Video Saliency Detection Using Bi-directional LSTM

Yang Chi[1] and Jinjiang Li[1,2]*

1 College of Electronic and Communications Engineering, Shandong Technology and Business University

Yantai, 264003 - CHN

[e-mail: chiyangsdibt@gmail.com]

2 College of Computer Science and Technology, Shandong Technology and Business University

Yantai, 264003 - CHN

[e-mail: lijinjiang@gmail.com]

*Corresponding author: Jinjiang Li

Received April 17, 2019; revised January 9, 2020; accepted March 16, 2020;

published June 30, 2020

Abstract

Significant detection of video can more rationally allocate computing resources and reduce the amount of computation to improve accuracy. Deep learning can extract the edge features of the image, providing technical support for video saliency. This paper proposes a new detection method. We combine the Convolutional Neural Network (CNN) and the Deep Bidirectional LSTM Network (DB-LSTM) to learn the spatio-temporal features by exploring the object motion information and object motion information to generate video. A continuous frame of significant images. We also analyzed the sample database and found that human attention and significant conversion are time-dependent, so we also considered the significance detection of video cross-frame. Finally, experiments show that our method is superior to other advanced methods.

Keywords: Visual saliency, Computer Graphics Deep learning, Deep two-way long-term short memory, Convolutional neural network

This research was supported by the National Natural Science Foundation of China (61772319, 61976125, 61976124), Shandong Natural Science Foundation of China (ZR2017MF049) and Yantai key research and development plan (2019XDHZ081).

http://doi.org/10.3837/tiis.2020.06.007 **ISSN : 1976-7277**

1. Introduction

In recent years, deep learning technology has been widely used in various fields such as classification, detection, recognition, retrieval, and speech processing. It has received extensive attention from academia and industry. At present, the commonly used deep learning networks include: AlexNet network, VGG network, GoogleNet network, ResNet network, Fully Convolutional Networks (FCN), Deconvolution Network (DN). Video saliency detection is a hot research direction in the field of computer vision. The main purpose is to realize the continuous extraction of motion-related saliency targets in video sequences by combining time and space information. Due to the variety of target motion patterns in the video sequence, complex scenes and camera motion, the video saliency detection has challenge.

At present, according to whether it is necessary to carry out training and learning, we divide the video saliency method into two methods based on the underlying cues and the learning-based methods. Among them, the video saliency detection method based on the underlying cues can be further divided into the method based on transform analysis. Based on the information theory method, the sparse representation method, the visual prior method and other methods, the learning-based method can be divided into two types: traditional learning method and deep learning method. The specific classification scheme is shown in **Fig. 1**. Recently, a new DNN-based idea has been proposed to detect image saliency, and the output data is directly obtained from the input end, thus avoiding preprocessing and feature extraction, and making the prediction result more accurate [1]-[7]. However, in the detection of video saliency, the application of DNN is really rare [8]-[10]. In fact, Cagdas [8] and others used the dual-stream CNN structure to take RGB frames and motion maps as inputs, and their work effectively combined CNN with video significance detection. Bazzani [9] and others trained the deep convolution 3D network to learn the characteristics of human attention by learning the LSTM network connected by the mixed density network to generate the saliency map of the Gaussian mixture distribution. Although these methods are based on DNN, they are not perfect enough for video saliency detection. They have the following disadvantages: (1) lack of sufficient data to train the DNN model; (2) cannot combine objects with motion information at the same time; (3) The significant difference caused by dynamic pixel conversion between consecutive frames of video has not received much attention. To improve the shortcomings of these methods, we propose a DNN-based bidirectional LSTM network approach to predict video saliency. Among them, the cyclic neural network is a building block of neurons connected to the input unit, the internal (or hidden) unit and the output unit, activated at time t, which can process the data in order. The output consists of a sequence of elements that are related to each other, and it processes only

one element at a time, so the output can be modeled. The RNN architecture performs well in processing and finding spatio-temporal data, especially for handling hidden patterns in audio and video. The RNN is performed sequentially in the data, and inputs are acquired during each time interval t, which inputs the previous hidden state S_{t1} and the new data X_t. The data is multiplied by the weight, the deviation is added and the activation function is entered. Due to the complicated calculation process, the impact of the initial input on the upcoming layers of data is negligible, causing the gradient problem to disappear. LSTM can effectively solve this problem. LSTM is mainly composed of storage unit, input gate, output gate and forget gate, it can maintain its status T_n, and the non-linear gating unit adjusts the information inflow/outflow unit. Because of the different variants of LSTM, such as the problem of processing continuous data, both the bidirectional LSTM and the multi-layer SLTM network are capable, so the researchers specifically emphasized this. In the proposed method, we analyzed the complex visual data patterns in each frame, which cannot be effectively identified using only a multi-layer LSTM network and a simple LSTM network.

In our proposed method, the characteristics of the video frame are analyzed for predicting pixel conversion of cross-frame video saliency. Using pre-trained AlexNet, the depth features of each frame in the video are extracted and then exploited in each forward and backward pass using DB-LSTM architecture to feature two layers for learning video frames. Since the video is divided into N time steps, the method can recognize the conversion of dynamic pixels in the long video. Because DB-LSTM has a high capacity for learning sequences and can effectively identify frame-to-frame feature changes while making small changes to the visual data of the video, this method is more suitable for video saliency detection. The rest of the structure of this paper is as follows: Part 2 outlines the work. Part 3 explains the framework we propose. Experimental results, technical assessments, and comparisons with other advanced methods are discussed in Section 4. Part 5 will combine this paper to look into the future research direction.

