Neurocomputing 332 (2019) 224–235

Contents lists available at ScienceDirect

Neurocomputing

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

Video summarization via spatio-temporal deep architecture

Sheng-hua Zhong[a,b,1], Jiaxin Wu[a,b,1], Jianmin Jiang[a,b,*]

[a] The National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China
[b] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

ARTICLE INFO

Article history:
Received 3 April 2018
Revised 4 October 2018
Accepted 18 December 2018
Available online 28 December 2018

Communicated by Dr. Yu Jiang

Keywords:
Video summarization
Convolutional Neural Network (CNN)
Class imbalance problem

ABSTRACT

Video summarization has unprecedented importance to help us overview current ever-growing amount of video collections. In this paper, we propose a novel dynamic video summarization model based on deep learning architecture. We are the first to solve the imbalanced class distribution problem in video summarization. The over-sampling algorithm is used to balance the class distribution on training data. The novel two-stream deep architecture with the cost-sensitive learning is proposed to handle the class imbalance problem in feature learning. In the spatial stream, RGB images are used to represent the appearance of video frames, and in the temporal stream, multi-frame motion vectors with deep learning framework is firstly introduced to represent and extract temporal information of the input video. The proposed method is evaluated on two standard video summarization datasets and a standard emotional dataset. Empirical validations for video summarization demonstrate that our model achieves performance improvement over the existing and state-of-the-art methods. Moreover, the proposed method is able to highlight the video content with the active level of arousal in affective computing task. In addition, the proposed frame-based model has another advantage. It can automatically preserve the connection between consecutive frames. Although the summary is constructed based on the frame level, the final summary is comprised of informative and continuous segments instead of individual separate frames.

1. Introduction

With the explosive growth of video data, there is increasing need to browse video content quickly [1–3]. Video summarization, which captures the attractive and representative information of the video, is an effective way to overview a large collection of videos [4]. Generally, video summarization can be divided into two categories: static video summarization and dynamic video summarization [4]. Static video summarization selects some important individual frames of the initial video as the final summary [5]. On the other hand, dynamic video summarization provides a more friendly browsing service for viewers [6,7]. It is comprised of informative and representative segments that keep motion information. Thus, in order to generate a good dynamic summary, visual appearance as well as temporal clue of the video should be well considered. In this paper, we propose a novel method for dynamic video summarization by making good use of temporal and spatial information of the video.

Previous work for dynamic video summarization have been studied in various perspectives. Chu et al. proposed a novel method to summarize a video by finding the shots that most frequently appeared among videos with the same topic [8]. They proposed a maximal biclique finding algorithm to find sparsely co-occurring patterns among thousands of irrelevant shots. Xu et al. used submodular maximization method based on gaze information to solve the summary problem [9]. They found that the gaze information of the wearers provided their intent and significantly helped the video summarization task. Zhang et al. tried to transfer summary structures from human-created summaries to unseen test videos [1]. They used semantic information about the video's genre to guide the transfer processing. Meng et al. formulated the video summarization task as a multi-view representative selection problem [10]. They selected visual elements that were representative of a video consistently across different feature modalities as the video summary by using the multi-view sparse dictionary selection with centroid co-regularization method.

Deep learning has achieved great success on computer vision and artificial intelligence [11–15]. Recent work on dynamic video summarization are also benefited from the progress in deep learning techniques. Gygli et al. used a supervised approach to learn the importance of the global characteristics in a summary by extracting deep features of video frames [16]. Yao et al. proposed a

* Corresponding author at: College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.
E-mail addresses: csshzhong@szu.edu.cn (S.-h. Zhong), jiaxin.wu@email.szu.edu.cn (J. Wu), jianmin.jiang@szu.edu.cn (J. Jiang).
[1] Sheng-hua Zhong and Jiaxin Wu contributed equally to this work.

https://doi.org/10.1016/j.neucom.2018.12.040
0925-2312/

pairwise deep ranking model that employed two-stream deep convolutional neural networks to generate the summarization of videos [17]. The final summary was comprised of those video segments with higher ranking scores. Zhou et al. proposed a video summarization method based on deep reinforcement learning [18]. The video summarization task was formulated as a sequential decision-making process and they developed a deep summarization network (DSN) to predict a probability for each video frame. The final summary was generated based on the probability, which indicated how likely a frame was selected.

In light of the recent successes of the deep learning on video content analysis, we propose our dynamic video summarization model based on a two-stream deep learning architecture. The two-stream network has been proved to be an effective architecture to recognize human actions [19,20]. However, we are aware of only one existing work constructed video summarization model via the two-stream network [17]. In the spatial stream, we follow these existing work [17,19,20] to use the RGB images to represent the appearance. In the temporal stream, most currently deep learning methods for video analysis used optical flow [19], dense trajectories [17] to represent the motion information. Although these techniques can detect and extract movement precisely, the temporal information is not exactly equal to the amplitude of all the movements between adjacent frames. In fact, some subtle changes detected by optical flow are often resulted from the illumination change or unsteady small-disturbance in the environment [21]. While precise movements obtained by optical flow techniques are useful for recognition of subtle actions, they may not provide reliable information for video summarization. In video summarization task, we believe that only the consistent motion of the prominent object with sufficient level of amplitude should be popped out as the temporal information for consideration of summarization. As the independent calculation in each frame pair has a high computational cost, optical flow and dense trajectory methods are often computationally expensive. Based on these considerations, we use multi-frame motion vectors (MV) with deep learning framework to represent and extract the motion information for video summarization. Motion vectors, which represent movement patterns of different image blocks, can be obtained from most of video files directly. It has been successfully utilized on action recognition task [22,23].

Video summarization is naturally a classification problem. One of the most important aims in video summarization task is to predict whether a frame should be in the final summary or not. Although recently, some deep learning models are proposed for video summarization task [16,17,24,25]. However, in these existing work, one important character of video summarization has never been seriously considered. That is, human beings tend to select a small subset of videos to be the summarization. This means that video summarization consists of generating a short summary of a video, which can either be a static summary or a dynamic summary [26]. In other words, the number of frames in the final summary is much less than the remaining frames. This character is recognized as the class imbalance problem [27]. Via checking all the public benchmark datasets for video summarization, it is found that no matter whether they are proposed for static or dynamic video summarization, all of them are imbalanced, as a matter of fact. Fig. 1 shows the average percentage of users selected summaries in 25 videos of the standard video summarization dataset SumMe [28]. The blue bar shows the proportion of the number of final summary in the whole video length, and users tend to take about 13% of the whole video as the video summary. If the class comprising the frames from the summary is treated as the positive class, and the class containing the remaining frames is thought of the negative class, then the data in these two categories are not balanced. This problem can be recognized as the imbalanced class distribution problem in machine learning and data mining, which causes seriously negative effects on the performance of learning methods. While there exists some work on class imbalance problem with deep learning networks [29–32], we are the first trying to solve this commonly existing problem in video summarization. In our proposed approach, we introduce a novel two-stream deep learning architecture with the cost-sensitive learning to handle the class imbalance problem.

The rest of this paper is organized as follows. Section 2 briefly reviews the representative work on imbalanced class distribution problem. In Section 3, we propose a novel framework and underlying algorithm in detail. In Section 4, we provide a series of experiments to validate the proposed method on standard datasets, and finally the conclusions are drawn in Section 5.

2. Related work

The class imbalance problem has been recognized as crucial in machine learning and data mining because such a problem is encountered in a large number of domains [31]. In classification, when the distribution of the training data among classes is uneven, the majority classes generally dominate the learning algorithm, whilst it makes the data from the minority classes difficult to be recognized [32]. Several existing research work focused on the class imbalance problem with deep learning networks [29–33], and they tried to solve the presence of underrepresented data and severe class distribution skews to improve the performance of the proposed algorithm [27].

The existing methods in tackling the class imbalance problem can be mainly divided into two groups: data resampling [29,31,34,35] and cost-sensitive learning [30,33]. The former group seeks to change the training data distribution to learn good classifiers for the majority and minority classes, usually by under-sampling and over-sampling techniques. The cost-sensitive learning operates at the algorithm level by adjusting misclassification costs for the majority and minority classes.

On the one hand, many research work tried to use data resampling technique to solve class imbalance problem [29,34,35]. Chawla et al. introduced an over-sampling method (SMOTE) which involved creating synthetic minority class examples for class imbalance problem [34]. They showed that a combination of their proposed method of over-sampling the minority class and under-sampling the majority class could achieve better classifier performance than only under-sampling the majority class or varying the loss ratios in Ripper or class priors in Naive Bayes. He et al. presented a novel adaptive synthetic sampling approach (ADASYN) for learning from imbalanced data sets [35]. They used a weighted distribution for different minority class examples according to their level of difficulty in learning. Jeatrakul et al. combined the synthetic minority over-sampling technique (SMOTE) and complementary neural network (CMTNN) together to handle the problem of classifying imbalanced data [29]. They compared the proposed method with several classical classification algorithms and the experimental results showed that the combined method could improve the performance of the class imbalance problem.

