to recognize human actions without high-level semantic representation. For this reason, RNN with Long-Short Term Memory (LSTM) neuron [33][34][35], is purposed for the temporal feature representation of sequential data such as videos and audio.

1.1. The Limitation of Existing Algorithms

Although the recent researches for human action recognition have achieved a great success, the existing methods are still struggling with two contradictions as follows:

Static appearance redundancy & **Structural information missing.** RGB data have too much redundancy between consecutive frames since two adjacent frames are very much alike in static appearance. The static appearance redundancy disturbs the representation of motion appearance and structural information, which are also significant clues for action representation. In contrast with the rich static appearance in RGB data, structural information is always in an inferior position as it's difficult to figure out the spatial position of different objects and complicated backgrounds. However, structural information is in fact extremely important for video understanding for human vision. Accordingly, it's a tough problem to balance the static appearance, the motion appearance and the structural information in a spatiotemporal feature representation.

Massive learning framework & **Overfitting problem.** Semantic gap [36] between low-level vision features and high-level semantics is an obstinate obstacle to video understanding. Extensive experiments demonstrate that it is really difficult to have an insight into patterns and semantics among the spatiotemporal features. Many algorithms try to solve this problem by complicating learning frameworks and designing elaborate models. However, the training process of massive framework can easily result in the overfitting problem [37] and additional pre-trained models have to be introduced to improve the generalization ability. Therefore, it is important to deal with the contradiction between the massive learning framework and the overfitting problem.

1.2. Proposed Framework

Our motivation is to simulate human vision system which recognizes human actions from the view of static appearance, optical flow and spatial depth information. Various feature representations for static and motion appearance have been proposed but there are few method at present which considers the spatial depth information to enhance the feature representation of human actions. Obviously, it is difficult to directly extract spatial depth information from a single RGB frame. Fortunately, the videos contain many meaningful information among consecutive frames, and we find that the dynamic objects have richer motion information than static backgrounds, which means that the corresponding motion trajectories around the dynamic objects are more dense than the static backgrounds. And if we segment the videos into several supervoxels, the regions with dynamic objects will have more motion trajectories than the regions with static backgrounds. In this way, we can give a greater weight to these regions with more motion trajectories and it also can be considered as the spatial depth information synthesized by motion trajectories and video segmentation. In addition, the consecutive frames are always similar, and they have too much static appearance redundancy which disturbs motion representation. And it is of vital importance to synthesize a new data to highlight the motion appearance and reduce the static appearance redundancy. On the other hand, human actions are commonly combined by several related acts and it's unreasonable to recognize a complex action without analyzing the correlation among these related acts. From this view, we expect to build a bottom-up framework including low-level spatiotemporal feature extraction and high-level information mining. The low-level spatiotemporal feature extraction is specific for a single act in short video snippets. And the high-level information mining focus on figuring out pattern and semantics among the related acts.

In summary, we propose two solutions, which are also the highlights of our work, from the view of spatial-optical data organization and sequential learning framework as follows:

Spatial-optical data organization To develop a discriminative spatiotemporal feature for action recognition, we propose a novel data organization which is a creative thought to eliminate the static appearance redundancy, enhance the spatial hierarchical information and highlight the motion appearance by introducing video segmentation, motion trajectories and optical flow.

Sequential learning framework Different from the existing methods, we propose a bottom-up sequential learning framework to overcome the semantic gap by two stage: low-level spatiotemporal feature extraction and high-level information mining. The low-level spatiotemporal features are extracted from both spatial and temporal dimension by using a two-stream deep learning model with a lightweight learning framework. And then, an efficient model for sequential information mining is introduced to explore patterns and semantics among the related acts of consecutive video snippets.

Figure 1 shows the detailed framework of our proposed method. First, we segment videos into several supervoxels and associate them with motion trajectories to enhance spatial depth information. And then we introduce optical flow to synthesize the spatial-optical data according to the spatial-optical data organization. After the above preprocessing, the spatial-optical data, along with the RGB data, are input into a sequential learning framework from the low-level spatiotemporal feature extraction to the high-level sequential information mining. The main idea of the low-level spatiotemporal feature extraction is to extract the discriminative features of single acts in each video snippet by a lightweight deep learning model. At last, a RNN based on two stacked LSTM layers is introduced for mining pattern and semantics among the single acts of consecutive video snippets.

The rest of this paper is organized as follows. We review the related works from the

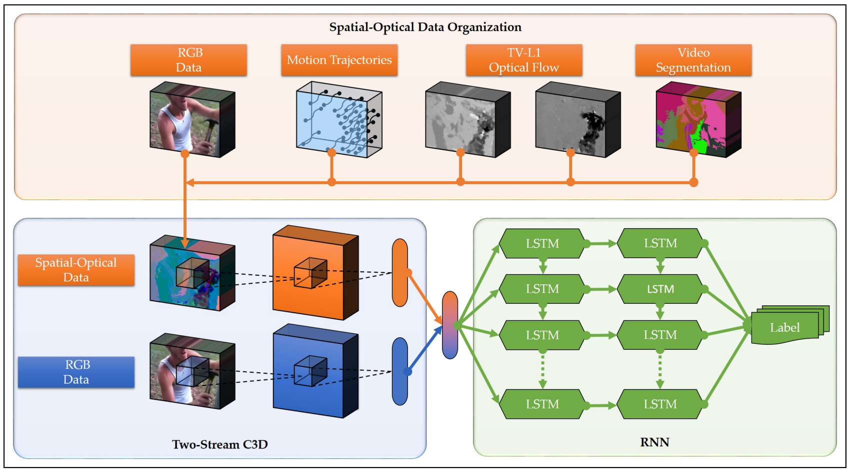

Figure 1: Overview of our approach. First, we synthesize a new data based on motion trajectories, optical flow, and video segmentation according to the spatial-optical data organization. And then a two-stream C3D network is introduced to extract the spatiotemporal features of single acts in spatial-optical data and RGB data respectively. At last, a RNN model based on two stacked LSTM layers is presented for mining patterns and semantics among the spatiotemporal features of single acts

views of the hand-crafted features and the deep-learned features in section 2. Section 3 introduces the spatial-optical data organization and the sequential learning framework. Experiments are presented in section 4 and conclusions are made in section 5.

2. Related Work

The researches of action recognition have achieved remarkable success in using the hand-crafted features and the deep-learned features. The hand-crafted features make use of various feature extraction methods such as optical flow, orientation histograms, frequency domain transformation and sparse representation. On the other hand, deep-learned features are commonly extracted by hierarchical learning architectures. And the recent experiments demonstrate that the deep-learned features outperform the hand-crafted features in many fields of computer vision.

2.1. Hand-crafted Features

Recently, many hand-crafted feature descriptors have been proposed and show a good performance in terms of efficiency and accuracy. For instance, local feature descriptors such as SIFT and SURF have been proven to be effective for image processing as these features are robust to illumination changes and background clusters. However, most of them ignored the temporal information, and therefore many researches pay more attention to the spatiotemporal information extraction by extending the image processing techniques to videos. For instance, by extending Harris detector to videos, Laptev proposed space-time interest points to describe the human actions. And 3DSIFT [22] extended the SIFT features into temporal dimension for a better space-time feature representation. However, these features only consider the local details of the human appearance, and they ignored the temporal information among consecutive frames.

In the meanwhile, global video descriptors based on optical flow have been successful in encoding both static appearance and motion information. These approaches are able to take the advantage of both spatial and temporal information by the hand-crafted features such as HOG, HOF, Motion Boundary Histogram (MBH) and spatio-temporal interest points (STIP), in a dense grid or around dense trajectories. These hand-crafted features are then encoded in order to generate a global video-level descriptor through bag of words (BoW) or fisher vector (FV) for further classification.

