Neurocomputing 444 (2021) 319–331

Contents lists available at ScienceDirect

Neurocomputing

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

DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition

Jun-Yan He [a,b], Xiao Wu [a,b,*], Zhi-Qi Cheng [a,b], Zhaoquan Yuan [a,b], Yu-Gang Jiang [c]

[a] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[b] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, China
[c] School of Computer Science, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 14 October 2019
Revised 16 January 2020
Accepted 5 May 2020
Available online 25 November 2020

Keywords:
Human action recognition
Deep learning
Convolutional neural network
Long-range temporal
LSTM

ABSTRACT

Although deep learning has achieved promising progress recently, action recognition remains a challenging task, due to cluttered backgrounds, diverse scenes, occlusions, viewpoint variations and camera motions. In this paper, we propose a novel deep learning model to capture the spatial and temporal patterns of human actions from videos. Sample representation learner is proposed to extract the video-level temporal feature, which combines the sparse temporal sampling and long-range temporal learning to form an efficient and effective training strategy. To boost the effectiveness and robustness of modeling long-range action recognition, a Densely-connected Bi-directional LSTM (DB-LSTM) network is novelly proposed to model the visual and temporal associations in both forward and backward directions. They are stacked and integrated with the dense skip-connections to improve the capability of temporal pattern modeling. Two modalities from appearance and motion are integrated with a fusion module to further improve the performance. Experiments conducted on two benchmark datasets, UCF101 and HMDB51, demonstrate that the proposed DB-LSTM model achieves promising performance, which outperforms the state-of-the-art approaches for action recognition.

1. Introduction

Human actions in videos are characterized by spatio-temporal evolution of appearance governed by motions. Visual appearances and action dynamics are two critical and complementary aspects for action recognition. Several visual representations have been proposed to recognize human actions in videos, by using various features and learning algorithms, including space–time pattern templates [1], optical flow [2], spatio-temporal interest points [3–7], and motion trajectories based representations [8,9]. Recently, deep learning has achieved remarkable success in many areas, including object detection, scene analysis, event detection, image generation, and so on [10–12]. It has been introduced to boost the performance of action recognition in videos [13–15], especially with the very deep Convolutional Neural Networks (CNN) (e.g., GoogLeNet [12], ResNet [16]).

Current deep learning methods for action recognition (e.g., [14,13,17,18]) usually apply existing CNN architectures for static images (e.g., AlexNet [10], ResNet [16]) to learn action representations in videos. They are commonly performed for a short video interval within a couple of frames, less than one second. Such kind of motion dynamics is referred as the *short-term* temporal pattern. However, typical human actions such as cooking and dancing, as well as repetitive actions such as walking and playing piano often last for hundreds of frames in several seconds. The motion dynamics are regarded as *long-range* temporal pattern. In this paper, long-term and long-range refer to the same thing. Actions in a video are composed of a sequence of continuous frames, which usually contain multiple subactions. As shown in Fig. 1, a man plays a flic-flac on the ground contains several key snippets, hand-stand on the ground, lying on the ground, somersault, and back handspring. It is meaningful to locate and identify the critical subactions, so that the action can be correctly recognized.

Recently, explorations of action recognition based on long-term temporal have been conducted in [19–21,14,13,22]. To model the long-term temporal properties, a straightforward way is to extend the 2D CNN to 3D spatio-temporal convolutions [19,23,24]. Unfortunately, the improvement is unsatisfactory and expensive computation cost is involved, due to the dense sampling and limited capability of learning motion dynamics. Recurrent Neural Network (RNN), especially Long Short-Term Memory (LSTM) [20,21,25] have

* Corresponding author at:School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China.
E-mail addresses: wuxiaohk@home.swjtu.edu.cn (X. Wu), zqyuan@swjtu.edu.cn (Z. Yuan), ygj@fudan.edu.cn (Y.-G. Jiang).

https://doi.org/10.1016/j.neucom.2020.05.118
0925-2312/

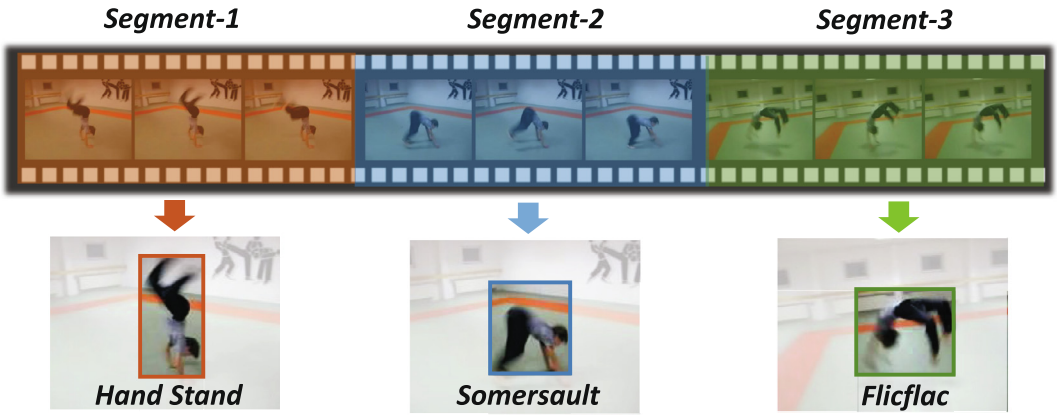

Fig. 1. An action usually refers to a long-range temporal process, which is composed of a sequence of subactions, such as hand-stand on the ground, somersault, and back handspring.

been widely used to capture the long-term evolution of actions in the video clips. However, the pipeline of CNN + LSTM is complex for training and the problem of gradient vanishing limits the construction of a deep and powerful LSTM network. To accelerate the efficiency of model training, a sampling scheme is proposed by dividing the video into several snippets, sampling the spatial and temporal domain data, and then aggregating the scores of each snippets. Although it is easy to implement and helpful for training a robust model, the latent temporal relationship of the segments is totally ignored, which is critical for the long-term action recognition.

To overcome the drawbacks of existing methods, a long-range temporal model for human action recognition is novelly proposed in this paper, which comprehensively integrates the spatial, short-term as well as long-term temporal patterns of human actions. The framework of the proposed approach is illustrated in Fig. 2. The video frames and optical flow maps are first sampled and fed into the Sample Representation Learner (SRL) to generate the visual features and short-term temporal features. To extract the long-range temporal feature, an efficient and effective training strategy is constructed by sparse sampling with a sampling stackfor temporal relationship learning. A Densely-connected Bi-directional LSTM (DB-LSTM) network is novelly proposed to capture the long-range temporal properties and to alleviate the problem of gradient vanishing. It models the bi-directional temporal patterns of actions from both the forward and reverse directions. In addition, LSTM units are stacked and combined with the dense skip-connections, offering the extra channels for signal transmission. The two modalities induced from visual appearance and motion dynamics are further integrated with a multi-scale fusion scheme to adapt different durations of actions, improving the accuracy and robustness of the proposed model. The major contributions of this paper are summarized as follows:

- A deep learning model based on long-range modeling is novelly proposed for action recognition, which captures the global appearance and local motion dynamics, meanwhile integrates visual appearance and long-range temporal dynamics of human actions.
- In order to capture the long-range temporal correlations, a concise abstraction is constructed with a sparse sampling strategy. The global and dynamic local features are enhanced by the stack representation learner.
- A densely-connected Bi-directional LSTM (DB-LSTM) network is novelly proposed to capture the long-range temporal pattern in forward and backward directions, which effectively alleviates the problem of gradient vanishing, strengthens the feature propagation, encourages feature reuse, and substantially reduces the number of parameters.
- Experiments conducted on UCF101 and HMDB51 benchmark datasets demonstrate that the proposed DB-LSTM model outperforms the state-of-the-art approaches for action recognition.

This paper is organized as follows. Section 1 gives a brief overview of related works and introduction of the proposed methods. Sections 3 and 4 elaborate the proposed Sample Representation Learner and DB-LSTM, respectively. Section 5 describes the experimental setup and performance comparison. Finally, this paper is concluded with a summary.

2. Related work

2.1. Action recognition

Action recognition is a fundamental task in computer vision community. Different approaches have been proposed to recognize the actions in videos, by using various features and learning algorithms. Comprehensive surveys on action recognition can be found in [26,27]. Traditional methods are based on the hand-crafted features, such as dense trajectory [28,18] and improved trajectory [8,9] to describe the motion dynamics of the actions. To effectively utilize the local features and consider the long-term temporal of the actions, the encoding methods are proposed to encode the features within the whole video. Fisher Vector (FV) [29] and Vector of Locally Aggregated Descriptors (VLAD) [30] are classic encoding methods, which share the merits of locality and simplicity, but may lack the semantic and discriminative capacity. A sparse temporal sampling strategy and long-range supervision learning framework [31] is proposed for the efficient and effective learning of long-term actions in videos.