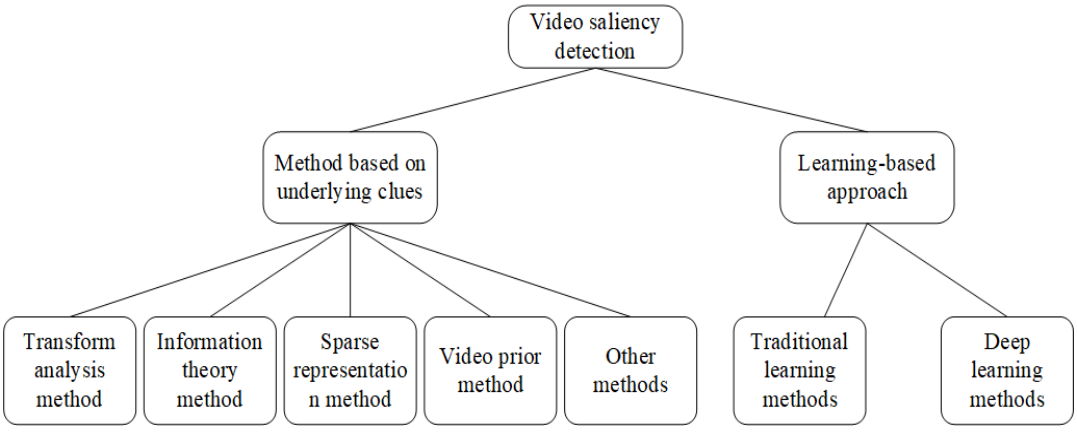

Fig. 1. Classification diagram of video saliency detection method

2. Related Work

In the past middle age, researchers have proposed many methods based on manual and deep networks to predict video saliency. In the video saliency detection method based on the underlying cues, Hou et al [11] proposed a simple and fast saliency detection method based on Fourier transform, called Spectral Residual (SR). The method uses the amplitude spectrum of the spectral residual to measure the significance of the image, and obtains better detection results. After experimental analysis, Guo et al [12]-[13] found the phase spectrum of the Fourier transform (Phase spectrum of Fourier Transform (PFT) can obtain better saliency detection results and reduce the computational complexity of the algorithm. The value of each pixel is represented as a quaternion composed of intensity, color and motion characteristics, and then the quaternion is utilized. The Phase Quaternion Fourier Transform (PQFT) calculates the spatiotemporal saliency of the video sequence. The experimental results show that the proposed method has better detection performance and is more robust to white noise, meeting the requirements of real-time processing, and is especially suitable for engineering applications. Cui et al [14] extended the SR method to the video field and proposed a fast motion saliency detection method, Temporal Spectral Residual. First, the significant target is automatically separated from the background using the time spectral residuals on the X-T and Y-T planes of the video segment, respectively. Then, the noise is suppressed using a threshold selection mechanism. Finally, the voting results are used to correct the experimental results. Unlike traditional complex background modeling, this method is based only on Fourier spectrum analysis and has good real-time performance. In the feature fusion strategy, Xi et al [15] extended the background prior in image saliency detection to the video domain, and proposed a video saliency target detection algorithm based on spatiotemporal background prior. Firstly, super-pixel segmentation is performed for each frame of video, and motion information is obtained by using SIFT stream estimation. Then, background prior information is extracted from time and space angles respectively, and the final time background prior is obtained after fusion. For the spatial background prior information, the traditional image processing method is still adopted. For the time background prior, the author adopts the energy function optimization method. Liu et al [16] used superpixels as the basic processing unit to obtain video saliency results by combining temporal saliency maps and spatial saliency maps. The method body performs calculations at the super-pixel level and assists with frame-level single-frame data as a global feature. In the learning-based video saliency detection method, recent deep learning is widely used in video saliency detection. Wang et al [10] proposed a video saliency target detection algorithm based on full convolutional network. This method proposes a new data enhancement method, which simulates according to the existing large number of labeled image data. Video training data is generated, which enables the algorithm network to learn a variety of saliency

information, preventing over-fitting problems caused by limited training data. With a large amount of training data (including 150K composite video sequences and real video data), The network has fully learned the significant clues of space and time, and can obtain accurate space-time significant estimation results. The deep learning-based approach can rely on enough data for a lot of pipeline work for feature extraction, so it can more accurately identify hidden patterns in the visual data of the video. But in another convenience, because deep learning requires a lot of data to train and requires high processing power, we consider the complexity and accuracy of the system when designing the algorithm, and the trans-pixel significant conversion of each frame of the video. Analysis is performed and spatio-temporal information is maintained through convolutional connections. In order to achieve better predictive video saliency, we intelligently combine CNN and LSTM due to its latest achievements in visual and temporal data.

3. Proposed framework

In this section, we discuss the main body of the experiment in detail and introduce its main structure. We combine a Convolutional Neural Network (CNN) with a deep bidirectional LSTM network, to process video data through this network architecture (see **Fig. 3** for network structure).

3.1 Database and analysis

We analyzed the LEDOV database used (which contains video material with at least one object, different video content, high quality and stable pixels, see **Fig. 2**), and found that people pay attention to video saliency mainly with objects. The moving objects and the moving parts in the objects are highly correlated. Therefore, the structure combining CNN and DB-LSTM is proposed. The model of CNN structure adopts the improved VGG19 network. The alternation of one convolutional layer and the largest pooling layer of our model, followed by three fully connected layers, and the activation function using Relu, in order to reduce the number of network training parameters, the entire convolution network uses a $\mathbf{3}\times\mathbf{3}$ size convolution. Unlike traditional convolutional LSTM, we consider prior knowledge based on saliency: center prior and sparse prior. Center a priori means that people tend to pay attention to the center when watching videos or pictures. To this end, we propose a central deviation:

$$
\begin{aligned}
I_l^t &= \sigma((H_l^{t-1} \circ Z_i^h) * W_i^h + (F^t \circ Z_i^f) * W_i^f + B_i), \\
A_l^t &= \sigma((H_l^{t-1} \circ Z_a^h) * W_a^h + (F^t \circ Z_a^f) * W_a^f + B_a), \\
O_l^t &= \sigma((H_l^{t-1} \circ Z_o^h) * W_o^h + (F^t \circ Z_o^f) * W_o^f + B_o), \\
G_l^t &= \tanh((H_l^{t-1} \circ Z_g^h) * W_g^h + (F^t \circ Z_g^f) * W_g^f + B_g), \\
M_l^t &= A_l^t \circ M_l^{t-1} + I_l^t \circ G_l^t, \quad H_l^t = O_l^t \circ \tanh(M_l^t)
\end{aligned}
\tag{1}
$$

