IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2019, accepted May 12, 2019, date of publication May 15, 2019, date of current version May 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916989

Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network

YUAN YUAN[1,2], (Senior Member, IEEE), HAOPENG LI[1,2], AND QI WANG[1,2], (Senior Member, IEEE)

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[2]Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Qi Wang (crabwq@nwpu.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002202, in part by the National Natural Science Foundation of China under Grant U1864204 and Grant 61773316, in part by the State Key Program of National Natural Science Foundation of China under Grant 61632018, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, and in part by the Project of Special Zone for National Defense Science and Technology Innovation.

ABSTRACT In this paper, a novel neural network named CRSum for the video summarization task is proposed. The proposed network integrates feature extraction, temporal modeling, and summary generation into an end-to-end architecture. Compared with previous work on this task, the proposed method owns three distinctive characteristics: 1) it for the first time leverages convolutional recurrent neural network for simultaneously modeling spatial and temporal structure of video for summarization; 2) thorough and delicate features of video are obtained in the proposed architecture by trainable three-dimension convolutional neural networks and feature fusion; and 3) a new loss function named Sobolev loss is defined, aiming to constrain the derivative of sequential data and exploit potential temporal structure of video. A series of experiments are conducted to prove the effectiveness of the proposed method. We further analyze our method from different aspects by well-designed experiments.

INDEX TERMS CRNN, CRSum, Sobolev loss, spatiotemporal modeling, video summarization.

I. INTRODUCTION

The amount of video data has been increasing exponentially for a few decades on account of the blossom of video media and a variety of video recording devices, such as digital video cameras, surveillance cameras, cell phones, drive recorders, etc. Statistically, 300 hours of video are uploaded to YouTube every minute on average. It takes approximately 50 years to watch all the uploaded videos in a single day. Large amounts of video data lead to two major problems: 1) the difficulty of retrieve valuable information conveyed by videos; 2) the extremely heavy burden of data storage. A plain idea to cope with those problems is that we generate a short version of the video that contains the important frames [1]–[9], subshots [4], [10]–[14], events [15]–[19] and objects [20], [21] that turn up in the original video. The short version is called as *video summary*.

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

A video summary of high quality would distill the crucial information from the original video and summarize it into a short watchable synopsis [4]. Three forms of summary are widely used for the video summarization task: keyframe [1]–[9], key subshot [4], [10]–[14] and time-lapse [13]. Among the three forms of summary, keyframe is most widely used since it emphasizes the exact key points in video and the generating process is simple. In this paper, we choose keyframe to form our summary.

For the last two decades, lots of research in automatic video summarization have been carried out. There has been a growing interest in the study of summarization techniques. Those techniques can be roughly classified into unsupervised approaches [2], [5], [15], [17], [18], [20]–[25] and supervised approaches [1], [3], [4], [6]–[14]. Unsupervised approaches depend on the picking rules defined by human. They have apparent disadvantages due to the complexity of the problem and the limitation of human cognition. On the other hand, supervised approaches utilize human annotations as guide

to learn the inner structure of a video and set reasonable scores to all the frames of videos. They have evident superiority compared to those unsupervised ones. In this paper we propose a supervised method for video summarization and consider it as a nonlinear regression problem.

As a considerably complicated task, video summarization needs thorough understanding of videos from their appearance to semantic meanings. In recent years, *deep neural network* (DNN) has been widely applied to many fields for its powerful capability of representation and fitting. There is a trend of summarizing videos using DNN. Two kinds of typical DNN, *convolutional neural network* (CNN), and *recurrent neural network* (RNN), are commonly employed in complicated tasks. It is worth mentioning that a novel architecture of DNN called *convolutional recurrent neural network* (CRNN) is brought up to handle some sequential-related tasks. The effectiveness of CRNN has been proven by experiments [26]–[28].

Traditional DNN based methods for video summarization usually have five drawbacks. 1) They use pretrained two-dimension (2D) CNNs (VGGNet [29] or GoogLeNet [30], etc) to extract features that are originally applied for the task of object recognition. It is not wise to directly employ those features to out task. 2) The output features of above mentioned 2D CNNs only contain the spatial information of each frame, and so the relation between frames are not exploited. 3) Those features are commonly known as deep features (i.e., semantic meanings), which means the shallow ones are always ignored in those methods. Such practice leads to a less comprehensive descriptions of frames. 4) The process of those methods is two-stage and does not coincide with human cognition which simultaneously receives and understands images. 5) MSE loss and cross entropy loss are commonly used to train those networks, indicating that the temporal structure in a video is not taken into consideration. A more comprehensive loss is required to bond with video summarization task.

Out of these drawbacks, we propose a novel method (**CRSum**) for video summarization in this paper. The **main contributions** of this paper are as follows:

- We set a new state-of-art-for video summarization as measured by F-measure on two public datasets. Our method integrates feature extraction, temporal modelling and summary generation into an end-to-end architecture.
- We for the first time leverage convolutional recurrent neural network for simultaneously modelling spatial and temporal structure of video for summarization, which coincides with human cognition and leads to better results.
- Thorough and delicate features of video are obtained in the proposed architecture by trainable three-dimension (3D) CNNs [31] and feature fusion. These features are especially learned for the task so as to improve the accuracy of our method.
- A new loss function is defined, aiming to constrain the derivative of sequential data and exploit potential temporal structure of video.

The rest of this paper is organized as follows. Section 2 elaborates typical work on video summarization and CRNN. Section 3 expounds our method of CRSum as well as Sobolev loss. Section 4 is the experiment part of this paper, where the effectiveness of our method is demonstrated and further analysis of our method is carried out. In Section 5, we make our conclusion.

II. RELATED WORK

As we explained in Section 1, video summarization techniques fall primarily into two categories, unsupervised approaches and supervised approaches. We first list some typical work in each category to make a brief exhibition of the progress in this field.