With the rapid progress of deep learning, many works on action recognition have been explored to learn deep features and design effective architectures for video action recognition. Two-stream fusion model [14] is a novel framework for action recognition, which incorporates spatial and temporal networks. The color images and the optical flow maps are modeled with CNN as the spatial and temporal patterns, respectively, which are then fused to obtain the final result. Due to the effectiveness of the two-stream architecture, several variances are developed. A hybrid deep learning framework for video classification is proposed in [32], which can model static spatial information, short-term motion, as well as long-term temporal clues in the videos. To fuse

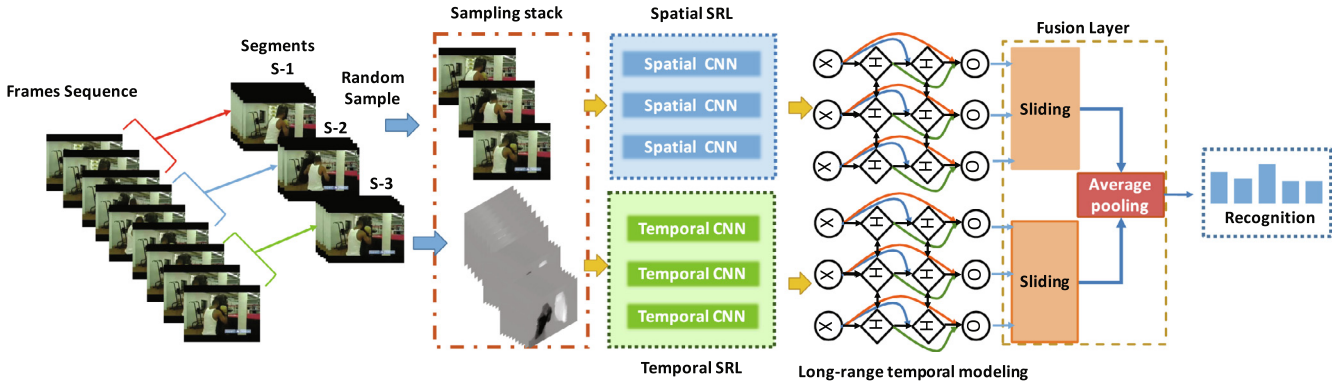

Fig. 2. The framework of the proposed Densely-connected Bi-directional LSTM (DB-LSTM) model. The input video is first extracted frames and optical flow maps, and then fed into the Stack Representation Learner (SRL) to produce the feature stack. The feature stack is utilized to model the temporal pattern. Finally, a fusion layer is employed to integrate the spatial and temporal clues to predict the final result.

these two features from spatial and temporal modules, a regularized fusion network is used to learn the latent relationships between spatial and temporal clues in the same action. A practical method called Temporal Segment Network (TSN) based on the random sampling scheme is proposed in [13], which is a simple but effective architecture for action recognition. To further improve the performance of TSN, a multi-scale sampling and fusion framework is proposed in [33]. With the help of posture estimation, the methods [34–37] are proposed to improve the performance of action recognition.

Since video is composed of a sequence of frames, 3D convolution is a more intuitive way to capture the temporal dynamics within a short period of time. 3D CNN (C3D) [19] is an end-2-end method jointly modeling the spatial and temporal information. However, C3D induces a large number of parameters and computation cost due to the additional kernel dimension. Inflated 3D ConvNet (I3D) [24] and Pseudo-3D Residual Networks (P3D) [23] are proposed to reduce the scale of the network, by applying extra convolutional operations on the temporal domain instead of utilizing a 3D kernel directly. In addition, I3D also integrates the two-stream architecture to construct a two-stream model based on the I3D backbone network, achieving significant improvement. To further improve the performance, C3D and two-stream architecture are combined in [38]. From the aforementioned works, we can find that recent studies focus on boosting the discrimination of deep features and the fusion methods to improve the performance of action recognition.

2.2. Temporal modeling

There are two types of temporal modeling for action recognition, i.e., short-term and long-range. The short-term temporal modeling is mainly based on the neighbor frames within the videos. The most widely used feature is optical flow [39], which is the apparent motion pattern of objects in a visual scene caused by the relative motion between an observer and a scene. It has been applied to deep CNN based models as the temporal domain in prior works [14,17], integrated with the visual and spatial domains.

To take the advantages of both the temporal and spatial attention models, VideoLSTM is proposed in [25], which integrates the spatial and motion guided attention to model the spatio-temporal properties of video actions. Different from RNN-based frameworks, 3D-CNN [19] is convolutional based method for long-range temporal modeling, which extracts features from both spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. Although existing action recognition approaches have shown impressive performance, the long-range modeling remains a critical problem worth further exploration. Temporal modeling is also widely studied in NLP. A long-range modeling method is proposed in [40], which constructs densely connected LSTM modules for sentence classification.

3. Sample representation learner

Sample Representation Learner (SRL) is designed to extract the long-range temporal patterns, which consists of two components, *sampling stack* and *representation learner*. The sampling stack is utilized to construct a stack structure to capture the long-range temporal relationship, while the representation learner is employed to produce the features of the stack.

3.1. Sampling stack

Actions are usually embedded in sequential scenes of videos, from which most of them represent the main topic of actions. However, the start and end parts of a video have high probabilities to be the scene changing and transition parts, which are not closely related to the action. Examples are illustrated in Fig. 3.The first row is an action about eating. However, the first scene is a woman sitting in front of a table, which only has a weak relationship with eating. Meanwhile, the last scene is similar to an interview, which is not correlated to the previous scene of eating. Similarly, for the second and third rows, they refer to the actions of eating and drinking, respectively. From these examples, we can see that not all scenes are closely relevant to the actions, especially for actions in long-range videos. Therefore, in order to meet the demand of long-range temporal modeling, the sampling and training strategy is needed to be well designed.

Since the videos contain redundant information, a practical way for action recognition is to sample multiple frames/clips from videos. To tradeoff the efficiency and effectiveness, a *sampling stack* is constructed by randomly extracting frames from fixed-length segments of a video and then concatenating them to form a new sequence of segments.

The Sampling stack sample only one frame from each segment randomly. Then, a frame stack according to the sequential of the segments is constructed. Specifically, given a video having n frames, $v = \{i_0, \ldots, i_i, \ldots, i_{n-1}\}$, where i_i refers to the i_{th} frame of the video. Video v is evenly divided into k segments, in which the length of segments is $m = n/k$. For each segment s_i, one sampled frame sf_i is randomly selected. Meanwhile, an optical flow map of_i is randomly extracted from five consecutive frames. The

Fig. 3. Three action examples: the first two actions are eating and the last one is drinking. The action video always contains more than one scene. However, the additional scene is not related to the action, misleading the recognition.

sampled frames and optical flow maps for all segments are then catenated to form the sample stack of segment s_i, which includes the visual and motion tracks, respectively. An example of the sampled frame and optical flow map is illustrated in Fig. 4. In the testing stage, a video is densely sampled to cover all shots of the video.

The sampling stack is a concise and abbreviated abstract of the whole video, which extracts the representative samples within the video and contains the long-range temporal information. This scheme guarantees the sequential data sampling and reduces the redundancy of the neighbor segments. With this strategy, more coherent contextual information can be included to facilitate the modeling of actions. Therefore, the segment based random sampling strategy is an easy but effective solution to prevent the model from over-fitting in the training stage.

The prior work [13] conducts the similar sparse sampling. Unfortunately, it only samples the frames from the segments, which are then directly fed into the deep CNN to train the model, ignoring the long-range temporal relationship between the sampling frames. Different from it, we construct a stack structure which samples data from all segments of a video. This design not only captures the comprehensive information of a video, but also retains the temporal relationship of all segments. Since the action changes over time, sampling frames from different segments can capture the motion changes, which is critical for video action recognition.

3.2. Representation learner

Once the sampling stack is constructed, a *representation learner* is adopted to learn the spatial and short-term temporal pattern. For the spatial branch, the sampled frames are directly fed into the representation learner to model the visual appearance. To capture the temporal information of actions, optical flow maps are then modeled by the representation learner as the spatial branch.

Considering the performance and efficiency, Densely Connected Convolutional Networks (DenseNet) [41] is adopted as the branches of the representation learner to extract the features of actions.

Sample Representation Learner (SRL) is formulated as:

$$SRL(s, w) = \omega(g(x_0, w), g(x_1, w), \ldots, g(x_{m-1}; w))$$

where $g(x; w)$ means the backbone networks, x is the sampled unit, which is from either the visual or temporal branch, and w refers to its weights.

Due to the existence of redundant information in videos, it is easy to lead the model over-fitting. To overcome this problem, the Dropout layer [42] is adopted to regularize the learner, which is set to 0.8. It effectively boosts the capability and robustness of the model. Once the features are extracted, they are fed into the long-range temporal modeling modules to capture the spatial, short-term and long-range temporal patterns.

