Visual Saliency Based Aerial Video Summarization By Online Scene Classification

Jiewei Wang*,†, Yunhong Wang*,†, Zhaoxiang Zhang*†

*Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, Beihang University, Beijing, China
†State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

Abstract—Compared with traditional video summarization approaches, aerial video summarization is a new and challenging issue for its particular characteristics. Aerial video data is a massive data stream, without pre-edit structures such as sports or news video data, lack of camera motion such as zoom and pan. On account of these characteristics, we proposed a novel approach for summarization. First, we extract GIST features for each frame as the holistic scene representation. Then, we divide aerial video into temporal segments representing a visual scene using on-line clustering method by examine GIST features of each frame only once. Finally, we select several key frames from each scene for summarization according to visual saliency index (VSI) of each frame computed from their visual saliency map. In the paper, we proposed new criterion for estimation of temporal segmentation of streaming video. Experimental observations show the success of our approach on aerial video summarization.

Keywords—Aerial video summarization; visual attention; online clustering; scene classification; saliency;

I. INTRODUCTION

Aerial video is widely used in military for observing enemy activities and commercial world for monitoring resources such as forests and crops. As the amount of aerial videos increases, there is an urgent need for effective methods of aerial video data transmission and storage. Due to the large amount of redundant data between adjacent frames in video data, it is not necessary to transmit or store every frame. Accordingly we use the video summarization method to merely transmit or store those most informative frames that capture the major elements in aerial video, so as to remove the redundant information in aerial video.

Although video summarization has recently become an active and popular research field, but considering the limitation of the storage, computation and communication capabilities in computing systems on aerial vehicle, existing approaches about video summarization cannot work. So we propose a novel method in the paper that can summarize aerial video while collecting them. Due to constrain of our system, aerial video that we deal with have three important characteristics compared to traditional video.

1) Without pre-edit structure. Current literatures have mainly focused on sports or news videos that conform to well-defined structures and characteristics [1], in which they can easily segment video clips into different shots according to the edit boundary, and then extract key frame from each shot for summarization. But the scene content in aerial video is continuous and changing because of the movement of the camera.

2) Lack of camera motion. Literatures for summarizing video recordings captured by personal digital cameras have been proposed [2]. Although these clips are unstructured, they can still rely on cues related to the camera operation such as zoom, pan and etc. On the contrary, aerial videos have fewer types of these camera operations.

3) Being massive data stream. Existing literature [1], [2] about video summarization generally implement their work by off-line methods on short video clips collected and stored in the database. But in our work, aerial video data is a kind of massive data stream.

Our primary contribution in this paper is a robust and novel aerial video summarization scheme that can satisfy the requirement of on-board system. Since general framework of video summarization segment video into shots according to edit boundaries, duo to first two characteristics of aerial video, there is no such obvious boundaries. Accordingly, we divide aerial video into temporal segments by scene classification. But existing approaches about scene classification generally use off-line method. Duo to the third characteristic of aerial video, we propose a novel online clustering method for scene classification. In addition, we select key frames from each divided temporal segment according to our method combing visual saliency index (VSI) and inhibition of return mechanism.

This paper is organized in the following way: Section 2 gives some existing works dealing with video summarization. In Section 3, we divide aerial video into temporal segments by scene classification. The process of key frame
Video summarization refers to creating an excerpt of a digital video, which must contain high priority entities and events from the video and exhibit reasonable degrees of continuity with little repetition. The challenge in video summarization is how to effectively extract certain content of the video while preserving the essential message of the original video.

Video summarization is not a new problem but there is no literature about aerial video summarization. Due to the structure of the video content, techniques in automatic video summarization can be categorized into two major approaches: structured video summarization [3] and unstructured video summarization [2].

A structured video usually consists of scenes, and each scene includes one or more shots. A shot is an uninterrupted segment of video frame sequence with static or continuous camera motion, such as sports or news videos. The structures facilitate summarization for such videos. To structured videos, they usually adopt the shot boundary detection method [3] to temporally partition videos into shots.

In contrast, there are huge amount of unstructured videos such as personal video clips captured by digital cameras. Because these clips are unstructured, it can only rely on cues related to the camera operator’s general intents, such as camera and object motion descriptors. A video clip is segmented into parts based on major types of camera motion (pan, zoom, etc.) [2].

Unfortunately, aerial videos have neither well-defined structure nor camera motions. The only information we can utilize is the scene context (urban, ocean, desert and etc.). A number of experimental studies have demonstrated that we integrate enough information about the meaning of a scene in less than 200ms. In fact, we recognize its gist as quickly and as accurately as a single object. So, GIST feature is proposed specifically for scene recognition tasks by Oliva and Torralba [4] [5]. In this paper, we use GIST feature to temporally partition videos into segments.