On the other hand, other existing work focus on cost-sensitive learning [30,33]. Shen et al. trained a cost-sensitive deep neural network to jointly optimize the class dependent costs and the neural network parameters. Specifically, a new loss function, named positive-sharing loss, in each subclass shared the loss for the whole positive class, was proposed to learn the parameters [30]. Khan et al. proposed a cost-sensitive (CoSen) deep neural network to automatically learn robust feature representations for both the majority and minority classes [33]. The proposed method was applicable to both binary and multiclass problems without any

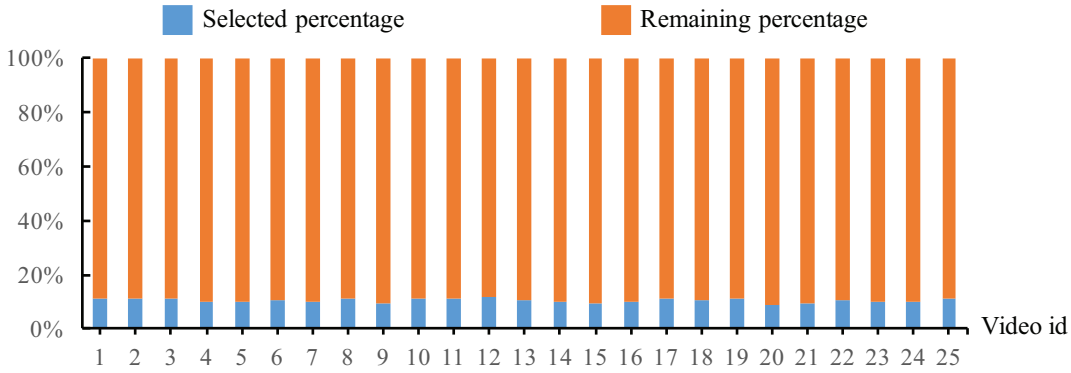

Fig. 1. The average percentage of users selected summaries in 25 videos of SumMe datasets.

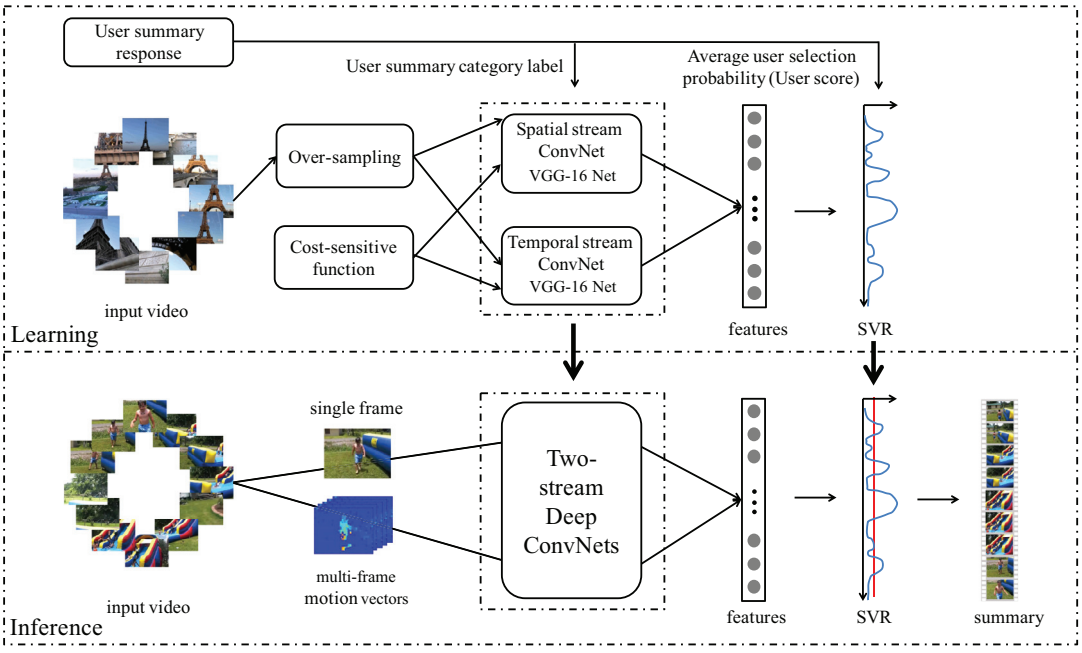

Fig. 2. The two-stream framework for video summarization.

modification. They conducted experiments on six image classification datasets and the results showed that the proposed method significantly outperformed the baselines.

Besides, there are some existing work trying to combine the data resampling technique and cost-sensitive method to enhance deep feature representations [31,32]. In 2006, Zhou et al. empirically studied the effect of data resampling in training cost-sensitive neural networks [31]. In 2016, Huang et al. investigated the combination of the data resampling technique and cost-sensitive method in face attribute classification task and edge detection task [32]. The representation learned by their approach showed significant improvements over previous methods on vision classification tasks that exhibited imbalanced class distribution.

3. Imbalanced video summarization

In this paper, we propose a novel dynamic video summarization method based on a two-stream deep learning architecture. Fig. 2 shows a visual scheme of the proposed video summarization via spatio-temporal deep learning model (VSST). Besides the summary results selected by each subject as the ground truth for classification, most of the datasets also provide user scores for each frame or each shot. One kind of user score is the average user selection probability [28]. Another is the score directly defined by subjects [36]. To fully exploit these two kinds of information, our learning model contains the classification objective function and regression objective function. In the learning scheme, we first construct a two-class classification model based on spatio-temporal deep learning architecture. The over-sampling method is conducted to handle the imbalanced class distribution problem in training data of the video summarization task. New balanced data with their corresponding summary category labels are then input to train a cost-sensitive two-stream deep network to extract the features with better discriminative ability. Then, these features with their corresponding summary probabilities are fused together as the input of support vector regression (SVR) to train an effective regression model and predict the highlight probability/score for each frame. In the inference scheme, the learnt VGG-16 models are used to extract features from the input data, and the learnt SVR is utilized to predict the highlight score for each frame based on the combined feature. Finally, we select the frames to construct the final video summary according to their predicted probabilities/scores.

In the following, we first describe how the over-sampling algorithm works for the class imbalance problem in video summarization. We then introduce the two-stream deep learning architecture with cost-sensitive learning, and finally, we briefly describe the

SVR-based highlight prediction to complete our proposed deep-learning based dynamic video summarization.

3.1. Over-sampling to balance class distribution

Over-sampling is an effective method to address the class imbalance problem. This technique changes the training data distribution such that the costs of the examples are conveyed by the appearance of the examples [31]. In simple words, over-sampling resamples the minority class until it has as many instances as the majority class [31]. There are many effective over-sampling methods such as SMOTE [34] and ADASYN [35]. In our paper, owing to the lower computational cost, we simply utilize data augmentation technique [20] for over-sampling.

Video summarization can be formulated as a two-class task. The class comprising the frames from the summary is the minority class and the class containing the remaining frames is the majority class. Let N_α be the number of training data in the majority class and N_β be the number of training data in the minority class. In the video summarization task, N_β is less than N_α. After the over-sampling stage, the minority class will have N^*_β training data, and it makes $N^*_\beta = N_\alpha$.

The detailed procedure of the over-sampling algorithm for a video is described in Algorithm 1. Specially, we use a corner-cropping strategy [20] to create the cropped version of original training samples.

3.2. Two-stream deep ConvNets for imbalanced feature learning

As shown in Fig. 2, our proposed method includes two-stream deep ConvNets to extract spatial and temporal information for videos.

In each stream, VGG-16 [37] is exploited to extract effective features for video frames. The architecture of this convolutional neural network is $C64 - C64 - C128 - C128 - C256 - C256 - C256 - C512 - C512 - C512 - C512 - C512 - C512 - F4096 - F4096 - F2$, which contains thirteen convolution layers (denoted by C with the number of neurons) and three fully-connected layers (denoted by F with the number of neurons).

Algorithm 1 Over-sampling algorithm with the corner-cropping strategy for a video.

Input:
The original set contains all video frames from the input video, S;
Output:
The balanced set for the input video, S^*;

```
Split S into S_α and S_β. S_α contains the majority class video frames while S_β contains the minority class video frames;
Calculate n_α and n_β. n_α and n_β are the number of video frames in S_α and S_β;
Let n*_β = n_β;
Put all original training examples (S) in S*.
while n*_β < n_α do
    for each video frame s_i in S_β do
        Generate a cropped image from s_i using corner-cropping strategy, and put them into S*.
        n*_β = n*_β + 1.
        if n*_β == n_α then
            break;
        end if
    end for
end while
return S*;
```

In the learning procedure, the balanced video data with their corresponding category labels are input to train each stream. In the spatial stream, we follow the existing work to use the RGB image from each frame as the input. In the temporal stream, different from existing methods in extracting optical flow or dense trajectories, we use the multi-frame motion vectors between frames as the input to convey the temporal dynamics. In the inference stage, the output of second fully-connected layer in each stream generates a 4,096-dimensional vector. The resulting two 4096-dimensional representations of each video frame are fused together to form the input to the subsequent support vector regression algorithm to predict the video summary score of the current frame.