4. Long-range modelling with DB-LSTM

Owing to the capability of learning long-term dependencies in a recurrent manner, Recurrent Neural Network (RNN), especially Long Short-Term Memory (LSTM) [20] has been widely used in many deep neural networks for learning to classify, process and predict time series. It is also applied for action recognition. In the following subsections, we will briefly review LSTM and its extensions, followed by the new proposed DB-LSTM.

4.1. LSTM and Its extensions

4.1.1. LSTM

Long Short-Term Memory (LSTM) [20] is an improved version of RNN. LSTM is developed to deal with the problems of long-term dependencies of traditional RNN. Therefore, LSTM has been widely used to capture the long-term evolution of actions in video clips [43,44]. A typical LSTM unit consists of an input gate i_t, a forget gate f_t, an output gate o_t as well as a candidate cell state g_t, which is illustrated in Fig. 5(a), where x_t and h_t are the input sample and hidden unit of the t_{th} time step, respectively.

4.1.2. Bi-LSTM

For some applications, current status is closely related to previous context, and is also correlated with forthcoming context. Therefore, Bi-directional LSTM (Bi-LSTM) [45] is proposed to process data in both directions with two separate hidden layers, which are then fed forwards to the same output layers. The forward and backward operations are labeled with red and green arrows in horizontal in Fig. 5(b), respectively, in which b_i and f_i indicate the backward and forward hidden units, respectively. The blue and yellow lines represent the input and the output of the forward LSTM. For action recognition, the temporal pattern is created with the development of actions, and reverse direction of temporal pattern also represents the same actions, which is an additional latent information for action recognition.

4.1.3. Res-LSTM

As the evolution of CNNs, it has been proven that a deeper network brings better performance. Unfortunately, for LSTM network, stacking multiple layers also lead to network training unstable. The gradient vanishing always occurs, since LSTM is consist of the *sigmoid* activation function, which also activates the neuron to saturation, leading the network unstable. Motivated by the modern design of CNN, Residual LSTM (Res-LSTM) is proposed in [46]. The structure of Res-LSTM is illustrated in Fig. 5(c). Res-LSTM is based on the architecture of residual learning, which utilizes

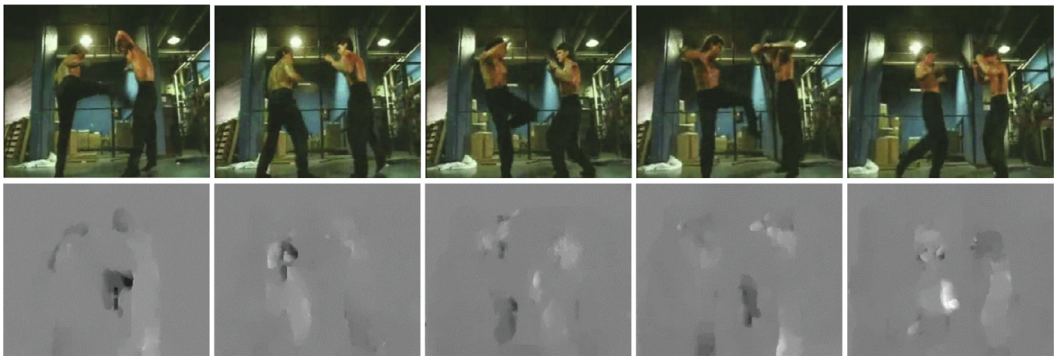

Fig. 4. The sampled frames and optical flow maps extracted in spatial and temporal domains, respectively.

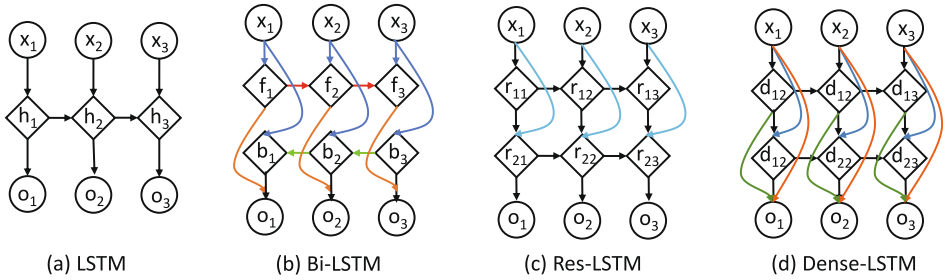

Fig. 5. LSTM and its variants, Bi-directional LSTM (Bi-LSTM), Residual LSTM (Res-LSTM) and Densely Connected LSTM (Dense-LSTM).

skip-connections to offer an extra channel for signal transmitting, which are labeled in blue in Fig. 5(c). The gradients of LSTM can be back propagated by the skip-connections, making it possible to build up an effective multiple-layer LSTM network. It improves the potential capability of the network.

4.1.4. Dense-LSTM

Inspired by the densely connected network [41], we design a naive version of the Densely-connected LSTM (Dense-LSTM), in which the skip-connections are adopted within the densely connected network blocks. Its structure is illustrated in Fig. 5(d), where d_{lt} denotes the densely connected hidden unit of the l_{th} layer at the t_{th} time step. The yellow, blue and green lines represent the skip-connections between different densely connected layers. Different from the residual learning structure, the element-wise additive operation is replaced by the concatenation, which avoids the performance degradation caused by the direct gradient back-propagation.

4.2. Temporal modeling with DB-LSTM

To take the advantage of the densely connected architecture and the bi-directional temporal property, Densely-connected Bi-directional LSTM (DB-LSTM) network is novelly proposed to model the temporal pattern of the actions. DB-LSTM layers are connected not only between two neighbor layers but also connected with other layers. The information can be passed forward and backward simultaneously, which is illustrated in Fig. 6. The structure of DB-LSTM alleviates the problem of gradient vanishing, strengthens the feature propagation, encourages feature reuse, and substantially reduces the number of parameters. It boosts the spatial and short-term temporal feature representation. As illustrated in Fig. 6, DB-LSTM is constructed by dual layers of LSTM networks, which are linked with densely connected skip-connections. These two LSTM networks model the long-range temporal patterns of actions in forward and reverse directions. The outputs of them at each time step are concatenated as one. Each DB-LSTM block is defined as follows:

$$\overleftrightarrow{d}_t = \left[\overrightarrow{d}_t, \overleftarrow{d}_t\right]$$

where $\overleftrightarrow{d}_t$ refers to the t_{th} output of DB-LSTM. $\overrightarrow{d}_t$ and $\overleftarrow{d}_t$ denote the outputs of the t_{th} time step of forward and reverse directions of LSTM networks, respectively, which are combined with dense skip-connections. $[,]$ denotes the concatenation operation. $\leftarrow$ and $\rightarrow$ indicate the forward and backward directions of output d, respectively, which are determined by the direction of the input sequence. The output of the l^{th} layer of the LSTM block d_t^l at the t^{th} time step is determined by previous layers, which is defined as follows:

$$d_t^l = \left[h_t^l\left(\left[d_t^0, d_t^1, \ldots, d_t^i, \ldots, d_t^{l-1}\right]\right), x_t\right]$$

where $\left[d_t^0, d_t^1, \ldots, d_t^i, \ldots, d_t^{l-1}\right]$ refers to the concatenation of the extracted features of previous layers. $H^l(X)$ denotes the l_{th} layer of LSTM networks, where

$$X = \{x_0, x_1, \ldots, x_{t-1}\}$$

is the input of the LSTM layer, which consist of t time steps of features.

$$\left[h_t^l\left(\left[d_t^0, d_t^1, \ldots, d_t^i, \ldots, d_t^{l-1}\right]\right), x_t\right]$$

denotes the concatenation of the output of prior LSTM layers, and the input feature x_t at the t_{th} time step. h_l^t represents the l_{th} layer of LSTM at t_{th} time step The output of the first layer is $d_t^0 = \left[h_t^0(x_t), x_t\right]$, since there is no prior layer.