The representative work of hand-crafted features is IDT proposed by Wang *et al.* [18]. It is said that dense sampling feature points outperform sparse interest points based on detection algorithm in recent evaluations. The authors densely sample a set of feature points and eliminate the points in homogeneous image areas as these areas have no worth to track. And then the selected points are tracked with the help of median filters. Note that both static trajectories and suddenly large displacement trajectories are ignored for getting a robust feature extraction. And HOG, HOF and MBH descriptors are computed along the dense trajectories for a robust spatiotemporal feature representation. A bag-of-features framework or support vector machine approach are applied to evaluate these features at last.

2.2. Deep-learned Features

However, the hand-crafted features have limited discrimination for various datasets because they are limited by their algorithm framework and parameter settings. At the present time, the rise of deep learning techniques provides many new train of thoughts for action recognition. The deep learning techniques have boosted classification precision and discriminative capability by means of establishing models from large numbers of images.

Convolutional neural network, as the representative work of the deep learning techniques, has widely applied for human action recognition. Some researches focus on improving the traditional CNN in order to make it appropriate for video representation. For example, Two-stream ConvNet architecture proposed by Simonyan and Zisserman [38] is based on two separate ConvNets which are specific for spatial and temporal feature extraction. Laptev *et al.* [29] propose a pose-based CNN feature which aggregates static and motion appearance along human body parts. In the meanwhile, 3-dimension convolutional kernel also provides a remarkable improvement for video representation. Ji *et al.* [21] propose a 3D convolutional neural network (3D CNN) by performing 3D convolutional operation to extract spatiotemporal feature. Different from the 3D CNN, which uses a human detector and head tracking to segment human subjects in videos, C3D Network proposed by Tran *et al.* [27], which is similar to our proposed method, directly input the whole video frames without any preprocessing to extract global features in the long term videos. However, they treat the IDT and C3D as two individuals and do not consider the sequential information among consecutive video snippets. Yue-Hei Ng *et al.* [39] introduce several tricks to aggregate the CNN features of the long term videos with the help of feature pooling and this method also achieves a good result on action recognition. However, the above methods only consider the spatiotemporal features but ignore the high-level sequential information in videos and the results of the above methods are just marginally better than the baseline that extracts the spatial features in a single frame. Temporal Segment Networks (TSN) proposed by Wang *et al.* [30] combines a sparse temporal sampling strategy and video-level supervision to enable efficient and effective learning using the whole action video and Two-Stream Inflated 3D Con-vNet (I3D) [31] making it possible to learn seamless spatiotemporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters.

On the other hand, another successful deep learning technique for the long video representation is RNN, which are designed for sequential feature learning specifically. And recently, the combination of CNN and RNN such as Long term Recurrent Convolutional Neural Networks (LRCN), proposed by Donahue, *et al.* [40] performs well in dealing with action recognition. They proposed a hierarchical framework including a five-layer CNN and a two-layer RNN to extract spatiotemporal features and sequential information for complex actions. Similarly, Yue-Hei Ng *et al.* [39] employ a recurrent neural network takes input the output from the final CNN layer at each consecutive video frame, where the CNN outputs are processed forward through time and upwards through five layers of stacked LSTMs. It achieves a good performance as it discovers the long-range temporal relationships over long periods of a video (120 frames). Srivastava *et al.* [41] proposed an unsupervised learning for video representations by using a high-level representations based on a pretrained CNN model and a RNN-LSTM encoder-decoder framework. Ma, *et al.* [35] propose TS-LSTM which follows the intuition of the temporal segment by Wang *et al.* [30] and divides the sampled video frames into several segments, and an LSTM layer is used to extract the embedded features from all segments. Liu, *et al.* [42] and Mahasseni *et al.* [43] introduce spatiotemporal LSTM network to skeleton-based action recognition as LSTM is effective in dealing with the sequential information of human action.

Different from the above methods, we propose a new data organization, spatial-optical data organization, to highlight the discriminative spatiotemporal features in a single video snippet. It is a creative thought to eliminate the static appearance redundancy, enhance the spatial hierarchical information and highlight the motion appearance for human action representation. And for further processing, we input the spatial-optical data into a bottom-top sequential learning framework to explore the high-level information among the video snippets. The following two sections will introduce the proposed spatial-optical data organization and sequential learning framework in details.

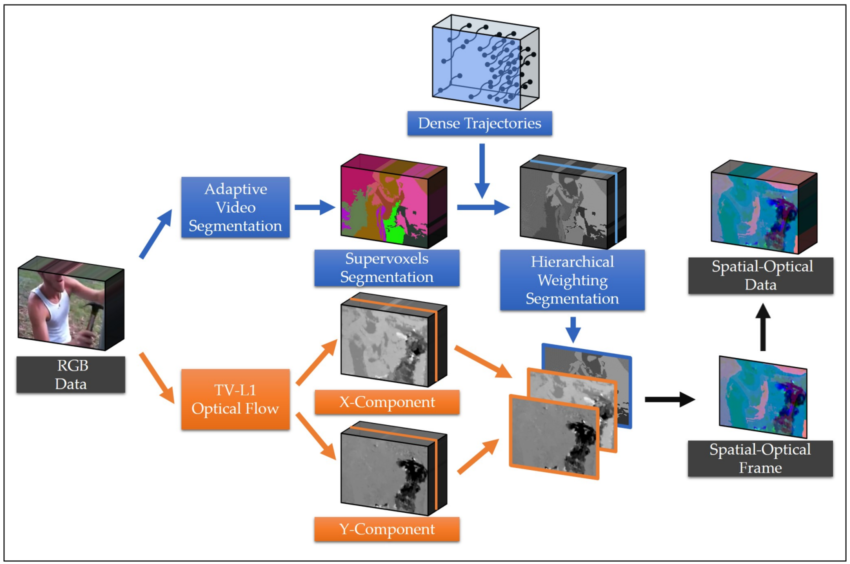

Figure 2: Overview of the spatial-optical data organization. First, we obtain hierarchical weighting segmentation by associating supervoxels segmentation with motion trajectories. And then, the X-component and Y-component of TV-L1 optical flow are extracted from RGB data for motion appearance. At last, we reorganize hierarchical weighting segmentation and TV-L1 optical flow according to spatial-optical data organization.

3. Spatial-optical Data Organization

In this section, a novel data organization is presented by associating hierarchical weighting segmentation with optical flow to eliminate the static appearance redundancy, enhance the spatial hierarchical information and highlight the motion appearance. The details are presented in the following three parts: adaptive video segmentation, hierarchical weighting segmentation and spatial-optical data synthesis. The details are presented in Figure 2.

3.1. Adaptive Video Segmentation

Motivated by image segmentation, which has been widely used to distinguish between foreground and background in a still images [44], video segmentation is designed to segment the videos into serval parts by exploring the spatiotemporal information. Compared with the image segmentation, the video segmentation is more robust to distinguish the dynamic object from the static background since it considers motion appearance between consecutive frames. A representative work, hierarchical graph-based video segmentations proposed by Grundmann *et al.* [45] has achieved a good performance. It builds upon on Felzenszwalb and Huttenlocher's graphed based algorithm [46] for image segmentation. Xu *et al.* [47] proposed a principled steaming approximation for the hierarchical video segmentation which outperforms the others. Motivated by the above methods, we propose an adaptive video segmentation by constraining both region size and segmentation scale to improve the video segmentation.