where σ and tanh are the activation functions of sigmoid and hyperbolic tangent. $\{W_i^h, W_a^h, W_o^h, W_g^h, W_i^f, W_a^f, W_o^f, W_g^f\}$ and $\{B_i, B_a, B_o, B_g\}$ denote the kernel parameters of weight and bias at the corresponding convolutional layer; I_1^t, A_1^t and O_1^t are the gates of input, forget and output for frame t; G_1^t, M_1^t and H_1^t are the input modulation, memory cells and hidden states. F_{st}^t means at frame t, with CNN features as input, $\{Z_i^h, Z_a^h, Z_o^h, Z_g^h\}$ and $\{Z_i^f, Z_a^f, Z_o^f, Z_g^f\}$ are 2 sets of random masks for the hidden states and input features before convolutional operation.

Fig. 2. Examples of each type of video in the video database. Left to right video belongs to daily activities, sports, social events, artistic performances, man-made objects and animals

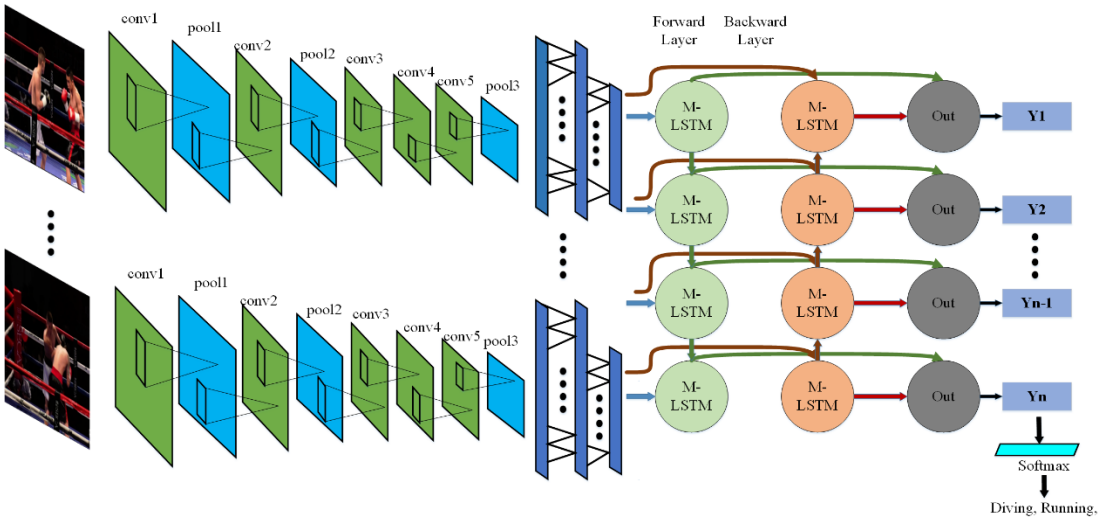

Fig. 3. Proposed DB-LSTM network framework for video saliency detection

Unlike normal Dropout, the dropout rates for all pixels in Center-bias Dropout are not the same, but based on a Center-bias map. In simple terms, the dropout rate of a central area pixel can be much lower than the dropout rate of the bounding area. Sparse a priori means

that there is a certain amount of sparsity in the human eye, and most existing algorithms ignore this sparsity and produce an overly dense saliency map. To this end, we designed a loss function based on sparsity.

$$L_{ss} - \mathrm{Conv}LSTM = \frac{1}{T}\sum_{i=1}^{T}\eta \cdot f_s \cdot D_{KL}(S_l^i, G_i),$$
$$f_s = \sum_i Hist_g(i)\log\frac{2Hist_g(i)}{Hist_g(i)+Hist_s(i)} + \sum_i Hist_s(i)\log\frac{2Hist_s(i)}{Hist_g(i)+Hist_s(i)} \quad (2)$$

In this loss function, not only the difference between the saliency map and the human eye focus map is calculated, but also the difference of the gray histogram distribution is calculated, so that the sparsity of the output saliency map tends to be true during the training process.

3.2 Recurrent neural network

RNN was introduced to analyze hidden order patterns in chronological and spatial order data [17]. In continuous video data, many frames constitute the movement of visual content, which constitute a sequence of frames that can help to understand the meaning of the continuous motion. In the long-term sequence, although the RNN can explain, the earlier input will be forgotten. This leads to a gradient disappearance problem, and an RNS with an LSTM structure can solve it. It is also a special type of RNN consisting of an input gate, an forgetting gate, and an output gate. You can learn long-term dependencies and control the pattern recognition of their sequences. During training, an S-type unit adjusts the door to learn its opening and closing [18]. In the operation explained following equation is performed in units LSTM, where X_t is the input of time t, and ft is the forgetting gate at time t, it can clear the information in the storage unit when necessary and allows the previous frame of the cleared information to be retained in the storage. The next new information is stored in the output gate O_t. G is calculated from the state of the current frame at time t and the state of the previous frame s at time t-1, which has an activation function tanh. We use the activation function tanh and the memory cell C_t to calculate the hidden pattern in the RNN sequence. Since the video saliency detection only requires the final result and does not require the output of the LSTM network intermediate process, we used the softmax classifier to determine the final state of the RNN network. A single LSTM unit does not recognize complex sequences, especially when a large amount of video data is being input. Thus, a plurality of stacking units LSTM long processing video data dependencies.

3.3 Multilayer LSTM

We have found through research that as the number of layers in the neural network model increases, the performance of deep neural networks increases. In order to enable the RNN to capture higher levels of sequence information [19], we superimpose the two LSTM layers in the two networks. Normally, the data is first transmitted to the single layer for activation and processing in the RNN. Then output, but for video saliency detection, we need to consider its chronological problem, so we have to put the data on multiple layers for processing. After the LSTM layer is stacked, the current layer in the RNN sends its hidden state to the next layer as the input to the layer, which is the same for each layer, greatly improving the efficiency of processing timing problems. The structure of the multilayer LSTM is shown in **Fig. 4**.