Unsupervised approaches of video summarization focus on designing the picking rules and how to quantify importance, or introducing various factors into their model [2], [5], [15], [17], [18], [20]–[25]. Reference [15] takes the notion of story into consideration to link sub-events when summarizing an egocentric video and introduces a comprehensive object function to quantify the quality of selected sub-shots. Reference [2] uses singular value decomposition (SVD) to derive the refined feature space for clustering and to define a metric for measuring the visual content contained in each cluster. Reference [5] proposes a generative architecture based on variational recurrent auto-encoders (VAE) and generative adversarial network (GAN) for unsupervised video summarization to select a subset of key frames. The architecture in [5] consists of a summarizer and a discriminator. The summarizer plays a role as an adversary of the discriminator, which is trained to maximally confuse the discriminator. Reference [16] proposes a new framework for video summarization ,where the features of human visual system are introduced to discriminate the perceptual significant events and eliminate perceptual redundancy. Those methods emphasize the fancy criteria which are well-defined to simulate the perception of human and their achievements are distinguished in the field of video summarization.

Supervised approaches of video summarization have drawn much attention to researchers in recent years in virtue of the access to big data and the development of deep learning [1], [3], [4], [6]–[14]. Reference [10] performs a semantically-consistent temporal segmentation on category-specific videos and assign an importance score to each segment by pre-trained SVMs, then generates a summary according to high scores. Reference [1] uses web-images as a prior to facilitate the process of creating summaries of user-generated videos and creates a crowd-sourcing based automatic evaluation framework to evaluate the results. Reference [11] uses a supervised approach to learn a comprehensive object function which takes interestingness, representativeness

and uniformity of selected segments into consideration. Reference [3] presents a Bag-of-Importance model to identify the importance of each local feature and extracts representative frames with more important local features. Reference [8] proposes a probabilistic model for diverse sequential subset selection named seqDPP which heeds the inherent sequential structure of videos and overcomes the deficiency of standard determinantal point processes (DPP). More flexible and powerful features are also used to represent frames in [8]. DPP are also exploited in [4], where the authors propose a novel method to summarize videos based on bidirectional LSTM, and take the diversity of selected frames into consideration using DPP. Besides, [4] mediates the demand that LSTMs require a large number of annotated samples by augmenting the training data with domain adaption, which reduces the data distribution discrepancy among datasets. Reference [6] proposes a novel supervised learning technique to select frames for video summarization by learning nonparametrically to transfer summary structures from training videos to test ones and generalizes the method to shot level summary. Reference [12] segments the video into superframes and select them according to their interestingness scores to form the summary of the video and introduces a benchmark (SumMe) that allows for automatic evaluation of video summarization methods. Reference [7] proposes a static video summarization method named VSUMM based on color feature extraction from video frames and k-means clustering algorithm. Besides, two annotated datasets, OVP and YouTube, are introduced in [7]. Reference [13] presents a highlight detection model by combining two deep convolutional networks architectures on spatial and temporal stream, followed by the pairwise deep ranking model for the training of each DCNN structure. Reference [14] combines semantic attributes and visual features as representations of frames, and uses bundling center clustering (BCC) to cluster video segments and pick a certain length of segments to generate summary. Generally speaking, supervised approaches perform better than those unsupervised since the summarizer can learn the deeper structure of video guided by human annotations.

In addition to exhibiting the work on video summarization, we refer to a few successful applications of CRNN. The idea of combining CNN and RNN has been proposed in other tasks, such as text recognition [26], video classification [27], human activity recognition [32]–[34], keyword spotting [28] and so on. Reference [26] considers an image with texts as a sequence, using CRNN to extract features and model sequential structure in the same time, greatly improving the performance of text recognition in natural scene. Reference [27] uses AlexNet and GoogLeNet to extract feature of each frame, and those features are processed forward and upwards through time and LSTMs respectively, where a softmax layer is added after LSTMs to predict the video class at each step. Reference [32] develops a sequential vector of locally aggregated descriptor (VLAD) to combine with CRNN architecture, which achieves good performance on UCF10 [35] and HMDB51 [36]. Based on VLAD, [33] proposes trajectory pooling and line pooling to address the problem in action recognition that the networks used are relatively shallow, which is the state-of-the-art method on UCF101. Reference [34] presents a generic deep framework for multimodal wearable activity recognition, which is base on CNN and LSTM, and the experiments demonstrate the satisfying efficacy of the framework. Reference [28] employs CRNN to exploit local structure and long-range context, making great progress in the task of key spotting. Overall, CRNN reveals its tremendous power for semantic-and-sequential-related tasks, which makes us conceive the thought of introducing CRNN into video summarization. We are the first to utilize CRNN in video summarization to the best of our knowledge. Moreover, there exist several spatiotemporal modeling methods that do not use CRNN but achieve great performance.

III. THE PROPOSED METHOD

In this section, we elaborate our approach of video summarization. First we explain the task mathematically and define some notations. Then we give a brief introduction to CRNN and 3D CNN. Next, we propose our distinctive architecture, viz., CRSum. At last, we show the limitation of traditional loss for video summarization and present a novel loss function called ''Sobolev loss''.

A. PROBLEM FORMULATION

A video is composed of consecutive frames. In this paper, a video is considered as an ordered set of frames and it is denoted as

$$\boldsymbol{V} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n\},$$

where $\boldsymbol{f}_i$ represents the i-th frame of the video and n is the number of frames. The purpose of video summarization by keyframe is to select a subset of $\boldsymbol{V}$, which has shorter length and contains almost all the important frames in $\boldsymbol{V}$. To achieve this purpose, a typical way is to assign an importance score to every frame and pick those frames with high scores. Apparently, this task can be divided into importance score prediction and summary generation.

