Future Generation Computer Systems 93 (2019) 583–595

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A long video caption generation algorithm for big video data retrieval

Songtao Ding [a], Shiru Qu [a], Yuling Xi [a], Shaohua Wan [b,1,*]

[a] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[b] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China

HIGHLIGHTS

- A long video segmentation algorithm is proposed based on the detection of STIPs.
- The dynamic clustering algorithm is adopted to construct the interesting segments.
- We detect the keyframe by directly constructing the region of interest.
- Our LSTM model is also influenced by the rules of attention mechanism.
- We provide experimental results for different stages and get good performance.

ARTICLE INFO

Article history:
Received 16 August 2018
Received in revised form 9 October 2018
Accepted 29 October 2018
Available online 14 November 2018

Keywords:
Big data
Superframe segmentation
STIP selection
KeyFrame selection
LSTM
Video captioning

ABSTRACT

Videos captured by people are often tied to certain important moments of their lives. But with the era of big data coming, the time required to retrieval and watch can be daunting. In this paper, novel techniques are proposed for the application of long video segmentation, which can effectively shorten the retrieval time. The motion extent of long video is detected by the improved of the spatio-temporal interest points (STIPs) detection algorithm. After that, the superframe segmentation of the filtered long video is performed to gain the interesting clip of long video. In the selection of keyframes, the region of interest is constructed by the use of the STIP already obtained on the video clips, and the saliency detection of these regions of interest is utilized to screen out video keyframes. Finally, we generate the video captions by adding attention vectors to the traditional LSTM. Our method is benchmarked on the VideoSet dataset, and evaluated by the BLEU, Meteor and Rouge.

1. Introduction

Semantic description of long videos is critical in the applications like video data retrieval, automatic driving, visual impaired people self-care, video surveillance, elderly and children care. Rapid development of internet technology and the popularity of multimedia devices have boomed the video capture, and approximately 100 h of video resources are uploaded to YouTube every minute. However, these considerable videos are not given accurate captions, which will not be conducive to searching required videos. Automatic generation of accurate video captions and capture of the primary information in videos can effectively solve these problems.

Presently, most researches are primarily conducted for short video or video clips description, but videos in practical applications sometimes are up to tens of minutes and the amount of information contained stays very large. How to segment video sequences accurately and reduce the unnecessary computation in the light of long video contents are the difficulty of this research. Sah et al. [1] proposed identifies interesting segments from long videos using image quality and behavioral performance. Then the key frames are extracted from these interesting segments. Another way of coping with the challenge is keyframe detection, where keyframes are selected such that they best overview the video [2–4]. Ejaz [3] put forward a keyframe extraction method based on variation of video contents, and papers [5,6] cluster low-level features or objects to obtain key-frames. Others resort to web priors to find important frames within a video [7,8].

After the video key frames are obtained, it is necessity to generate video captions. It is a challenging problem due to the dependence of global information from time and the unboundedness of the spacial domain [9,10]. Kojima et al. [11] designed a heuristic method for video object recognition, and specified the rules for producing verbs and prepositions. Following their work, some succeeding approaches [12–14] apply this mechanism to video datasets with a large number of objects and complex cases.

As the development of deep learning technology [15–17], especially the related research of recurrent neural networks (RNNs), has been boosted, video semantic description has been markedly advanced. Mao [18] proposed a multimodal recurrent neural network

* Corresponding author.
E-mail address: shaohua.wan@ieee.org (S. Wan).
[1] Member, IEEE.

https://doi.org/10.1016/j.future.2018.10.054
0167-739X/

(m-RNN) model, combining deep CNNs with RNNs to solve the problems of image caption generation and retrieving creatively. Vinyals [19] put forward a neural image caption (NIC) model. Compared with the m-RNN model, it abandoned the traditional RNNs and adopted the long short-term memory (LSTM) network to solve the problems of gradient vanishing and gradient exploding.

The attention mechanism allows our algorithm to focus on elements, parts or details of a visual environment that we might otherwise have missed. Therefore, it is necessary to introduce the attention mechanism into the process of the generation of video semantic. The attention mechanism was proposed in the field of image recognition first, and Google DeepMind [20] used the attention mechanism for image object detection. Then, [21] completed machine translation tasks by employing the attention mechanism, and they were the first to apply the attention mechanism to the Natural Language Processing (NLP) field. After the success of attention mechanism used in machine translation [22], attention has been proposed to apply to image caption generation [23].

Our proposed method takes advantage of recent advances in textual summaries [24,25], image caption generation [19,26], 3D structure analysis [27], video segmentation [28] and video caption generation [29,30] to summarize long videos. The paper is organized as follows- Section 2 introduces the related work, Section 3 describes the methodology in different stages and Section 4 gives results of experiments.

2. Related work

Semantic summarization of long videos has been largely driven by advancements in identification of interesting segments, automatic selection of video key-frames, and image caption generation. [31] proposed a video semantic summarization approach that generates story from an egocentric video. Given a long input video, their approach selects a short subshots from a longer egocentric video and depicts the essential events. In addition, using keyframes to represent important segments of video has proven to be a very effective step in video summarization. [3] present a motion continuity detection algorithm in order to define a visual attention score. [32] utilized spatial saliency of frame level to detect keyframes.

Different researchers will choose different keyframes due to different preferences. Some choose frames with higher contrast and sharpness [33] some are interested in colors [34], others place emphasis on people and object contents in videos [6]. Facial information is of great importance in keyframe selection [35], however, not every human face appears in every video, thus human face is also an unstable condition.

Leveraging recent improvement of object recognition allows us to build natural language generation systems, although these are limited in their expressivity. Kulkarni et al. [36] proposed a system that automatically generated a natural language description of images by using statistical information collected from the analysis on large amounts of text data and recognition algorithms. Kuznetsova et al. [37] presented a tree-like structure and made use of image titles from the web for image description generation. Although these methods can describe the content of an image initially, they are heavily hand-designed and rigid when it comes to text generation. In order to resolve this problem, researchers used deep learning models to improve the performance of still image caption generation. Kiros et al. [38] proposed a neural language model which can be adopted to retrieve images with given complex sentence queries. Vinyals et al. [19] firstly introduced LSTM units and a deep recurrent generation model for image captioning. Karpathy et al. [39] developed a combination of CNN and bidirectional RNN, casting images and sentences into the embedding space. [40] introduced time information into video semantic description and then put the information into a language model. In addition they also use a spatial temporal 3-D convolutional neural network representation of the short temporal dynamics.

In this paper, the video captions are generated by long video segmentation, key-frame filtering and language model. We evaluate our methods on the MSR-VTT dataset, marked results are acquired with the application of YouTube2Text dataset and certain egocentric videos. The novel contributions of this paper comprise:

(1) A long video segmentation algorithm is proposed based on the detection of spatio-temporal interest points (STIPs), which are applied to replace the traditional optical flow method. Accordingly, the object detection is made to be more robust, and the computation is simplified.
(2) The dynamic clustering algorithm and selective spatio-temporal interest points are adopted to construct the interesting segments, and thereupon the key frames are extracted from these segments. With the testing result in first part, we can cut down on the computing time if we detect the keyframe by directly constructing the region of interest.
(3) We add attention rate to specific regions by adding attention vectors to the traditional LSTM model. This paper proposes a video semantic model based on attention mechanism. We add attention rate to specific regions by adding attention vectors to the traditional LSTM model.
(4) We also provide experimental verification results for different stages of interest point detection, long video segmentation, and key frame selection.

3. Methodology

Our proposed approach consists of four main components:

(1) It is a primary task to detect and remove the redundant frames by using STIPs in the stage of processing video frames.
(2) After redundant video frames have been screened, the long video is segmented by using non-linear combination of different visual elements.
(3) In the stage of key frame selection, the region of interest is constructed by using the STIPs that is obtained in the previous part directly. Saliency detection is implemented on these region of interest to choose suitable key frames.
(4) After obtaining key frames, we propose an LSTM variant model that is combined with attention mechanism. The model is not only affected by long-term information, but also influenced by the rules of the attention mechanism.

The following is a detailed introduction of each module.

3.1. An improved STIP detector for redundant video frame detection

In this paper, the interest points are optimized by improving the STIPs detection algorithm, and then properties of STIPs are used to support the long video segmentation. The STIP detection is a method based on local features, which is easily affected by camera motion, moving background and illumination changing in video sequences. The result of traditional interest point detection contains an amount of useless background information. About 82% of the interest points belong to the background and require background suppression. Thus only 18% of the interest points belong to valid objects [41]. A large number of useless interest points not only interfere with the detection of valid points of interest, but also increase the amount of calculation.

To overcome this problem, an improved Harris–Laplace algorithm is proposed to detect the spatial interest points, and the difference method is applied to discretize the corner operator [42]. Then spatio-temporal constraints are used to implement background suppression. When the effective STIPs are under the

threshold, related videos are identified as duplicate content required to be deleted.