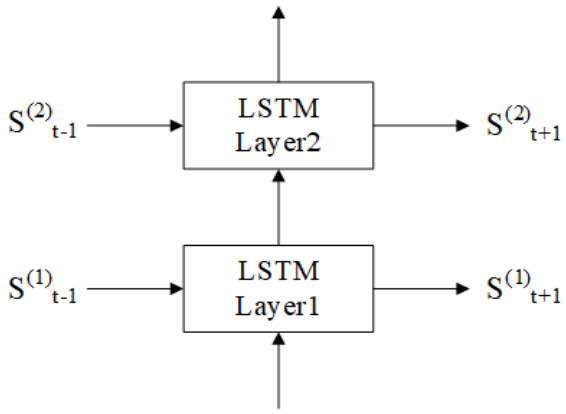

Fig. 4. Multilayer LSTM

3.4 Bidirectional LSTM

In bidirectional LSTM, the previous and subsequent frames in the sequence are directly related to time t, which together determine the output of t [20]. **Fig. 5** shows how the two-way LSTN works. The two-way RNN is very simple, with two RNNs stacked together. The two RNNs are oriented one after the other, extracting their hidden state to calculate the combined output. Our multi-layer LSTM has two hierarchies of forward and reverse transfer. **Fig. 6** shows the external network structure of the bidirectional LSTM, which is transmitted to the bidirectional RNN after data input, and then the hidden state through forward propagation and back propagation is coupled to the output layer. To verify the results, we used backpropagation to calculate weights and deviations on the output layer and separate 20% of the data in the data set. We used cross entropy when verifying the error calculation of the data. To minimize its cost, we used a random optimization with a learning rate of 0.001 [21] to control it. Since we have a bidirectional LSTM processing layer, we calculate the previous

frame and the next frame at time t1 to get the output frame at time t. Due to its mechanism of calculating output, our proposed method is more efficient than other state-of-the-art methods.

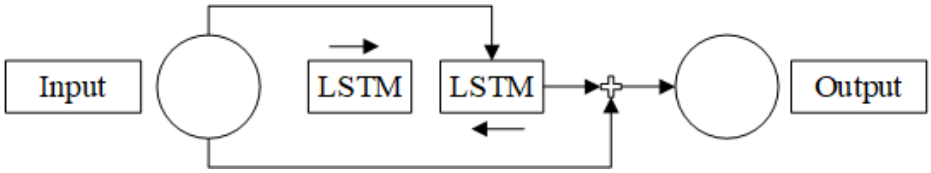

Fig. 5. Working principle of bidirectional LSTM

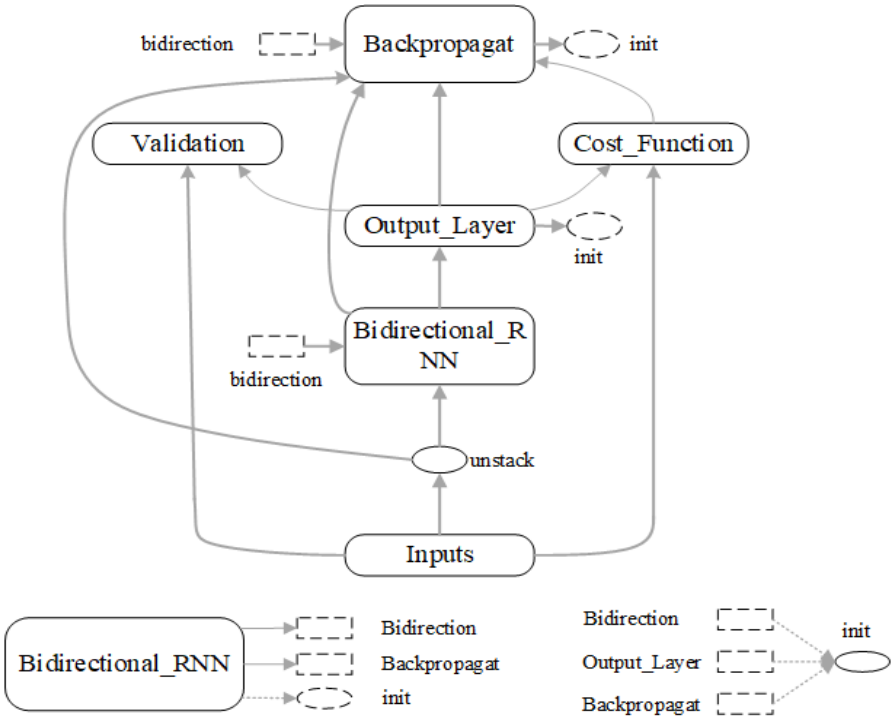

Fig. 6. Two-way LSTM network structure

3.5 Training process

To train the CNN, we use the loss function based on the Kullback-Leibler (KL) divergence to iterate over the parameters. When using the deep neural network training model to predict video saliency, KL divergence is more accurate and efficient than other indicators, which is proved by Huang [3]. Using the significant map as the probability distribution, we can measure the KL divergence between the fine saliency map S_f of CNN and the ground truth value G.

$$D_{KL}(G, S_f) = G \log \frac{G}{S_f} \tag{3}$$

The smaller the KL divergence, the higher the accuracy of the significance prediction. In addition, we found that the object area is related to the saliency area, so we calculated the KL divergence between the CNN's coarse graph Sc and the ground truth value G, and used it as a helper function to train the CNN. Then, we minimized the loss function below to train the CNN model.

$$L_{OM-CNN} = \frac{1}{1+\lambda} D_{KL}(G, S_f) + \frac{\lambda}{1+\lambda} D_{KL}(G, S_c) \tag{4}$$

λ is a hyperparameter that controls the weight of two KL divergence. We pre-trained CNN on AlexNet and then initialized the rest of the parameters with the Xavier initialization program [22]. At the same time, in order to ensure consistent results of the two-way LSTM training, we edited the video material used for training and divided them into segments of the same length. In addition, in order to extract the temporal and spatial characteristics of each video clip at frame T, we fixed the parameters of the CNN. As shown below, the average KL divergence on the T-frame is defined by the loss function of 2C-LSTM.

$$L_{2C\text{-}LSTM} = \frac{1}{T} \sum_{i=1}^{T} D_{KL}(S_l^i, G_i) \tag{5}$$

The saliency map finally generated by 2C-LSTM is $\{S_l^i\}_{i=1}^{T}$, and the ground truth value of the video saliency map is $\{G_i\}_{i=1}^{T}$. We use Xavier to initialize the kernel parameters of each LSTM unit.