In this paper, we propose to use multi-frame motion vectors as the input of the temporal stream to convey the movement of objects (or scenes) across frames. Motion vectors, which represent movement patterns of different image blocks, can be obtained from most of video files directly. We assume that the motion vectors of frame t are denoted as $\mathbf{M}_t$. A multi-frame motion vectors input can be seen as a set of displacement vector fields $\mathbf{M}_t$ between the pairs of consecutive frames t and $t+1$. Formula (1) shows the construction of multi-frame motion vectors $\mathbf{T}_t \in \mathbb{R}^{w\times h\times 2L}$ for frame t. In this equation, $\mathbf{M}_t$ denotes the motion vectors of frame t. $\mathbf{M}^x_t$ and $\mathbf{M}^y_t$ are the horizontal and vertical components of $\mathbf{M}_t$. To represent the motion across a sequence of frames, we stack these two components crossly of L consecutive frames as formula (1) to form a total of $2L$ input channels. L is the stacking length.

$$\begin{cases} \mathbf{T}_t(2k-1) = \mathbf{M}^x_{t+k-1} \\ \mathbf{T}_t(2k) = \mathbf{M}^y_{t+k-1} \end{cases}, 1 \le k \le L \tag{1}$$

Considering that the VGG-16 ConvNet has a fixed size input, we sample $\mathbf{T}_t$ to be a $224 \times 224 \times 2L$ sub-volume and treat it as the input of temporal steam.

The cost-sensitive learning is proposed to handle the class imbalance problem in feature learning. It directly operates at the algorithm level by adjusting misclassification costs for the majority and minority classes. In the following, we describe how we define the learning objectives in our model.

Given a training set which contains m sample: $\{x^{(i)}, y^{(i)}\}_{i=1}^m$, where $x^{(i)}$ is the i^{th} sample and $y^{(i)} \in \{0, \ldots, k, k+1, \ldots, K\}$ is its class label. $K+1$ is the number of classes in this dataset. Suppose S_α denotes the majority set and S_β denotes the minority set. If $y^{(i)} \le k$, then $x^{(i)}$ is a majority sample, else if $y^{(i)} > k$, then $x^{(i)}$ is a minority sample. The probability that the label of the $x^{(i)}$ is equal to j can be given by:

$$P^{(i)}_j = \frac{\exp(a^{(i)}_j)}{\sum_{l=0}^{K} \exp(a^{(i)}_l)} \tag{2}$$

where $a^{(i)}_j$ is the output of the unit j in the last layer of the fully-connected sub-networks for $x^{(i)}$. The output of the last fully-connected layer is then fed into a $(K+1)$-way softmax which aims to minimize the following loss function:

$$J = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=0}^{K} \mathbf{1}(y^{(i)} = j) \log P^{(i)}_j\right] \tag{3}$$

where $\mathbf{1}(\bullet)$ is the indicator function. In the case of the standard softmax loss function, it tries to penalize the classification error for each class equally. In video summarization task, to predict the label of a positive class (minority class) sample to be negative is a more critical error than the opposite case. Thus, our new loss function assigns higher misclassification costs to the cases that predict a minority class to be a majority class. Further, this setting is consistent with the construction of the loss function in cost-sensitive learning for imbalanced data. To counter the adverse effects of

imbalanced data, cost-sensitive learning is often applied, which assigns higher misclassification costs to the minority class than to the majority [30,32,33]. Correspondingly, a novel loss function is defined as follows:

$$J^* = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=0}^{k}\left(\mathbf{1}(y^{(i)}=j)\log P_j^{(i)}\right)\right] - \frac{\lambda}{m}\left[\sum_{i=1}^{m}\sum_{j=k+1}^{K}\left(\mathbf{1}(y^{(i)}=j)\log P_j^{(i)}\right)\right] \tag{4}$$

where λ is the misclassification cost for the minority classes, and it is greater than 1. With this new loss function, we can assign higher misclassification costs to the cases that predict the label of a minority sample to be a majority label.

3.3. Highlight prediction via support vector regression

A version of support vector machine for regression was proposed by Drucker et al. [38]. This method is called support vector regression (SVR), which has been widely reported to achieve good performances in many computer vision and machine learning problems.

In this paper, SVR algorithm is constructed to predict the highlight value for each video frame. In the learning scheme, as shown in Fig. 2, the features of two-stream ConvNets with the corresponding average user selection probability (we can also call it as user score) are combined together as the input of SVR. In the inference scheme, the learnt SVR is used to predict the probability/score for each frame based on its feature. As a result, we select the frames to construct the final video summary according to their predicted probabilities/scores, and the final summary is comprised of those video frames with highest M percentage of the predicted probabilities/scores. For SVR, we use the standard toolbox LIBSVM [39]. The Radial Basis Function (RBF) is selected as the kernel function, and a grid search is run to find the optimal parameter settings.

4. Experiments

In this section, we first describe the experimental setting we utilize for the evaluations. Secondly, we compare the video summarization results obtained by our method with several state-of-the-art methods on three standard datasets: the SumMe dataset [28], the TVsum dataset [36] and the Continuous LIRIS-ACCEDE dataset [40]. The comparison results on each dataset are shown separately in Sections 4.2–4.4. In Section 4.2, we step by step evaluate two key stages of our proposed method on the SumMe dataset. We firstly provide the classification accuracy of our proposed imbalanced two-stream network, and then we visualize the video summary generated by our VSST approach. Next, we investigate the influence of different summary length M on the summarization results and the distribution of user summaries. Finally, we compare the efficiency of the motion vectors and optical flow on feature extraction. In Section 4.3, we evaluate our proposed method on a category-based video summarization benchmark dataset. We compare our results with two state-of-the-art dynamic video summarization models. The comparison results demonstrate that our proposed method can generate similar video summaries to subjects' on the TVsum dataset. In Section 4.4, we apply our proposed model on an emotional dataset. The experimental result shows that our model is able to highlight the video content as consistent as the active level of arousal in affective computing task.

4.1. Experimental settings

In this paper, we evaluate the performance of the proposed models on three standard datasets: the SumMe dataset [28], the TVsum dataset [36] and the Continuous LIRIS-ACCEDE dataset [40]. While SumMe and TVsum are two benchmark datasets in video summarization with multiple human-annotated summaries for each video, the continuous LIRIS ACCEDE is a standard annotated emotional dataset.

In our experiments, we evaluate automatic summarization results (A) by comparing them to the human-created summaries (B) and report the F-measure values to measure the performances of compared methods. This metric has been widely used in current work for video summarization [8,16,28], which is defined as follows:

$$F = \frac{2\times p\times r}{p+r}, \tag{5}$$

$$p = \frac{\#matched\quad pairs}{\#frames\quad in\quad A}\times 100\%, \tag{6}$$

$$r = \frac{\#matched\quad pairs}{\#frames\quad in\quad B}\times 100\%. \tag{7}$$

where p is the precision and r is the recall. In this paper, we report the Mean F-measure and the Nearest-Neighbor F-measure (NN-F-measure) by comparing the predicted summaries with the ground truth summaries. The Mean F-measure is the average value of the F-measure for all subjects. It is given by:

$$\bar{F} = \frac{1}{N}\sum_{i=1}^{N} F_i \tag{8}$$

where N is the number of subjects, and F_i is the F-measure for subject i. The NN-F-measure is the maximum of F_i, and it is given by:

$$F_{max} = \max_i(F_i) \tag{9}$$

This metric is used to evaluate the performance of the proposed method based on the most similar summary from all viewers. We use the standard toolbox proposed by Gygli et al. [28] to evaluate our performance on SumMe and CLA dataset, and we utilize the evaluation code provided by Zhang et al. [24] on TVsum dataset.