According to [47], a long video can be considered as several non-overlapping consecutive video snippets $\mathcal{V} = \{V_1, V_2, V_3, ...\}$, and the corresponding video segmentation are presented as $\mathcal{S} = \{S_1, S_2, S_3, ...\}$. Assuming that each segmentation S_i is only related to the corresponding video snippet V_i, the previous video snippet V_{i-1} and its previous video snippet segmentation S_{i-1}. As a result, a directed Bayesian-like network using Markov chain is introduced into the video segmentation as:

$$\begin{aligned} M(\mathcal{S}|\mathcal{V}) = M^1(S_1|V_1) + M^1(S_2|V_2, S_1, V_1) + ... \\ + M^1(S_m|V_m, S_{m-1}, V_{m-1}), \end{aligned} \tag{1}$$

where $M(\cdot|\cdot)$, $M^1(\cdot|\cdot)$ are the segmentation model for the whole video and the single video snippet respectively. The optimum segmentation can be considered as a minimization problem according to Eq. 1. However, the models do not provide an explicit energy function and it's difficult to find a global optimum directly. Assuming that the segmentation result S_i for the current subsequence V_i never influences segmentation results for previous subsequences, it only influences the next subsequence V_{i+1}. The above problem can be solved by a greedy solution as follows:

$$\begin{aligned} S^* &= \{S_1^*, ..., S_m^*\} \\ &= \arg\min_{S_1,...,S_m} \left\{ M^1(S_1|V_1) + \sum_{i=2}^{m} M^1(S_i|V_i, S_{i-1}, V_{i-1}) \right\}, \\ &= \left\{ \arg\min_{S_1} M^1(S_1|V_1) + \arg\min_{S_2} M^1(S_2|V_2, S_1, V_1) + \right. \\ &\quad \left. ... + \arg\min_{S_m} M^1(S_m|V_m, S_{m-1}, V_{m-1}) \right\}, \end{aligned} \tag{2}$$

Therefore, we can obtain S_1 based on V_1, then fix S_1 and obtain S_2 based on S_1, V_1, and V_2. Finally, we can get the global solution until the last S_m is completed. Note that, each hierarchical segmentation S_i is hierarchical segmentation of subsequence V_i, s.t. $S_i = \{S_i^1, S_i^2, ..., S_i^h\}$ where S_i^j is the j_{th} layer of h hierarchical segmentation layers. Continuing with a similar Markov assumption that S_i^j only depends on S_i^{j-1}, S_{i-1}^j, and S_{i-1}^{j-1} and the hierarchical segmentation S_i^* can be represented as follows:

$$\begin{aligned} S_i^* &= \arg\min_{S_i} M^1(S_i|V_i, S_{i-1}, V_{i-1}) \\ &= \arg\min_{S_i^2,...,S_i^h} \sum_{j=2}^{h} M^2(S_i^j|V_i, S_i^{j-1}, S_{i-1}^{j-1}, S_{i-1}^j, V_{i-1}), \end{aligned} \tag{3}$$

where $M^2(\cdot|\cdot)$ is the conditional segmentation model for each layer of the video snippets. And then we transform the minimization problem into the graph segmentation by introducing a graph $\mathbf{G} = \{\mathbf{N}, \mathbf{E}\}$ over the spatiotemporal video volume with a 26-neighborhood in 3D space-time for two consecutive video snippets V_{i-1} and V_i and the first layer of the edge weights are direct color similarity of voxels. According to Eq. 3, the hierarchical segmentation S_i can be inferred by the hierarchical segmentation result S_{i-1} layer by layer.

Specifically, each supervoxel in the $j-1_{th}$ layer S_i^{j-1} and S_{i-1}^{j-1} is a node $n \in \mathbf{N}$ represented by a histogram of Lab color of all voxels in the supervoxel and each edge weight $w(e \in \mathbf{E})$ is χ^2 distance of two histograms of the corresponding supervoxel. According to [46], a region R is defined as a subset of the whole connected nodes. When the edge weight of two neighbor regions is smaller than their internal variation, then these two regions are merged. The internal variation of a region is defined as $RInt(R) = Int(R) + \tau/|R|$, where $|R|$ the number of voxels in region R and $Int(R) = \max_{e \in MST(R)} w(e)$. $w(e)$ is the edge weight in the Minimum Spanning Tree(MST) of graph G and τ is a parameter to limit local region size. For a video, the final segmentation is a set of hierarchical video segmentation $\mathcal{S} = \{S^1, S^2, ...\}$ according to different parameter τ.

However, too many regions may disturb the line between foreground and background while too few regions may lose the details of motion appearance. Unfortunately, the constant parameter τ have limited abilities to maintain a proper segmentation scale as it is sensitive to various situation. So we introduce a new constraint λ from the view of appearance similarities to limit the global segmentation scale adaptively by $N(S^{(i-1)}) < \lambda < N(S^{i})$ where $N(\cdot)$ is the region number of the corresponding segmentation and S^{i} is the objective segmentation.

3.2. Hierarchical Weighting Segmentation

The existing video segmentation techniques only segment the videos and assign the segment regions with random values but they do not consider the relationship among motion appearance and spatial information. Motivated by Wang *et al.* [18], we present a new thought to associate the adaptive video segmentation with the motion trajectories for highlighting the hierarchical structure information of dynamic objects and static backgrounds.

As we know, the dynamic objects have more motion trajectories than the static backgrounds and we normalize the video segmentation according to their motion trajectory density. According to [18], we densely sample the tracking points on a grid spaced by W pixels and track these points $P_t = (x_t, y_t)$ at frame t by median filtering in a dense optical flow field $w = (u_t, v_t)$ as follows:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|_{(x_t, y_t)}, \tag{4}$$

where M is a median filter kernel. Moreover, we eliminate the trajectories with sudden large displacement to get rid of drifting and the points in homogeneous image areas which have no structure to track based on the criterion [48], that if the smaller eigenvalue of the points' autocorrelation matrix is too small the points are trivial. Then we can get a set of trajectories $T = \{t_1, t_2, ..., t_K\}$ for a video and each trajectory t_i can be represented as a set of tracking points $t_i = \{t_i^1, t_i^2, ..., t_i^L\}$ where L is the length of the trajectory t_i.

Intuitively, the region with higher trajectory density contains richer motion information, so we assign each region R_i, that belongs to the adaptive video segmentation S, according to its corresponding trajectories density D_i instead of random assignments:

$$\begin{aligned} D_i &= \frac{\sum_{k=1}^{K} \{t_k | t_k \in R_i\}}{|R_i|}, \\ S_{hw} &= \{R_1^*, R_2^*, ..., R_{N(S)}^*\} \\ &= \left\{ \frac{D_1}{\max{(D_i)}}, \frac{D_2}{\max{(D_i)}}, ..., \frac{D_{N(S)}}{\max{(D_i)}} \right\}, \end{aligned} \tag{5}$$

where $|R_i|$ is the number of voxels in region R_i. In this case, we associate the adaptive video segmentation with the motion trajectories and the regions with more dynamic objects will have a greater weight. Compare to the traditional video segmentation, the hierarchical weighting video segmentation S_{hw} highlights the motion appearance and enhances the spatial hierarchal information.

3.3. Spatial-optical Data Synthesis

The hierarchical weighting video segmentation is capable of providing the hierarchical structure information between dynamic objects and static backgrounds. However, it has limited abilities to describe the detailed motion appearance so we further introduce TV-L1 optical flow [49] into the proposed spatial-optical data organization.