First and foremost, we use Gaussian sliding window to reverse the entire image, and the scale-space representation of a 2D image in different space could be expressed through the convolution of the image and the Gauss Core.

$$L(x, y, \sigma) = G(x, y, \sigma) \otimes I(x, y) \tag{1}$$

Where $L(x, y, \sigma)$ signifies scale-space, I denotes input video frame, and $G(x, y, \sigma)$ is a Gaussian kernel function with scale factor σ. Model based on the 2D Harris corner detection as second-order moments matrix, which use second-order moments matrix of multi-scale as formula:

$$M = \mu(x, y, \sigma_I, \sigma_D) \tag{2}$$

Where x and y represent the pixels coordinate of video frame, σ_I is the integral scale and σ_D represents differential scale. In normal conditions $\sigma_D = s \times \sigma_I$ (In our article, $s = 0.6$). Using Formula (2) to calculate the first set of interest points C_σ, where σ is the spatial scale.

$$C_\sigma = detM - \alpha \cdot (traceM)^2 \tag{3}$$

Where α is a constant, the range of value $(0.4 - 0.6)$ is used to control the number of interest points. When the initial interest points are obtained, it is also necessary to suppress the surrounding interest points. We use the neighborhood suppression label (NSL), and select the central point taking the neighborhood points under evaluation. Similar to [43], an impact factor θ_σ is computed.

$$\theta_\sigma(X, X_{u,v}) = |\cos(\theta_\sigma(X) - \theta_\sigma(X_{u,v}))| \tag{4}$$

Where $\theta_\sigma(X)$ and $\theta_\sigma(X_{u,v})$ are the gradients of points. u and v define the neighborhood range. If $\theta_\sigma(X)$ and $\theta_\sigma(X_{u,v})$ tend to be identical, the impact factor attains to the maximum. If the gradient is orthogonal the impact factor reaches a minimum. For each central point $C_\sigma(X)$, we define a total weight vector $t_\sigma(X)$ as the sum of gradient values as

$$t_\sigma(X) = \int \int_\Omega C_\sigma(X_{u,v}) \times \theta_\sigma(X, X_{u,v}) du dv \tag{5}$$

Where Ω is the coordinate range. The neighborhood suppression factor $\beta\{0.8 \sim 1.6\}$ and an operator $C_{\sigma,\beta}(X)$ is defined as formula:

$$C_{\sigma,\beta}(X) = f(C_\sigma(X) - \beta \times t_\sigma(X)) \tag{6}$$

Where $f(X) = X$, when $X \geq 0$. The temporal constraints also play an important role in the detection of interest point. We consider two consecutive frames at a time and remove the common interest points, since static interest points do not contribute to any motion information.

$$P^T_{\sigma,\beta} = C^T_{\sigma,\beta} / C^T_{\sigma,\beta} \cap C^{T-1}_{\sigma,\beta} \tag{7}$$

Where $C^T_{\sigma,\beta}$ denotes the set of interest points at time T, $P^T_{\sigma,\beta}$ represents the set of non-static interest points at time T; The static interest points are removed from T to $T-1$ frame. In Fig. 1, the final performance of our interest points detection is demonstrated.

Ultimately, these STIPs are used to detect the repeated video frames in long videos. As shown by Fig. 2, When the number of effective interest points are under the threshold T(the frames in the red box), these video frames are identified as duplicate content required to be deleted. Arising from a high repeatability of the video frame, deleting invalid frames would not affect the expression of the video contents.

3.2. Video frame segmentation

After removing a large number of redundant video frames by STIPs, we combine low-level information, such as the image quality with high-level features to segment candidate video frames. We compute an interestingness score by using non-linear combination of fractions including Attention (A), Contrast (C), Sharpness (E), Colorfulness (S) and Facial Impact (F). Finally, the boundary of long video is determined by the interestingness score.

Contrast: For the calculation of the contrast score C, we refer to the method of [33]. Each frame in long video is converted to luminance, and then resampled to $64 \times width$ after lowpass filtering. The score C is the standard deviation of luminance pixel.

Sharpness: Similar to [33], The frames are converted to luminance, then divided up into 10×10 equally spaced regions. For each region, the standard deviation of luminance pixels is calculated on center area. In order to get more accurate result, we calculate three times and each of the three times a random shift is added. The sharpness score E is calculated the maximum of those standard values of deviation.

Colorfulness: Resemble to [34], In color psychology, color tones is important, every video frame is converted to HSV color space. We compute the pairwise Euclidean distances between the geometric centers c_i of each cube i, after conversion to HSV space. The color space is divided into 64 blocks and every block has four equal partitions. The colorfulness measure is computed as follows:

$$S = emd(D_1, D_2, d(a, b)), 0 \leq a, b \leq 63 \tag{8}$$

$$d(a, b) = || \, rgb2luv(c_a) - rgb2luv(c_b) \, || \tag{9}$$

Where D_1 is generated as the color distribution of image such that for each of 64 sample points, the frequency is 1/64. $D2$ is computed from the given image by finding the frequency of occurrence of color within each of the 64 cubes.

Attention: Similar to [3], we predict the human attention score based on spatial and temporal saliency. The formula for Attention Score is

$$A = \alpha \times m + (1 - \alpha) \times v \tag{10}$$

Where $\alpha = 0.7$, m denotes the super frame motion, and v represents variance.

Facial impact: We aim at detecting faces in a frame and use them as features for superframe segmentation. Ptucha et al. [35] reported the importance of facial content in image processing and described a method for image segmentation. We detect faces using the algorithm of [28] and following the rules from [35]. Face exists in each video frame would be assigned a score and the sum of all scores in each frame is reported as a facial impact score, F.

At first, the size of a face F_{size} is normalized to the size of the frame,

$$F_{size} = \frac{W^2_{face}}{W_{frame} \times H_{frame}} \tag{11}$$

Fig. 3 demonstrates the effect of various image feature on the superframe segmentation in the experiment. Different colors represent contribution of features — Attention, Contrast, Sharpness, Colorfulness and Facial Score.

Empirical testing has shown that Attention, Contrast, Colorfulness and Sharpness are essential feature elements for video segmentation. Facial information is of great importance, however, not every human face appears in every video, thus an influence factor η is added to Facial score. The final measure of superframe cut interestingness score is computed as:

$$I_{score} = A \times C \times S \times E + \eta(F) \tag{12}$$

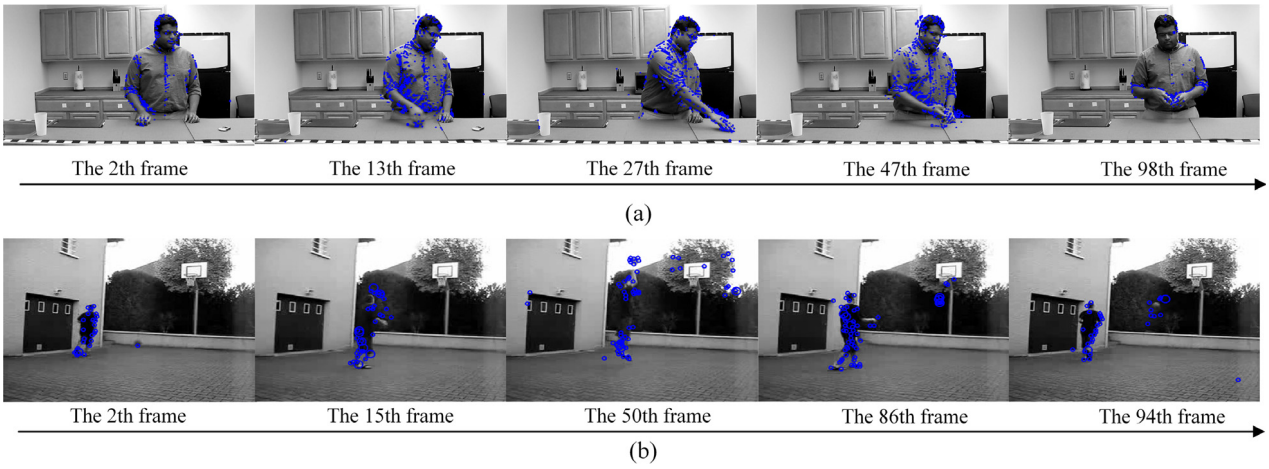


Fig. 1. Two examples from (a) Answering Phone of Hollywood 2 dataset and (b) playing basketball of YouTube dataset, demonstrate the performance of our STIPs detector.

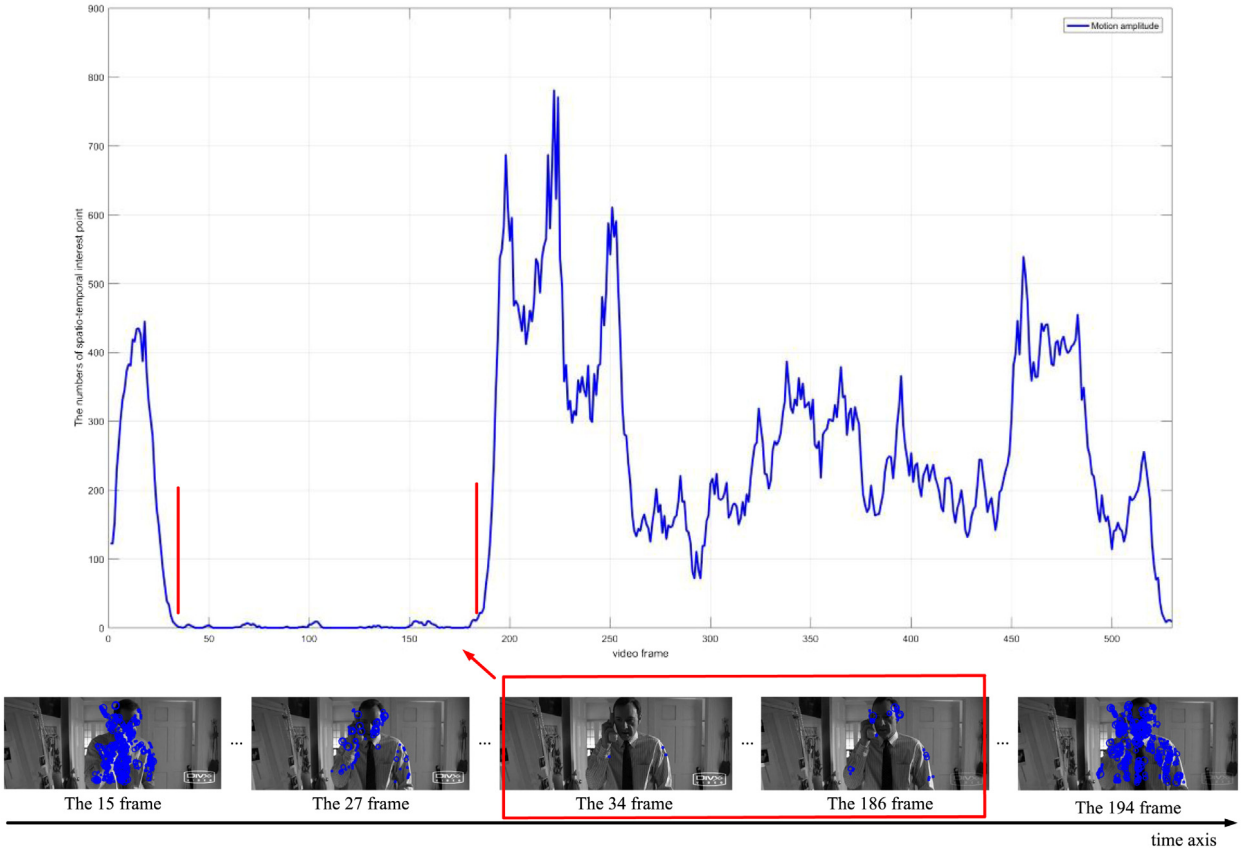


Fig. 2. Detection of change of long video content by using STIPs, the frames in the red part will be deleted.