Based on the statistical analysis by Gygli et al. [28], the length of the final summary is about 15% of the original video. For our models, we set the summary length $M = 15$ in most of the experiments. We follow the setting in existing work [1] to randomly select 80% of videos for training, while the remaining videos are used for testing. In the first part of our proposed method, we construct a two-class classification model based on the spatio-temporal deep learning architecture. The class comprising the frames from the summary is treated as the minority class (positive class). The class containing the remaining frames is set as the majority class (negative class). Therefore, in our algorithm, k is equal to 0 and K is equal to 1. While the misclassification cost λ is set to 1.1, we have also tested its sensitivity. When the value is changed from value 1.1 to value 2, our proposed model achieves consistent performance improvement over other models, and when the lambda is equal to 1.1, the model reaches the best performance. We believe that our over-sampling technique does improve the imbalanced situation in advance, thus we do not need a high penalization rate in this stage. In this paper, we train an effective VGG-16 network for video summarization by learning from the previous practices in [20]. Firstly, we pre-train VGG-16 network on the ImageNet dataset [41]. Secondly, we use a small learning rate (0.001) compared with the learning rate used in the standard two-stream network [19]. Thirdly, we use data argument techniques to avoid the problem of

over-fitting. We also set dropout rates equal to 0.9 for the fully connected layers. Our models are implemented using the video extension version [20] of the Caffe toolbox [42] on a Tesla K80 GPU. We use the widely used toolbox to extract optical flow [43,44]. Further, we follow the algorithm proposed by Zhang et al. to obtain motion vectors [45] and stacking length L is set to 10, the same setting as in [19,20]. All the statistical experiments are repeated for five times, and the average results are reported.

4.2. Video Summary prediction on the SumMe dataset

In this Section, we have conducted many experiments to demonstrate the effectiveness of our proposed method on the SumMe dataset. SumMe consists of 25 videos covering different real-world topics, such as holidays, accidents, and sports. Each video was annotated with more than 15 different user summaries. It has 390 reference summaries in total. The annotation stage was processed in a controlled environment, where participants were asked to create their own summary for a given video. We find subjects' responses are various within the same video. The diversity and variety of the video contents and the subjects' responses make this dataset a challenging benchmark for video summarization. As SumMe is a widely used standard benchmark dataset for video summarization, more than seven models have been validated in this dataset, including: Exemplar-based Subset Selection (ESS) [1], Learning Submodular Mixtures of Objectives (LSMO) [16], Creating Summaries from User Videos (CSUV) [28], Summarizing Web Videos using Titles (SWVT) [36], Video MMR [46], Video Summarization with Long Short-term Memory (dppLSTM) [24], Unsupervised Video Summarization with Adversarial LSTM Networks (SUM-GAN) [25] and so on.

To evaluate the effectiveness of our proposed imbalanced network, we firstly compare our results with a random baseline as well as the state-of-the-art models of video summarization, including Video Representation Clustering based Video Summarization (VRCVS) [5], ESS [1], LSMO [16], CSUV [28], SWVT [36], Video MMR [46], dppLSTM [24] and SUM-GAN [25]. For the random baseline, we randomly select $M = 15$ percentage of video sequences as the final summary. Considering the fact that VRCVS is a recent cluster-based static video summarization model, we provide two versions of VRCVS for comparisons, i.e. VRCVS and VRCVS-shot. VRCVS directly represents the final summary as individual separated frames, and VRCVS-shot is an extension of the original VRCVS, which constructs the final summary with the shots containing those individual frames. The video shots in this method are obtained via a superframe segmentation algorithm [28]. In our experiments, we also provide the comparisons with some state-of-the-art dynamic video summarization models, such as ESS [1], SWVT [36], LSMO [16], CSUV [28], Video MMR [46], dppLSTM [24] and SUM-GAN [25]. For those models, we follow the parameter settings provided by their work. Besides, for the comparison, we also provide different versions of the proposed methods based on the spatio-temporal deep architectures (VSST). These methods include VSST-OP, VSST-MV, VSST-RGB, VSST-RGB&MV and VSST-Imbalance. Among them, VSST-OP, VSST-MV, and VSST-RGB are with one-stream deep architecture. VSST-OP, VSST-MV, and VSST-RGB indicate the methods that use optical flow, motion vectors, and RGB image as the input of the one-stream ConvNet, respectively. VSST-RGB&MV is model with a two-stream learning structure, including RGB images as the input of the spatial stream and multi-frame motion vectors as the input of the temporal stream. VSST-Imbalance uses imbalance technique to handle the class imbalance problem in video summarization, which can be seen as the imbalanced version of VSST-RGB&MV. All of these proposed models are evaluated on frame level. The first M% frames with the higher predicted scores are selected to construct the summary results. Since most of compared methods were produced on shot level, we also provide a shot-level version of VSST-Imbalance (VSST-Imbalance-shot) for fair comparisons. We follow the existing work [24,25] to generate shot-level summary result. The videos are initially temporally segmented into disjoint intervals using kernel temporal segmentation (KTS) [47]. The final summary is comprised of those segments with highest predicted scores. The predicted score of a segment is equal to the average score of the frames in that interval.

The comparison results are shown in Table 1 with the average Mean-F-measure (AMF) and the average NN-F-measure (ANF). ESS [1] and LSMO [16] were supervised methods based on deep features while CSUV [28] and video MMR [46] were unsupervised methods based on hand-crafted features. DppLSTM was also based on deep learning architecture using long short-term memory (LSTM) [24]. Here, we report the best performances of their method. SUM-GAN was proposed by Mahasseni et al., which utilized generative adversarial framework (GAN) for video summarization based on the long short-term memory network (LSTM) [25]. SUM-GAN$_{sup}$ is the supervised version proposed in their paper. Generally, the deep learning based methods [1,16,24,25] outperform the classical models [28,46]. It can be seen that the performances of the dynamic video summarization techniques are better than those of the static video summarization methods [5]. The proposed imbalance-based method achieves the best AMF and ANF. Compared with the random baseline, the proposed model achieves more than twice of the corresponding values in the evaluation metrics. In addition, the performances of nearly all the proposed models (VSST-RGB, VSST-MV, VSST-RGB&MV, VSST-Imbalance and VSST-Imbalance-shot) are also better than those state-of-the-art models (CSUV, LSMO, ESS, SWVT, dppLSTM and SUM-GAN$_{sup}$), which confirms that the proposed method could capture most of the attractive and representative contents from video sequences. Although our two-stream deep ConvNets are constructed based on VGG-16, which is not the most innovative deep networks, compared with the model based on LSTM or GAN, our architecture achieves the best performance. The experimental results also indicate that the models with two-stream learning structure (VSST-RGB&MV and VSST-Imbalance) are better than those one-stream methods (VSST-RGB, VSST-OP and VSST-MV). In these one-stream models, the performance of VSST-OP is worse than that of VSST-MV, although motion vectors cannot represent the motion information as precisely as optical flow. According to film theorists, motion is highly expressive able to evoke strong emotional responses in viewers [48,49]. In fact, studies by Detenber et al. [49] and Simmons et al. [50] concluded that an increase of motion intensity on the screen causes an increase in the audiences arousal. The analysis of the relationship between motion intensity and user summaries is conducted. We investigate the distribution of subject summaries with the increase of motion intensity in term of motion vectors in SumMe dataset. We find the average motion intensity of the frames, which are selected by half of subjects, is more than 1.7 times higher than the corresponding values of all frames in the videos of SumMe. The experimental results in Table 1 support that the multi-frame motion vectors are effective to capture this kind of temporal information than optical flow.

We also explore other deep ConvNets such as the residual network [51] on our proposed architecture. Table 1 shows the performances generated by our one-stream model VSST-RGB implemented by different deep architectures including ResNet-18-RGB and ResNet-50-RGB. To ensure the fairness of the comparison, we obtain the results from the standard residual network [51] and the residual network with our setting, i.e. the dropout rate and the learning rate, and the best performances of them are given in Table 1. From these results, it is easily observed that although ResNet-18-RGB has the similar number of layers with VSST-RGB, the AMF and ANF of it are less than the proposed VSST-RGB.

Table 1
The performance comparison of our proposed methods with other models on SumMe dataset. '—-' denotes that the result is not reported in existing papers.

		Method	AMF (%)	ANF (%)
Unsupervised methods	Baseline	Random	14.3	28.6
	Existing static methods	VRCVS [5]	1.0	0.5
		VRCVS-shot	14.9	40.4
		CSUV [28]	23.4	39.4
		Video MMR [46]	—-	26.6
	Existing dynamic methods	SWVT [36]	26.6	—-
		LSMO [16]	—-	39.7
		ESS [1]	—-	40.9
		dppLSTM [24]	17.7	42.9
		SUM-GAN$_{sup}$ [25]	—-	43.6
Supervised methods	Other deep architectures	ResNet-18-RGB	26.5	44.6
		ResNet-50-RGB	29.5	45.8
		VSST-OP	23.0	39.9
		VSST-RGB	32.0	53.4
		VSST-MV	35.2	53.8
	Proposed methods	VSST-RGB&MV	35.4	56.3
		VSST-Imbalance	35.5	57.7
		VSST-Imbalance-shot	26.1	54.2

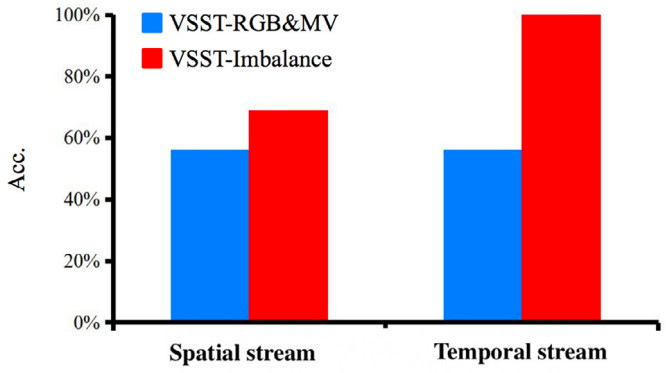

Fig. 3. The classification accuracies of two versions of the proposed methods on SumMe dataset.