In general, given two image consecutive frames I_0 and I_1, the main objective of optical flow is to find out the proper disparity maps U_x and U_y of vertical and horizontal direction to estimate the motion of pixels in two consecutive frames I_0 and $I_1 : (\Omega \subseteq \mathcal{R}^2)$. And the disparity map U_x and U_y of vertical and horizontal direction can be considered as the minimization of the sum of an image-based error criterion and a regularization force as follow:

$$\begin{aligned} E = \int \Big\{ &\mu | I_0(x, y) - I_1(x + U_x(x, y), y + U_y(x, y)) | \\ &+ |\nabla U_x| + |\nabla U_y| \Big\} d_\Omega, \end{aligned} \tag{6}$$

where μ weights between the data fidelity and the regularization force. After that, we combine the disparity maps U_x and U_y of vertical and horizontal direction with the proposed hierarchical weighting segmentation to synthesize a novel data organization. The proposed spatial-optical data organization generates the new synthetic data and

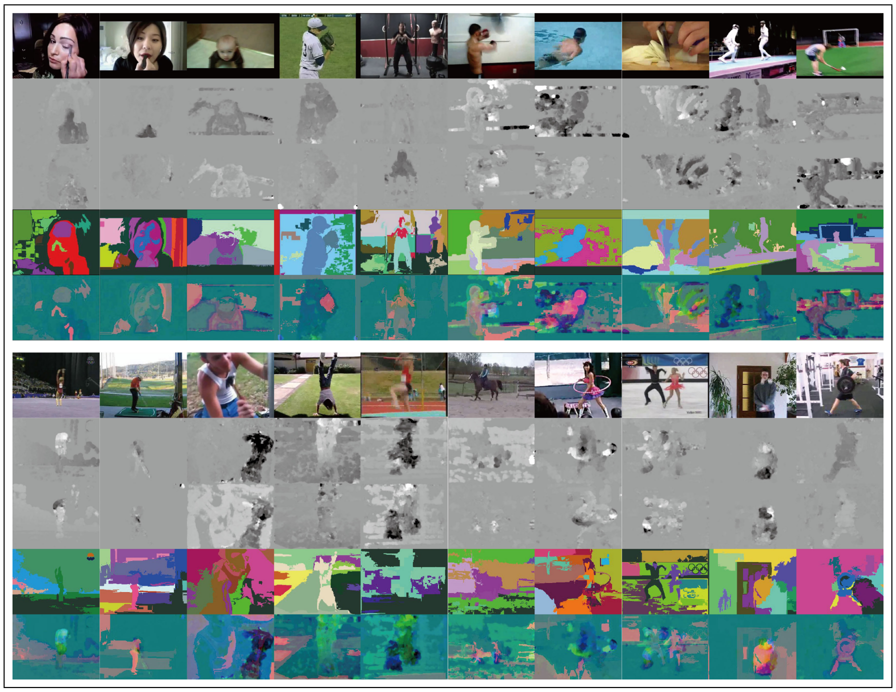

Figure 3: The first row is the frames of RGB data and the images in the second and the third rows are the corresponding components of optical flow. The images in the fourth row are from adaptive supervoxel segmentation and the last row presents the frames of spatial-optical data.

they are same as the RGB data format which has three channels. The first two channels are filled with the displacement maps U_x and U_y and the last channel is filled with the hierarchical weighting segmentation S_{hw}.

Figure 3 presents the sample frames of the RGB data, the optical flow, the adaptive video segmentation and the spatial-optical data. It is obvious that motion trajectories in the supervoxels with rich motion appearance are much denser than the static backgrounds. The motion appearance of the spatial-optical data is more salient than the RGB data and extensive experiments also show that the proposed spatial-optical data organization have a better performance for human action representation.

4. Sequential Learning Framework

Notwithstanding the spatial-optical data organization provides a robust representation to highlights the motion appearance and structural information, it's still a tough problem to extract a discriminative spatiotemporal feature in a long-term video. To solve this problem, we divide the long-term videos into several consecutive video snippets and introduce a two-stream C3D network to extract the spatiotemporal features of RGB data and spatial-optical data in each video snippet. Different from the existing methods, the proposed two-stream C3D network only forces on spatiotemporal features of each single act in short video snippets. Each spatiotemporal feature is a small representation of a single act and a complex action can be represented by the combination of several related spatiotemporal features.

On the other hand, semantic gap between low-level vision feature and high-level semantics is a major obstacle in computer vision. Accordingly, figuring out patterns and semantics among the related spatiotemporal features in the short video snippets is significant to solve the semantic gap in action recognition. Traditional methods based on shallow classifiers such as SVM only work for simply repetitive actions such as running, walking but it has limited abilities for complex actions such as weightlifting, high jump and javelin throwing as these actions have more complex patterns and semantics. Deep learning techniques have a better performance than shallow classifiers and recent deep learning techniques are in favor of making deeper and deeper models to get a better performance. However, most of them have to face a more serious overfitting problem of massive framework. To overcome the above methods, we introduce a RNN model with two stacked LSTM layers as it has achieved a good performance in processing sequential data such as speech. Different from the existing methods, this RNN model is only for pattern and semantics mining among spatiotemporal features. In this way, the proposed method can get a better performance for complex action and long-term videos. The details are given in the following parts.

4.1. Spatiotemporal Feature Learning

As we know, a discriminative spatiotemporal feature plays an important role for video representation. And the CNN based features have show their great performance in the recent works, but it only extracts the features from the spatial dimension. C3D Network, which is an improved CNN feature based on 3D convolutional kernel, is proposed to extract the spatiotemporal features of the video snippets. As a result, we introduce a two-stream C3D network to extract the spatiotemporal features, and the corresponding features can be considered as a space-time representation for the simple acts. In the following parts, we firstly introduce the 3D convolutional kernel and the related convolutional operation as the base. And then the global architecture of the two-stream C3D network will be presented in detail.

3D convolution kernel Traditional convolution kernel is 2-dimension of $k \times k$ which only convolutes the single image from spatial dimension. The details of 2-dimension convolutional operation is presented in Eq. 7 and v_{ij}^{xy} is the pixel value at position (x, y) of feature map v_{ij} which is the j_{th} feature map in the i_{th} layer:

$$v_{ij}^{xy} = f\left(\sum_{m}\sum_{p=0}^{k-1}\sum_{q=0}^{k-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} + b_{ij}\right), \tag{7}$$

where $f(\cdot|\cdot)$ is an activation function of sigmoid in general and b_{ij} is the bias for the current feature map. And m indexes over the set of feature maps in the $(i-1)_{th}$ layer connected to the current feature map. w_{ijm}^{pq} is the weight at the position (q, p) of kernel connected to the m_{th} feature map.

According to [27], the 3D convolutional kernel extends to temporal dimension by convoluting with a cube kernel. The value v_{ij}^{xyt} at the position (x, y, t) of feature cube can be presented as follow:

$$v_{ij}^{xyt} = f\left(\sum_{m}\sum_{p=0}^{k-1}\sum_{q=0}^{k-1}\sum_{r=0}^{d-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(t+r)} + b_{ij}\right), \tag{8}$$

where the size of kernel is $k \times k \times d$. With the help of cube kernels, the receptive field of the feature map is extended to temporal dimension. And each simple act in a single video snippet can be represented by the local spatiotemporal features from both spatial and temporal dimension.

Network Architecture The original RGB data have so much static appearance redundancy that it disturbs the motion appearance for video representation. To solve this problem, we introduce a more efficient data organization, the spatial-optical data organization, which enhances the spatial hierarchical information by associating the video segmentation with dense trajectories and highlight the motion appearance by the optical flow. However, it's difficult to directly integrate the spatial-optical data into the spatial-optical feature extraction framework. Motivated by [38], a two-stream C3D network is presented for the RGB data and the videos of the spatial-optical data organization respectively. And the corresponding 3D ConvNet features are fused in two ways for further sequential information mining.

Specifically, we firstly resize the video frames into 128×171 and split each video into non-overlapped 16-frame chips. And then these video snippets will be randomly cropped with a size of $3 \times 16 \times 112 \times 112$ as the input. The detailed architecture of C3D nerwork is presented in Table 1. The c3d1 is the first 3D convolution layer and the p3d1 is the first 3D pooling layer. We follow the settings proposed in [27] and fix all the convolutional kernels to $3 \times 3 \times 3$ and all the pooling layer are max pooling of which the cube kernel size is $2 \times 2 \times 2$. These kernels are capable of decreasing the size of feature maps from both temporal and spatial dimensions. The number of the feature maps which can be considered as the outputs of convolution layers are 64, 128, 256, 256, 256. At last, there are two full connection layers and one softmax layer for calculating the loss function and further classification. In this way, the 3D ConvNet feature extraction can provide the spatiotemporal information including both the static appearance from the RGB data and the motion appearance from the spatial-optical data for the simple acts in each video snippet.