A long video is divided into some interesting segments through the STIPs detection and video segmentation. These selected super-frame cuts define the more valuable fragments in the original video which are used for video caption generation.

3.3. Key frame selection framework

Video key frame extraction is to select, in the light of certain rules, and denote corresponding contents. In the selection of key frames, the region of interest is constructed by using the STIPs that are obtained in the previous part, and finally the saliency detection of these regions of interest is performed to filter video key frames.

In the first place, the mean shift algorithm is adopted to assemble the interest points in a video frame to obtain the location of all clustering centers. In an N-dimensional space where a candidate interest point has been generated, a random point is selected, and then this point is used as a center, and h as a radius to generate an N-dimensional sphere. All the interest points and the center of the circle within the sphere would generate vectors respectively. The vectors are all formed with the center of interest as the starting point and ending at interest points on the sphere. Finally, these vectors results are added as the Meanshift vector of the area.

$$M_h = \frac{1}{K} \sum_{x_i \in S_k} (x_i - x) \tag{13}$$

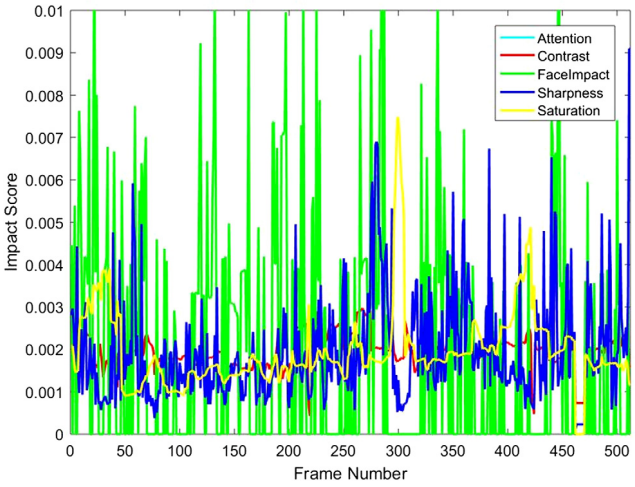


Fig. 3. Impact scores for super-frame cuts in our test video. X-axis is the frame number and Y-axis is the normalized impact score.

S_k is a set of interest points being consistent with the following relations on the spherical region of radius h:

$$S_h(x) \equiv \{y : (y-x)^T(y-x) < h^2\} \tag{14}$$

As exhibited in Fig. 4(a), four pedestrians are in the video, and the dynamic mean drift clustering is implemented to grasp the center of those pedestrians. After completing the clustering of interest points, we shall start to establish the interest region. The Euclidean distance between each interest point and its clustering center point is firstly to be acquired, and the point beyond a certain range will be removed. In this experiment, k-nearest neighbor algorithm is added. In addition, when the distance between two interest points pertaining to the same clustering center outstrips a threshold, this interest point is perceived as an interference point and is to be removed. Where the better range is 0.4 0.6. Too small will cause the area of interest coverage to be insufficient to increase the difficulty of recognition and reduce the recognition accuracy.

As shown in Fig. 4(d), when a cluster center of interest points in accordance with object is certain, the Euclidean distance d_{x_i} between each of the point of interest and the cluster center is calculated. After completion of multi-target center point distance determination and nearest neighbor discrimination, the construction of a region of interest is started. In the vertical direction, d_{x_1} and d_{x_2} represent the largest distance from the center of the target, and their sum is the height of the region of interest. Similarly, in the horizontal direction of the target, select two points with the largest distance from the target center as the width of the region of interest. Thus, the previously determined target center point is the center point of the rectangle. h and w are the height and width respectively, the constructed of regions of interest are green boxes shown in Fig. 4(c). Experiments have proved that the rectangular constructed by the algorithm can quickly and accurately wrap the object in the moving object. The method has 2 points to note,

(1) The number of cluster centers is not pre-set, whereas it is dynamically changed in keeping with the number of targets in the video;

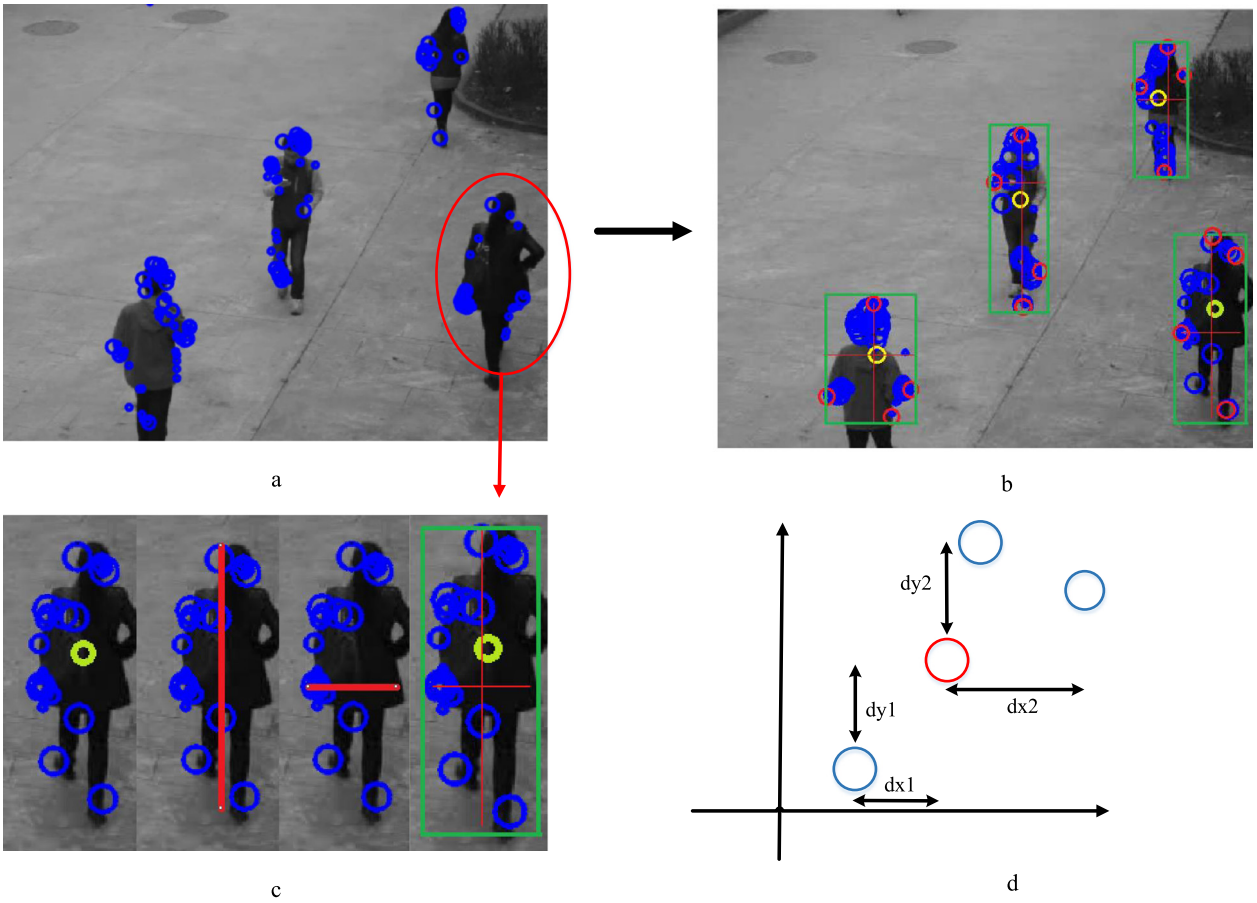


Fig. 4. These candidate interest points are generally attached around the object contour and have good adaptability to light changes, object occlusion, and background disturbances. Using this feature, a clustering center of points of interest of the corresponding object is generated.

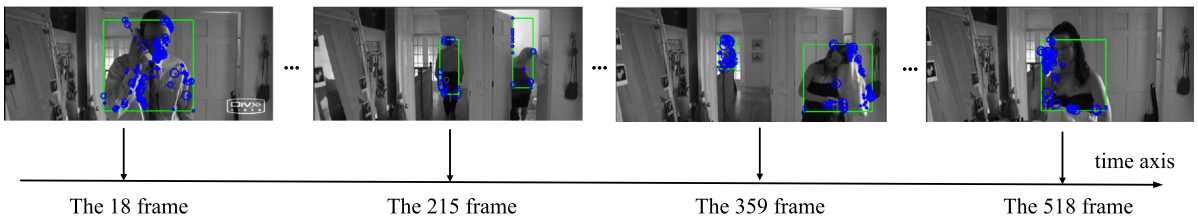


Fig. 5. Using the constructed region of interest for saliency detection.