To capture the comprehensive long-range temporal pattern, we only use the output of the last time step of DB-LSTM, which is rep-

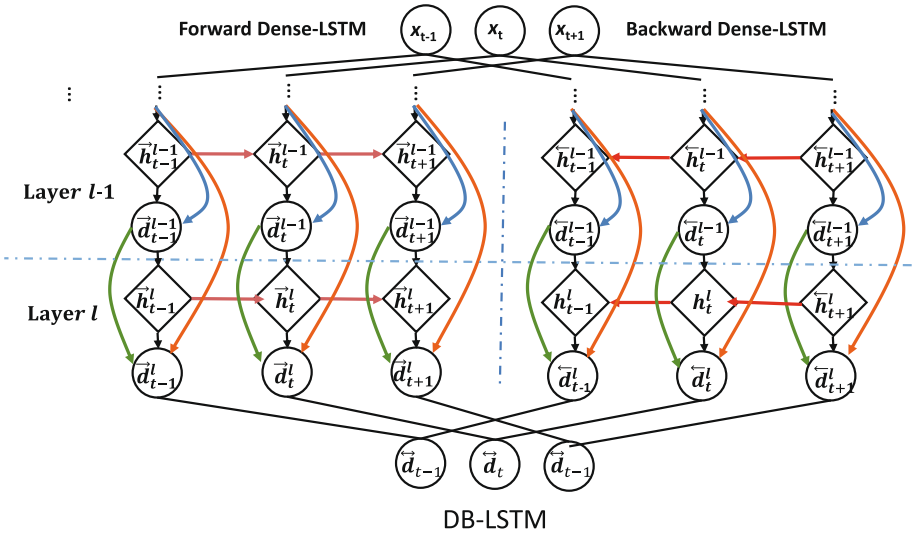

Fig. 6. The architecture of the proposed DB-LSTM.

resented as $\varphi(S, F, W_S, W_L)$, where S is a sampling stack, F refers to the backbone convolutional network, and W_S and W_L denote the weights of backbone networks of SRL and DB-LSTM, respectively. A cross-entropy function is utilized to compute the loss, whose objective function is defined:

$$\Psi(y, \varphi) = -\sum_{j=1}^{C} y_i \left(\varphi_i - \log \sum_{j=1}^{C} \exp \varphi_i \right)$$

where C refers to the total number of action categories, and y_i refers to the label of the i_{th} input video.

4.3. Fusion

The scores produced by the long-range temporal modeling are incorporated with the fusion layer. For the outputs of all time steps, a multi-scale sliding window is utilized to fuse the scores. For the window size of $q \in \{1, 2, 4, 8, 16\}$, the fusion is defined:

$$W_q = \frac{1}{(K-q)q} \sum_{s=1}^{K-q} \sum_{t=s}^{s+q} D_t^{bi}$$

where K refers to the number of time steps, which is the same as the segments, and s represents the start time step of the sliding window. To solve the problem of different time durations of actions, the average fusion of multi-scale sliding windows is applied.

5. Experiments

To evaluate the performance of the proposed DB-LSTM, experiments are conducted to compare with the state-of-the-art methods. The ablation studies are also performed to assess the effect of individual component independently.

5.1. Dataset and evaluation metric

Dataset: All experiments are evaluated on two benchmark datasets for human action recognition, HMDB51 [31] and UCF101 [47], which are illustrated in Fig. 7. HMDB51 dataset contains 6,766 video clips from 51 action categories, which covers most of the daily actions, such as hug, climbing and throwing ball. UCF101 dataset has 13,320 video fragments from 101 action categories, which has large action diversity. There exist large variations in camera motion, object appearance, pose, object scale, and so on. The source video clips of UCF101 are all from the YouTube, which are user generation content. The resolution of all clips is 320 × 240. The video quality of UCF101 is much better than HMDB51.

Performance metric: Similar to other works [13,48,15], the recognition accuracy is used as the performance metric. *Acc@K* is the percentage of actions whose truly matched videos are ranked in the top K results. To be consistent with other prior works, the top-1 accuracy is employed to evaluate the performance.

5.2. Implementation details

In order to verify the effectiveness of our model, three CNNs, i.e., BN-Inception [49], ResNet [16] and DenseNet [41] are treated as the baselines. Models pre-trained on ImageNet are used to initialize branches of SRL. The models on target dataset are fine-tuned to determine the weights of SRL. Due to its capability of adaptive updating the weights, Adam optimization strategy is chosen to update the weights of SRL. The trained SRL is utilized as the feature extractor to produce the features of sampling stacks for long-range temporal pattern modeling. To construct the temporal module, the layer of DB-LSTM is set to 3 in this paper. The random horizontal flipping and random-size cropping are applied to improve the generalization of the model. All the experiments are conducted on a workstation with an I7 processor and four NVIDIA TITAN-XP GPUs. Due to the limitation of the GPU memory, the batch sizes for backbone networks of BN-Inception, ResNet and DenseNet are set to 64, 32 and 32, respectively.

5.3. Model ablation studies

To better understanding the proposed DB-LSTM model, we conduct ablation experiments to verify the contribution of each component and latent spatial and temporal relationships.

5.3.1. Effect of segments

We first verify the effect of segments. Different combinations of segment strategies on training and testing are tested, and the best performance of combinations is listed in Table 1. In the training stage, sparse sampling strategy is adopted, which divides the videos into 3, 5 and 7 segments, respectively. Since there are only 1 to 4 shots in the videos from HMDB51, so 7 segments have already covered all shots. Accordingly, different dense sampling strategies are applied to perform the detection. All the experiments

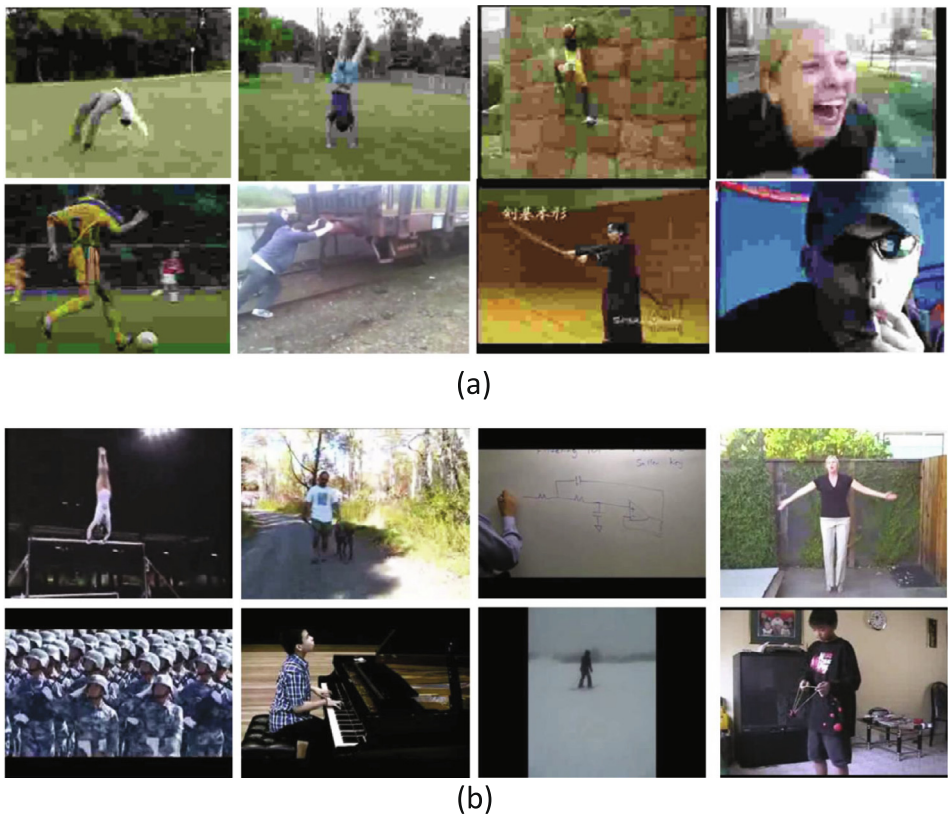

Fig. 7. Examples of HMDB51 (a) and UCF101 (b) datasets.

Table 1
Effect of segments for RGB and Optical Flow domains (the best performance is highlighted in bold).

# of segments		RGB	Optical Flow
Training	Testing		
3	15	52.7%	63.1%
5	25	53.7%	63.2%
7	28	54.5%	**63.5%**

are conducted with BN-Inception backbone network and original Bi-LSTM on the HMDB51 [31] dataset for fair comparison.

As can be seen, the segment strategy has a substantial impact on the performance. In the spatial domain, more segments will have better performance. It achieves the best performance when the training segments are set to 7 and the testing is 28. The dense sampling captures more information of the action, and the noise can be avoided by DB-LSTM to model the video-level temporal pattern.

5.3.2. Effect of temporal modeling

To evaluate the effect of temporal modeling, we compare the performance of LSTM, Bi-LSTM, Res-LSTM, Res-Bi-LSTM, Dense-LSTM and the proposed DB-LSTM for video-level temporal modeling. All experiments are conducted based on the backbone network BN-Inception. The training and testing segments are set to 7 and 28, respectively, as indicated in last subsection.

In addition, we compare with two state-of-the-art works, Two-Stream [50] and TSN [13], which do not consider the long-range temporal modeling. Since the original Two-Stream does not provide the ablation results on HMDB51, we copy its results from TS-LSTM [50], which is its baseline. The ablation results are not offered by TSN, so we reproduce the results by using the open source code.[1] For fair comparison, BN-Inception is utilized for SRL as the backbone network.

The results are listed in Table 2. Overall, the long-range temporal modeling based on LSTM outperforms TSN and Two-Stream. The improvement among different solutions is not significant on HMDB51. The densely connected structure demonstrates better performance than the residual learning methods. DB-LSTM achieves the best performance, in which dense connections contribute a lot. DB-LSTM can automatically optimize the connections in the training stage by tuning the connected weights.