Moreover, owing to the contribution of deeper layers, the performance of ResNet-50-RGB is better than ResNet-18-RGB, but it is still worse than ours.

Next, we report the classification accuracies (Acc.) of the two versions of our models, including VSST-RGB&MV and VSST-Imbalance in Fig. 3. In the learning scheme, the balanced data with their corresponding summary category labels from SumMe dataset are input to train the two-stream deep ConvNet, and the cost-sensitive learning is utilized in the two-stream network. From Fig. 3, it is clear that the imbalance-based method obtains higher accuracies on both spatial stream and temporal stream in SumMe dataset.

Fig. 3 indicates that the proposed imbalanced deep model has already achieved a very high accuracy. This very high accuracy, however, does not necessarily result in a very high final AMF/ANF score. This is because that the high accuracy in the classification task only means the model can predict whether a frame should be in the final summary or not. But to achieve a high value of AMF/ANF, the model requires precisely predicting the selection of each frame similar to most of the subjects. Unfortunately, for different subjects, their responses often vary even within the same video. Even to the same subject, the ranges of the responses for different videos also fluctuate. Thus, high classification accuracy is not equivalent to high AMF/ANF score.

To visualize the predicted results of a given video, we present a sample of the predicted result of the proposed model for the video "Jump" from SumMe dataset in Fig. 4. As seen, this video sequence depicts the jump procedure including preparation, jumping and landing stages. The first row of Fig. 4 describes the average possibility of each frame whether it would be selected as summary based on all subjects' selections. We can also call this probability as the score for each frame. In the following three rows show the prediction scores generated by three versions of our models: VSST-RGB&MV, VSST-Imbalance-shot and VSST-Imbalance, respectively. The last row shows the final automatic summary of this video. From this example, we can find that the predicted scores of our methods are very similar to the ground truth of all subjects, and our final summary covers all the main stages in the action "Jump". Furthermore, a comparison of the second and the third or fourth rows of Fig. 4 reveals the influence of the class imbalance issue on video summarization. We can see that fast fluctuations exist from 300 to 400 frames in the prediction score of VSST-RGB&MV. We speculate it is due to the class imbalance problem in video summarization, as this fluctuations phenomenon does not happen in the average selection of the video. From the fourth row, we can see that the proposed VSST-Imbalance method could handle this issue well. VSST-Imbalance can detect the landing stage of "Jump" (from 940 to 950 frames), which was not in the summarized results from the average selection of the video. But this stage is also an important component in the action "Jump". In addition, by the comparison of our proposed frame-based model and shot-based model, we can easily observe that our frame-based model can automatically preserve the connection between consecutive frames. Although the summary is constructed based on frame level, the content of it is coherent. The final summary is comprised of informative and continuous segments that keep motion information instead of individual separate frames. We believe this is another important advantage of our method.

The mismatch between our selection and user summaries in Fig. 4 (from 940 to 950 frames) inspires us to investigate the distribution of the user summaries in the different locations of the target video. Fig. 5 shows the experimental result. We divide the video into two groups: the first $\delta\%$ and the last $(100-\delta\%)$. We then calculate the percentage of user summaries in each group. In this figure, each bar corresponds to a value of δ. We report 21 values of δ, which are 0, 5, 10,..., 90, 95, 100. From the last three bars, we find most of the subjects are prone to assign less attention in the last 10% of the videos. The reason is that the landing part of the action "Jump" has not been selected in the subjects' response (Fig. 4).

We also investigate the impact of different summary lengths M on SumMe dataset. Based on the statistical analysis by Gygli

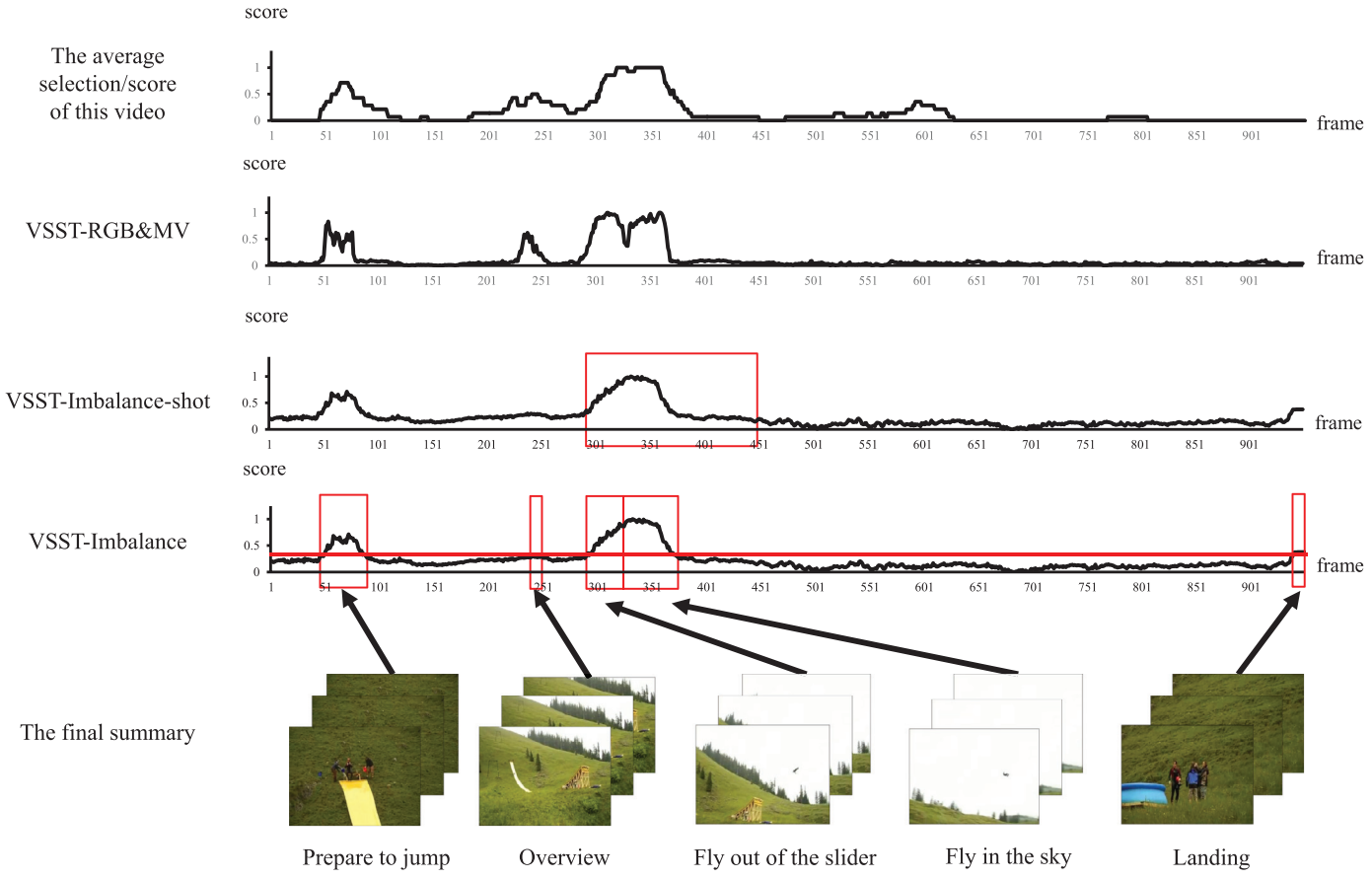

Fig. 4. A sample of the predicted result on video "JUMP" of SumMe database.