In order to make that process of generating clear, a concept in mathematical statistics is adopted. Given a random variable X obeying certain population and $\forall\alpha \in (0, 1)$, x_α is α-upper quantile for variable X if

$$P\{X \geq x_\alpha\} = \alpha. \tag{1}$$

Using upper quantile, the process of creating a summary by importance scores can be easily described. Assuming a video $\boldsymbol{V}$ is labeled with scores $\boldsymbol{s} = \{s_1, s_2, \cdots, s_n\}$, we treat $\boldsymbol{s}$ as a sample from certain population of discrete values. $\forall\alpha \in (0, 1)$, the α-upper quantile of $\boldsymbol{s}$, s_α, can be calculated by definition. A subset of $\boldsymbol{V}$ is created as follows,

$$\boldsymbol{V}_s{}^\alpha = \{\boldsymbol{f}_i \in \boldsymbol{V} | s_i > s_\alpha, i = 1, 2, \ldots, n\}. \tag{2}$$

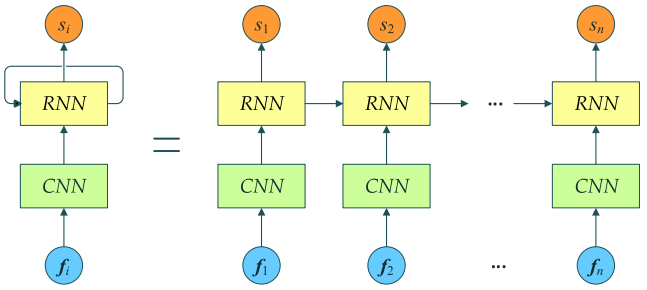


FIGURE 1. The basic CRNN structure. The compact structure of CRNN is in the left. Given a sequence of frames, CRNN directly figures out the scores of them. A certain frame is processed by 2D CNN first. The CNN outputs the features denoted by a vector. Then the features are sent to RNN. RNN combines the information of the current step and that of the last step, outputting the score of the current frame. The expanded structure is in the right, which demonstrates how CRNN works. The CNN/RNN blocks share the same state.

V_s^{α} is the generated summary of V based on scores s with synopsis ratio α.

B. CONVOLUTIONAL RECURRENT NEURAL NETWORK

As aforementioned, the essential of video summarization is quantifying importance. Approaches purely based on manually defined criteria have obvious weaknesses, because different people could share distinctive points of view and the human understanding of this task has limitations. A wise idea to deal with that problem is to let the machine learn the rules by the auxiliary of human. DNN reveals its power to handle sophisticated tasks due to its distinguished representation and fitting ability. In our method, we use CRNN, a special architecture of DNN, to achieve the task.

A basic structure of CRNN is exhibited in Fig. 1. Compared with traditional DNN, CRNN owns several distinctive advantages: 1) It directly learns from sequence of frames and labels, requiring no detailed annotations. 2) It has the same properties of learning informative representations directly from original frames as 2D CNN does, requiring neither hand-craft features nor pretraining steps. 3) It has the same properties of modelling temporal structure of video as RNN does. 4) It is naturally capable of handling videos in arbitrary lengths. 5) It is an end-to-end trainable network, making it easy to train and evaluate. Mathematically, the basic CRNN is presented as

$$s_i = RNN(CNN(\boldsymbol{f}_i), s_{i-1}), \quad i = 1, 2, \ldots, n, \tag{3}$$

which is an iterative procedure. CRNN combines the advantages of CNN and RNN, making itself quite effective for spatiotemporal modelling and applicable to video understanding.

C. 3D CONVOLUTIONAL NEURAL NETWORK

Previous methods utilize 2D CNNs to extract features of each frame. Spatial information contained in frame is easily obtained. However, video summarization task needs comprehensive understanding of the whole video instead of each single frame. 3D convolutional neural networks has been drawing attention to many researchers recently. It shows the power of extracting effective spatiotemporal features of video

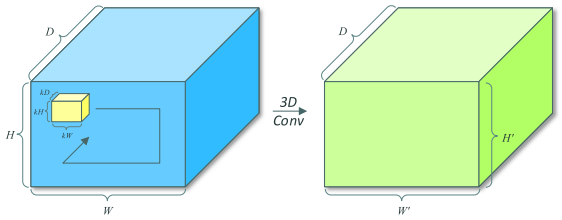


FIGURE 2. The process of 3D convolution. The blue cube is the input video volume. The yellow cube represents the 3D kernel that slides spatially and temporally in the video. The green cube is the output 3D feature. In this paper, we use padding to ensure the output feature has the same depth as the input volume does.

in many tasks, such as action recognition [37], [38], action similarity labeling, scene recognition [31], video classification [39], etc.

3D CNNs use 3D kernels to operate convolutions on the video cube spatially and temporally as shown in Fig. 2. It takes the whole video as input and outputs the features of video instead of features of each frame. Empirically, 3D CNNs achieve comprehensive understanding of video.

D. CRSUM FOR VIDEO SUMMARIZATION

Based on CRNN and 3D CNNs, we build our distinctive CRSum. Owning all the advantages of CRNN and 3D CNNs, it is especially designed for video summarization. The structure of CRSum is demonstrated in Fig. 3. As shown in Fig. 3, it is composed of 3D CNNs, RNNs and a multi-layer perceptron (MLP), taking the whole video as input and importance scores as output.