4. Experiment

In this section, we will use the proposed method to verify the performance of predictive video saliency. Introduce our experimental setup and the network structure used in Section A. The accuracy of the significance predictions of LEDOV and the other two public databases is compared in Sections B and C.

4.1 Configure experimental parameters and network structure

In the experiment, we used the LEDOV dataset, which included training (436 videos), verification (41 videos) and tests (41 videos) for a total of 538 videos, which we randomly divided into three groups. To train DB-LSTM, we clip 456 videos into 24,685 video blocks. All videos have 16 frames and allow 10 frames to overlap for data enhancement. Among the

methods we propose, CNN is the main source of image representation and classification. After the video samples are input, each individual frame is extracted by CNN, and DB-LSTM is used to find continuous information between them. The parameters used in feature extraction are pre-trained by the CNN model on the ImageNet [23] dataset [24] The AlexNet network consists of five convolutional layers, three pool layers and three fully connected layers, each with a weight and using the linear rectification function ReLU as the activation function. The block of an input step of the RNN is composed of the characteristics of each frame, and the time interval block is fed back to the RNN.

4.2 LEDOV Data set evaluation

We analyzed the LEDOV database and found that the focus on video saliency is related to the temporal correlation of successive frames. We measured the CC value, which represents the linear correlation coefficient of the fixed graph between two consecutive frames, so that this correlation can be better quantified. If the fixed map of the current frame and the previous frame is Gc and Gp, from this you can calculate the CC value of the average fixed map on the video:

$$CC = \frac{1}{|V_c|}\sum_{c\in V_c}\frac{1}{|V_c|}\sum_{p\in V_p}\frac{Cov(N(G_c),N(G_p))}{Std(N(G_c))\cdot Std(N(G_p))},$$
$$N(G_c) = \frac{G_c - Mean(G_c)}{Std(G_c)} \tag{6}$$

Among them, all frames in the video are represented by V_c , and consecutive frames before the c frame are represented by V_p. Cov(), Std(), Mean()are covariance, standard deviation and averaging operators. **Fig. 7** shows the CC values of four types of objects in video significance detection. From the figure we can see that the CC value is much higher in time consistency than the single baseline. Explain that there is a high degree of temporal correlation in successive video frames. When we increase the distance between the two frames before and after, the time correlation of attention becomes smaller. Therefore, we can further verify its long- and short-term dependencies by focusing on video frames.

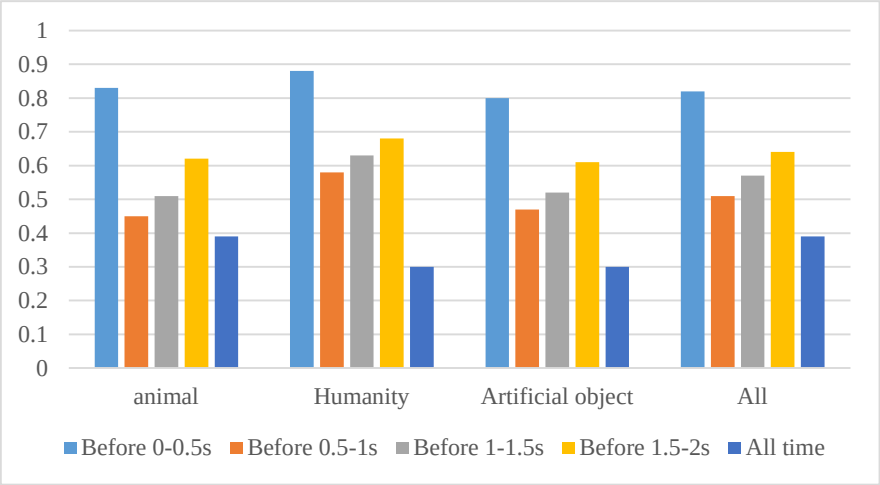

Fig. 7. CC results related to time attention

During the experiment, the significance of the determination result of video detection accuracy metrics are the following four: area under the receiver operating characteristic curve(AUC), normalized scan path significance(NSS), CC and KL divergence. The larger the AUC, NSS and CC values, the more accurate the prediction results, and the smaller KL divergence means better saliency prediction. We compared the other nine most advanced video saliency methods(GBVS [25], PQFT [13], Rudoy [26], OBDL [27], SALICON [3], Xu [28], BMS [29], SalGan [7], Lai [30]). We initialize the comparator and initialize the memory unit and hidden state by zero. From this **Table 1** can see that our method performs much better in all four metrics than the other methods. More specifically, our method improves the AUC, NSS, CC and KL by at least 0.04, 0.63, 0.09 and 0.43. We found that SALICON [3] and SalGAN [7], two DNN-based methods, have higher accuracy than other conventional methods. This shows that the significance of the significance-related functions automatically learned by DNN is more efficient than the manual method. Next, we turn to a comparison of video saliency predictions of subjective results. We demonstrate the saliency maps of the nine randomly selected videos in the test set, which are detected by us and nine other methods. In this figure, one frame is selected for each video. It can be seen from **Fig. 8** that our method is able to locate a significant area well, closer to the human ground map. In contrast, most other methods fail to accurately predict areas that attract attention.