Layer	c3d1	p3d1	c3d2	p3d2	c3d3	p3d3	c3d4	p3d4	c3d5	p3d5	fc6	fc7
kernel-size	3×3	2×2	3×3	2×2	3×3	2×2	3×3	2×2	3×3	2×2	-	-
kernel-depth	3	2	3	2	3	2	3	2	3	2	-	-
spatial-stride	1	2	1	2	1	2	1	2	1	2	-	-
temporal-stride	1	1	1	2	1	2	1	2	1	2	-	-
channel	64	64	128	128	256	256	256	256	256	256	2048	2048

Table 1: The architecture of C3D network with five-layer 3D ConvNet.

There are two types of features, F_{rgb} and F_{so}, extracted from RGB data and spatial-optical data, extracted from the $7th$ full connection layer of C3D network and we combine these features in two ways as the input of sequential information mining. The new synthetic features can be represented as $F_{sum} = \lambda F_{rgb} + (1 - \lambda) F_{so}$ and

$F_{con} = \{F_{rgb}; F_{so}\}$. We further analyze the above features in section 5 and the experiments demonstrate that the synthetic features outperform the state of the art.

4.2. Sequential Relationship Mining

After getting the low-level spatiotemporal features from the C3D feature extraction framework, we explore the correlations among the consecutive video snippets. Recently, shallow classifiers such as SVM has achieved a good result for images processing in many fields of computer vision. Some researches follow this pattern to treat each video snippet as an individual one by averaging scores across the whole videos to recognize human actions. However, the results are not very desirable because of ignoring the sequential correlation and the semantical information of single acts among consecutive video snippets. To solve this problem, we introduce deep learning techniques, namely RNN-LSTM, to explore the useful patterns and the high-level semantics among the low-level spatiotemporal features instead of the shallow classifiers as the RNN-LSTM is designed for sequential data mining specially. In the following parts, we firstly introduce the LSTM neuron as the base and then the details of the RNN-LSTM are presented.

The LSTM neuron Traditional RNN has its own drawback referred to in the literature as the vanishing gradient problem. It can be difficult to train them to learn long-range dependencies. To solve this problem, researchers proposed LSTM neuron which is capable of learning long-range temporal dependencies. As visualized in detail in Eq. 9, each LSTM unit maintains one memory cell c_t and three gates which are input gate i_t, output gate o_t and forget gate f_t. Figure 4 is the architecture of the LSTM neuron where the input is the output of the previous layer x_t and the current hidden state h_{t-1}, and the current output is z_t which is also the input x_{t+1} for next layer. Specifically, The input gate decides whether or not to consider the current input x_t and the forget gate allows to selectively forget its pervious memory c_{t-1}. Likewise, the output gate decides how much of the memory to transfer to the next hidden state z_t.

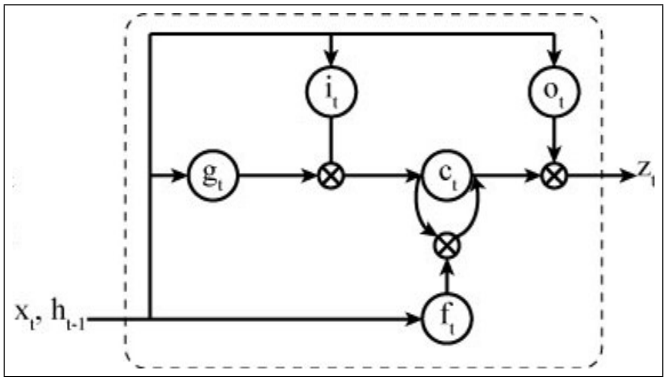

Figure 4: The structures of the LSTM neuron

The details can be present as follows:

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \\
g_t &= \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\
h_t &= o_t \odot \phi(c_t),
\end{aligned} \tag{9}$$

where i, f, o, c are corresponding to input gate, forget gate, output gate and cell activation vectors. σ is the sigmoid nonlinearity function, ϕ is the hyperbolic tangent nonlinearity function and $\odot$ donates element-wise multiplication.

Network Architecture Motivated by [50], which translates videos to natural language, the RNN is introduced to explore the sequential information among the single acts of consecutive video snippets. Specifically, we input the features F_{sum} and F_{con} into a RNN model with two stacked LSTM layers which follows the architecture setting proposed by Donahue *et al.* [40] as the two stacked LSTM layers has a better performance than one or four stacked LSTM layers. Each layer has 256 LSTM neurons and each LSTM neuron is related to the current input and the pervious hidden layer's state.

5. Experiments

We experiment with our proposed method on the state-of-the-art dataset, UCF101 and HMDB51, which are widely used for evaluating action recognition algorithm as they have adequate videos and various actions. The experiments focus on two points as follows:

Spatial-optical data organization We firstly compare the hand-crafted features such as SIFT, HOG and IDT with our method under the condition of traditional classifier SVM. It demonstrates that the spatial-optical data organization can provide a better feature representation beyond the state-of-the-art methods since it highlight the motion appearance and reduce the influence of the static appearance redundancy.

Sequential learning framework On the other hand, we experiment with our sequential learning framework which is designed for exploring sequential patterns and high-level semantical information among video snippets. The proposed framework extract the local space-time features of the simple acts in each video snippet and explore the high-level semantical information among consecutive video snippets. Compared with the existing methods which extract features from single frames, this framework can handle long term videos as well. And the experiments demonstrate that the sequential learning framework have achieved a remarkable improvement in recognizing complex human actions.

5.1. Datasets

UCF101: It has over 13,000 videos categorized into 101 action types and these videos are taken from the YouTube website. The videos are subject to different viewpoints, camera motion and video qualities. In each action type, there are 25 groups and the videos in the same group are similar.

HMDB51: The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips. It is collected from various sources, mostly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos. The actions categories can be grouped in five types: general facial actions, facial actions with object manipulation, body movements with object interaction, and body movements for human interaction.

Method	UCF101		HMDB51	
	Train	Test	Train	Test
F_{rgb}+SOFTMAX	99.5%	50.1%	99.9%	25.4%
F_{of}+SOFTMAX	95.3%	55.7%	96.3%	52.2%
F_{so}+SOFTMAX	96.9%	62.7%	99.3%	58.7%

Table 2: The comparison of the features extracted from scratch on UCF101 and HMDB51 in terms of accuracy.

5.2. Baselines

We compare the proposed method with the state-of-the-art techniques with lightweight learning framework including both hand-crafted features and deep-learned features such as IDT [18], Multi-shIp Feature Stacking (MIFS) [51], LRCN [40], Dynamic Image Network (DIN) [52], C3D [27], Beyond Short Snippets (BSS) [39], Trajectory-pooled Deep-convolutional Descriptor (TDD) [53], and Two-stream ConvNet [38]. All of the above methods have lightweight framework and good performance for action recognition. IDT and MIFS are representative work of hand-crafted features with shallow framework, and these methods do not need pre-trained models. However, they both need very complicated hand-crafted model for multiple scales and sequential information. The deep learning techniques such as DIN, TDD, and Two-stream ConvNet directly introduce a 5-layer CNN model for their feature extraction, and pre-train their models on ImageNet [54] dataset because the training process of large ConvNet can easily result in overfitting problem [37]. C3D use a 5-layer 3D ConvNet with a pre-trained model on Sports-1M. LRCN and BSS use the combination of a 5-layer 2D ConvNet and a RNN model and they are pre-trained on ILSVRC [55] dataset.