Three characteristics of the CNNs in CRSum are worth mentioning. 1) We use 3D CNNs instead of 2D CNNs to extract spatiotemporal features directly from video for video summarization. Previous methods extract features of each frame by 2D CNNs, which means the obtained features only contain the spatial information of each frame. By 3D CNNs, the local temporal dependencies are effectively exploited so that our network is capable of perceiving short-term structure in video. 2) Two 3D CNN blocks are utilized in our model. One has 3 layers for extracting shallow features while another one has 8 layers for deep ones. The shallow CNN is responsible for extracting low-level features. The deep CNN is applied to digging up the high-level information, i.e., semantic meanings, in video. 3) The parameters in our 3D CNNs are learnable while other methods use fixed deep features as the descriptors of frames. Hence the outputs of our 3D CNNs change adaptively according to different tasks. By the learnable features, the proposed network explores solutions in a broader solution space and generates summaries that contain more meaningful context. Further, both CNN blocks use (3, 5, 5) kernels with stride (1, 2, 2) and padding (1, 2, 2).

As the depth of the 3D feature map is the same as the length of video, we slice it along the depth dimension to obtain the feature of each frame. Learnable deep feature and shallow one are fused by concatenation before being sent into RNNs, which makes a more comprehensive representation of video. CRSum involves both kinds of feature so that a thorough understanding of video is obtained.

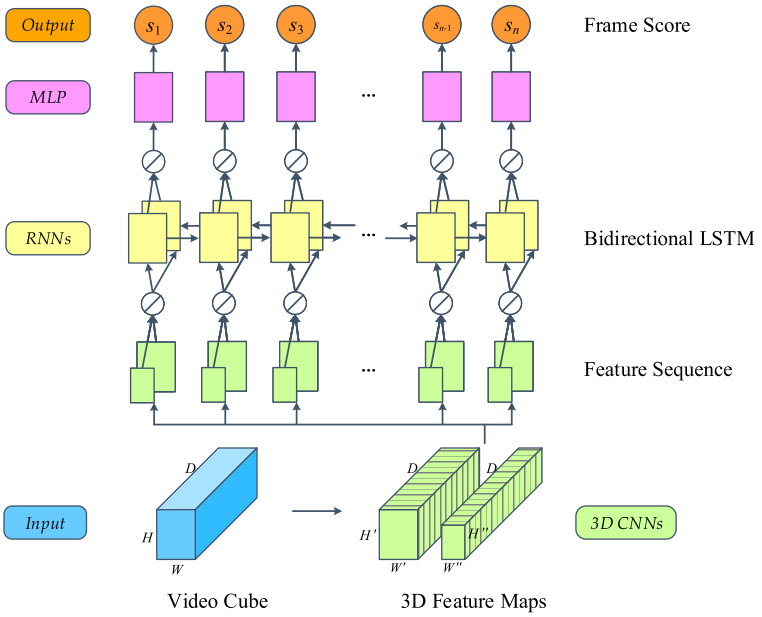


FIGURE 3. The structure of CRSum. It takes the whole video cube as input. First, the 3D CNNs extract deep and shallow features from the video (we use padding to ensure that the depth of the 3D feature map is the same as the length of video). Then the sliced 3D feature is sent to RNNs. The RNNs process the features and output a vector per step. At last, the vector is converted to a scaler (important score) by a MLP.

A bidirectional *long short-term memory* (LSTM) [40] block is utilized to model the sequential structure implied in video. A simple RNN struggles to learn long-term dependencies in video [41]. But LSTM [42], a special kind of RNN, is explicitly designed to avoid that problem. LSTM reveals its powerful ability of modelling such long-term dependencies in many fields. Moreover, compared with normal LSTM, a bidirectional LSTM owns one more block responsible for processing the inverse temporal information so as to fully exploit deeper informative structure in sequential data.

After passing through the RNNs, the features are converted by the MLP (with one hidden layer and activated by Sigmoid function) to a scalar ranging from 0 to 1, i.e., the importance score. Moreover, the MLP increases the representation and fitting ability of CRSum. The number of hidden units of LSTMs and the size of hidden layer of MLP are both 256.

E. SOBOLEV LOSS

A significant problem in machine learning is how to design an appropriate loss function coinciding with a specific task. For video summarization, the most widely used loss function are MSE loss and cross entropy loss.

Cross entropy loss models the task as a binary classification problem: whether to choose a certain frame. However, the information of frame is not fully exploited by cross entropy, making the summary quality relatively poor [4]. Besides, MSE loss has evident shortcomings for our task either, which is explained as follows.

Mathematically, MSE loss of continuous form quantifies the distance between two functions in L_2 space. Namely, assuming $s, t \in L_2(\Omega)(\Omega \subset \mathbb{R}^d)$, the MSE loss of s (predicted score) and t (target) is

$$MSE(s,t) = \|s-t\|^2_{L_2(\Omega)} = \int_\Omega (s-t)^2 \mathrm{d}\Omega. \tag{4}$$

As shown in Fig. 4, we consider t as the target curve while s_1, s_2 are two regression results of t. According to their conditions, we derive

$$MSE(s_1,t) = MSE(s_2,t), \tag{5}$$

which means s_1 and s_2 are equally fine estimations of t under MSE criterion. However, we prefer s_1 because it has the same trend as t does. In other words, the derivatives of them are closer, directly making the generated summary more accurate (the summary from s_1 overlaps more with that from t than s_2 does as demonstrated in Fig. 4).