Table 1. The mean (standard deviation) of the prediction accuracy of all the test videos in the video database with our other 9 methods.

	Our	GBVS [25]	PQFT [13]	Rudoy [26]	OBDL [27]	BMS [29]	SALG [7]	XU [28]	SALI [3]	Lai [30]
AUC	**0.91** (0.05)	0.84 (0.06)	0.70 (0.08)	0.80 (0.08)	0.80 (0.09)	0.76 (0.09)	0.81 (0.06)	0.83 (0.06)	0.85 (0.06)	0.87 (0.07)
NSS	**2.96** (0.87)	1.54 (0.74)	0.69 (0.46)	1.45 (0.64)	1.54 (0.84)	0.98 (0.48)	2.19 (0.59)	1.47 (0.47)	2.05 (0.87)	2.33 (0.45)
CC	**0.58** (0.13)	0.32 (0.13)	0.14 (0.08)	0.32 (0.14)	0.32 (0.16)	0.21 (0.09)	0.43 (0.09)	0.38 (0.11)	0.44 (0.13)	0.49 (0.11)
KL	**1.27** (0.42)	1.82 (0.39)	2.46 (0.39)	2.42 (1.53)	2.05 (0.74)	2.23 (0.39)	1.75 (0.32)	1.85 (0.30)	1.84 (0.42)	1.70 (0.45)

4.3 Evaluation of other databases

To evaluate the generalization capabilities of our approach, we compared the video saliency detection performance of our and other nine methods on all of SFU [31] and DIEM [32] videos, which are two widely available videos online. database. During the experiment, our test videos were from the SFU and DIEM databases, and the DB-LSTM model used for video significance detection was from the LEDOV training set. **Table 2** gives the results of our average and AUC, NSS, CC and KL for SFU and DIEM for the other nine methods. From this table we can see that the detection accuracy of our method is still better than the other nine methods, which is especially evident in the SFU database. In particular, the four indicators of AUC, NSS, CC and KL have at least 0.05, 0.46, 0.10 and 0.28 improvements. In addition, in **Fig. 9**, we selected some consecutive frames of pictures from a test video to test the video saliency. As shown, our method is closer to the ground truth than the other 9 methods. This means that our method has a good generalization ability in video saliency prediction, and we have greatly improved compared with the more advanced methods.

Fig. 8. We randomly selected 9 video saliency detection maps from the test set in the database, and each selected video only shows the results of one frame. We compared the other 9 methods and ground truth values.

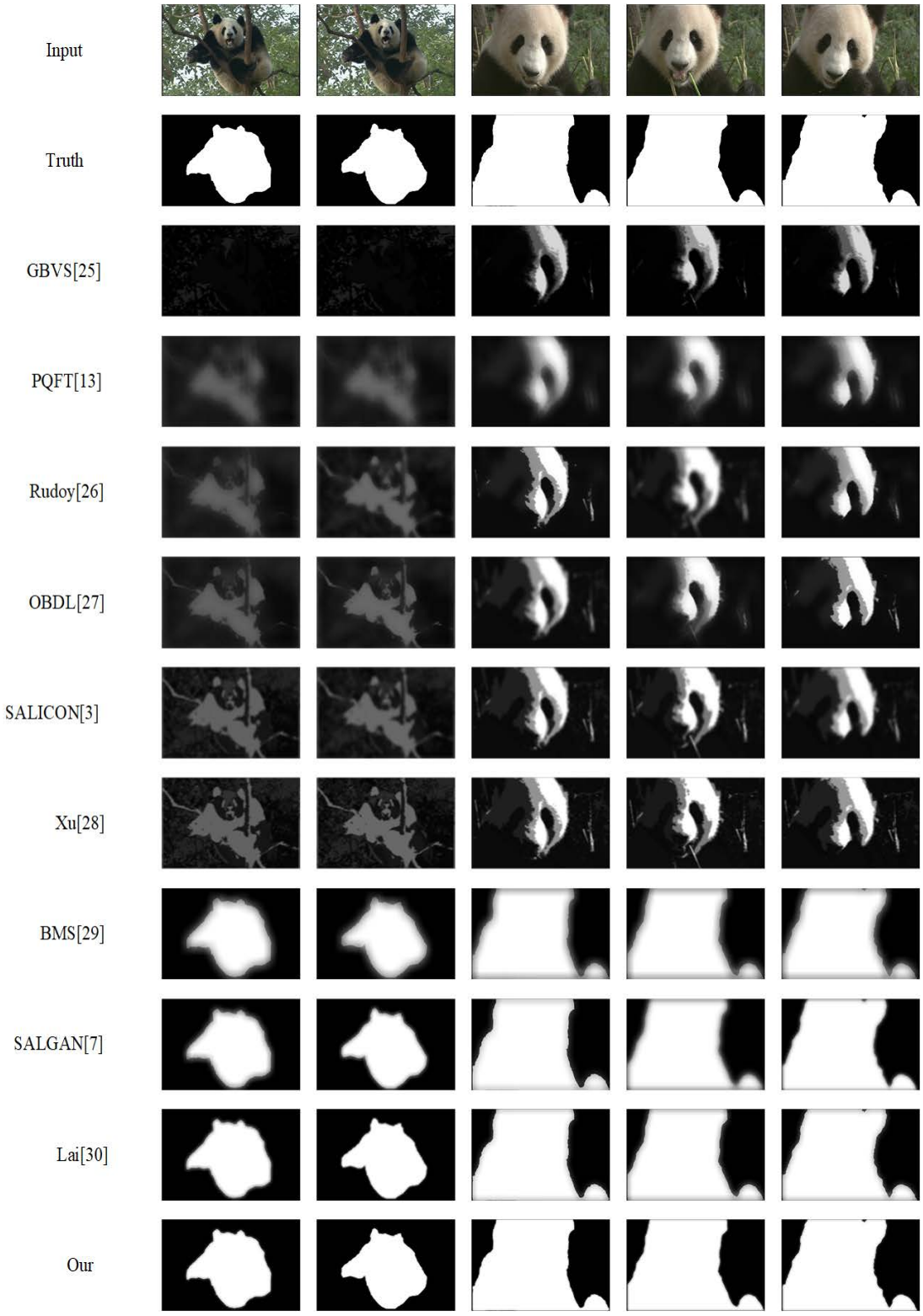

Fig. 9. A continuous frame video saliency of a single test video selected from the database. We compared the other 9 methods and the ground truth.