For sake of comparison to the above methods, we introduce a two-stream C3D network consist of two 5-layer 3D ConvNets and a 2-layer RNN network with LSTM units for RGB videos and spatial-optical videos. Similar with [27], the C3D network for RGB videos is pre-trained on Sport-1M [15] dataset. However, we do not pre-train the other C3D network for optical flow and spatial-optical videos as they have a good generalization ability and provide a robust spatiotemporal feature description.

The experiments also demonstrate that the proposed spatial-optical data has a great performance for action representation even without pre-trained model.

5.3. Implementation Details

Data inputs. We use three data including RGB videos, TVL1 optical flow videos, and spatial-optical videos as the input of the 3D ConvNets. The C3D-like network for RGB data use a 5-layer 3D ConvNet pre-trained by Sport-1M as base network, and the C3D-like networks for TVL1 optical flow and spatial-optical data are directly trained by UCF101 and HMDB51 datasets without any pre-trained model. For all architectures we follow each convolutional layer by a batch normalization layer and a ReLU activation function. For the extraction of optical flow and warped optical flow, we choose the TVL1 optical flow algorithm [49] implemented in OpenCV with CUDA. We threshold the absolute value of motion magnitude to 50 and rescale the horizontal and vertical components of the optical flow to the range [0, 255]. The spatial-optical data is presented by associating hierarchical weighting segmentation with optical flow as Section 3. During training, we resize the smaller video side to $128 \times 171 \times 16$ pixels as in [27], then randomly cropping a $112 \times 112 \times 16$ patch for RGB, TVL1 optical flow and spatial-optical videos. During test time the models are applied convolutionally over 16 frames taking center crops, and we average predictions from the output of a 2-layer RNN network on 5 video snippets which are sampled equally across the whole video.

Hyper-parameter optimization. Training on videos used standard Adam Optimizer with learning rate set to 1e-4 in all cases. The weight decay is set to be 5×10^-5 and the momentum is 0.9. The batch sizes for both C3D networks are 24 and the parameter λ for feature fusion is set to $1/3$ as default. We trained models of RGB data on both UCF101 and HMDB51 for up to 10k steps as the Sport-1M-pretrained model provides a good generalization ability. The training step of spatial-optical data is set to 20k and the training step of optical flow data is set to 30k as the optical flow data is more difficult to train [30]. The whole training time on UCF101 is around 7 hours for RGB, 15 hours for spatial-optical data and 20 hours for optical flow with 1 TITANX GPU.

5.4. Feature Comparison

The features in Table 2 are extracted from scratch on the UCF101 and HMDB51 dataset with a 5-layer 3D ConvNet which is same as the architecture presented in Section 4.1. We input a stack of RGB frames, optical flow frames and spatial-optical frames into the same spatiotemporal feature extraction framework. Note that, the last layer of the feature extraction framework in training step is a softmax layer, which could be considered as a simple classifier and it has only 101 neurons for calculating the loss function. And we can loosely think of the video belong to the corresponding type of the maximum neuron. Without pre-trained models of large scale datasets such as Sports-1M and extracting random short snippets of the whole video as inputs, the accuracy of RGB videos are not satisfactory. The first row in Table 2 lists the features F_{rgb} extracted from the RGB videos in UCF101 and HMDB51, and they only achieve 50.1$\%$ and 25.4$\%$ accuracy on testing set, but they get almost 100$\%$ accuracy on the training set. Obviously, it is struggling with the overfitting problem because of the static appearance redundancy between consecutive frames and the absence of spatial hierarchical information. The feature F_{of} extracted from the TVL1 optical flow videos takes advantage of motion appearance, and it performs better than F_{rgb}. However, the feature F_{of} is more difficult to train as it contains only optical flow frames which lose the structure information and detailed appearance. The training step for RGB videos and spatial-optical videos is 10k. The training step for optical flow is up to 30k, and the detailed experiments have bee presented in Section 5.3.

In this condition, the discrimination of the spatiotemporal features extracted from the proposed spatial-optical data is much more noticeable. Different from the existing methods, we come up with a new thought that improving the spatiotemporal feature representation before feature extraction and description by the spatial-optical data organization. It is obvious that the feature F_{so}, extracted from spatial-optical data, provides a better action representation than F_{rgb} and F_{of} as it highlights both motion appearance and structural information. Most importantly, it eliminates the static appearance redundancy in the RGB data and this makes it more efficient to describe the spatiotemporal feature. Even without elaborate classifiers and pre-trained models, the proposed data organization is capable of providing a remarkable improvement for action representation. The experiments demonstrate that the combination of the spatial-optical data organization and the C3D feature extraction framework has made the accuracy increased by nearly 7%. And this experiment result is a striking proof for our argument that motion appearance and structural information play significant roles in human action recognition, and the static appearance redundancy between consecutive frames is an obstacle for the spatiotemporal feature representation.

5.5. Traditional Classifier

In this part, we compare the proposed method with the state-of-the-art hand-crafted features based on SVM classifier. The hand-crafted features can be represented as a set of trajectories with different descriptors. The first three methods in Table 3 use HOG, HOF and MBH descriptors to extract statistical histograms around dense trajectories, and the accuracy are about 75% on UCF101 and 48% on HMDB51. As the combination of the above three features, the accuracy of IDT increases to 84.7% and 57.2%. And these methods are the classical hand-crafted features which have been widely used in various computer vision fields, but they have limited discrimination to process complex data such as human action recognition. Note that, the fifth one is a deep-learned feature based on 2-stream ConvNet (2S) which combines spatial ConvNet and temporal ConvNet and it achieves a state-of-the-art results which are 88.0% on UCF101 and 59.4% on HMDB51. The experiments show that the single hand-crafted feature have limited discriminative representation, and the multiple hand-crafted feature combination makes a remarkable improvement but it still falls behind the deep-learned features in human action representation.

Compared with the existing methods, our method improves the accuracy by spatial-optical data organization and C3D feature extraction framework. The features extracted from the RGB data have achieved 79.0% accuracy on UCF101 dataset and 61.2% accuracy on HMDB51 dataset based on the combination of the C3D feature extraction framework and the SVM classifiers. Our method has a similar performance as the representative hand-crafted feature which combines the MBH descriptors and the SVM classifiers. It is obvious that the Sport-1M pre-trained model improves the generalization ability especially for HMDB51. The features F_{of} and F_{so} are directly extracted

Model	UCF101	HMDB51
HOG+SVM	72.4%	40.2%
HOF+SVM	76.0%	48.9%
MBH+SVM	80.8%	52.1%
IDT+SVM	84.7%	57.2%
2-stream+SVM	88.0%	59.4%
F_{rgb}+SVM	79.0%	61.2%
F_{of}+SVM	75.7%	57.4%
F_{so}+SVM	85.3%	62.9%
F_{con}+SVM	82.1%	60.6%
F_{sum}+SVM	87.6%	64.1%

Table 3: Comparison of different feature representation based on SVM classifier.

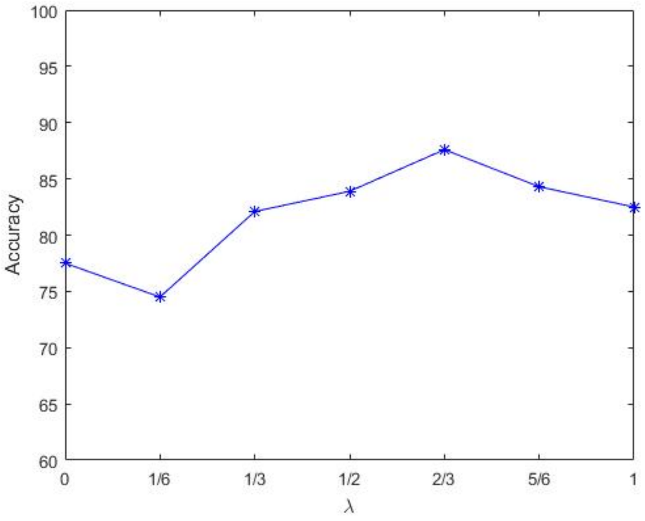

Figure 5: Action recognition accuracy based on different feature combination. The vertical ordinate is the accuracy of recognizing the human action and the horizontal is λ, where $F_{sum} = \lambda F_{rgb} + (1 - \lambda)F_{so}$.

from TVL1 optical flow videos and the spatial-optical videos so they do not benefit from pre-trained model on Sport-1M. The feature F_{of} does not perform well because of the loss of appearance details and structure information. However, the feature F_{so} extracted from the proposed spatial-optical videos improves the performance of action representation as it takes the advantage of eliminating static appearance redundancy and highlighting the motion appearance. It is an important evidence that motion appearance plays a more important role than static appearance in human action representation.