In order to constrain the derivatives of sequential data, we introduce a novel gradient-based content loss function named Sobolev loss. The proposed loss is based on Sobolev space [43] which is a function space equipped with a norm that is a combination of L_2-norms of the function itself and its derivatives. Rigorously, $H(\Omega)$ is Sobolev space if

$$H(\Omega) = \left\{ u \in L_2(\Omega) \middle| \frac{\partial u}{\partial x_k} \in L_2(\Omega), k = 1, 2, \ldots, d \right\}, \tag{6}$$

and the norm in $H(\Omega)$ is defined as

$$\begin{aligned} \|u\|_{H(\Omega)} &= \left(\|u\|^2_{L_2(\Omega)} + \sum_{k=1}^{d} \left\| \frac{\partial u}{\partial x_k} \right\|^2_{L_2(\Omega)} \right)^{\frac{1}{2}} \\ &= \left[\int_\Omega u^2 d\Omega + \sum_{k=1}^{d} \int_\Omega \left(\frac{\partial u}{\partial x_k} \right)^2 \mathrm{d}\Omega \right]^{\frac{1}{2}}. \end{aligned} \tag{7}$$

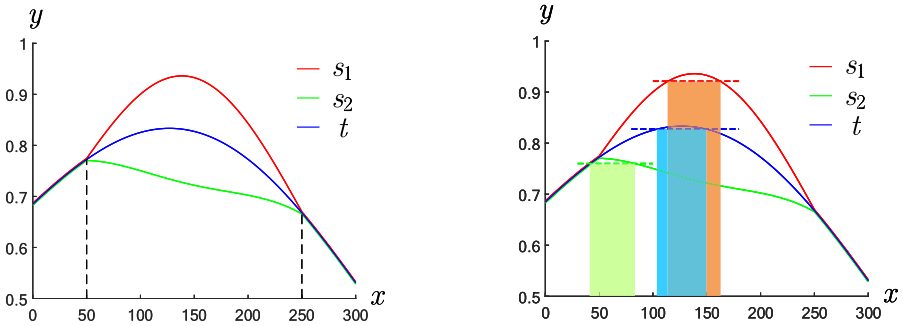


FIGURE 4. A toy example to show the limitation of MSE. The left figure includes three functions, s_1, s_2 and t, drawn by synthesis data. We force them to satisfy $s_1(x) \equiv s_2(x) \equiv t(x), x \in [0, 50] \cup [250, 300]$ and $s_1(x) - t(x) \equiv t(x) - s_2(x), x \in [50, 250]$. It is to easy to prove $\|s_1 - t\|^2_{L_2([0,300])} = \|s_2 - t\|^2_{L_2([0,300])}$. The right one demonstrates the process of generating summaries using different regression results.

The norm defined in $H(\Omega)$ includes terms corresponding to derivatives, indicating that the distance between two functions under Sobolev criterion is affected by their derivatives. A fitting result is better when the distance between the original data and the distance between the derivative data are shorter simultaneously. Hence, this criterion quite agrees with our motivation.

Therefore, we define a novel loss function named Sobolev loss based on Sobolev space as follows,

$$SobLoss(s, t) = \|s - t\|^2_{H(\Omega)}. \tag{8}$$

MSE loss considers frames as unrelated individuals and ignores their dependencies. Compared to MSE loss, Sobolev loss includes the constraint of derivatives which relate frames with their neighbors. So the local temporal dependencies are modeled in our loss function. Sobolev loss is more applicable to sequential data because it models the temporal structure more sophisticatedly by introducing a term corresponding to the derivatives, leading to more precise predictions.

Particularly, when $s = \{s_i\}_{i=1}^n$ and $t = \{t_i\}_{i=1}^n$ are discrete,

$$SobLoss(s, t) = \frac{1}{n}\sum_{i=1}^{n}(s_i - t_i)^2 + \frac{1}{n-1}\sum_{i=1}^{n-1}[(s_{i+1} - s_i) - (t_{i+1} - t_i)]^2. \tag{9}$$

It is Eq. (9) that we use to train CRSum. Video summarization needs temporal modelling of video. So we involve the sequential character in the loss function.

IV. EXPERIMENTS

A. EXPERIMENT SETUPS

In this section, we conduct a series of experiments to demonstrate the effectiveness of the proposed method. First, we explain the experiment setups. Then, the proposed method is evaluated and compared with several existing methods on two public datasets. In addition, we carry out further analysis of our method from different aspects by well-designed experiments.

1) DATASETS

Two public datasets for video summarization are commonly used: SumMe [12] and TVSum50 [25]. SumMe includes 25 videos, each of which is annotated by 15 to 18 people. They give segment summaries for each video and according to the segments, every frame is assigned an important score computed as the ratio of selections over views. The lengths of videos in SumMe vary from dozens of seconds to a few minutes and the topics include sports, sceneries, stuffs in life, etc., captured by egocentric, moving or static cameras. TVSum50 includes 50 videos collected from YouTube, each of which is annotated by 20 people. Each uniform length subshot of videos is assigned 20 importance scores and the ground truth score is the average of them. We use the scores of subshots as the frame scores. Video lengths in TVSum50 range from 2 to 10 minutes and 10 categories are covered. VTW [49] is a larger dataset that is originally proposed for video captioning. It contains 18,100 videos and 2,529 of which are annotated with shot-based video highlights, which are transferred into importance scores for evaluation.

2) EVALUATION METRICS

Plainly, a generated summary is deemed to be better when it is more similar to the ground truth one. We use F-measure to quantify the similarity of two summary. Given a generated summary $\boldsymbol{V_s}^\alpha$ and the ground truth summary $\boldsymbol{V_g}^\alpha$, the precision (P) and the recall (R) are defined as

$$P = \frac{|\boldsymbol{V_s}^\alpha \cap \boldsymbol{V_g}^\alpha|}{|\boldsymbol{V_s}^\alpha|}, \quad R = \frac{|\boldsymbol{V_s}^\alpha \cap \boldsymbol{V_g}^\alpha|}{|\boldsymbol{V_g}^\alpha|}, \tag{10}$$

where $|\cdot|$ is the counting measure of finite set. The F-measure is the harmonic mean of P and R, denoted as

$$F = 2 \cdot \frac{P \times R}{P + R}, \tag{11}$$

where F is defined to be 0 when $P = R = 0$. Obviously, $F = 0$ when $\boldsymbol{V_s}^\alpha$ and $\boldsymbol{V_s}^\alpha$ have no common element, and $F = 1$ when $\boldsymbol{V_s}^\alpha$ and $\boldsymbol{V_s}^\alpha$ are the same. It is F-measure that we use to quantify the performance of a certain method.