5.3.3. Effect of fusion

We compare the fusion methods based on mean, max and the sliding window schemes, which are listed in Table 3. Generally, max pooling has a slightly better performance than mean pooling, meanwhile the fusion with sliding window outperforms other methods. This is because the sliding window fusion utilizes multiple scales of time windows to adaptively capture the actions over different time spans. On the contrary, the min and max fusion is based on the fixed length of time steps, which will divide a complete action into several snippets, or mix some irrelevant frames into the fusion procedure, causing the degraded effectiveness.

5.3.4. Effect of backbone network

The backbone network is a crucial part for human action recognition. In general, a powerful backbone network for image classification can bring better performance. Nevertheless, the better backbone network often brings the problem of over-fitting as indicated in [13], since it can fit to the mistaken context in the video. In

[1] https://github.com/yjxiong/tsn-pytorch.

Table 2
Performance comparison on temporal modeling (the best performance is highlighted in bold).

Model	RGB	Optical Flow
Two-Stream [50]	50.4%	59.7%
TSN [13]	51.9%	62.3%
LSTM [20]	52.2%	63.1%
Bi-LSTM [51]	52.9%	64.3%
Res-LSTM	53.1%	62.5%
Res-Bi-LSTM	53.2%	63.7%
Dense-LSTM	53.8%	63.8%
DB-LSTM	**53.5%**	**65.9%**

addition, as action recognition is affected by the visual and temporal aspects, it cannot be directly enhanced by the backbone network. To explore the impact of backbone network, we conduct the experiments of DB-LSTM based on BN-Inception and DenseNet-161. The results are listed in the last two lines of Table 3 and the third part of Table 4. As we can see that it overcomes the problem of over-fitting induced from the high-performance backbone network DenseNet-161, obtaining the best performance.

5.3.5. Case study

Two examples and their corresponding action detection scores are illustrated in Fig. 8. We compare the results of LSTM, Bi-LSTM, Res-Bi-LSTM, and DB-LSTM. The first video describes a scene that a man is holding a cup in his hand. At the beginning, it is hard to judge the action, probably they are talking. Later, a cup is appeared and then this man drinks it with a straw. It is a long-range action which cannot be recognized by visual appearance only. It is falsely detected as 'eating' by other methods. However, our proposed method can correctly detect it as 'drinking'. Similarly, the second scene describes the scene that a man runs toward a woman and then hug, which is a long-range action. It is not easy to recognize the action simply determining by a single frame or a span of shot. 'Sword' is ranked higher than 'hug' for other methods. Fortunately, our propose method accurately recognize it as 'hug', since it possesses the capability long-range modeling and is boosted by the dense skip-connections and Bi-direction temporal modeling.

5.4. Comparison with the state-of-the-art methods

After exploring the impact of stack sampling strategy and the video-level temporal modeling, we will compare the performance of the proposed DB-LSTM method with the state-of-the-art methods for action recognition, including traditional hand-crafted feature based methods and deep learning based approaches.

5.4.1. Baselines

Traditional hand-crafted feature based methods include:

- **DT + MVSV** [52] utilizes Multi-View Super Vector (MVSV) to encode the dense trajectory and other motion feature descriptors, which is composed of relatively independent components derived from a pair of descriptors.
- **iDT + FV** [8] corrects the dense trajectories descriptor to improve the representation of motion dynamics by applying camera motion estimation.
- **iDT + Hybrid-SV** [53] proposes a hybrid super vector to encode the output of the bag of visual words to boost the discrimination of features.
- **MoFAP** [54] captures the rich spatial–temporal structures of videos at multiple levels of granularity by a hierarchy of mid-level action elements.

Deep learning based methods are:

- **Two-Stream** [14] integrates the spatial and motion dynamics, which employs two branches of CNNs to model the RGB and optical flow maps, respectively.
- **FstCN** [55] is factorized convolutional networks with a sequential architecture of spatial and temporal convolutional layers, which can handle 3D signals more effectively.
- **Conv-Two-stream** [48] fuses the spatiotemporal information both in the feature and scores level to boost the performance.
- **TS-LSTM** [50] utilizes a temporal segment RNN and Inception-style temporal ConvNet to model the spatiotemporal dynamics.
- **TSN** [13] is a framework based on long-range temporal structure modeling, which combines a sparse temporal sampling strategy and video-level supervision for video-based action recognition.

Table 4
Performance comparison with state-of-the-art methods. (the best performance is highlighted in bold).

Model	UCF101	HMDB51
DT + MVSV [52]	83.5%	55.9%
iDT + FV [8]	85.9%	57.2%
iDT + Hybrid-SV [53]	87.9%	61.1%
MoFAP [54]	88.3%	61.7%
Two-stream [14]	88.0%	59.4%
FstCN [55]	88.1%	59.1%
Conv-Two-stream [48]	92.5%	65.4%
TSN [13]	94.0%	68.5%
TS-LSTM [50]	94.1%	69.0%
TLE [15]	95.6%	71.1%
S-TPNet + iDT [17]	96.0%	74.8%
IF-FTN [58]	96.2%	74.8%
HAF + BoW/FV [59]	–	82.5%
PA3D + I3D [60]	–	82.1%
DB-LSTM (BN-Inception)	94.2%	72.4%
DB-LSTM (DenseNet-161)	96.1%	73.7%
DB-LSTM + SSPF(DenseNet-161)	**96.5%**	**75.1%**
DB-LSTM (DenseNet-161 + kinectis)	97.1%	79.5%
DB-LSTM (I3D + Kinectis)	**97.3%**	**81.2%**

Table 3
Performances of different fusion methods (the best performance is highlighted in bold).

Model	Backbone	Mean	Max	Window
LSTM	BN-Inception	71.1%	71.3%	71.6%
Bi-LSTM	BN-Inception	71.2%	71.4%	71.7%
Res-LSTM	BN-Inception	71.3%	71.5%	71.7%
Res-Bi-LSTM	BN-Inception	71.5%	71.6%	71.9%
Dense-LSTM	BN-Inception	71.6%	71.7%	72.3%
DB-LSTM	BN-Inception	72.2%	71.4%	72.4%
DB-LSTM	DenseNet-161	73.2%	73.3%	**73.7%**

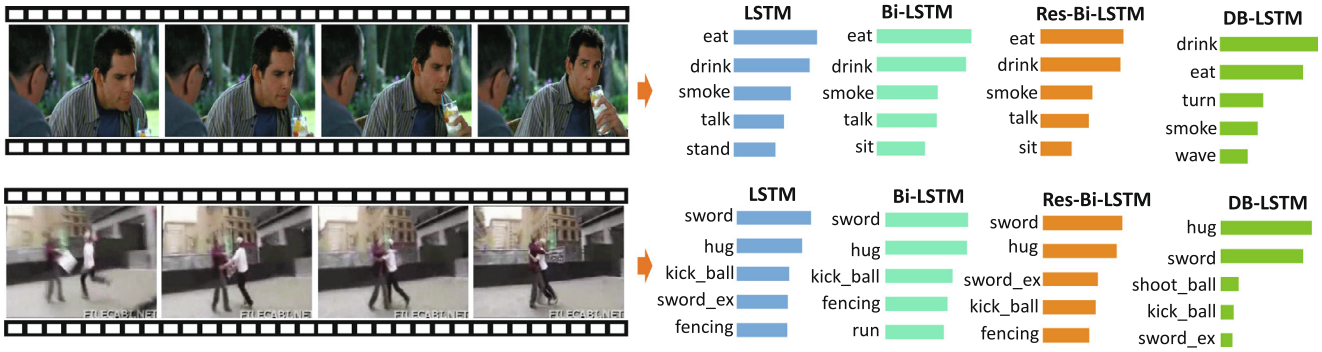

Fig. 8. Video examples and their corresponding detection scores of action recognition with different temporal modeling modules, including the original LSTM, Bi-LSTM, Res-Bi-LSTM and DB-LSTM.

- **TLE** [15] is a new compact feature representation, which captures the appearance and motion throughout entire videos, learning the semantics and a discriminative feature space.

There are several recent studies [24,38,56] evaluated on the HMDB51 and UCF101 datasets, which achieve more than 80% accuracy on HMDB51 dataset. However, these methods are pre-trained on a large video action dataset such as Sports-1M [57] or Kinectis [24]. For fairness, so we do not compare with these methods.

5.4.2. Performance comparison

The performance comparison is listed in Table 4, in which the first 4 methods are traditional hand-crafted feature based approaches, and the rests are deep learning based ones. From Table 4, we can see that traditional methods have relatively good performance in UCF101, which is around 85%. Unfortunately, the overall performance is pretty poor in HMDB51, below 62%, since HMDB51 dataset is more challenging than UCF101 dataset. The hand-crafted features, such as DT, iDT are shallow features. Although these features are encoded to reduce the noises and capture the latent relationship, they are not discriminative enough for accurate action recognition. The generalization ability of the hand-crafted features is limited with the shallow features, which cannot model the spatiotemporal information under the complex backgrounds and rapidly changed action dynamics.