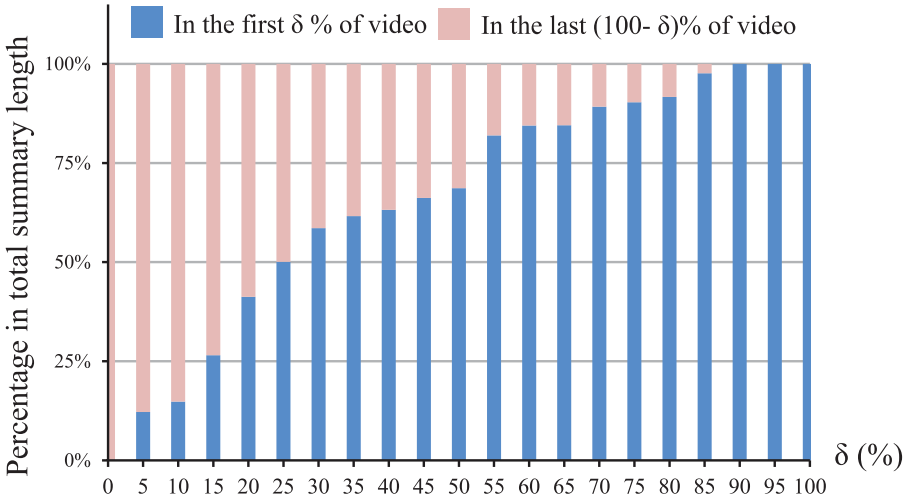

Fig. 5. The distribution of the user summaries in the different locations on the video. δ indicates the location where we split the video and it ranges from 0 to 100. For example, when $\delta = 50$, the blue bar gets about 68% and the red bar achieves about 32%. It means that subjects are prone to assign about 68% of the summary result in the first 50% of the videos.

et al. [28], the length of the final summary is about 15% of the original video. In our work, we set the summary length M to be 15 percentage of the whole video sequence. In the following, we provide the performance results for a range of models, including VSST-RGB&MV, VSST-Imbalance, VSST-OP and VRCVS-shot, for which different values of M are applied. VRCVS-shot is an extension of the static summarization model VRCVS [5], which constructs the final summary with the shots containing those individual frames summarized by VRCVS. The others are three versions of our proposed method. VSST-OP is with one-stream deep architecture using optical flow as the input. VSST-RGB&MV is with two-stream learning structure, including RGB images as the input of the spatial stream and motion vectors as the input of the temporal stream. VSST-Imbalance is the imbalanced version of VSST-RGB&MV. Fig. 6(a) shows the values of the average Mean-F-measure of these four methods when M varies from 5 to 25. Fig. 6(b) shows the value of the average NN-F-measure when M varies from 5 to 25. From these figures, we can see that

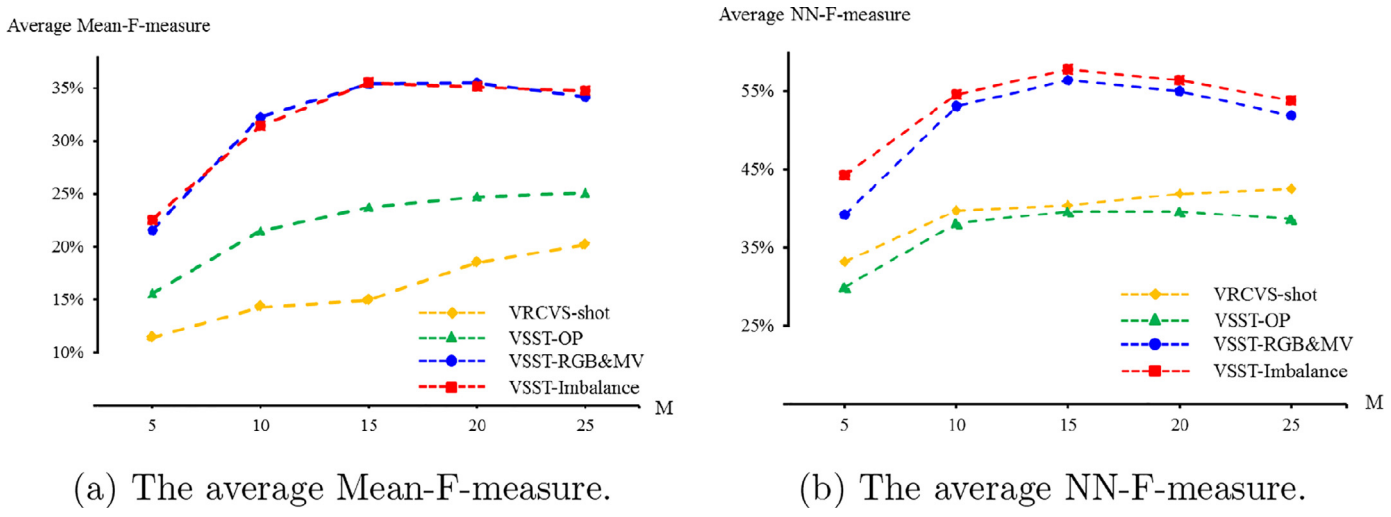

Fig. 6. Performance comparison with different summary length M on SumMe dataset.

Table 2
Efficiency comparison of different feature extraction methods on SumMe dataset.

Method	Average speed (fps)	STD
VSST-MV	71.06	0.01
VSST-OP	1.86	0.40

VSST-RGB&MV and VSST-Imbalance achieve the best performances when the summary length $M = 15$. They outperform the static model and one stream model in all different values of M. VRCVS-shot gains the best results when the summary length $M = 25$. VSST-OP achieves the best average mean-F-measure when $M = 25$ and the best average NN-F-measure when $M = 15$.

Finally, we compare the efficiency of feature extraction on SumMe dataset in Table 2. In VSST-MV, the motion vectors are extracted as the temporal information. In VSS-OP, we follow the existing work to calculate and obtain the optical flow as the temporal information. Table 2 shows the average speed and the standard deviation of different methods. The average speed of motion vectors extraction is about 71.06 frames per second (fps). This speed is almost 40 times faster than the process of optical flow. Taking into consideration the large number of frames in videos, this difference matters and presents a significant advantage for practical application of video summarization. Therefore, the selection of motion vectors instead of optical flow reduces the computational cost of our model.

4.3. Video summary prediction on the TVSum dataset

The TVSum dataset is a category-based benchmark for dynamic video summarization proposed by Song et al. [36]. This dataset is commonly used in video summarization [24,25,36]. It contains 50 videos downloaded from YouTube in 10 categories defined in the TRECVid Multimedia Event Detection (MED). The length of the videos varies from 2 to 10 min. Videos represent various genres, including news, documentaries and user-generated content. This dataset provides 20 user-annotated summaries as well as a shot-level important score for each video. And each shot has a uniform length of 2 s. Thus, in our SVR process, we also uniformly subsample the videos of TVsum to 2 fps by following the setting of existing work [24]. Then, for the training data, we assign the shot-level score to each input frame. After SVR prediction, each test frame in the same interval has the identical predicted score.

In this section, we conduct the comparisons using the random baseline as well as the state-of-the-art models of video summarization, including Video Representation Clustering based Video Summarization (VRCVS) [5], Summarizing Web Videos Using Titles (SWVUT) [36], Video Summarization with Long Short-term Memory (dppLSTM) [24] and Unsupervised Video Summarization with Adversarial LSTM Networks (SUM-GAN) [25]. SWVUT is a title-based dynamic video summarization method [36]. Song et al. collected an extra set of images to learn the visual concepts from a video title. They utilized these image search results to find visually important shots later. Zhang et al. applied LSTM technique to model the variable-range temporal dependency among video frames [24]. They believed that LSTM was helpful to derive both representative and compact video summaries. In their experiments, two extra static video summarization databases were adopted as their training data, and dppLSTM was one of their proposed method which achieved the best performance on TVSum dataset. SUM-GAN is a recent dynamic video summarization model based on the advanced deep learning architecture (GAN) [25]. Here, we report the best performances of their proposed methods on TVSum which utilized augmented data for training. For the random baseline, we randomly select $M = 15$ percentage of video sequences as the final summary. Since all of the compared methods were evaluated on shot-level, we provide different shot-level versions of the proposed methods including our one-stream model (VSST-MV-shot), and our two-stream models (VSST-RGB&MV-shot and VSST-Imbalance-shot). We report the comparison results in Table 3.

Table 3 shows the video summarization performance with the Mean-F-measure and the NN-F-measure on TVSum dataset. Obviously, all dynamic video summarization methods outperform the static method (VRCVS) and the random baseline. The deep learning methods (dppLSTM, SUM-GAN and VSST) achieve higher AMF and ANF than the classical method (SWVUT). Although we do not utilize any spatial information in the experiments, our proposed one-stream model based on MV (VSST-MV-shot) is still competitive with the LSTM and GAN based models, and our two-stream models (VSST-RGB&MV-shot and VSST-Imbalance-shot) gain higher AMF and ANF on TVSum dataset.

4.4. Video affective computing on the Continuous LIRIS-ACCEDE dataset

Affective video content analysis aims to automatically recognize emotions elicited by videos [40]. It has a large number of related applications, such as mood-based personalized content

Table 3
The performance comparisons using the average F-measure on tvsum dataset. '—-' denotes that the result is not reported in existing papers.