(2) Appropriate methods are required to remove the interference points near the target, otherwise the region of interest shall be too large;

As denoted by Fig. 5, merely region within the green box will be used to detect key frame, which markedly reduces the amount of computation.

In this paper, salient region detection based on global contrast is introduced to obtain saliency fractions of video frames. A video frame is divided into different regions, then each region is assigned a color saliency fraction. We calculate the saliency fraction by measuring the color contrast between the region and other regions of video frame,

$$S(f_k) = \sum D(f_k, f_i) \tag{15}$$

Where $D(f_k, f_i)$ is the measurement of the color distance in the two regions of the space $L * a * b$. The color distance of the two regions is:

$$S(f_k) = D(f_k, f_1) + D(f_k, f_2) + \cdots + D(f_k, f_N) \tag{16}$$

Where N represents the number of pixels in video frames. The frames belonging to the same video clips are to be compared concerning the saliency.

$$F^* = arg_{f \in (f_p, f_q)} max \sum S(f_k) \tag{17}$$

In contrast with the saliency of different frames f_p, f_q, the final selected key-frame F^* is obtained.

Fig. 6 is the overview of video caption generation algorithm based on attention mechanism.

3.4. Video caption generation framework

After obtaining the key-frame, we pass through the CNN model pre-trained on the ImageNet and get the video captioning. Recent advances in video clip captioning mainly concentrate on the neural network model, while the most successful model is LSTM that not only inherits the benefits of the RNN model, but also solves the problem of gradient explosion in the reverse propagation process.

Through quickly scanning the global image, human vision obtains the target area needing special emphasis, which is generally considered as the focus of attention. Then, more attention resources are invested on this area with the aim to obtain more details about the target while suppressing other useless information. Modulated by this attentional control system, the brain selectively amplifies or filters the sensory information [42]. So we introduce the visual attention mechanism to assist language model to generate more accurate captions. Our attention model generates a caption Y with a sequence of 1-of-K encoded words.

$$Y = \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_C}\} \tag{18}$$

Where K is the size of our vocabulary and C is the length of the caption. Each word y_i is a K-dimensional probability. We use Faster R-CNN as the backbone net for a faster and more principled implementation. Additionally, we use VGG-19 as our main feature extractor. We firstly extract feature vectors from key frame with VGG-19 architecture, then use the region of interest that has been constructed as the attention region and obtain attention regions B'.

$$B' = \{B'_1, \ldots, B'_h\}, h \leq 20 \tag{19}$$

In Faster R-CNN, the lower convolutional layer contains relatively lower-level cues including colors and shapes for proposal judgment, making both useful for our purposes. This allows the language model to selectively pay attention on certain regions of an image. Referring to the strategy in paper [44], we extract features from a lower convolutional layer. This allows the decoder to selectively focus on certain parts of an image by weighting a subset of all the feature vectors. So, we extract features from C5 feature (last convolutional output of 5rd-stage).

$$\boldsymbol{a} = \{\boldsymbol{a_1}, \ldots, \boldsymbol{a_v}\} \tag{20}$$

The extractor gets $v = 196$ vectors, each of these representing a corresponding part of image is 512-dimension.

$$d = \{d_1, \ldots, d_h\} \tag{21}$$

We found sites corresponding to attention regions in original images on the convolution layer via the mapping relation. The region d represents the attention regions on the convolution layer *con*5_3. As shown in Fig. 6, the region d_i corresponds to attention region b_i on original image by using the convolution mapping relationship of CNN. Then a Gauss filter is applied to construct a weight matrix α_0 on the convolution layer *con*5_3. We use the geometric center of attention region d_i as the origin to assign the weights and in the meantime adjust parameter setting on the account of different shapes of region. Ultimately, region with maximum values becomes the attention region.

$$\alpha_0 = (G(x, y), \{d\}) \tag{22}$$

For each location i on *con*5_3, the attention mechanism generates a weight $\alpha_{t,i}$ which can be interpreted either as the probability that location i is the right place to focus on.

When time step $t = 0$, in such circumstances, attention vector r_0 is totally up to image feature a and attention-based weight value matrix α_0.

$$\boldsymbol{r_0} = \sum_{i=1}^{L} \alpha_{0,i} \times \boldsymbol{a_i} \tag{23}$$

When time step $t > 0$, the attention vector r_t is calculated by computing a weighted annotation vector $\{a_i\}, \{\alpha_i\}$ as proposed by [21]. This corresponds to feeding in a α weighted into the system.

$$\boldsymbol{r_t} = \sum_{i=1}^{L} \alpha_{t,i} \times \boldsymbol{a_i} \tag{24}$$

The weight $\alpha_{i,t}$ dimension as $L = 196$ records the focus obtained by each pixel site of image. The attention r_t is not only related to the current image content, but also affected by the previous hidden state. The weight α_t of each annotation vector

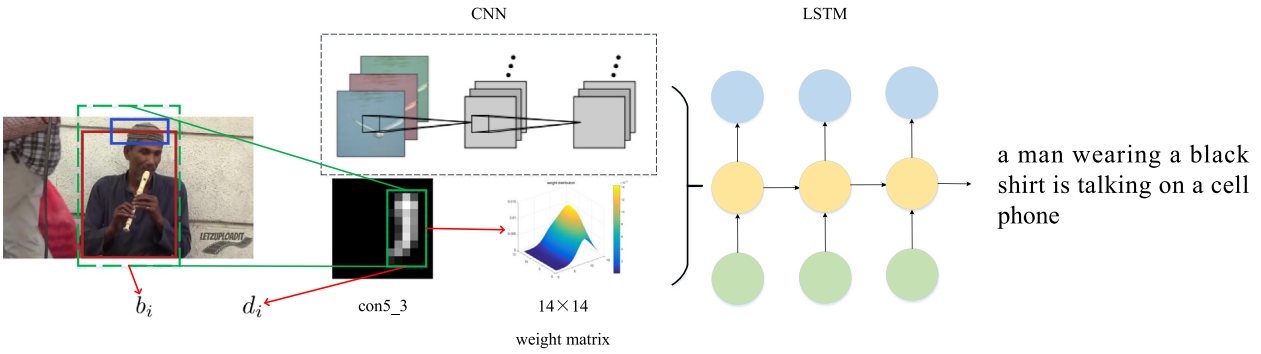


Fig. 6. In the experiment, we took the relative central site in this region on the convolution layer as attention central point, utilized Gaussian Function to calculate weight matrix and ultimately conducted normalization processing on weight value to obtain the weight matrix of attention. The Gaussian distribution on the convolution layer represents attention distribution conditions.

a_i is computed by an attention model f_{att}, for which we use a multilayer perceptron conditioned on the previous hidden state m_{t-1} [44]. The m_t state varies as the output LSTM advances in its output sequence: where attention model looks next depends on image features and the sequence of words that has already been generated.

$$e_{ti} = f_{att}(\boldsymbol{a_i}, \boldsymbol{m_{t-1}}) \tag{25}$$

$$\alpha_{t,i} = \frac{exp(e_{ti})}{\sum_k^L exp(e_{tk})} \tag{26}$$

On that basis, the key-frame feature vectors are transmitted sequentially into a LSTM network to specifically generate a description. Once all frames are processed, a begin of sentence keyword is fed into the network, triggering word generation until and end of sentence keyword is produced.

In addition, Fig. 7 explicitly indicates our variation model. An attend gate is added to endow cell with the focus information and to determine matters to be emphasized. The cell incorporates three inputs, with m_{t-1} and m_t denote hidden state at time t and $t-1$ respectively, x_t refers to a vocabulary vector at time t, and x_0 is image features originated from CNN initially. Attention r_t is selected as the feature vector of the RoI in the key-frame, and, on that basis, the attention of the specific area of the image is increased. c_{t-1} and c_t represent the states of memory cell at time $t-1$ and t.

LSTM cell is defined as follows:

$$\boldsymbol{i_t} = \sigma(W_{ix}\boldsymbol{x_t} + W_{im}\boldsymbol{m_{t-1}} + W_{ir}\boldsymbol{r_t} + \boldsymbol{b_i}) \tag{27}$$

$$\boldsymbol{f_t} = \sigma(W_{fx}\boldsymbol{x_t} + W_{fm}\boldsymbol{m_{t-1}} + W_{fr}\boldsymbol{r_t} + \boldsymbol{b_f}) \tag{28}$$

$$\boldsymbol{o_t} = \sigma(W_{ox}\boldsymbol{x_t} + W_{om}\boldsymbol{m_{t-1}} + W_{or}\boldsymbol{r_t} + \boldsymbol{b_o}) \tag{29}$$

$$\boldsymbol{c_t} = \boldsymbol{f_t} \odot \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \odot h(W_{cx}\boldsymbol{x_t} + W_{cm}\boldsymbol{m_{t-1}} + W_{cr}\boldsymbol{r_t} + \boldsymbol{b_c}) \tag{30}$$

$$\boldsymbol{m_t} = \boldsymbol{o_t} \odot \boldsymbol{c_t} \tag{31}$$

$$\boldsymbol{p_{t+1}} = softmax(\boldsymbol{m_t}) \tag{32}$$

Where $\odot$ denotes the product of a value derived from output gate taking on hidden state, and the weights W and the bias b represent the training parameter matrices. σ signifies element-wise sigmoid function and h indicates the tanh function. The probability shall be generated by softmax to be distributed over all words. Finally, the video keyframes are input into the language model in chronological order, and the corresponding video captions are generated.

4. Experiments

4.1. Evaluating spatio-temporal interest point

A human action recognition experiment in complex scenes based on STIP has been designed in order to test performance of improved STIP detector. For benchmark testing, we use the video database KTH [45]. A good recognition rate is obtain in some other datasets such as UCF [46], YouTube action dataset [47]. Table 1 demonstrates the recognition rates of our method and the performance of our detector clearly.