And we also compare different fusion strategies where we report the average accuracy on the first split of UCF101. The experiments presented in Table 3 also illustrate that the feature $F_{con} = \{F_{rgb}; F_{so}\}$ of concatenation strategy performs considerably lower than the feature $F_{sum} = \lambda F_{rgb} + (1 - \lambda) F_{so}$ of sum strategy. The feature F_{sum} increases the accuracy of human action recognition to 87.6% on UCF101 dataset and 64.1% on HMDB51. Since this, as well as the high result of sum strategy, suggests that simply computing a weighted average between F_{rgb} and F_{so} is already a good fusion technique, and similar results are also declared in [56] and [40]. Furthermore, we report results of F_{so} with different parameter λ in Figure 5. Since the feature F_{so} extracted from spatial-optical videos outperforms the feature F_{con} extracted from RG-B videos, weighting the C3D network of the spatial-optical data higher unsurprisingly can get a better performance. In general, the proposed spatial-optical data organization has improved the action representation by highlighting the motion appearance and eliminating the static appearance redundancy.

5.6. Deep Learning Classifier

The shallow classifier has achieved a good performance in many computer vision fields but it still has a limited discrimination for complicated actions since the lack of semantical information and sequential correlation for sequential data. Compared with traditional SVM classifiers, RNN can discover more meaningful semantics and pattern as it considers the context among consecutive video snippets. The methods listed in Table 4 are the state-of-the-art technologies of action recognition. Wang *et al.* proposed IDT to describe videos by dense trajectory description and it is an outstanding feature for action representation and it is widely used for video understand. Lan *et al.* [51] proposed a feature enhancing technique named MIFS and it also achieves a good results. The above two representative methods of hand-crafted features outperform the existing hand-crafted features for action recognition. However, they are limited by the discrimination of shallow classifiers, and more researchers turn their attention to deep learning techniques to get a better recognition accuracy. For example, Donahue *et al.* [40] proposed a general framework called LRCN which is based on CNN and RNN for video understanding. And the C3D network proposed by Tran *et al.*[27] achieves 85.2% accuracy for action recognition with the help of the pre-trained model on Sports-1M [57] and the optical flow. Two-stream ConvNet architecture proposed by Simonyan and Zisserman is a competitive with the state of the art as it incorporates spatial and temporal networks. Ng *et al.* [39] increase the accuracy to 88.6% by taking the benefits of a pre-trained ImageNet model and fine-tuning on Sports-1M videos. Recently, Bilen *et al.*, [52] propose a novel deep learning neural network, DIN, which is similar with our motivation, to introduce a compact video representation into deep-learned feature, and it achieved a good performance in video understanding. Another representative work based on both hand-crafted features and deep-learned features is proposed by Wang *et al.* [53] and it achieved a remarkable improvement as it shares the merits of features from IDT and CNN and enhances the robustness by spatiotemporal normalization and channel normalization.

Note that, TDD which takes the merits of both hand-crafted features and deep-learned features outperform the existing methods. And many researches point out that the combination of deep learned features and hand-crafted features can get a better result in various applications. The primary cause of this problem is that the hand-crafted features pay more attention to dynamic appearance such as optical flow and motion trajectories but the deep-learned features only focus on searching the most discriminative parts. However, the deep-learned features might be misled by the static appearance redundancy that are discriminative but not available for action representation. And this may cause a problem that the features of static appearance have overfit the training data but the features of motion appearance still struggle with underfitting problem. To solve this problem, many researches try to introduce pre-trained models or additional constrains to reduce the negative effect of static appearance overfitting problem.

Different from the existing methods, we propose a novel technique to solve the problem from the view of eliminating static appearance redundancy and enhancing motion appearance for action representation by spatial-optical data organization and sequential learning framework. The methods listed in Table 4 are the state-of-the-art technologies with lightweight framework (AlexNet-like architecture). The two fused features, F_{sum} and F_{con}, achieve 90.9% and 88.6% accuracy on UCF101 dataset and they achieve 62.3% and 65.7% accuracy on HMDB51 dataset. It is obvious that the spatial-optical data organization makes a great improvement as it considers both motion appearance and spatial hierarchical information. More importantly, the bottom-up sequential learning framework extracts discriminative spatiotemporal features and provides a great ability to find out high-level semantics and useful patterns.

5.7. Discussion

Last but not least, many recent work try to get a better accuracy of action recognition by massive learning framework and complicated models. As presented in Table 4, Convolutional Two-stream Network Fusion (C2S) [56] and Very Deep Two-stream Convents (D2S) [58]. Although these methods have increased the accuracy by 1 or 2 percent, they have to use massive learning framework with ten more layers, such as Inception and VGG-16 architecture, which make the framework much more difficult to train. The recent work such as Temporal Segment Networks (TSN [30]) and Inflated 3D ConvNet(I3D [31]) are good practices to design effective ConvNet architectures for action recognition. TSN combines a sparse temporal sampling strategy and video-level supervision to enable efficient and effective learning using the whole action video. I3D Inflates 2D Inception-V1 ConvNets into 3D, bootstraps 3D filters from 2D filters which are pre-trained on ImageNet. The fundamental cause of these extra efforts with massive learning frameworks and elaborate pre-trained models is the redundancy of static appearance and the lack of sequential data mining. The above methods directly use raw data and have limited ability to extract the spatiotemporal feature of action representation.

Different from the above methods, we propose the spatial-optical data organization

Model	UCF101	HMDB51	Architecture
IDT (*Wang et al. 2013.*)	85.9%	57.2%	None
MIFS (*Lan et al. 2015.*)	89.1%	65.4%	None
LRCN (*Donahue et al. 2015.*)	82.9%	-	AlexNet
C3D (*Tran et al. 2015.*)	85.2%	-	AlexNet-like
2S ConvNet (*Simonyan et al. 2014.*)	88.0%	59.4%	VGG-M
BSS (*Ng et al. 2015.*)	88.6%	-	AlexNet
DIN (*Bilen et al. 2016.*)	89.1%	65.2%	AlexNet
TDD (*Wang et al. 2015.*)	90.3%	63.2%	VGG-M
D2S (*Wang et al. 2015.*)	91.4%	-	VGG-16
C2S (*Feichtenhofer et al. 2016.*)	92.5%	65.4%	VGG-16
RGB-I3D (*Carreira et al. 2017*)	84.5%	49.8%	Inception
Flow-I3D (*Carreira et al. 2017.*)	90.6%	61.9%	Inception
2S-I3D (*Carreira et al. 2017.*)	93.4%	66.4%	Inception
TSN (*Wang et al. 2016.*)	93.5%	-	Inception
F_{con}+lstm	88.6%	62.3%	AlexNet
F_{sum}+lstm	90.9%	65.7%	AlexNet

Table 4: Comparison of the proposed method to the state of the art.

which is our main contribution, and the novel data organization try to improve the action representation by highlighting the motion appearance and eliminating the static redundancy before feature extraction by 3D ConvNets. Our method try to improve the feature representation by simplifying and highlighting motion appearance and explore global semantics and patterns by a lightweight learning framework with less convolutional layers for sequential data. We believe that the discriminative spatiotemporal feature extracted from the spatial-optical data and the robust sequential learning framework with lightweight learning framework can handle various applications of video understanding. We will consider to introduce more effective architecture like I3D and TSN to improve the action representation of the proposed spatial-optical data organization in future work.