The deep learning based methods achieve much better performance than traditional ones. The accuracy is nearly over 90% and 60% on UCF101 and HMDB51, respectively, since the deep features significantly improve the spatial and temporal representation. Two-stream is the first model based on the two-stream structure which models both the spatial and temporal information. It has better performance on UCF101, but not satisfactory result on HMDB51. Two-stream is a simple CNN based framework utilizing only optical flow descriptor and raw RGB data. The backbone network of Two-stream is a small scale of CNN which is hard to model the complex pattern. $F_{st}CN$ obtains 88.1% and 59.1% on UCF101 and HMDB51, respectively, which are not competitive results. This is because $F_{st}CN$ utilizes only the color modality without the optical flow. The performance is better than the original 3DCNN but lower than other improved two-stream liked methods. Conv-Two-Stream is a two-stream liked method, in which two modalities (RGB and optical flow) are integrated into the framework. Compare to the original two-stream, the 3D convolutional operation and 3D pooling are introduced to fuse the temporal dynamics and spatial appearance. The results also show that the fusion strategy is effective in improving the accuracy of human action recognition. TSN and TS-LSTM are also two-stream structure models which are based on the BN-Inception backbone network. They achieve around 94.1% and 69.0% top-1 accuracy on UCF101 and HMDB51, respectively, significantly boosting the performance. Deep features, especially the temporal modeling using optical flow map with CNN, significantly improve the representation of motion dynamics. The improved backbone network BN-Inception is a critical factor for the performance. BN-Inception has a much deeper network than VGG, and the multi-size kernel and the batch normalization operation also help to build a more robust and deeper network to obtain better performance. Although TS-LSTM combines the temporal segment network and LSTM, there is no obvious improvement. This also demonstrates that original LSTM cannot model the long-range temporal properties effectively. TLE adopts a deep encoder to encodes the deep two-Stream features and integrates the temporal information in the encoding module, which combines the merits of deep learning and the temporal feature encoding, achieving the best performance so far (71.1% top-1 accuracy in HMDB51).

The latest state-of-the art method on HMDB51 is [3], which achieves more than 82% accuracy. However, it is pre-trained on the large scale human action dataset kinetics [24]. This dataset contains more than 400K videos, which boosts the performance significantly. Furthermore, Kinetics contains part of the videos of HMDB51, making the Kinetics pre-trained models have improved performance on HMDB51. Moreover, it is a very computational complex method. Four modalities for data, i.e. I3D feature, IDT, RGB, and optical flow are combined as a single input. Besides, four feature encoding methods, Fisher Vector (first-order), Fisher Vector (second-order), BoW and HAF are adopted for feature encoding. Although the performance is significantly improved, the complexity of computing is over ten times compared to our method. This is the main reason that why the performances of these methods [60,59,61,62] have large improvements compared to ours. Actually, the only state-of-the-art model on HMDB51 without pre-trained on Kinetics is IF-TTN [58], which achieves 74.8% accuracy on HMDB51. This model leverages the multi-level features for temporal learning, integrating the low-level spatial dynamics and rich high-level semantic information. Nevertheless, it will cost more computing resources.

For a fair comparison, we also modify our model with multi-level feature-based temporal learning, which is illustrated in Fig. 9. It is because the high-level features carry rich semantic information but lose most of the spatial dynamics, so that the low-level features can be treated as a carrier of fast changing dynamics and the high-level ones carry slow motion dynamics [63]. Motivated by recent works [17,63], we upgrade our methods, where both the high-level and low-level features are integrated with our DB-LSTM. A long-range temporal learning module based on SlowFast Feature Pyramid (SFFP) is constructed, which is illustrated in Fig. 9. Three levels of the features are utilized as the input of SFFP. Features from low-level ones contain rich motion dynamics, which is beneficial for temporal learning. However, the noisy data in the low-level features are negative factor for the semantic

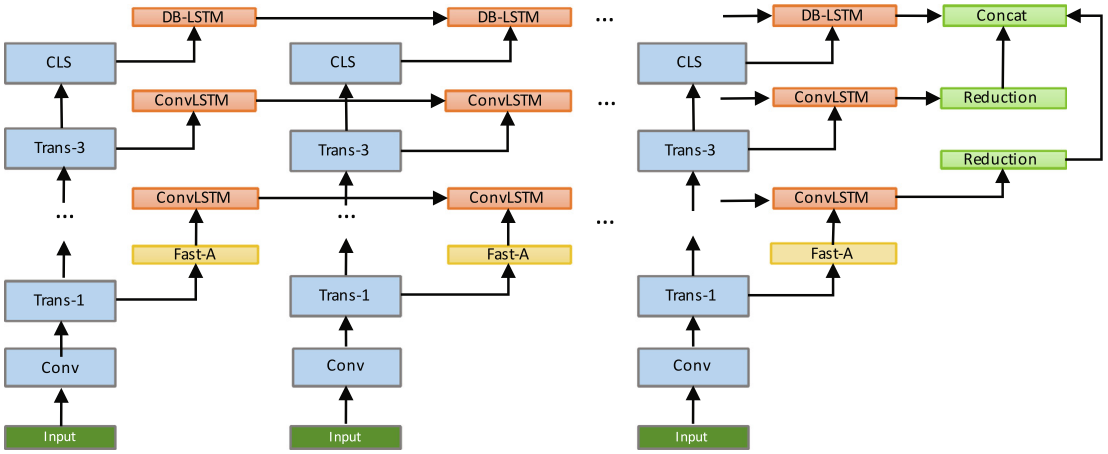

Fig. 9. The structure of the proposed SlowFast Feature Pyramid.

level learning. To address this issue, a Fast Attention (Fast-A) module is deployed before Conv-LSTM to filter the noisy data. FA is constructed by a weight learning branch and main branch. The weight learning branch is fast downsampled by two convolutional layers, and the global average pooling is adopted to learn the channel-wise attention weight. To address this issue, a Fast Attention (Fast-A) module is deployed before Conv-LSTM to filter the noisy data. FA is constructed by a weight learning branch and main branch, which is illustrated in Fig. 10. The weight learning branch is fast downsampled by two convolutional layers, and the global average pooling is adopted to learn the channel-wise attention weight. Furthermore, to fuse the output of three levels, the reduction module is designed to reduce the dimension of the feature map. Convolutional, ReLU and Batch Normalization layers are combined as a ConvBlock. Three and two ConvBlocks are stacked for the Trans-1 level and Trans-3 level, respectively. To reduce the computation resource, the ConvBlocks are downsampled eight times and twice and applied to the results of Trans-1 and Trans-3 layer, respectively. Finally, global average pooling is adopted to obtain the feature vector, which is fed into an FC layer to learn the spatial–temporal pattern of the action. Once the multi-level features are produced, they are concatenated as a feature vector to perform the final classification. The results of SFFP are listed in Table 4, which demonstrate the effectiveness of SFFP. On the other hand, the I3D backbone proposed in [64] and kinetics pre-trained model are also the main idea for recent state-of-the-art methods [60–62,59]. I3D backbone captures the spatial temporal. We also conduct the experiments on the kinetics pre-trained model with I3D backbone. Because the inputs of I3D are frame sequence, we sample 5 frames sequentially from each segment. The results are added to the manuscript, which has competitive performance as current methods.

Our proposed DB-LSTM model achieves competitive result compared to existing state-of-the-art approaches (without Kinectics pretrain models) on both UCF101 and HMDB51 datasets. It is a powerful evidence that long-range temporal pattern is crucial for action recognition. For UCF101 dataset, our method with BN-Inception network has competitive performance with other recent methods, such as TSN and TS-LSTM, which have the same backbone network. For the more challenging HMDB51, DB-LSTM improves the performance, almost 5% improvement compared to TSN and TS-LSTM. When DensNet-161 acts as the backbone network, DB-LSTM achieves the best performance in both datasets, i.e., 96.1% for UCF101 and 73.7% for HMDB51.

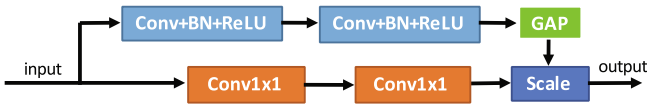

Fig. 10. The structure of Fast Attention module.