After we segment videos into scenes, the next step is summarization. Key frame extraction is a common way that implements video summarization by selecting a set of summary key frames to represent video sequences. The most simple method is time-sampled the video at pre-defined intervals, and key frames were taken from a set location within the shot or segment. Another way is clustering method, the frames within the shot or segment were divided into several clusters, and key frames were taken from each cluster. The drawback to most of these approaches is that key frames extraction usually depends on low level features. To bridge the semantic gap, [6] proposed an adaptive video summarization method that allows for the extraction of the most meaningful key frames using visual attention clues. In [7], they proposed an effective approach to video indexing based on importance ranking using user attention model.

A number of experimental studies have demonstrated that primates have a remarkable ability to interpret complex scenes in real time. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing, most likely to reduce the complexity of scene analysis. This selection appears to be implemented in the form of a spatially circumscribed region of the visual field, the so-called visual attention. According to this mechanism, [8] proposed a computational model to analog the visual attention of human. [9] and some other paper have been optimized Itti’s work.

III. TEMPORAL SEGMENTATION

Since aerial videos are captured by cameras on aerial vehicle, so the visual scene in them is continuous but changing all the time. Consequently we cannot use bottom visual features such as edit boundary, camera motion, and object motion et al. to divide aerial video into temporal segments (similar to the notion ‘shot’ in edited videos). The only information we can utilize is the scene context (urban, ocean, desert and etc.) of aerial video. We want to utilize some features (GIST feature in this paper) to infer where the aerial vehicle are and what the camera is looking at, and then segment videos into different scenes.

A. GIST feature extraction

GIST feature is proposed specifically for scene recognition tasks by Oliva and Torralba [4] [5]. A number of studies represent a scene by segment and analyze individual objects or regions. But Oliva’s work have suggested that we can define features correlated with scene properties without having to specify individual objects within a scene, just as we can build face templates without needing to specify facial features such as nose, mouse et al.

To compute texture features (Fig. 2), we implement Gabor filters on RGB sub-bands. Each image location is represented by the output of filters tuned to different orientations and scales. We use a steerable pyramid with 3 scales and totally 20 orientations applied to each color sub-band. So the total number of sub-bands is N=20*3=60. The local representation of the output sub-band images i is given by $v_i(x)$.

To capture global image properties, while keeping some spatial information. Therefore, we take the mean value of the magnitude of the local features averaged over large spatial regions:

$$gist_i(x) = \sum_{x'} |v_i(x')|w(x' - x)$$

(1)

where $w(x)$ is the averaging window. The resulting representation $gist_i$ is down-sampled to have a spatial resolution
of $M \times M$ pixels (here we use $M = 4$). Thus, $gist$ composed of $gist_i$ has size $D = M \times M \times N = 960$.

Gist computation speed is less than 1ms/frame. Low computation costs of extracting GIST features facilitate effective usage in real-time video analysis such as our work.

**B. Online clustering**

Recently, the data generation rates in some data sources (customer click streams, telephone records, aerial video data and etc.) become faster than ever before. This rapid generation of continuous streams has challenged our storage, computation and communication capabilities in computing systems. Over the past few years, several approaches based on data stream mining have been proposed. A data stream is defined as a massive unbounded sequence of data elements continuously generated at a rapid rate. Due to this reason, it is impossible to maintain all elements of a data stream [10].

According to the definition above, aerial video is a kind of data stream. Traditional data mining methods cannot be directly applied to the data stream domain due to characteristics of aerial video. For this reason, we proposed a novel data stream processing methods which satisfy constrain of on-board system. Our methods base on the following three principles. First, each data element should be examined at most once to analyze a data stream. Second, memory usage for data stream processing should be confined finitely for new data elements are continuously generated in the data stream. Third, newly generated data elements should be processed as fast as possible to produce the up-to-date analysis result of a data stream, so that the result can be instantly utilized upon request.

1) *Sliding window:* The inspiration behind sliding window is that the user is more concerned with the analysis of most recent data streams. Thus the detailed analysis is done over the most recent data items and summarized versions of the old ones [11].

Compared with common data stream, aerial video data has a excellent characteristic. Since visual context in aerial video is continuous, so the corresponding data of adjacent frames in feature space is adjacent either. For this reason, we believe that those frames which are adjacent in time space and similar in GIST feature space should be clustered into one scene. On the contrary, even though two frames are similar in GIST feature space, if they are far apart in time space, we believe these two frames are in different scenes.

Due to this characteristic of aerial video, we just maintain one window to preserve GIST features of several latest aerial video frames that probably represent a scene.

2) *Scene boundary:* The characteristic of aerial video make us could focus on recent data, at the same time, it brings huge difficult to detect scene boundary. In another word, it is hard to decide when to move the sliding window.