		Method	AMF (%)	ANF (%)
Unsupervised methods	Baseline	Random	14.4	29.2
	Existing static methods	VRCVS [5]	4.9	6.0
		VRCVS-shot	24.7	34.0
	Existing dynamic methods	SWVT [36]	50.0	
		dppLSTM [24]	58.7	78.6
		SUM-GAN$_{sup}$ [25]	61.2	—-
Supervised methods	Proposed methods	VSST-MV-shot	58.2	81.0
		VSST-RGB-shot	62.0	83.9
		VSST-RGB&MV-shot	**62.8**	83.8
		VSST-Imbalance-shot	**62.8**	**84.0**

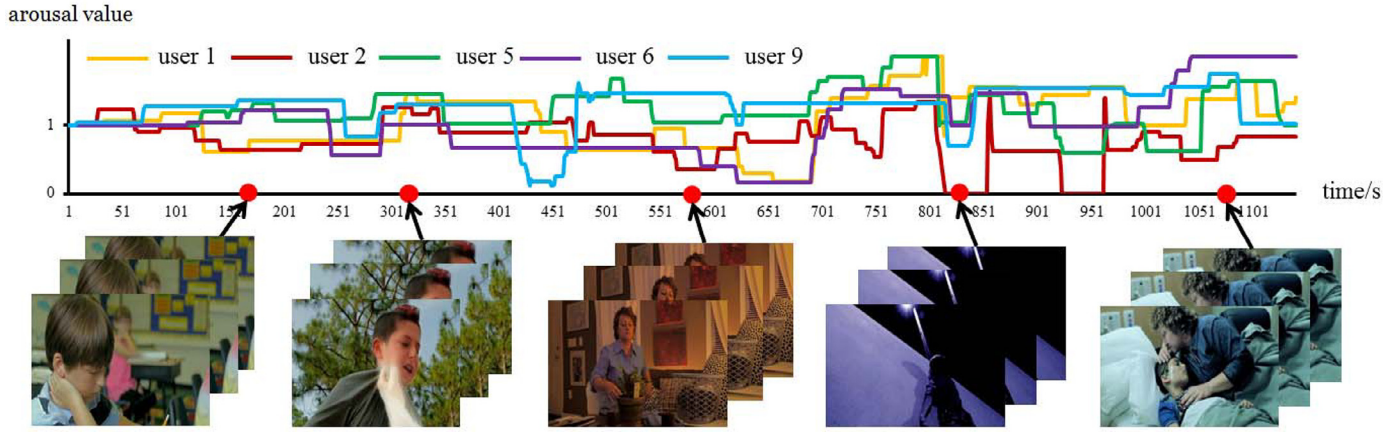

Fig. 7. A sample video called "Superhero" on CLA dataset. The different color curve reflects the arousal value for each viewer. The red point in the axis denotes the corresponding visual content.

delivery, video indexing, and video summarization. The affective level is an particularly important measure of the viewers' attitude toward video content. Hence, we believe an effective video summarization model should also be helpful to do the affective video content analysis.

In this section, we evaluate the performance of the proposed method for affective computing on the Continuous LIRIS-ACCEDE (CLA). CLA is an annotated emotional database for affective video content analysis [40]. It has valence and arousal self-assessments for 30 movies. The CLA covers several movie genres, such as comedy, animation, action, adventure, thriller, documentary, romance, drama and horror. The total length of the movies in this dataset is 7 h, 22 min, and 5 s. Annotations were collected from ten participants ranging in age from 18 to 27. The annotation process aimed at continuously collecting the self-assessments of arousal and valence that viewers feel while watching the movies. CLA uses the well-known 2D valence-arousal, in which arousal scale measures the intensity of the emotion. It means the video contents with high arousal parts are more attractive and memorable than others. Hence, in this experiment, we try to explore the performance for emotion prediction of the proposed method, and the arousal value is treated as the ground truths for our evaluation.

Fig. 7 shows a sample video called "Superhero" on the CLA dataset with the corresponding arousal values of five different viewers. The different color curve reflects the value of arousal index for each viewer who watched this video, and the red point on the axis denotes the corresponding visual content in this video. This video depicts a sad story about a little boy. The little Jeremy is a shy boy with a vivid imagination. Unfortunately, he was diagnosed with Leukemi. His mother wanted him to be brave and build a superhero in his imagination. From this figure, we can find the arousal value is changing with the content of this movie.

Table 4
The performance comparisons using the average F-measure (AF) on CLA dataset.

	Method	AF (%)
Baseline	Random	13.32
Proposed methods	VSST-RGB&MV	32.13
	VSST-Imbalance	**54.28**

When Jeremy was bullied by other kids in classroom (160th to 165th s), most of the viewers started to have a relatively high level of arousal. When the boy thought of his fantastical hero and fought back (305th to 310th s), all of the viewers were in high spirits. In the middle of the video sequence, when his mother was folding laundry, all of the viewers maintained a stable state of arousal. After several days, Jeremy fell ill, and he dreamed of himself falling down from a building in his coma. In this dream, he was hanging out of the building, but his superhero failed to save him (820th to 825th s), and all viewers were in relatively low spirits. In the end, the little boy was not able to overcome his illness, and his mother said goodbye to her little child with tears from 1070th to 1075th s. If we observe the curve of arousal, we can also find that the viewers were associated with a visible emotional change in this process. We want to investigate that if our model can predict the arousal of the video.

By applying our effective VSST-RGB&MV and VSST-Imbalance models to this emotional dataset, we carried out another phase of experiments to compare the proposed methods with the random baseline. For the random baseline method, we randomly select $M = 15$ percentage of video sequences as the final summary. The ground truths of the videos are generated depending on their arousal value. The experimental results are displayed in Table 4,

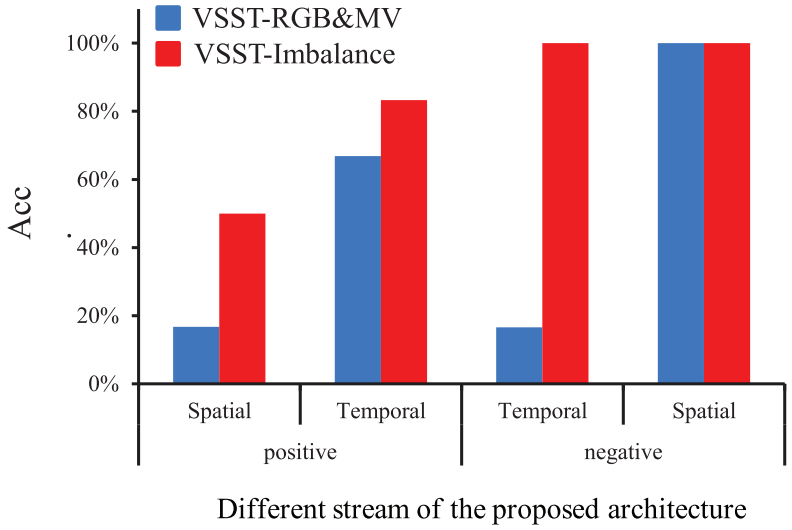

Fig. 8. The validation accuracy of the spatial and temporal streams our proposed methods on positive and negative classes in CLA dataset.

in which the average F-measure (AF) is reported and it shows the similarity between the method and the ground truths. From the results listed in Table 4, it can be seen that the performances of the proposed methods are much better than the random baseline, and our imbalanced model is quite similar to the arousal value of the videos. These results indicate that the proposed method has a potential for affective computing as well as other related applications.

To investigate the effectiveness of our proposed imbalanced two-stream network, we provide the classification accuracy (Acc.) of two versions of our methods: VSST-RGB&MV and VSST-Imbalance on CLA dataset in Fig. 8, and it is shown on negative and positive classes separately. It is known that, in the classical machine learning, the classifiers usually try to minimize the number of errors they will make in dealing with data. This setting is valid when the costs of different errors are equal [31], and as a result, the class imbalance problem causes severely negative effects on the performance of learning methods. In the Fig. 8, the blue bars represent the classification accuracy achieved by VSST-RGB&MV, and the red bars represent the corresponding values achieved by the proposed imbalanced model VSST-Imbalance. From Fig. 8, it is seen that our proposed imbalanced networks improve the validation accuracy of two-stream ConvNets by about 20% in the positive class. And in the negative class of the temporal stream, we can also see that with the help of over-sampling and cost-sensitive learning technique, there is a significant improvement. It supports that the proposed method is effective in addressing the class imbalance problem.

5. Conclusions and future work

In this paper, we propose a novel dynamic video summarization model based on deep learning architecture. While the over-sampling algorithm is conducted to balance the class distribution on training data, and the two-stream ConvNets with the cost-sensitive learning is proposed to handle the class imbalance in feature learning. The novel deep learning architecture for video highlight prediction contains two information streams. In the spatial stream, RGB images are used to represent the appearance of video frames, and in the temporal stream, multi-frame motion vectors are introduced to extract temporal information of the input video.