5.4.3. Result analysis

Overall, the performance on UCF101 is pretty good while the results on HMDB51 are not satisfactory. Obviously, due to the effectiveness of our model, the majority of the categories has more than 80% accuracy. Confuse matrix is an effective tool to analyze the results. Since the confusion matrix on the whole dataset is too large, we only illustrate the categories whose accuracies are lower than 60% in Fig. 11, which are easily to be incorrectly recognized. Therefore, only the categories with less than 50% Top-1 accuracy are selected, such as the sword, climb stair, wave, whose performance is 33%, 40%, 13%, respectively. The confusion matrix shows that the "sword exercise", "draw sword" and "sword" are hard to distinguish. t is because the spatial details of them are very similar, and their differences in the motions are small, which may be ignored by CNN. It means that this is a fine-grain level action recognition, which is still a challenge for existing models.

Two correctly detected and two falsely detected examples of DB-LSTM and their corresponding detection scores are illustrated in Fig. 12. Actions of "draw sword" and "cartwheel" are correctly recognized. The detection scores for each frame of the sampling stack of "draw sword" are pretty high, even for the first frame where the person is sitting on the grass. It is because the action of drawing sword is closely associated with the person sitting, which is learned by the model, and no other actions contain the sitting person. The scores of the second row are changed dramatically. The fourth frame is hand standing, which is very similar to flic-flac. For actions "walking" and "swing baseball", our model cannot correctly detect them since it is very confusion. The scene of walking only contains a person and no other critical context or background information is available. It is falsely detected as "running". The swing baseball is very similar to throwing in the motion dynamics. The spatial modeling can capture the context details, such as the sports clothes and the playground. This video is falsely detected as "throwing". In essence, the visual difference between these two actions is minor, which is also a problem for this dataset.

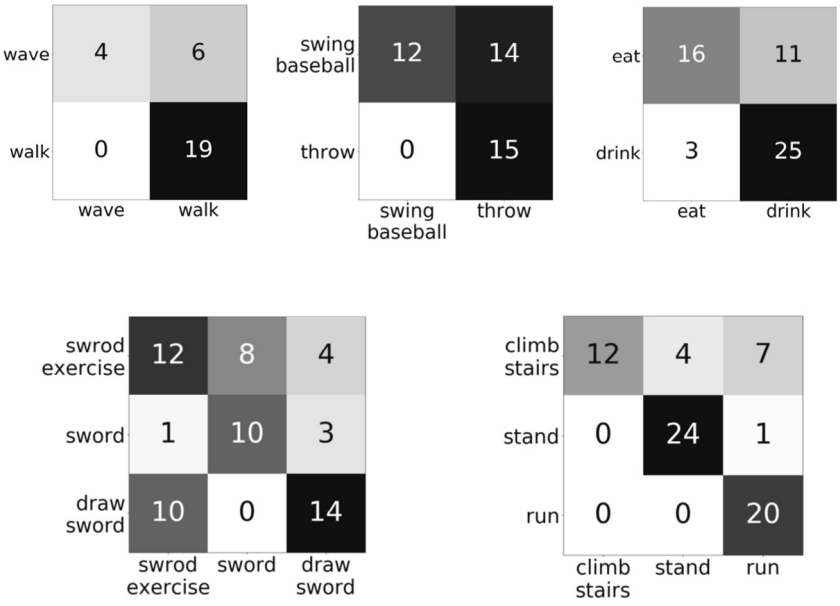

Fig. 11. Representative results produced by DB-LSTM on HMDB51 are construct the confuse matrix.

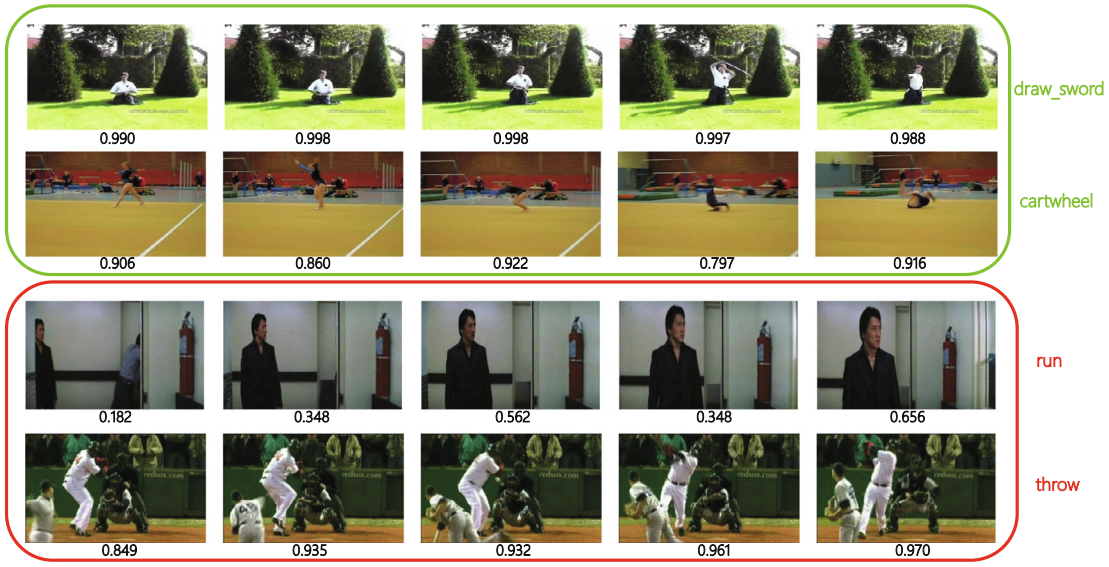

Fig. 12. Two correctly detected (in green) and two falsely detected (in red) examples of DB-LSTM and their corresponding detection scores.

5.5. Efficiency analysis

Although the proposed DB-LSTM method is more complex compared to TSN and other methods, it is efficient. All the experiments are conducted on a workstation with an I7 processor and four NVIDIA TITAN-XP GPUs. This model is based on the BN-Inception backbone network, which costs around 12 h for training on HMDB51 dataset. It has similar time cost with Two-Stream and TSN models. If the proposed model is based on DenseNet-161, which is several times deeper than the BN-Inception network, it will cost around 24 h for training. For a larger dataset, the time cost is proportional to the scale of the dataset. Our method has similar efficiency with the compared methods. Therefore, the training and testing time is within a reasonable range.

6. Conclusion

In this paper, we novelly propose a Densely-connected Bi-directional LSTM (DB-LSTM) to capture the visual and temporal patterns of human actions, exploring a variety of insights around the long-range temporal pattern modeling. The contribution of this work mainly comes from the proposed DB-LSTM framework, which integrates the merits of dense network, bi-directional modeling, and short-term as well as long-range temporal modeling. The experiments conducted on two publicly available datasets UCF101 and HMDB51, demonstrate that the proposed model outperforms the state-of-the-art approaches for action recognition. Although promising performance has been achieved, there are still many issues to be further explored. In the future, we will explore poten-

tial solutions on the sampling strategy, the architecture of temporal module and the score fusion strategy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the National Key Research and Development Project, National Natural Science Foundation of China (Grant No: 61772436), Sichuan Science and Technology Program (Grant No.: 2020YJ0207), Foundation for Department of Transportation of Henan Province (2019J-2-2), and the Fundamental Research Funds for the Central Universities (A0920502052001-3).