TABLE 1. Performance (F-measure) of several video summarization methods. The best and the second best results on two datasets are in bold and underlined respectively.

Methods	Supervised	DNN	SumMe	TVSum50	VTW
TVSum [25]			——	0.500	——
LiveLight [24]			0.384	0.477	——
VSUMM [7]	√		0.335	0.391	——
DR-DSN [44]		√	0.414	0.576	——
Framework [45]	√		0.431	0.527	——
Transfer [6]	√		0.409	0.541	——
dppLSTM [4]	√	√	0.386	0.547	0.443
H-RNN [46]	√	√	0.421	0.579	0.465
SUM-GAN [5]	√	√	0.417	0.563	——
re-seq2seq [47]	√	√	0.423	0.572	<u>0.480</u>
SASUM [48]	√	√	<u>0.453</u>	**0.582**	——
CRSum (Ours)	√	√	**0.473**	<u>0.580</u>	**0.483**

3) EXPERIMENT DETAILS

For each dataset, we use 5-fold cross-validation strategy to train and evaluate our model. Each video is sub-sampled every 60 frames in consideration of the similarity between adjacent frames, and resized to 256 × 256 for the limitation of GPU memory. The synopsis ratio is 15% as in [4], [12], [25]. As for optimization, we use mini-batch Adam algorithm [50] with initial learning rate 10^{-3} and $\boldsymbol{\beta} = (0.9, 0.999)$. The batch size is set to 8. The proposed model is trained for 10000 epochs. The learning rate is halved every 2000 epochs. Our method is implemented by PyTorch on NVIDIA GeForce GTX 1080 Ti.

B. MAIN RESULTS

The proposed method is compared with nine previous typical video summarization methods on SumMe and TVSum50. We use common **canonical setting** [4] to conduct those methods and report the results in Table 1. The compared methods are different in many ways. LiveLight [24] and TVSum [25], DR-DSN [44] are unsupervised methods, while VSUMM [7], dppLSTM [4], SUM-GAN [5], Transfer [6], Framework [45], H-RNN [46], SASUM [48] (the state-of-the-art method), re-seq2seq [47] and our method are supervised methods of video summarization. In addition, DR-DSN [44], dppLSTM, H-RNN, SUM-GAN, SASUM, re-seq2seq [47] and our method utilize DNN to predict the importance scores for summary generation, while other methods formula the task as optimization problems.

As shown in Table 1, our method achieves the best performance on SumMe and the second best on TVSum50. On SumMe, CRSum surpasses SASUM by 4.4%, which is a significant improvement on video summarization, while SASUM surpasses our method by only 0.3%, which can be ignored in practice. It is worth mentioning that SumMe contains relatively less videos, leading to the lack of training samples when conducting experiments on it. But CRSum still outperforms other methods, which indicates that it is capable of learning the inner structure of video from small amounts of samples and has strong generalization ability. We compare our method with three typical methods on VTW. The results show that our method surpass previous methods. The experiments on VTW indicate that our method is also capable of dealing with large-scale datasets. Examples of generated summaries by our method is shown in Fig. 5. As show in Fig. 5, our method is capable of selecting frames with high scores while maintaining diversity in summaries, which directly demonstrates the effectiveness of our method.

C. FURTHER ANALYSIS

The effectiveness of the proposed method is evident according to above experiments. Furthermore, we discuss more details to analyze our method from different aspects.

How important is Sobolev loss?

In order to clarify the effectiveness of Sobolev loss, we extend the definition of Sobolev loss as follows,

$$SobLoss_\lambda(s,t) = \frac{1}{n}\sum_{i=1}^{n}(s_i - t_i) + \frac{\lambda}{n-1}\sum_{i=1}^{n-1}[(s_{i+1} - s_i) - (t_{i+1} - t_i)]^2, \tag{12}$$

where λ is introduced to control the effect of derivatives. The extended Sobolev loss equals the original one when $\lambda = 1$ and equals MSE loss when $\lambda = 0$. λ is set to different values to illustrate how Sobolev loss affects the performance. The results are shown in Fig. 6 ($\log 0 \stackrel{\text{def}}{=} -\infty$).

CRSum trained by Sobolev loss with appropriate values of λ surpasses that trained by pure MSE loss (i.e., $\lambda = 0$). It proves the effectiveness of the proposed loss function. Sobolev loss models temporal structure of video, while MSE loss does not take the temporal structure into consideration. However, a bigger λ does not achieve a better result necessarily. Sobolev loss with $\lambda = 100$ actually leads to poorer F-measure than MSE loss does on TVSum50 as shown in Fig. 6. The best performance is achieved when $\lambda = 1$ on both datasets.

What kind of feature matters?

In CRSum, two 3D CNN blocks are utilized to extract features from video. CNN_1 is composed of three convolutional layers and aims to extract shallow features, while CNN_2 is responsible for extracting deep features, viz., semantic meanings. We conduct a series of experiments to dig out the potential significance of different kinds of feature. To make this study clear, we define CRSum(i, j) as variants of CRSum, where

$$i = \begin{cases} 1, & CNN_1 \text{ is pretrained on UCF101} \\ 0, & \text{Otherwise,} \end{cases} \tag{13}$$

$$j = \begin{cases} 0, & CNN_2 \text{ is abandoned} \\ 1, & \text{Otherwise.} \end{cases} \tag{14}$$

CRSum(0, 0) is the baseline model. Obviously, CRSum(0, 1) is the original model. Performance of CRSum(i, j) are shown in Table 2.