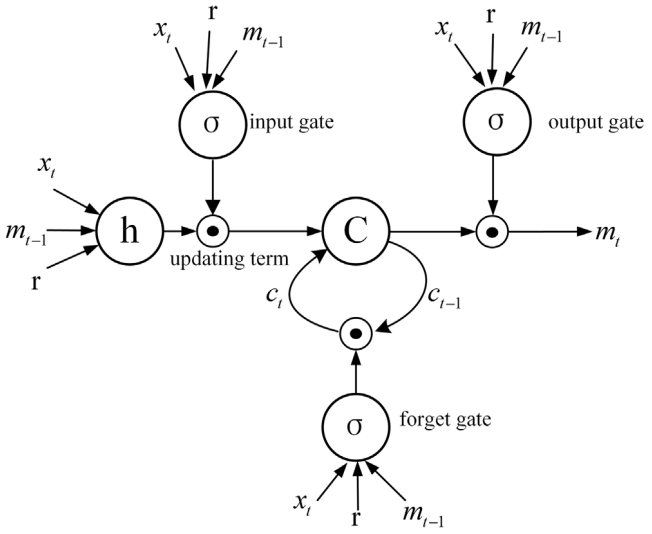


Fig. 7. Our LSTM model.

Table 1
Human action recognition accuracies for KTH, UCF and YouTube.

Method	KTH	UCF	YouTube
Kim et al. [48]	95.33	–	–
Wu et al. [49]	95.10	88.76	–
Lin et al. [50]	93.43	89.38	–
Gilbert et al. [51]	94.5	–	–
Jhuang et al. [52]	91.70	87.80	–
Dollar et al. [53]	81.17	–	–
Our Method	95.58	89.48	82.91

4.2. Evaluating super-frame selection

We use the SumMe Dataset [28] and the Hollywood Dataset [54] to evaluate the effectiveness of our algorithm in superframe selection. The SumMe Dataset consists of 25 videos. The Hollywood Dataset contains 3,669 samples including 12 action categories and 10 scenes, and all of the samples are drawn from 69 Hollywood films.

Fig. 8. Identification of interesting segments from the long video. The test video has a low resolution rate and the background is cluttered, while the scene is constantly changing.

Table 2
Feature evaluation of complex streetscape on SumMe dataset.

Feature	Mean rank	Top-1	Top-2	Top-3
Contrast	0.358	2	2	2
FaceImpact	0.336	1	3	5
Sharpness	0.428	2	4	4
Saturation	0.438	4	5	9
Attention	0.395	1	4	6
Boundary	0.423	2	3	4

Table 3
Feature evaluation on SumMe dataset.

Feature	Top-1	Top-2	Top-3
Contrast	7	8	12
FaceImpact	1	9	11
Sharpness	3	6	11
Saturation	6	8	10
Attention	3	7	9
Boundary	1	6	12

As shown in Fig. 8, Video content shows human activities in the complex outdoor environment by a egocentric camera. Table 2 show the Feature evaluation score corresponding to Fig. 8.

The Contrast and FaceImpact have the lowest scores, but the top-3 shows the balanced nature. Although the Faceimpact was the most important factors in [35], the real outcome can be affected greatly by video clarity and shooting angle.

The Table 3 is an ablation analysis across the 6 features that was performed on all 25 videos. We ranked the contributions of the 6 features, and counted the number of top-1, top-2, and top-3 for each of the 6 features in 25 videos. The top-1 shows that contrast and sharpness have made important contributions to video segmentation. The top-3 shows that all features play a significant role in super-frame selection.

To further test the validity of interest points algorithm, we also conduct comparative experiment. As shown in Fig. 9(a), the black curve signifying motional variation from video frame of 34 to 180 is nearly 0, which implies that these video segments are redundant. After the filtering by STIP, as seen in 9(b), the number of video frames are decreased to 390. The rest of frames could express the primary content of the whole video.

Finally, after the superframe is segmented, the number of interesting segments is reduced from 10 to 7. Since the key frames are in respective correspondence with the segments of interest, the workload of subsequent key frame selection and language model processing is reduced.

Table 4
Evaluation scores for key frame selection.

Camera	Scene	12-neighbor	24-neighbor
egocentric	Cooking	0.50	0.66
egocentric	bike polo	0.54	0.69
moving	Talk show	0.45	0.63
moving	jumps	0.51	0.67
static	car over camera	0.68	0.82
static	Family monitoring	0.66	0.81

As indicated in Fig. 10, the long video that contains more than 23,276 frames is filtered by STIPs algorithm and retains only less than 800 candidate frames.

As indicated in Fig. 11, the long video that contains more than 7373 frames is filtered by STIPs algorithm and retains only less than 1700 candidate frames.

4.3. Evaluating key-frame selection

We use one SumMe Dataset and two Hollywood Dataset to evaluate the effectiveness of our algorithm in key-frame selection. The SumMe Dataset consists of 25 videos, and The Hollywood Dataset contains a total of 3669 samples of 12 action categories and 10 scenes, all samples extracted from 69 Hollywood movies. In the video sample, the expression, posture, dress, camera motion, illumination change, occlusion and background etc. vary drastically, which is similar with real scenes, thus, it is challenging to analyze and identify human behaviors.

The ground-truth key-frames of the videos are manually selected by four students with image processing background. Precision and Dissimilarity are selected as indicators for measuring the quality of our key-frames. Our key-frame is considered correctly if it is within N-neighborhood of a ground truth frame. Table 4 reports the ratio of selected key frames from different neighborhood that match with ground truth.

Finally, the reasonable segmentation of interest obtained by video segmentation can greatly improve the detection accuracy of key frames. Even if you randomly select a video frame in a reasonable segment of interest, it can represent a key frame.

4.4. Evaluating video caption generation

To evaluate our proposed captioning model, we use the MSCOCO 2014 captions dataset [55] in our experiments. For MSCOCO, we

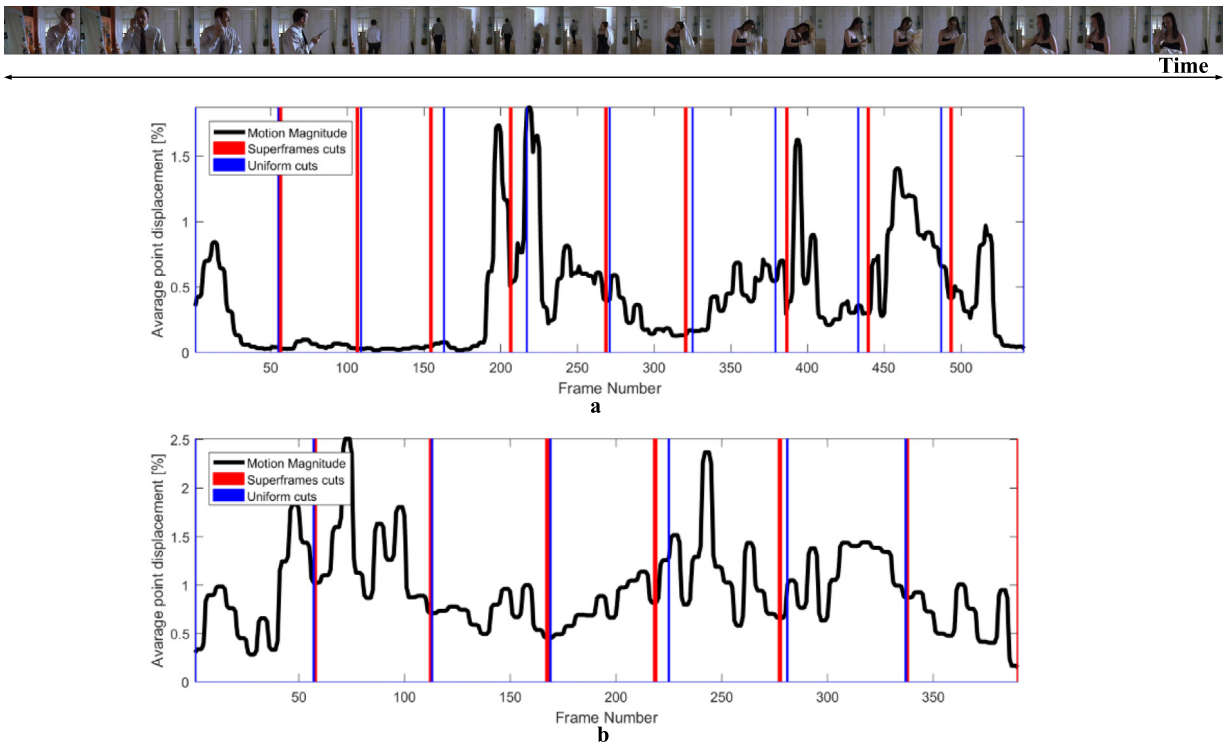


Fig. 9. The black trace shows the motion extent of video, the blue lines represent evenly spaced boundaries and red lines represent the final segmentation results. The video contains 631 frames and threshold value of interest points is set at 50.