Table 2. The mean (standard deviation) of the detection accuracy of our and other methods in the SFU and DIEM databases.

SFU										
	Our	GBVS [25]	PQFT [13]	Rudoy [26]	OBDL [27]	BMS [29]	SALG [7]	Xu [28]	SALI [3]	Lai [30]
AUC	**0.82** (0.08)	0.72 (0.08)	0.62 (0.10)	0.74 (0.09)	0.75 (0.11)	0.76 (0.09)	0.71 (0.08)	0.67 (0.09)	0.65 (0.07)	0.77 (0.08)
NSS	**1.48** (0.67)	0.93 (0.49)	0.33 (0.36)	0.85 (0.47)	0.98 (0.66)	0.48 (0.62)	0.97 (0.41)	0.52 (0.33)	0.84 (0.60)	1.02 (0.49)
CC	**0.56** (0.16)	0.45 (0.16)	0.13 (0.16)	0.35 (0.16)	0.43 (0.22)	0.39 (0.23)	0.44 (0.13)	0.26 (0.12)	0.45 (0.14)	0.46 (0.14)
KL	**0.68** (0.25)	1.54 (0.20)	0.99 (0.28)	0.98 (0.37)	1.06 (0.34)	1.13 (1.77)	1.36 (0.26)	1.18 (0.21)	1.65 (0.43)	0.96 (0.26)
DIEM										
	Our	GBVS [25]	PQFT [13]	Rudoy [26]	OBDL [27]	BMS [29]	SALG [7]	Xu [28]	SALI [3]	Lai [30]
AUC	**0.87** (0.09)	0.82 (0.10)	0.72 (0.12)	0.81 (0.12)	0.76 (0.15)	0.80 (0.12)	0.81 (0.08)	0.81 (0.12)	0.83 (0.07)	0.86 (0.09)
NSS	**2.27** (1.18)	1.23 (0.84)	0.88 (0.73)	1.42 (0.85)	1.28 (1.05)	1.70 (1.06)	1.36 (0.09)	1.36 (0.82)	1.53 (0.89)	1.82 (0.73)
CC	**0.50** (0.22)	0.31 (0.19)	0.20 (0.15)	0.35 (0.21)	0.30 (0.23)	0.37 (0.20)	0.36 (0.08)	0.36 (0.18)	0.28 (0.14)	0.42 (0.14)
KL	**1.33** (0.58)	1.67 (0.51)	1.76 (0.47)	2.36 (2.08)	2.80 (1.61)	1.69 (0.61)	1.70 (0.10)	1.99 (1.16)	1.81 (0.45)	1.54 (0.44)

4.4 Performance analysis of significant predictions.

Since our proposed video saliency detection framework utilizes CNN's frame-level depth features and can be processed through DB-LSTM. The ability to extract CNN features from video frames and provide video frames to DB-LSTM helps identify complex frames to build hidden sequence patterns in features and reduces redundancy and complexity. Next, the DB-LSTM network is used to learn the sequence information between the frame features. Our proposed method is capable of learning long-term sequences, and we have stacked multiple layers in the DB-LSTM network to give it enough depth for forward and backward propagation. Since this allows analysis of video frame features at specific time intervals, it is possible to better handle lengthy video. To prevent this from happening due to the high dimensionality of DB-LSTM, we analyzed the Bayesian loss in DB-LSTM. Experimental results show that low loss rates can cause inappropriate problems, resulting in reduced accuracy of significant predictions. In order to influence the loss rate, we trained different DB-LSTM models with implicit loss rate ph and feature loss rate pf. The trained model was tested on the LEDOV verification set. When both ph and pf were set to 0.25, Bayesian loss can result in a reduction of approximately 0.03 KL. However, as ph and pf increase from

0.25 to 1, the KL divergence rises sharply. Therefore, we set ph and pf to 0.25. They may adjust based on the amount of training data.

5. Conclusion and future work

In this paper, we propose a new method for video saliency detection using deep neural networks. Developed CNN and DB-LSTM network structures, and innovatively developed bidirectional LSTM, enabling CNN extracted features to be cascaded forward and backward in LSTM, whether for intra-frame significance or inter-frame significance. We can all make better predictions. The detection of video samples in the video database LEDOV shows that our method is successful. The experimental results also confirmed that the DNN-based method is superior to the other nine most advanced methods in terms of AUC, CC, NSS and KL video saliency detection indicators.

In the future, our goal is to focus on the video saliency detection of some natural scenes, because they do not have particularly prominent objects, which will be a big challenge. In addition, we intend to extend this work to potential applications in perceptual video coding. Our method is used to locate significant and non-significant regions in the video, and it is desirable to improve the coding efficiency of the video by eliminating perceptual redundancy present in non-significant regions. As a result, the number of bits required to encode and transmit video is reduced, greatly reducing the bandwidth requirements in video transmission.

References

[1] Matthias K, Lucas T, and Matthias B, “Deep gaze i: Boosting saliency prediction with fea-ture maps trained on imagenet,” *arXiv preprint arXiv:1411.1045*, 2014. Article (CrossRef Link).

[2] Srinivas SS K, Kumar A, and Venkatesh B, “Deepfifix: A fully convolutional neural network for predicting human eye fifixations,” *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, September, 2017. Article (CrossRef Link).

[3] Xun H, Chengyao S, Xavier B, and Qi Z, “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in *Proc. of the IEEE International Conference on Computer Vision*, pp. 262–270, December 7-13, 2015. Article (CrossRef Link).