6. Conclusion

In this paper, we proposed a novel method for action recognition from the view of the spatial-optical data organization and the sequential learning framework. 1) The spatial-optical data organization has a good performance for video representation and provides a discriminative feature for action recognition as it takes the advantages of motion appearance, motion trajectories and optical flow. The experiments show that the recognition accuracy using the spatial-optical data has increased by 12% with a lightweight C3D network. 2) On the other side, to overcome the problem of semantic gap between low-level features and high-level semantics, we propose a sequential learning framework based on the low-level spatiotemporal feature extraction and the high-level sequential information mining.

To further get improvements for the proposed method, we expect to apply a fast spectral clustering techniques based on nystorm to make it more efficient for video segmentation. We believe that a robust structural information extraction can make a remarkable improvement for video understanding and a good data input can also help the massive learning framework to get avoid of overfitting problem.

7. Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant 61773316, Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and the Open Research Fund of Key Laboratory of Spectral Imaging TechnologyChinese Academy of Sciences.

References

[1] Q. Wang, J. Wan, Y. Yuan, Locality constraint distance metric learning for traffic congestion detection, Pattern Recognition 75 (2018) 272–281.

[2] Q. Wang, J. Wan, Y. Yuan, Deep metric learning for crowdedness regression, IEEE Transactions on Circuits and Systems for Video Technology 10.1109/TCSVT.2017.2703920.

[3] D. Gerónimo, H. Kjellström, Unsupervised surveillance video retrieval based on human action and appearance, in: International Conference on Pattern Recognition, 2014, pp. 4630–4635.

[4] Y. Zeng, C. Tsai, W. Chang, Abnormal action warning on encrypted-coded surveillance video for home safety, in: International Conference on Multimedia and Expo Workshops, 2013, pp. 1–6.

[5] H. Lin, Y. Chen, Y. Chen, Robot vision to recognize both object and rotation for robot pick-and-place operation, in: International Conference on Advanced Robotics and Intelligent Systems, IEEE, 2015, pp. 1–6.

[6] F. Tango, M. Botta, Real-time detection system of driver distraction using machine learning, IEEE Transaction on Intelligent Transportation Systems 14 (2) (2013) 894–905.

[7] Y. Yuan, Z. Jiang, Q. Wang, Video-based road detection via online structural learning, Neurocomputing 168 (2015) 336–347.

[8] Q. Wang, J. Fang, Y. Yuan, Adaptive road detection via context-aware label transfer, Neurocomputing 158 (2015) 174–183.

[9] A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal localization and categorization of human actions in unsegmented image sequences, IEEE Transaction on Image Processing 20 (4) (2011) 1126–1140.

[10] I. Everts, J. C. van Gemert, T. Gevers, Evaluation of color spatio-temporal interest points for human action recognition, IEEE Transaction on Image Processing 23 (4) (2014) 1569–1580.

[11] H. Xu, Q. Tian, Z. Wang, J. Wu, A survey on aggregating methods for action recognition with dense trajectories, Multimedia Tools and Applications 75 (10) (2016) 5701–5717.

[12] S. Herath, M. T. Harandi, F. Porikli, Going deeper into action recognition: A survey, CoRR abs/1605.04988.

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

[14] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: A local SVM approach, in: International Conference on Pattern Recognition, 2004, pp. 32–36.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, T. Serre, HMDB: A large video database for human motion recognition, in: International Conference on Computer Vision, 2011, pp. 2556–2563.

[16] K. Soomro, A. R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, CoRR abs/1212.0402.

[17] I. Laptev, On space-time interest points, International Journal of Computer Vision 64 (2-3) (2005) 107–123.

[18] H. Wang, A. Kläser, C. Schmid, C. Liu, Dense trajectories and motion boundary descriptors for action recognition, International Journal of Computer Vision 103 (1) (2013) 60–79.

[19] Y. Zhao, Q. Wang, Y. Yuan, Action recognition based on semantic feature description and cross classification, in: IEEE China Summit & International Conference on Signal and Information Processing, ChinaSIP 2014, Xi'an, China, July 9-13, 2014, 2014, pp. 626–630.

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

[21] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence 35 (1) (2013) 221–231.

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

[23] H. Bay, T. Tuytelaars, L. J. V. Gool, SURF: speeded up robust features, in: European Conference on Computer Vision, 2006, pp. 404–417.

[24] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[25] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.

[26] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

[27] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[28] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1106–1114.

[29] G. Chéron, I. Laptev, C. Schmid, P-CNN: pose-based CNN features for action recognition, in: IEEE International Conference on Computer Vision, 2015, pp. 3218–3226.

[30] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. V. Gool, Temporal segment networks: Towards good practices for deep action recognition, in: 14th European Conference on Computer Vision, 2016, pp. 20–36.

[31] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4724–4733.

[32] Q. Wang, J. Gao, Y. Yuan, Embedding structured contour and location prior in siamesed fully convolutional networks for road detection, IEEE Trans. Intelligent Transportation Systems 19 (1) (2018) 230–241.

[33] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[34] V. Veeriah, N. Zhuang, G. Qi, Differential recurrent neural networks for action recognition, CoRR abs/1504.06678.

[35] C. Ma, M. Chen, Z. Kira, G. AlRegib, TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition, CoRR abs/1703.10667.

[36] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Transactions on Image Processing 27 (8) (2018) 3893–3903.

[37] B. Zhao, B. Huang, Y. Zhong, Transfer learning with fully pretrained deep convolution networks for land-use classification, IEEE Geosci. Remote Sensing Lett. 14 (9) (2017) 1436–1440.

[38] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.

[39] J. Y. Ng, M. J. 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, 2015, pp. 4694–4702.

[40] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, K. Saenko, Long-term recurrent convolutional networks for visual recognition and description, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[41] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised learning of video representations using lstms, in: International Conference on Machine Learning, 2015, pp. 843–852.

[42] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, G. Wang, Skeleton-based action recognition using spatio-temporal LSTM network with trust gates, CoRR abs/1706.08276.

[43] B. Mahasseni, S. Todorovic, Regularizing long short term memory with 3d human-skeleton sequences for action recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3054–3062.

[44] J. Han, R. Quan, D. Zhang, F. Nie, Robust object co-segmentation using background prior, IEEE Transaction on Image Processing 27 (4) (2018) 1639–1651.

[45] M. Grundmann, V. Kwatra, M. Han, I. A. Essa, Efficient hierarchical graph-based video segmentation, in: IEEE Conference on Computer Vision and Pattern, 2010, pp. 2141–2148.

[46] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 59 (2) (2004) 167–181.

[47] C. Xu, J. J. Corso, Evaluation of super-voxel methods for early video processing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1202–1209.

[48] J. Shi, C. Tomasi, Good features to track, in: IEEE Conference on Computer Vision and Pattern Recognition, 1994, pp. 593–600.

[49] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime tv-L^1 optical flow, in: Pattern Recognition, 2007, pp. 214–223.

[50] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. J. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, in: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1494–1504.

[51] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, B. Raj, Beyond gaussian pyramid: Multi-skip feature stacking for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 204–212.

[52] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, S. Gould, Dynamic image networks for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[53] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 4305–4314.

[54] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, 2009, pp. 248–255.

[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, F. Li, Imagenet large scale

visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[56] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[57] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. Li, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[58] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, Towards good practices for very deep two-stream convnets, CoRR abs/1507.02159.