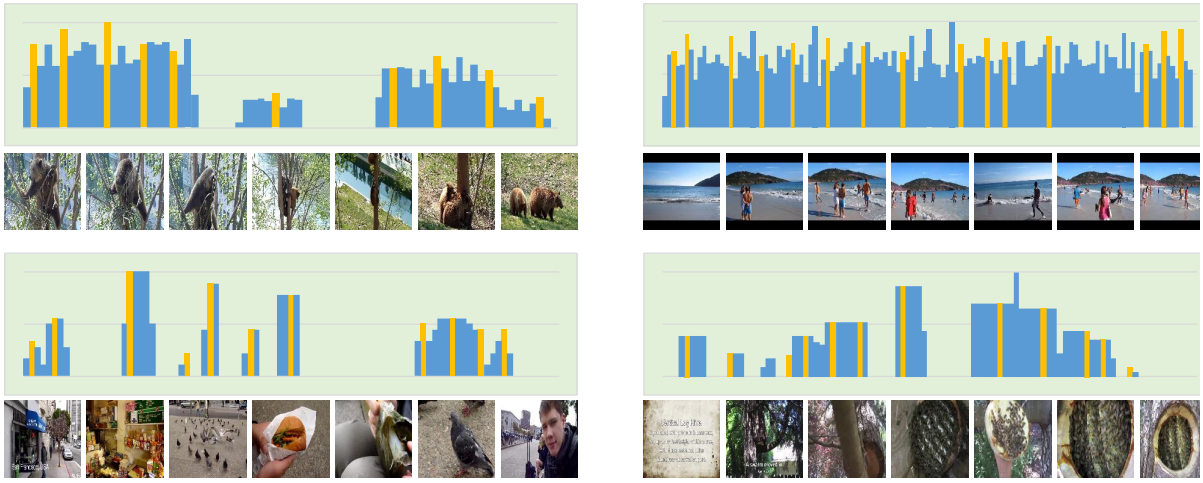


FIGURE 5. Four examples of generated summaries of the videos SumMe and TVSum50 by CRSum. The upper two are from SumMe, while the bottom two are from TVSum50. The blue bar indicates the ground truth importance score, and the orange bar indicates being chosen as summary.

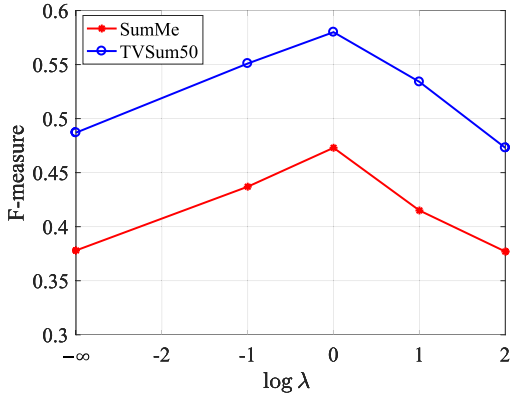


FIGURE 6. Performance of different λ on two datasets.

TABLE 2. Performance of variant models on two datasets.

Models	SumMe	TVSum50
CRSum(0, 0)	0.354	0.394
CRSum(1, 0)	0.442	0.558
CRSum(0, 1)	0.473	0.580
CRSum(1, 1)	**0.477**	**0.582**

As shown in Table 2, CRSum(0, 0) reaches the level of some previous supervised methods such as VSUMM [7], which proves the effectiveness of the proposed structure. CRSum(0, 0) achieves the poorest performance among four variants because the structure of CRSum(0, 0) is too simple to handle such sophisticated task. As for CRSum(1, 0), though the deep feature extractor (CNN_2) is removed from our model, the network still achieves respectable results which are just slightly poorer than those of the original CRSum. This phenomenon reveals a quite fact: the deep features are not necessary in video summarization and certain appropriate shallow features still work in this task. Besides, CRSum(1, 1) achieves the best performance as we expected because we give more prior to the network before training. However, the improvement by CRSum(1, 1) is not that remarkable compared to the original model.

V. CONCLUSION

In this paper, we present a novel method for video summarization. Our method utilizes CRNN, a special type of DNN, to adaptively extract spatiotemporal features and accurately predict the importance scores in an end-to-end fashion. In the proposed network (CRSum), the learnable 3D deep features and shallow features are fused to make comprehensive descriptions of video. Additionally, a new loss function (Sobolev loss) is defined for video summarization, aiming to exploit potential temporal structure of video and achieve better performance. We compare our method with several existing ones on two public datasets and the experiments prove its effectiveness. Moreover, we further analyze our method from different aspects.

REFERENCES

[1] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using Web-image priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2698–2705. doi: 10.1109/CVPR.2013.348.

[2] Y. Gong and X. Liu, "Video summarization using singular value decomposition," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Hilton Head Island, SC, USA, Jun. 2000, pp. 2174–2180. doi: 10.1109/CVPR.2000.854772.

[3] S. Lu, Z. Wang, Y. Song, T. Mei, and D. D. Feng, "A bag-of-importance model for video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, San Jose, CA, USA, Jul. 2013, pp. 1–6. doi: 10.1109/ICMEW.2013.6618454.

[4] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. 14th Eur. Conf., Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 766–782. doi: 10.1007/978-3-319-46478-7_47.

[5] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2982–2991. doi: 10.1109/CVPR.2017.318.

[6] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1059–1067. doi: 10.1109/CVPR.2016.120.

[7] S. E. F. Avila, A. P. BrandãoLopes, A. Luz, Jr., and A. A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, 2011. doi: 10.1016/j.patrec.2010.08.004.