References

[1] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.
[2] S.S. Beauchemin, J.L. Barron, The computation of optical flow, ACM Comp. Surv. 27 (3) (1995) 433–466.
[3] T. Lindeberg, I. Laptev, On space-time interest points, Intl. J. Comp. Vis. 64 (2–3) (2005) 107–123.
[4] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proc. ACM Int. Conf. Multimedia, 2007, pp. 357–360..
[5] G. Willems, T. Tuytelaars, L. Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Proc. Eur. Conf. Comp. Vis., 2008, pp. 650–663..
[6] J. Liu, M. Shah, Learning human actions via information maximization, in: Proc. IEEE Intl. Conf. Comp. Vis. Pattern Recognit., 2008, pp. 1–8..
[7] A. Abdulmunem, Y. K. Lai, X. Sun, 3d gloh features for human action recognition, in: Proc. Intl. Conf. Pattern Recognit., 2017..
[8] H. Wang, C. Schmid, Action recognition with improved trajectories, in: Proc. IEEE Intl. Conf. Comp. Vis., 2014, pp. 3551–3558..
[9] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2015, pp. 4305–4314..
[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105..
[11] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proc. Intl. Conf. Learn. Represent..
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2015, pp. 1–9..
[13] 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, ACM Trans. Info. Syst. 22 (1) (2016) 20–36.
[14] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Proc. Adv. Neural Inf. Process. Syst. 1 (4) (2014) 568–576.
[15] A. Diba, V. Sharma, V. G. L., Deep temporal linear encoding networks, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2017, pp. 1541–1550..
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comp. Vis. Pattern Recog. (2016) 770–778.
[17] Z. Zheng, G. An, D. Wu, Q. Ruan, Spatial-temporal pyramid based convolutional neural network for action recognition, Neurocomputing 358 (2019) 446–455.
[18] B. Lin, B. Fang, W. Yang, J. Qian, Human action recognition based on spatio-temporal three-dimensional scattering transform descriptor and an improved vlad feature encoding algorithm, Neurocomputing..
[19] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, in: Proc. Intl. Conf. Mach. Learn., 2010, pp. 495–502..
[20] S. Hochreiter et al., Long short-term memory, Neural. Computat. 9 (8) (1997) 1735–1780.
[21] 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. 39 (4) (2017) 677–691.
[22] M. Majd, R. Safabakhsh, Correlational convolutional lstm for human action recognition, Neurocomputing..
[23] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in: Proc. IEEE Intl. Conf. Comp. Vis., 2017, pp. 5534–5542..
[24] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2017, pp. 4724–4733..
[25] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, C.G. Snoek, Videolstm convolves, attends and flows for action recognition, Comp. Vis. Image Understand. 166 (2018) 41–50.
[26] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, ACM Comp. Surv. 43 (3) (2011) 1–43.
[27] Y. Kong, Y. Fu, Human action recognition and prediction: A survey, arxiv abs/1806.11230..
[28] A. Klaser, C. Schmid, Action recognition by dense trajectories, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2011, pp. 3169–3176..
[29] F. Perronnin, T. Mensink, J. Verbeek, Image classification with the fisher vector: Theory and practice, Intl. J. Comp. Vis. 105 (3) (2013) 222–245.
[30] H. Jgou, F. Perronnin, M. Douze, J. Snchez, P. Prez, C. Schmid, Aggregating local image descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell 34 (9) (2012) 1704–1716.
[31] H. Kuehne, H. Jhuang, R. Stiefelhagen, T. Serre, Hmdb51: A large video database for human motion recognition, in: Proc. IEEE Intl. Conf. Comp. Vis., 2011, pp. 2556–2563..
[32] Z. Wu, X. Wang, Y. G. Jiang, H. Ye, X. Xue, Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, in: Proc. ACM Int. Conf. Multimedia, 2015, pp. 461–470..
[33] B. Zhou, A. Andonian, A. Torralba, Temporal relational reasoning in videos, arxiv abs/1711.08496..
[34] D. C. Luvizon, D. Picard, H. Tabia, 2d/3d pose estimation and action recognition using multitask deep learning, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2018, pp. 5137–5146..
[35] Y. Tang, Y. Tian, J. Lu, P. Li, J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2018, pp. 5323–5332..
[36] J. Liu, G. Wang, L.Y. Duan, K. Abdiyeva, A.C. Kot, Skeleton-based human action recognition with global context-aware attention lstm networks, IEEE Trans. Imag. Process. 27 (4) (2018) 1586–1599.
[37] J. Zhu, W. Zou, Z. Zhu, Y. Hu, Convolutional relation network for skeleton-based action recognition, Neurocomputing..
[38] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2018, pp. 6450–6459..
[39] B.K.P. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1–3) (1981) 185–203.
[40] Z. Ding, R. Xia, J. Yu, X. Li, J. Yang, Densely connected bidirectional LSTM with applications to sentence classification, in: Proc. Natur. Lang.Process. Chinese Comput., 2018, pp. 278–287..
[41] G. Huang, Z. Liu, K. Q. Weinberger, V. D. Laurens, Densely connected convolutional networks, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2017, pp. 2261–2269..
[42] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, Computer Science 3 (4) (2012) 212–223.
[43] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1510–1517.
[44] 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. 39 (4) (2017) 677–691.
[45] A. Graves, N. Jaitly, A. R. Mohamed, Hybrid speech recognition with deep bidirectional lstm, in: Proc. IEEE Conf Autom. Speech Recog. Understand., 2014, pp. 273–278..
[46] A. Van Den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: Proc. Intl. Conf. Mach. Learn., 2016, pp. 1747–1756..
[47] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, arxiv abs/1212.0402..
[48] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2016, pp. 1933–1941..
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2016, pp. 2818–2826..
[50] C. Y. Ma, M. H. Chen, Z. Kira, A. G, TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition, arxiv abs/1703.10667..
[51] M. Schuster et al., Bidirectional recurrent neural networks, Trans. Signal Process. 45 (11) (1997) 2673–2681.
[52] Z. Cai, L. Wang, X. Peng, Y. Qiao, Multi-view super vector for action recognition, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2014, pp. 596–603..
[53] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition, Comp. Vis. Image Understand. 150 (2016) 109–125.
[54] L. Wang, Y. Qiao, X. Tang, Mofap: A multi-level representation for action recognition, Int. J. Comput. Vision 119 (3) (2016) 254–271.
[55] L. Sun, K. Jia, D. Y. Yeung, B. E. Shi, Human action recognition using factorized spatio-temporal convolutional networks, in: Proc. IEEE Intl. Conf. Comp. Vis., 2015, pp. 4597–4605..
[56] L. Wang, W. Li, W. Li, L. Van Gool, Appearance-and-relation networks for video classification, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2018, pp. 1430–1439..
[57] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.F. Li, Large-scale video classification with convolutional neural networks, in: Proc. IEEE Conf. Comp. Vis. Pattern Recog., 2014, pp. 1725–1732..

[58] K. Yang, J. Fu, X. Guo, Y. Lu, P. Qiao, D. Li, Y. Dou, IF-TTN: information fused temporal transformation network for video action recognition, CoRR abs/1902.09928..
[59] D.Q.H. Lei Wang, Piotr Koniusz, Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns, in: Proc. IEEE Intl. Conf. Comp. Vis., 2019, pp. 8698–8708..
[60] PA3D: Pose-Action 3D Machine for Video Recognition..
[61] J. Zhu, Z. Zhu, W. Zou, End-to-end video-level representation learning for action recognition, in: Proc. Intl Conf.on Patt. Recog., 2018, pp. 645–650..
[62] A.T.M.S.R. AJ Piergiovanni, Anelia Angelova, Evolving space-time neural architectures for videos, in: Proc. IEEE Intl. Conf. Comp. Vis., 2019, pp. 1793–1802..
[63] J.M.K.H. Christoph Feichtenhofer, Haoqi Fan, Slowfast networks for video recognition, in: Proc. IEEE Intl. Conf. Comp. Vis., 2019, pp. 6202–6211..
[64] Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset..

Jun-Yan He is pursuing his Ph.D. degree from School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. He received the B.Sc. degree in Software Engineering from Southwest Jiaotong University in 2013. He was interned at Alibaba DAMO Academy in 2019. His research interests include computer vision, artificial intelligence and intelligent transportation systems.

Xiao Wu received the B.Eng. and M.S. degrees in computer science from Yunnan University, Yunnan, China, in 1999 and 2002, respectively, and the Ph.D. degree in Computer Science from City University of Hong Kong, Hong Kong in 2008. Currently, he is a Professor and the Assistant Dean of School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. He was with the Institute of Software, Chinese Academy of Sciences, Beijing, China, from 2001 to 2002. He was a Research Assistant and a Senior Research Associate at the City University of Hong Kong, Hong Kong, from 2003 to 2004, and 2007 to 2009, respectively. He was with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, and at School of Information and Computer Science, University of California, Irvine, CA, USA as a Visiting Scholar during 2006 to 2007 and 2015 to 2016, respectively. He has authored or co-authored more than 70 research papers in well-respected journals, such as TIP, TMM, TMI and prestigious proceedings like CVPR, ICCV and ACM MM. He received the Second Prize of Natural Science Award of the Ministry of Education, China in 2016 and the Second Prize of Science and Technology Progress Award of Henan Province, China in 2017. His research interests include artificial intelligence, computer vision, and multimedia information retrieval.

Zhi-Qi Cheng received the B.S. degree and PhD degree in computer science from Southwest Jiaotong University, China, in 2014 and 2019, respectively. Currently, he is a post-doc research fellow at School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. He was with Alibaba Inc., Hangzhou, China as an intern from 2015 to 2016, with the City University of Hong Kong, Hong Kong ad a Research Assistant from 2016 to 2017, and with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA as a visiting scholar during 2017 to 2019. His research interests include computer vision, artificial intelligence and image/video computing.

Zhaoquan Yuan received the B.S. degree from the School of Computer Science and Technology, University of Science and Technology of China (USTC), and the Ph. D. degree in Pattern Recognition and Intelligent System from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Currently, he is an Assistant Professor at the School of Information Science and Technology, Southwest Jiaotong University. His research interests include multimedia information retrieval and deep learning.

Yu-Gang Jiang received the PhD degree in Computer Science from City University of HongKong in 2009 and worked as a Postdoctoral Research Scientist at Columbia University, New York during 2009-2011. He is currently a Professor and Dean at School of Computer Science, Fudan University, Shanghai, China. His research lies in the areas of multimedia, computer vision and AI security. His work has led to many awards, including the inaugural ACM China Rising Star Award, the 2015 ACM SIGMM Rising Star Award, and the research award for excellent young scholars from NSF China.