As we analyze before, for a given scene, there are two attributes which affect when to separate frames in the window and upcoming frames in the data stream into two scenes, GIST feature similarity and temporal distance.

We define $gistSim(x, y)$ as the GIST feature similarity between two video frames $x, y$. That is

$$ gistSim(x, y) = \frac{\sum_{i=1}^{D} \frac{\min(gist_x(i) - gist_y(i))}{\max(gist_x(i) - gist_y(i))}}{D} \quad (2) $$

where $gist_x(i)$ and $gist_y(i)$ are the $i$th component of GIST feature of frames $x, y$ respectively. $D$ is the size of GIST feature vector.

In order to separate frames far apart in time space into different scenes, we introduced a decreasing function of the temporal distance $tempSim(x, y)$ proposed by [1]. That is

$$ tempSim(x, y) = e^{-\frac{(x - y)^2}{\alpha^2}} \quad (3) $$

where $t_x$ and $t_y$ are coordinates of frames $x, y$ in time space respectively. We can change the final number of final clusters or scenes by adjusting $\alpha$.

Finally we define $Sim(x, y)$ which reflects the likelihood that two frames $x, y$ belong to one scene. That is

$$ Sim(x, y) = tempSim(x, y) \times gistSim(x, y) \quad (4) $$

We use sum squared error (SSE) to evaluate the scatter of the cluster composed of GIST features preserved in the sliding window. Let $\vec{\mu}$ is the mean vector of the cluster. We compute $SSE$ as:

$$ SSE = \frac{1}{D} \sum_{i=1}^{D} (gist_i - \bar{gist})^2 $$

where $gist_i$ is $i$th component of GIST feature.
\[ SSE = \sum_{x \in W} (1 - \text{Sim}(x, \pi))^2 \] (5)

When SSE is bigger than predefined threshold, which means the scene represented by GIST features in the sliding window has spanned a certain distance. We should divide the frames in the sliding window and upcoming frames into different scenes, and empty the sliding window. Fig.3 shows 7 temporal segments resulted from our online clustering.

**IV. Saliency Based Key Frame Extraction**

After dividing aerial video into temporal segments, there are many approaches to extract key frames. The most common way is clustering method, the frames within one shot or segment were divided into several clusters, and key frames were taken from each cluster. The drawback to most of these approaches is that key frames extraction usually depends on low level features, and users wouldn’t be able to retrieve video according to high level concepts. To bridge the semantic gap, in [6] they described a new visual saliency index (VSI) descriptor based on a visual saliency model. With VSI, we can automatically evaluate the value of each frame in the visual attention aspect. So, the extracted key frames will be most aligned with a human’s perception.

Itti et al. [8] present a saliency based computational model for visual attention. The fundamental idea is based on human vision characteristics that objects with features are distinct from their surroundings. When viewers watch a video sequence they will be attracted by the interesting objects presented in images, because human perception system is sensitive to the contrast of visual signals such as color, intensity and texture. We show two images and their saliency maps in Fig.4.

Let \( S_i \) is the \( i \)th pixel’s value in the generated saliency map according to saliency model proposed by Itti et al. We define VSI of a frame according to every pixel’s position and brightness in the saliency map as follows:

\[
VSI = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot S_i
\] (6)

\[
w_i = \exp\left(\frac{|p_i - p_{\text{center}}|}{2\sigma^2}\right)
\] (7)

For viewers usually pay more attentions to the region nearest the center of frame, we use a Gaussian falloff weight \( w_i \) with variance \( \sigma^2 \). \( p_i \) and \( p_{\text{center}} \) are the position of \( i \)th pixel and frame center respectively. As we can see from Fig. 4, two areas indicated by red circles in (d) are nearer to the center of frame than those two in (b), so the VSI of the (d) is higher than that of (b).

As a VSI curve (Fig.5) is generated according to equation (6), we can select the frame with the maximum VSI as the key frame when only one key frame is required for each segment. However, long or large content segments usually
require the generation of multiple key frames to abstract them. Always using the highest attention key frames as the representative frames can lead to the extraction of similar and redundant key frames.

To avoid extracting redundant key frames, we used a scheme similar to inhibition of return proposed by Itti et al. [8]. When selecting key frames, we focus on frames having highest VSI score first, while all other frames are suppressed. After we chosen the most salient frame as key frame, then we set VSI scores of the frames nearby the key frame to zero. Therefore the next key frame will be chosen from frames far away from the former key frame, and consequently avoid extracting similar key frames.