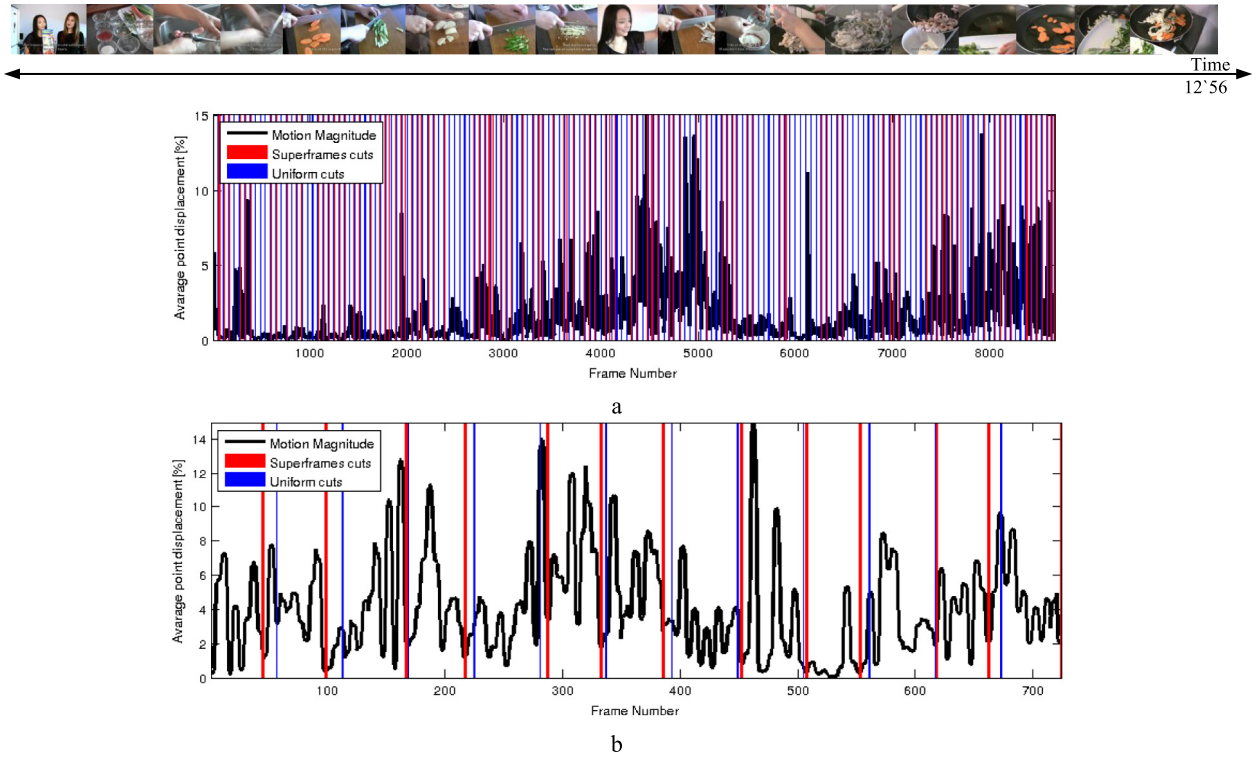


Fig. 10. The video frames that need to be calculated are reduced, and the boundary division is more reasonable.

use 82,783 images for training, 500 images for both validation and testing.

We also collected 160 images of traffic from the Internet and downloaded 1140 images from ImageNet. The bounding boxes are filtered by our attention model to obtain attention regions. The objects in the attention regions are the subjects of sentences. Based on attention regions, corresponding weights of image features on convolution layer con5 are acquired by our model. Finally, weights and features are both added to the language model to train.

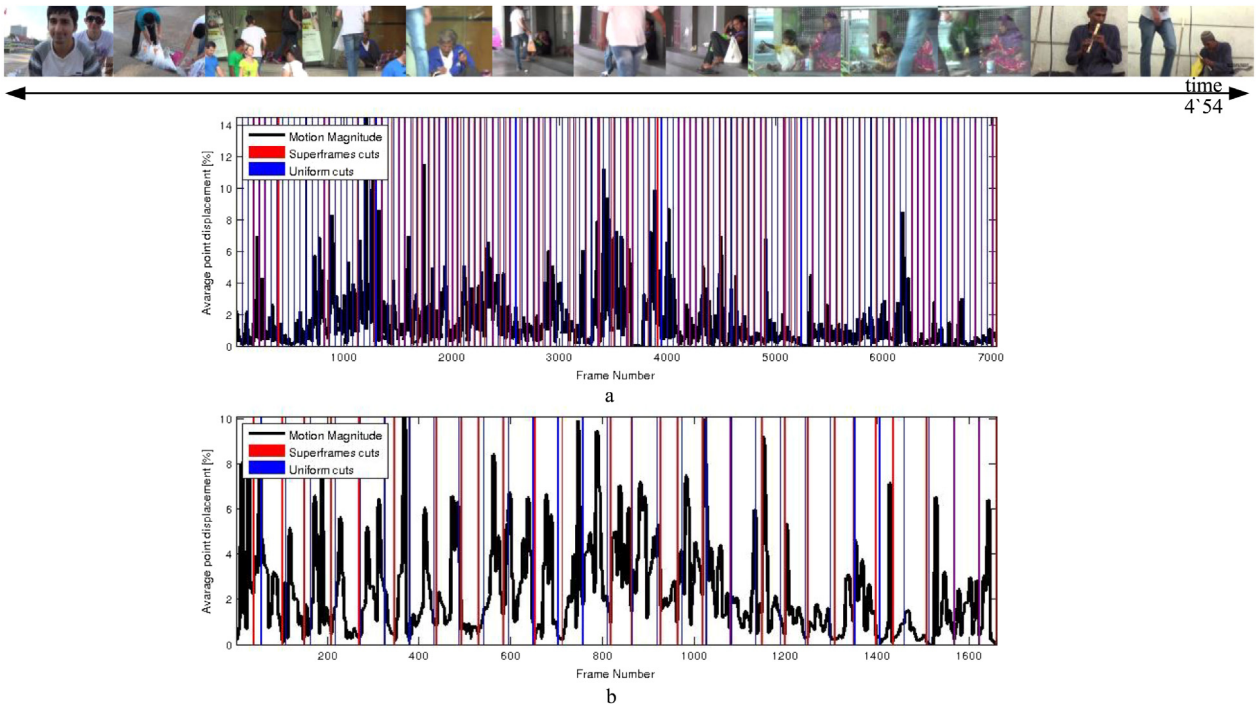


Fig. 11. Compared with the direct multi-frame averaging processing, the removal of redundant frames by judging the number of STIPs, which represent the motion amplitude, can better preserve the boundary information.

Table 5
Leaderboard of published image captioning models on the online MSCOCO test server.

Model	B1	B2	B3	B4
Google NIC [19]	0.713	0.542	0.407	0.309
MSR Captivator [60]	0.715	0.543	0.407	0.308
LRCN [15]	0.718	0.548	0.409	0.306
Hard-Attention [44]	0.705	0.528	0.383	0.277
ATT-FCN [23]	0.731	0.565	0.424	0.316
Review Net [61]	0.720	0.550	0.414	0.313
MSM [62]	0.739	0.575	0.436	0.330
SCST:Att2all [63]	0.781	0.619	0.470	0.352
Up–Down [64]	0.802	0.641	0.491	0.369
Human	0.663	0.469	0.321	0.227
Our Method	0.764	0.563	0.436	0.317

Table 6
Leaderboard of published image captioning models on the online MSCOCO test server.

Model	M	R	CIDEr
Google NIC	0.254	0.530	0.943
MSR Captivator	0.248	0.526	0.931
LRCN	0.247	0.528	0.921
Hard-Attention	0.241	0.516	0.865
ATT-FCN	0.250	0.535	0.943
Review Net	0.256	0.533	0.965
MSM	0.256	0.542	0.984
SCST:Att2all	0.270	0.563	1.147
Up–Down	0.276	0.571	1.179
Our Method	0.265	0.535	1.103

To evaluate caption quality, we use the standard automatic evaluation metrics, namely BLEU [56], METEOR [57], ROUGE [58] and CIDEr [59]. We also give detailed experimental analysis about uncertainty of image caption generation. Image captioning results of different algorithms on MS COCO are reported in Tables 5, 6. B1, B2, B3 and B4 are the BLEU score that uses up to n-grams.

To further prove the effectiveness of our algorithm in video captioning, experiments are performed on the MSR-VTT dataset [65]. The dataset is a massive video benchmark for video comprehension, which comprise 10000 video clips and 41.2 h in total.

As indicated in Fig. 12, the video is from the MSR-VTT dataset and the length is 12:56. Redundant frames exist multiple times in a video. In addition, the details of cooking are very important, the accurate object recognition and language description requirements are high. Sliced green vegetables should virtually be the shallot, and the food should be the potato. Therefore, accurately recognizing small objects in some complex videos remain to be challenging.

Fig. 13 is indicative of the video length 04:50, from the "The Tonight Show Starring Jimmy Fallon". The 5 key frames are selected from the 4000 frames to summarize the video content by using redundant frame detection and superframe cut selection.

As indicated in Fig. 14, the video is from the MSR-VTT dataset and the length is 04:54. Considering complex street scenes, varied objects and irrelevant obstruction, it is difficult for effective textual summarization of the video.

As indicated in Fig. 15, the video is from the Hollywood dataset and the length is 03:25. For fixed scene and simple relationship between objects, the use of interest point algorithm for long video segmentation and the region of interest for key frame detection can greatly reduce the amount of calculation with accurate description of the key frame.