In empirical validation, we evaluate our proposed method on two datasets. The experimental results demonstrate that the proposed methods produce video summary with better quality compared with the baseline methods as well as the other representative state-of-the-art models. In addition, extensive experimental results also support that our proposed method is able to predict the video content with high level of arousal in affective computing task. Further research can be identified as: (i) to integrate other imbalance techniques with our proposed method; (ii) to apply the proposed method to other video-based applications; (iii) to propose an end to end architecture for video summarization.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61502311, No. 61620106008), the Natural Science Foundation of Guangdong Province (No. 2016A030310053, No. 2017A030310521), the Shenzhen Emerging Industries of the Strategic Basic Research Project under Grant (No. JCYJ20160226191842793), and the Shenzhen high-level overseas talents program.

References

[1] K. Zhang, W. Chao, F. Sha, K. Grauman, Summary transfer: exemplar-based subset selection for video summarization, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[2] Y.J. Liu, C. Ma, G. Zhao, X. Fu, H. Wang, G. Dai, L. Xie, An interactive spiraltape video summarization, IEEE Trans. Multimed. 18 (7) (2016) 1269–1282.
[3] B. Plummer, M. Brown, S. Lazebnik, Enhancing video summarization via vision-language embedding, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[4] B.T. Truong, S. Venkatesh, Video abstraction: a systematic review and classification, ACM Trans. Multimed. Comput. Commun. Appl. 3 (1) (2007).
[5] J. Wu, S.-H. Zhong, J. Jiang, Y. Yang, A novel clustering method for static video summarization, Multimed. Tools Appl. 76 (7) (2017) 9625–9641.
[6] L. Zhang, Y. Gao, R. Hong, Y. Hu, R. Ji, Q. Dai, Probabilistic skimlets fusion for summarizing multiple consumer landmark videos, IEEE Trans. Multimed. 17 (1) (2015) 40–49.
[7] S.K. Kuanar, K.B. Ranga, A.S. Chowdhury, Multi-view video summarization using bipartite matching constrained optimum-path forest clustering, IEEE Trans. Multimed. 17 (8) (2015) 1166–1173.
[8] W.-S. Chu, Y. Song, A. Jaimes, Video co-summarization: video summarization by visual co-occurrence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3584–3592.
[9] J. Xu, L. Mukherjee, Y. Li, J. Warner, J.M. Rehg, V. Singh, Gaze-enabled egocentric video summarization via constrained submodular maximization., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2235–2244.

[10] J. Meng, S. Wang, H. Wang, J. Yuan, Y.P. Tan, Video summarization via multiview representative selection, IEEE Trans. Image Process. 27 (5) (2018) 2134–2145.
[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
[12] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[13] S.-H. Zhong, Y. Liu, B. Li, J. Long, Query-oriented unsupervised multi-document summarization via deep learning model, Expert Syst. Appl. 42 (21) (2015).
[14] S.-H. Zhong, Y. Liu, K.A. Hua, Field effect deep networks for image recognition with incomplete data, ACM Trans. Multimed. Comput. Commun. Appl. 12 (4) (2016) 52:1–52:22.
[15] S. Wu, S.-H. Zhong, Y. Liu, Deep residual learning for image steganalysis, Multimed. Tools Appl. 77 (9) (2018) 10437–10453.
[16] M. Gygli, H. Grabner, L. Van Gool, Video summarization by learning submodular mixtures of objectives, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[17] T. Yao, T. Mei, Y. Rui, Highlight detection with pairwise deep ranking for first-person video summarization, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[18] K. Zhou, Q. Yu, T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, Proceedings of the AAAI Conference on Artificial Intelligence, 2018.
[19] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the International Conference on Neural Information Processing Systems, 2014, pp. 568–576.
[20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, Towards good practices for very deep two-stream ConvNets, CoRR (2015). arXiv: 1507.02159
[21] S.-H. Zhong, Y. Liu, F. Ren, J. Zhang, T. Ren, Video saliency detection via dynamic consistent spatio-temporal attention modelling, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2013, pp. 1063–1069.
[22] V. Kantorov, I. Laptev, Efficient feature extraction, encoding, and classification for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2593–2600.
[23] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans Pattern Anal Mach Intell 40 (6) (2018) 1510–1517.
[24] K. Zhang, W. Chao, F. Sha, K. Grauman, Video summarization with long short--term memory, Proceedings of the European Conference on Computer Vision, 2016.
[25] B. Mahasseni, M. Lam, S. Todorovic, Unsupervised video summarization with adversarial LSTM networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[26] S.E.F. de Avila, A.P.B. Lopes, A. da Luz, A. de Albuquerque Araújo, Vsumm: a mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recogn. Lett. 32 (1) (2011) 56–68.
[27] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[28] M. Gygli, H. Grabner, H. Riemenschneider, L. Van, Creating summaries from user videos, Proceedings of the European Conference on Computer Vision, 2014.
[29] P. Jeatrakul, K.W. Wong, C.C. Fung, Classification of imbalanced data by combining the complementary neural network and smote algorithm, in: Proceedings of the International Conference on Neural Information Processing, 2010, pp. 152–159.
[30] W. Shen, X. Wang, Y. Wang, X. Bai, Z. Zhang, Deepcontour: a deep convolutional feature learned by positive-sharing loss for contour detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3982–3991.
[31] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Trans. Knowl. Data Eng. 18 (1) (2006) 63–77.
[32] C. Huang, Y. Li, C.C. Loy, X. Tang, Learning deep representation for imbalanced classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5375–5384.
[33] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Netw. Learn. Syst. PP (99) (2017) 1–15.
[34] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intel. Res. 16 (1) (2002) 321–357.
[35] H. He, Y. Bai, E.A. Garcia, S. Li, Adasyn: adaptive synthetic sampling approach for imbalanced learning, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 2008, pp. 1322–1328.
[36] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: summarizing web videos using titles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5179–5187.
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proceedings of the International Conference on Learning Representations, 2015.
[38] H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, V. Vapnik, Support vector regression machines, in: Advances in Neural Information Processing Systems 9, MIT Press, 1997, pp. 155–161.
[39] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intel. Syst. Technol. 2 (3) (2011) 27:1–27:27.
[40] B. Yoann, D. Emmanuel, C. Christel, C. Liming, Liris-accede: a video database for affective content analysis, IEEE Trans. Affect. Comput. 6 (1) (2015) 43–55.
[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[42] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R.B. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, Proceedings of the ACM International Conference on Multimedia, 2014, pp. 675–678.
[43] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L.V. Gool, Temporal segment networks: towards good practices for deep action recognition, Proceedings of the European Conference on Computer Vision, 2016.
[44] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime tv-l1 optical flow, in: Proceedings of the DAGM Conference on Pattern Recognition, 2007, pp. 214–223.
[45] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with enhanced motion vector CNNS, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2718–2726.
[46] Y. Li, B. Merialdo, Multi-video summarization based on video-mmr, in: Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services, 2010, pp. 1–4.
[47] D. Potapov, M. Douze, Z. Harchaoui, C. Schmid, Category-specific video summarization, Proceedings of the European Conference on Computer Vision, 2014.
[48] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE Trans. Multimed. 7 (1) (2005) 143–154.
[49] B. Detenber, R. Simons, G. G. Bennett Jr, Roll 'em!: the effects of picture motion on emotional responses, J. Broadcast. Electron. 42 (1) (1998) 113–127.
[50] R. Simons, B. Detenber, T.M. Roedema, J. Reiss, Emotion processing in three systems: the medium and the message, Psychophysiology 36 (5) (1999) 619–627.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

Sheng-hua Zhong received her Ph.D. from Department of Computing, The Hong Kong Polytechnic University in 2013. She worked as a Postdoctoral Research Associate in Department of Psychological & Brain Sciences at The Johns Hopkins University from 2013 to 2014. Currently, she is an Assistant Professor in College of Computer Science & Software Engineering at Shenzhen University in Shenzhen. Her research interests include multimedia content analysis, cognitive science, psychological and brain science, and machine learning.

Jiaxin Wu received her B.Sc. and M.S. in College of Computer Science and Software Engineering from Shenzhen University in 2015 and 2018. She is currently a research assistant in the Department of Computer Science, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her current research interests include video content analysis and deep learning methodology.

Jianmin Jiang received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 1994. He joined Loughborough University, Loughborough, U.K, as a Lecturer in computer science. From 1997 to 2001, he was a Full Professor of Computing with the University of Glamorgan, Wales, U.K. In 2002, he joined the University of Bradford, Bradford, U.K, as a Chair Professor of Digital Media, and Director of Digital Media and Systems Research Institute. In 2014, he moved to Shenzhen University, Shenzhen, China, to carry on holding the same professorship. He is also an Adjunct Professor with the University of Surrey, Guildford, U.K. His current research interests include image/video processing in compressed domain, computerized video content understanding, stereo image coding, medical imaging, computer graphics, machine learning, and AI applications in digital media processing, retrieval, and analysis. He has published over 400 refereed research papers. Prof. Jiang is a Chartered Engineer, a member of EPSRC College, and EU FP-6/7 evaluation expert. In 2010, he was elected as a scholar of One-Thousand-Talent-Scheme funded by the Chinese Government.