[8] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2069–2077. [Online]. Available: http://papers.nips.cc/paper/5413-diverse-sequential-subset-selection-for-supervised-video-summarization

[9] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool, "Query-adaptive video summarization via quality-aware relevance estimation," in *Proc. ACM Multimedia Conf. (MM)*, Mountain View, CA, USA, Oct. 2017, pp. 582–590. doi: 10.1145/3123266.3123297.

[10] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. 13th Eur. Conf., Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 540–555. doi: 10.1007/978-3-319-10599-4_35.

[11] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3090–3098. doi: 10.1109/CVPR.2015.7298928.

[12] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 505–520. doi: 10.1007/978-3-319-10584-0_33.

[13] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 982–990. doi: 10.1109/CVPR.2016.112.

[14] K. Sun *et al.*, "Learning deep semantic attributes for user video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, Jul. 2017, pp. 643–648. doi: 10.1109/ICME.2017.8019411.

[15] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2714–2721. doi: 10.1109/CVPR.2013.350.

[16] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Perceptual video summarization—A new framework for video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1790–1802, Aug. 2017. doi: 10.1109/TCSVT.2016.2556558.

[17] J. Nam and A. H. Tewfik, "Event-driven video abstraction and visualization," *Multimedia Tools Appl.*, vol. 16, nos. 1–2, pp. 55–77, 2002. doi: 10.1023/A:1013241718521.

[18] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: Still and moving video storyboard for the Web scenario," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 47–69, 2010. doi: 10.1007/s11042-009-0307-7.

[19] A. Hendel, D. Weinshall, and S. Peleg, "Identifying surprising events in video using Bayesian topic models," in *Detection and Identification of Rare Audiovisual Cues*. Berlin, Germany: Springer-Verlag, 2012, pp. 97–105. doi: 10.1007/978-3-642-24034-8_8.

[20] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From keyframes to key objects: Video summarization by representative object proposal selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1039–1048. doi: 10.1109/CVPR.2016.118.

[21] D. Liu, G. Hua, and T. Chen, "A hierarchical visual model for video object summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2178–2190, Dec. 2010. doi: 10.1109/TPAMI.2010.31.

[22] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," in *Proc. 2nd ACM Workshop Video Summarization (TVS)*, Vancouver, BC, Canada, Oct. 2008, pp. 144–148. doi: 10.1145/1463563.1463590.

[23] Y. Li and B. Mérialdo, "Multi-video summarization based on video-MMR," in *Proc. 11th Int. Workshop Image Anal. Multimedia Interact. Services (WIAMIS)*, Desenzano del Garda, Italy, Apr. 2010, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/5617655/

[24] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 2513–2520. doi: 10.1109/CVPR.2014.322.

[25] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing Web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 5179–5187. doi: 10.1109/CVPR.2015.7299154.

[26] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017. doi: 10.1109/TPAMI.2016.2646371.

[27] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4694–4702. doi: 10.1109/CVPR.2015.7299101.

[28] S. O. Arik *et al.*, "Convolutional recurrent neural networks for small-footprint keyword spotting," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Stockholm, Sweden, Aug. 2017, pp. 1606–1610. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1737.html

[29] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[30] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.

[32] Y. Xu, Y. Han, R. Hong, and Q. Tian, "Sequential video VLAD: Training the aggregation locally and temporally," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4933–4944, Oct. 2018. doi: 10.1109/TIP.2018.2846664.

[33] S. Zhao, Y. Liu, Y. Han, R. Hong, Q. Hu, and Q. Tian, "Pooling the convolutional layers in deep convnets for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1839–1849, Aug. 2018. doi: 10.1109/TCSVT.2017.2682196.

[34] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016. doi: 10.3390/s16010115.

[35] K. Soomro, A. R. Zamir, and M. Shah. (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild." [Online]. Available: https://arxiv.org/abs/1212.0402

[36] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 2556–2563. doi: 10.1109/ICCV.2011.6126543.

[37] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV)*, Venice, Italy, Oct. 2017, pp. 3154–3160. doi: 10.1109/ICCVW.2017.373.

[38] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-C3D: Temporal convolutional 3d network for real-time action recognition," in *Proc 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 7138–7145. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17205

[39] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 1725–1732. doi: 10.1109/CVPR.2014.223.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997. doi: 10.1109/78.650093.

[41] Z. C. Lipton, J. Berkowitz, and C. Elkan. (2015). "A critical review of recurrent neural networks for sequence learning." [Online]. Available: https://arxiv.org/abs/1506.00019

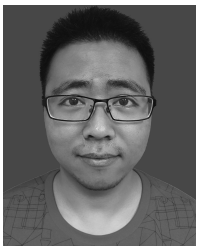
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[43] S. L. Sobolev, "On a theorem of functional analysis," *Mat Sbornik*, vol. 4, pp. 471–497, 1938.

[44] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 7582–7589. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16395

[45] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017. doi: 10.1109/TIP.2017.2695887.
[46] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc. ACM Multimedia Conf. (MM)*, Mountain View, CA, USA, Oct. 2017, pp. 863–871. doi: 10.1145/3123266.3123328.
[47] K. Zhang, K. Grauman, and F. Sha, "Retrospective encoders for video summarization," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 391–408. doi: 10.1007/978-3-030-01237-3_24.
[48] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video summarization via semantic attended networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 216–223.
[49] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 609–625. doi: 10.1007/978-3-319-46475-6_38.
[50] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

YUAN YUAN (M'05–SM'09) is currently a Full Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, Shaanxi, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals, such as IEEE TRANSACTIONS and *Pattern Recognition*, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.



HAOPENG LI received the B.E. degree in mathematics and applied mathematics from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2017, where he is currently pursuing the master's degree with the Center for OPTical IMagery Analysis and Learning (OPTIMAL). His research interests include computer vision and pattern recognition.

QI WANG (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.