Despite the rapid upgrading of image semantic model and fast improvements in its evaluation scores in recent years, there remains a big gap between human beings and language model in the following several aspects:

(1) The accurate identification of the target is the basis of image caption generation, and the recognition result will affect the accuracy of image captions. However, the language model still has a large gap from humans in terms of small object recognition and object recognition at lower resolutions;
(2) Through the accumulation of experience, human can infer the relationship between targets in an unfamiliar circumstance;
(3) Human descriptions are more vivid and diverse. Compared with language models, humans can give more elaborate and

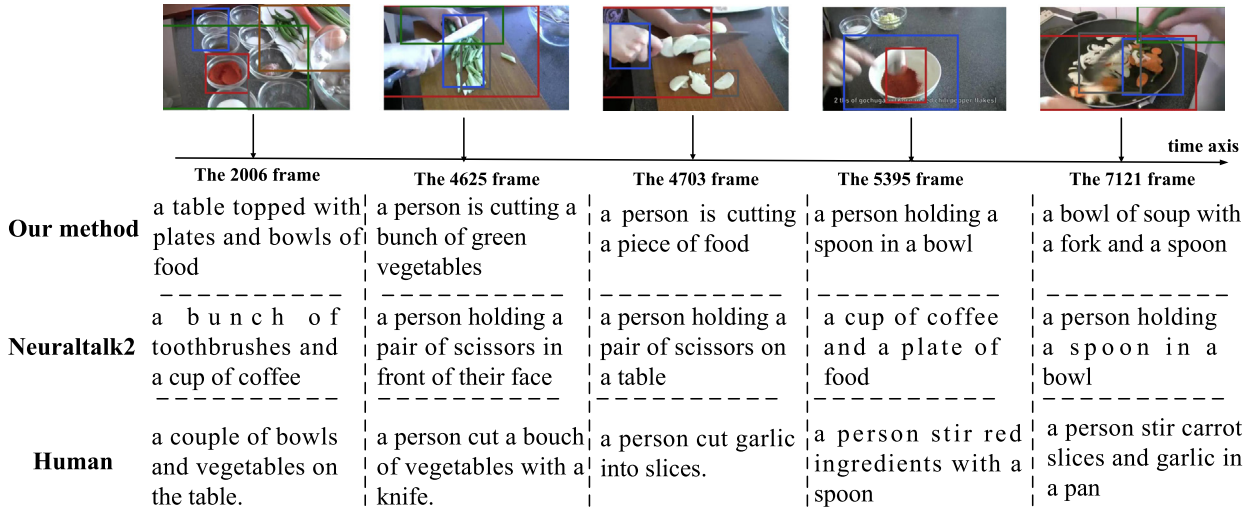


Fig. 12. Qualitative results of our method and the other two algorithms. The theme of the video is clear, which is a process of cooking presented to the audience.

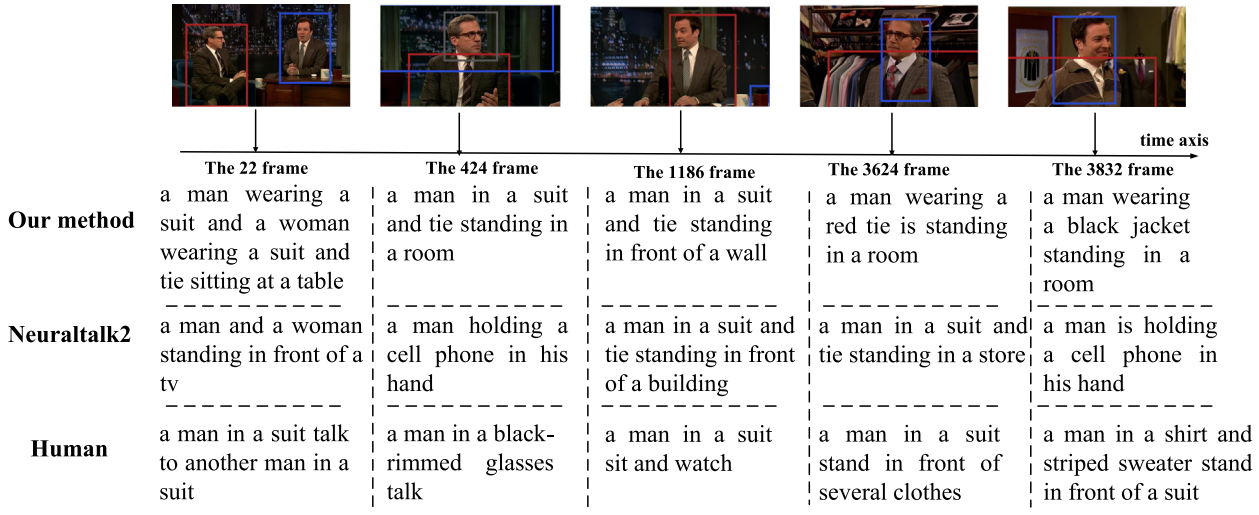


Fig. 13. Qualitative results of our method and the other two algorithms. The talk show is characterized by single scenes and characters, and the movements of characters are relatively simple. More information is included in the conversation, thus it is difficult to rely on the image to understand the video content.

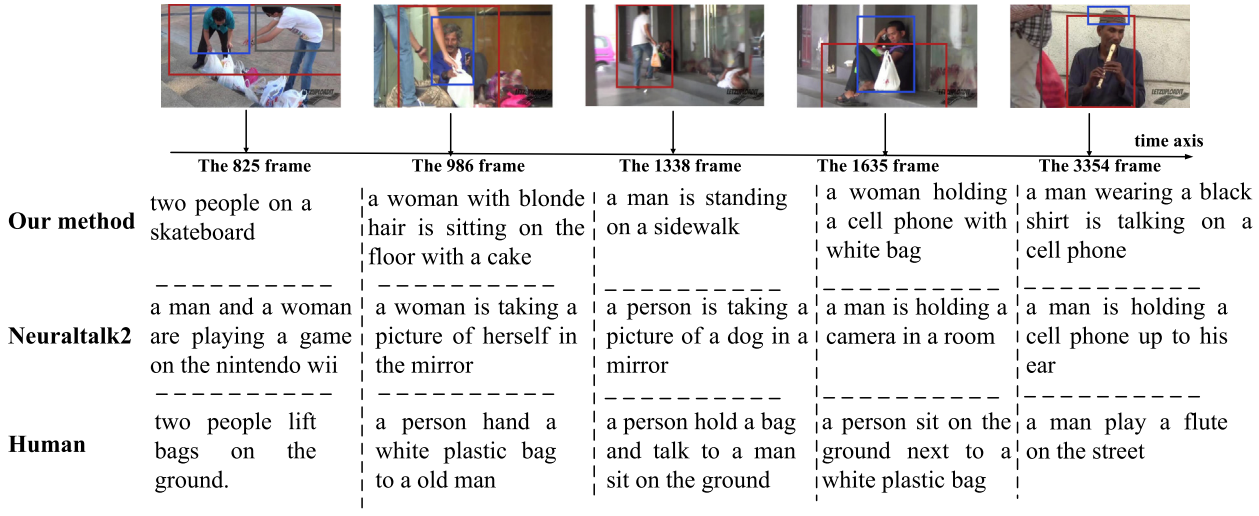


Fig. 14. Qualitative results of our method and the other two algorithms. By adding the attention mechanism, the language model can selectively describe complex scenes containing multiple objects.

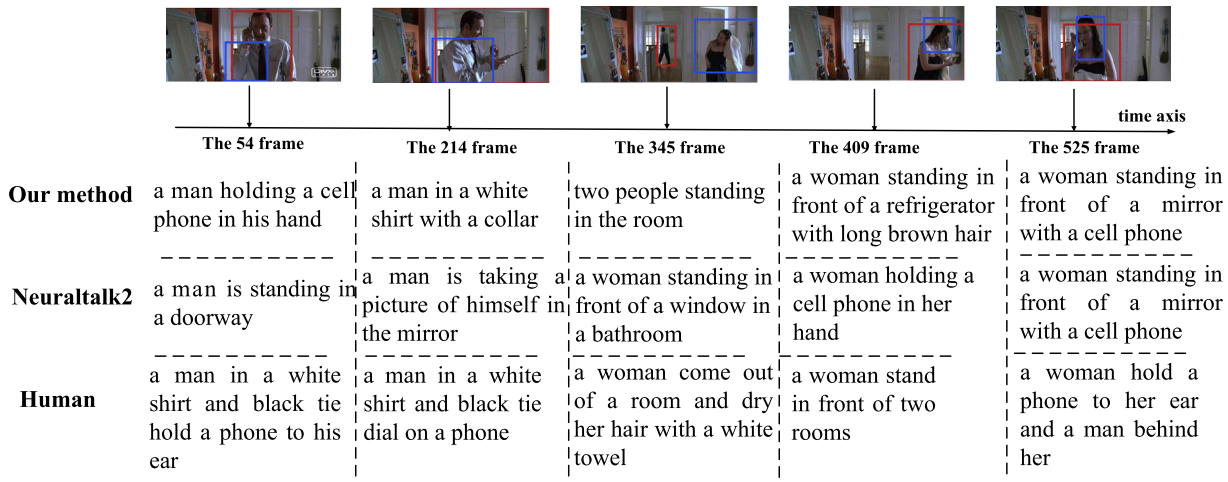


Fig. 15. Qualitative results of our method and the other two algorithms.

accurate descriptions of complex images and behaviors in the images;

5. Conclusion

This paper introduces a novel method for both video segmentation and semantic textual generation. Redundant video frame detection based on STIP and a novel super-frame segmentation are combined to improve the effectiveness of video segmentation. Keyframes from the most impactful segments are converted to video captioning by using the saliency detection and LSTM variant network. The purpose of introducing attention mechanisms is to select more crucial information to the current task from numerous information. In order to improve the accuracy of semantic summarization of long videos, more accurate object recognition network and datasets with detailed captions may be needed. The evaluation results indicates that our approach produces more reasonable video segments and accurate keyframes. Moreover, the language model has good performance on the MSCOCO test server.