V. EXPERIMENTAL RESULTS AND ANALYSIS

Since movie industry is an important application field of aerial video. In this section, we applied our proposed approach to several video clips chosen from very famous documentary in China, Rediscovering the Yangtze River. These are totally 15 video clips during from 30 seconds to 2 minutes long.

Since the results of temporal segmentation and key frame extraction both affect summarization, and then we evaluate the quality of them separately. The first experiment assesses the recall and precision of the detected scene boundaries. The correct scene borders are manually identified by human subjects. The second experiment is based on subjective evaluation. Since the quality of a video summary is subject to human perception, we carry out a user study experiment to quantitatively evaluate the informativeness (content coverage) and the enjoyability (perceptual quality) of each summary.

A. Temporal segmentation

We employ recall precision as the measure for performance evaluation. Let $N_c$ be the number of correctly detected scenes, $N_m$ the number of detected scenes by our approach, and $N_h$ the scene detected by human subjects. The recall and precision is defined as

$$\text{recall} = \frac{N_c}{N_h}$$

$$\text{precision} = \frac{N_c}{N_m}$$

Because scene boundaries between two segments are ambiguous, there is no convincing approach to obtain $N_h$. In this experiment, we invited 20 subjects to evaluate the scene boundary.

Because the adjacent frames in aerial video are very similar, so we cannot determine the scene boundaries exactly. For example, if one say frame $i$ is the scene boundary, there must be someone else argue that why not choose frame $i+1$ or $i-1$ as the boundary, since they are so similar.

Consequently, we adopt an approximate method to evaluate the scene boundary. First, we make the scene boundaries detected by proposed approach as reference scene boundaries, and then we make the center frames between adjacent reference scene boundaries as the reference scene boundaries. These students just have to judge if each of the reference scene boundaries be the real scene boundary. If more than 10 subjects think one reference scene boundary be real scene boundary, then we believe it’s the ground truth.

Table I shows the experimental results of temporal segmentation.

### Table I
RESULTS OF VIDEO TEMPORAL SEGMENTATION

<table>
<thead>
<tr>
<th>$N_c$</th>
<th>$N_m$</th>
<th>$N_h$</th>
<th>recall</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>87</td>
<td>123</td>
<td>102</td>
<td>85.29%</td>
<td>70.73%</td>
</tr>
</tbody>
</table>

B. Key frame extraction

To quantitatively investigate the performance of video summarization, we use two crierions for evaluation, informativeness and enjoyability [3]. Informativeness accesses the capability of maintaining content coverage while reducing redundancy. Enjoyability accesses the performance of the attention model in selecting perceptually enjoyable video frames for summaries.

In this experiment, each tested video has two associated summarized key frames, one with 5% of the original video number and the other with 10% of the original length. We also invited 20 subjects to access the quality of these video summaries. The subjects watched the videos from high to low skim ratio, i.e., 10%, 25%, and then the original video(100%), after watching a video, a subject is requested to assign two scores ranging from 0 to 100, in terms of informativeness and enjoyability.

Table II shows the experimental results of key frame extraction.

### Table II
RESULTS OF KEY FRAME EXTRACTION

<table>
<thead>
<tr>
<th>Skim Ratio</th>
<th>Informativeness</th>
<th>Enjoyability</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>91.75</td>
<td>77.85</td>
</tr>
<tr>
<td>10%</td>
<td>94.25</td>
<td>83.05</td>
</tr>
</tbody>
</table>

The experimental results are indeed encouraging. Our proposed temporal segmentation approach achieves 85.29% recall and 70.73% precision, that means our approaches are most aligned with a human’s perception. In the process of extracting key frames, by reducing 95% of the original video content, the overall informativeness drop only around 9%. By reducing 90% of the original video content, the overall informativeness drop only around 6%. This is not surprised since video data provides extra information.
results of Enjoyability show most testers feel enjoyable in the experiment.

C. Speed Efficiency

Because the two main processing of our proposed approach, GIST feature computing and saliency map computing, are both efficient, so we can get the saliency map from the original frame in less than 100ms, and GIST computation speed is less than 1ms/frame. So it is very impossible to get a system to run in real time.

VI. CONCLUSION

We have proposed a robust and novel aerial video summarization scheme that can satisfy the requirement of on-board system. First, we divide aerial video into temporal segments by online scene classification. Then we select key frames from each divided temporal segment according to our method combing visual saliency index (VSI) and inhibition of return mechanism. Our experiment results demonstrate that the proposed method yields a relatively robust performance. In our future research, we will do further video content analysis such as object detection utilize the key frame set obtained in this paper.

ACKNOWLEDGMENT

This work is funded by the National Basic Research Program of China (No. 2010CB327902), the National Natural Science Foundation of China (No. 60873158, No. 61005016, No. 61061130560) and the Fundamental Research Funds for the Central Universities.

REFERENCES