[4] Junting P, Elisa S, Xavier Giro-i N, Kevin M, and Noel E, “Shallow and deep convolutional networks for saliency prediction,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 598–606, 2016. Article (CrossRef Link).

[5] Xi L, Liming Z, Lina W, Ming-Hsuan Y, Fei W, Yueting Z, Haibin L, and Jingdong W, “Deepsaliency: Multi-task deep neural network model for salient object detection,” *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016. Article (CrossRef Link).

[6] Linzhao W, Lijun W, Huchuan L, Pingping Z, and Xiang R, "Saliency detection with recurrent fully convolutional networks," in *Proc. of European conference on computer vision. Springer*, pp. 825–841, September, 2016. Article (CrossRef Link).

[7] Junting P, Cristian Canton F, Kevin M, Noel E, Jordi T, Elisa S, and Xavier Giro-i N, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017. Article (CrossRef Link).

[8] Bak Ç Ş, Erdem A, Erdem E, "Two- stream convolutional networks for dynamic saliency prediction," *arXiv preprint arXiv:1607.04730*, vol. 2, no. 3, pp. 6, 2016. Article (CrossRef Link).

[9] Loris B, Hugo L, and Lorenzo T, "Recurrent mixture density network for spatiotempora visual attention," *arXiv preprint arXiv:1603.08199*, 2016. Article (CrossRef Link).

[10] Wenguan W, Jianbing S, and Ling S, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, January, 2017. Article (CrossRef Link).

[11] Xiaodi H and Liqing Z, "Saliency detection: A spectral residual approach," in *Proc. of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. Article (CrossRef Link).

[12] Chenlei G, Qi M, and Liming Z, "Spatiotemporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. Article (CrossRef Link).

[13] Chenlei G and Liming Z, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE transactions on image processing*, vol. 19, no. 1, pp. 185–198, August, 2009. Article (CrossRef Link).

[14] Xinyi C, Qingshan L, and Dimitris M, "Temporal spectral residual: fast motion saliency detection," in *Proc. of the 17th ACM international conference on Multimedia*, pp. 617–620, October, 2009. Article (CrossRef Link).

[15] Tao X, Wei Z, Han W, and Weisi L, "Salient object detection with spatiotemporal background priors for video," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3425–3436, November, 2016. Article (CrossRef Link).

[16] Zhi L, Xiang Z, Shuhua L, and Olivier Le M, "Superpixel-based spatiotemporal saliency detection," *IEEE transactions on circuits and systems for video technology*, vol. 24, no. 9, pp. 1522–1540, February, 2014. Article (CrossRef Link).

[17] Ken-ichi F and Yuichi N, "Approximation of dynamical systems by continuous time recurrent neural networks," *Neural networks*, vol. 6, no. 6, pp. 801–806, 1993. Article (CrossRef Link).

[18] Sepp H and Jurgen S, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, November, 1997. Article (CrossRef Link).

[19] Sak H, Senior A, Beaufays F, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. of Fifteenth annual conference of the international speech communication association*, 2014. Article (CrossRef Link).

[20] Atsunori O and Takaaki H, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks," *Speech Communication*, vol. 89, pp. 70–83, May, 2017. Article (CrossRef Link).

[21] Diederik P and Jimmy B, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. Article (CrossRef Link).

[22] Xavier G and Yoshua B, "Understanding the diffificulty of training deep feedforward neural networks," in *Proc. of the thirteenth international conference on artifificial intelligence and statistics*, pp. 249–256, 2010. Article (CrossRef Link).

[23] Jia D, Wei D, Richard S, Li-Jia L, Kai L, and Li F, "Imagenet: A large-scale hierarchical image database," in *Proc. of 2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, June 20-25, 2009. Article (CrossRef Link).

[24] Alex K, Ilya S, and Geoffrey E, "Imagenet classifification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012. Article (CrossRef Link).

[25] Jonathan H, Christof K, and Pietro P, "Graph-based visual saliency," *Advances in neural information processing systems*, pp. 545–552, 2007. Article (CrossRef Link).

[26] Dmitry R, Dan B, Eli S, and Lihi Z, "Learning video saliency from human gaze using candidate selection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1147–1154, June 23-28, 2013. Article (CrossRef Link).

[27] Sayed Hossein K, Nuno V, Ivan V, and Yufeng S, "How many bits does it take for a stimulus to be salient?" in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5501–5510, June 7-12, 2015. Article (CrossRef Link).

[28] Mai X, Lai J, Xiaoyan S, Zhaoting Y, and Zulin W, "Learning to detect video saliency with hevc features," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 369–385, November, 2016. Article (CrossRef Link).

[29] Jianming Z and Stan S, "Exploiting surroundedness for saliency detection: a boolean map approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 889–902, August, 2015. Article (CrossRef Link).

[30] Lai J, Mai X, and Zulin W, "Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm," *arXiv preprint arXiv:1709.06316*, 2017. Article (CrossRef Link).

[31] Hadi H, Mario J, and Ivan V, "Eye-tracking database for a set of standard video sequences," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 898–903, August, 2011. Article (CrossRef Link).

[32] Parag K, Tim J, Robin L, and John M, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cognitive Computation*, vol. 3, no. 1, pp. March 5–24, 2011. Article (CrossRef Link).

Yang Chi: He received the B.S. degree in computer science and technology from NanJing Tech University PuJiang Institute, NanJing, China in 2017. Currently, he is a M.S. degrees candidate in the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China. His research interests include computer graphics, computer vision, and image processing.

JINJIANG LI: He received the B.S. and M.S. degrees in computer science from Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, the Ph.D. degree in computer science from Shandong University, Jinan, China, in 2010. From 2004 to 2006, he was an assistant research fellow at the institute of computer science and technology of Peking University, Beijing, China. From 2012 to 2014, he was a Post-Doctoral Fellow at Tsinghua University, Beijing, China. He is currently a professor at the school of computer science and technology, Shandong Technology and Business University. His research interests include image processing, computer graphics, computer vision, and machine learning.