References

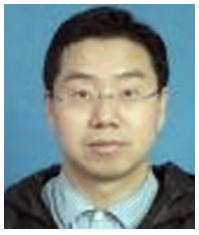
[1] S. Sah, S. Kulhare, A. Gray, et al., Semantic text summarization of long videos, in: Applications of Computer Vision, 2017 IEEE Winter Conference on, IEEE, 2017, pp. 989–997.
[2] W. Wolf, Key frame selection by motion analysis, in: Acoustics, Speech, and Signal Processing, 1996 ICASSP-96 Conference Proceedings. 1996 IEEE International Conference on, vol. 2, IEEE, 1996.
[3] N. Ejaz, I. Mehmood, S.W. Baik, Efficient visual attention based framework for extracting key frames from videos, Signal Process., Image Commun. 28 (1) (2013) 34–44.
[4] T. Huang, S. Mehrotra, Adaptive key frame extraction using unsupervised clustering, in: Image Processing, 1998 ICIP 98 Proceedings. 1998 International Conference on, IEEE, 1998, pp. 866–870.
[5] D. Avila, S.E. Fontes, A.P.B. Lopes, A. Arajo, VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognit. Lett. 32 (1) (2011) 56–68.
[6] Y.J. Lee, J. Ghosh, K. Grauman, Discovering important people and objects for egocentric video summarization, in: Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1346–1353.
[7] A. Khosla, R. Hamid, C. Lin, N. Sundaresan, Large-scale video summarization using web-image priors, in: Computer Vision and Pattern Recognition, IEEE, 2013, pp. 2698–2705.
[8] G. Kim, L. Sigal, E.P. Xing, Joint summarization of large-scale collections of web images and videos for storyline reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014, pp. 4225–4232.
[9] D. Li, J. Zhang, Efficient implementation to numerically solve the nonlinear time fractional parabolic problems on unbounded spatial domain, J. Comput. Phys. 322 (2016) 415–428.
[10] Li Dongfang, C. Zhang, Split newton iterative algorithm and its application, Appl. Math. Comput. 217 (5) (2010) 2260–2265.
[11] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions, Int. J. Comput. Vis. 50 (2) (2002) 171–184.
[12] M.W. Lee, A. Hakeem, N. Haering, S.-C. Zhu, SAVE: A framework for semantic annotation of visual events, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8.
[13] M.U.G. Khan, L. Zhang, Y. Gotoh, Human focused video description, in: Proceedings ofthe IEEE International Conference on Computer Vision Workshops, 2011, pp. 1480–1487.
[14] P. Hanckmann, K. Schutte, G.J. Burghouts, Automated textual descriptions for a wide range of video events with 48 human actions, in: Proceedings of the European Conference on Computer Vision Workshops and Demonstrations, 2012, pp. 372–380.
[15] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[16] S. Venugopalan, M. Rohrbach, J. Donahue, R.J. Mooney, T. Darrell, K. Saenko, Sequence to sequence - video to text, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4534–4542.
[17] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R.J. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, in: Proceedings of the North American Chapter of the Association for Computational Linguistics, 2015, pp. 1494–1504.
[18] J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille, Explain images with multimodal recurrent neural networks, Comput. Sci. (2014) 1090–1410.
[19] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[20] V. Mnih, N. Heess, A. Graves, Recurrent models of visual attention, in: International Conference on Neural Information Processing Systems, 2014, pp. 2204–2212.
[21] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, Comput. Sci. (2014) 1409.0473.
[22] M.T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, Comput. Sci. (2015) 4651–4659.
[23] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651-4659.
[24] G. Durrett, T. Berg-Kirkpatrick, D. Klein, Learningbased single-document summarization with compression and anaphoricity constraints, 2016, arXiv preprint arXiv:1603.08887.
[25] A. Nenkova, S. Maskey, Y. Liu, Automatic summarization, in: Proceedings ofthe 49th Annual Meeting ofthe Association for Computational Linguistics: Tutorial Abstracts of ACL 2011, Association for Computational Linguistics, 2011, p. 3.
[26] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014) arXiv:1409.1556.
[27] L. He, C. Chen, T. Zhang, H. Zhu, S. Wan, Wearable depth camera: monocular depth estimation via sparse optimization under weak supervision, IEEE Access (2018) 99.
[28] M. Gygli, et al., Creating summaries from user video, in: European Conference on Computer Vision, Springer, Cham, 2014, pp. 505–520.
[29] X. Hou, J. Harel, C. Koch, Image signature: Highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. (2012) 194–201.

[30] N. Ejaz, I. Mehmood, S.W. Baik, Efficient visual attention based framework for extracting key frames from videos, Signal Process., Image Commun. (2013) 34–44.
[31] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: Proceedings of the IEEE CVPR, 2013, pp. 2714–2721.
[32] X. Hou, J. Harel, C. Koch, Image signature: Highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 194–201.
[33] Y. Ke, X. Tang, F. Jing, The design of high-level features for photo quality assessment, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 1, IEEE, 2006, pp. 419–426.
[34] R. Datta, D. Joshi, J. Li, J.Z. Wang, Studying aesthetics in photographic images using a computational approach, in: ECCV2006, Springer, 2006, pp. 288–301.
[35] R. Ptucha, D. Kloosterman, B. Mittelstaedt, A. Loui, Automatic image assessment from facial attributes, in: IST/SPIE Electronic Imaging, International Society for Optics and Photonics, 2014, 90200C–90200C.
[36] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, Babytalk: Understanding and generating simple image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2891–2903.
[37] P. Kuznetsova, V. Ordonez, T.L. Berg, Y. Choi, TREETALK: composition and compression of trees for image descriptions, Trans. Assoc. Comput. Linguist. 2 (1) (2014) 351–362.
[38] R. Kiros, R. Salakhutdinov, R. Zemel, Multimodal neural language models, in: International Conference on Machine Learning, 2014, pp. 595–603.
[39] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
[40] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4507–4515.
[41] L. Cao, Z. Liu, T. Huang, Cross-dataset action detection, in: CVPR, 2010.
[42] D. Li, C. Zhang, J. Wen, A note on compact finite difference method for reaction–diffusion equations with delay, Appl. Math. Model. 39 (5–6) (2014) 1749–1754.
[43] B. Chakraborty, M.B. Holte, T.B. Moeslund, et al., A selective spatio-temporal interest point detector for human action recognition in complex scenes, in: Computer Vision, ICCV, 2011 IEEE International Conference on, IEEE, 2011, pp. 1776–1783.
[44] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. vSalakhudinov, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, Comput. Sci. (2015) 2048–2057.
[45] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Pattern Recognition, 2004, ICPR 2004, vol. 3, pp. 32–36.
[46] S. Khurram, Z.R. Amir, S. Mubarak, UCF101: A dataset of 101 human action classes from videos in the wild, in: CRCV-TR-12-01, November, 2012.
[47] J. Liu, Jingen, J. Luo, M. Shah, Recognizing realistic actions from videos, in: Computer Vision and Pattern Recognition, 2009 CVPR 2009 IEEE Conference on, IEEE, 2009, pp. 1996–2003.
[48] T. Kim, S. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: Computer Vision and Pattern Recognition, 2007 CVPR '07 IEEE Conference on, IEEE, 2007, pp. 1–8.
[49] X. Wu, W. Liang, Y. Jia, Incremental discriminative-analysis of canonical correlations for action recognition, Pattern Recognit. 43 (12) (2010) 4190–4197.
[50] Zhe Lin, Z. Jiang, L.S. Davis, Recognizing actions by shape-motion prototype trees, in: IEEE, International Conference on Computer Vision, IEEE, 2010, pp. 444–451.
[51] Andrew Gilbert, J. Illingworth, R. Bowden, Fast realistic multi-action recognition using mined dense spatio-temporal features, in: IEEE, International Conference on Computer Vision, IEEE, 2009, pp. 925–931.
[52] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: ICCV, 2007, p. 2, 6; Jhuang, H. et al., A biologically inspired system for action recognition, in: Proc IEEE Iccv, 2007, pp. 1–8.
[53] P. Dollar, et al., Behavior recognition via sparse spatio-temporal features, in: Joint IEEE International Workshop on Visual Surveillance and PERFORMANCE Evaluation of Tracking and Surveillance, IEEE, 2006, pp. 65–72.
[54] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2929–2936.
[55] T.Y. Lin, M. Maire, S. Belongie, et al., Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, Cham, 2014, pp. 740–755.
[56] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318.
[57] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: ACL-2005, 2005, pp. 228–231.
[58] C.Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004.
[59] R. Vedantam, C.Z. Lawrence, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
[60] H. Fang, S. Gupta, F. Iandola, et al., From captions to visual concepts and back, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.
[61] Z. Yang, Y. Yuan, Y. Wu, W.W. Cohen, R.R. Salakhutdinov, Review networks for caption generation, in: Advances in Neural Information Processing Systems, 2016, pp. 2361–2369.
[62] T. Yao, Y. Pan, Y. Li, Z. Qiu, T. Mei, Boosting image captioning with attributes, OpenReview 2 (5) (2016) 8.
[63] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, 2016, arXiv preprint arXiv:1612.00563.
[64] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and VQA, 2017, arXiv preprint arXiv:1707.07998.
[65] H. Fang, S. Gupta, F. Iandola, et al., From captions to visual concepts and back, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.

Songtao Ding received the M.S. degree from the School of Software Engineering, Beijing University of Technology, Beijing, in 2013. He is currently working toward the Ph.D. degree in the School of Automation, Northwestern Polytechnical University, China. His research interests include computer vision, object detection and image caption generation.

Shiru Qu received her Ph.D. degree in automatic control from Northwestern Polytechnical University, Xian, China, in 2002. She is currently a professor in the Department of Automation at Northwestern Polytechnical University. Her research interests are Computer Vision, Object detection and recognition, Image Processing.

Yuling Xi received the M.S. degree from the School of Automation, Northwestern Polytechnical University, China, in 2018. She is currently working toward the Ph.D. degree in the School of Computer Science, Northwestern Polytechnical University, China. Her research interests include computer vision and image caption generation.



Shaohua Wan received his Ph.D. degree from the School of Computer, Wuhan University in 2010. From 2016 to 2017, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. He is currently an Associate Professor in Zhongnan University of Economics and Law. His research interests include Internet of Things and computer vision.