2022 IEEE International Conference on
Systems, Man, and Cybernetics (SMC)

# NewsThumbnail: Automatic Generation of News Video Thumbnail

Jinyu Li[1], Shujin Lin[2,3,*], Fan Zhou[1], Ruomei Wang[1],
[1]*School of Computer Science and Engineering, National Engineering Research Center of Digital Life*
[2]*Guangdong Key Laboratory for Big Data Analysis and Simulation of Public Opinion*
[3]*School of Communication and Design*
*Sun Yat-sen University*, Guangzhou, China
[*]Corresponding author: linshjin@mail.sysu.edu.cn

2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-5258-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SMC53654.2022.9945444

***Abstract*—Reading news is an important way for people to obtain information. People can quickly sort out the context of events through a short news video. However, there are numerous news generated around the world every day. It's challenging to locate the interesting video. Thumbnails are often used as video covers and play an important role in displaying video content and driving views. Video owners can choose from individual images or elaborate thumbnails to upload to the site. But manually selecting from a large number of frames is time-consuming, and customizing thumbnails requires a high degree of expertise. Therefore, this paper proposes an automatic generation method of news video thumbnail, which can screen out semantically similar contents according to user query and combine them into a thumbnail. In order to facilitate the screening of graphic materials, we also propose a video content structuring method based on multiple cues, which can accurately segment the video into theme units. At the same time, we designed a visual system to display thumbnails and designed a user survey to investigate the performance of this method in news retrieval and understanding. Compared with peer methods, the thumbnails generated by our method can help users better understand the video content and locate the videos they are interested in.**

***Index Terms*—Video summarization, video thumbnail creation, image synthesis, video content structuring**

## I. Introduction

With the rapid development of short video and live broadcast platforms in recent years, everyone can record videos and post them on the Internet or open live broadcast on the platform at any time and place. The daily video production has seen a tremendous increase. News videos account for a large proportion, including anecdotes, diplomatic events and so on. Such videos often summarize one or several events and include their own comments, which contain rich visual and text content so that people can quickly understand the whole process of the event.

Thumbnails are one of the most common methods of summarizing video content and an important medium for displaying content in today's video sharing platforms. High-quality video thumbnails should meet the following characteristics: (1) The content is rich and the video content can be clearly presented. (2) Attractive and able to arouse user interest. At present, video websites support users to customize thumbnails for videos or automatically select them by the system. However, the thumbnails selected by the system are often only a certain frame in the video. There is a high probability that images of low quality or little information will be generated, which cannot fully display the content of news videos. Besides, customizing thumbnails requires highly specialized skills and is not applicable to most people. In this paper, we put forward a novel method of automatically generating thumbnail in news video called NewsThumbnail. We firstly divide the video into thematic units and extracts keyframes and key phrases. Next, we select highly relevant graphic materials according to user query. Finally, a clever layout method is used to integrate the graphic materials into a rich thumbnail that can reflect the content of the video.

The division of video thematic units is not an easy task. Video includes the frames and audio. Continuity and complex structure brought difficulties to people using the video data, so people consider turning complex abstract video stream into structured data, makes the contents of video indexable and easy to handle. In this paper, we propose a method of video content structuring to divide the video into a series of theme units from the channels of vision and audio, but there may still be some deviations in the results of the two channels. Thus we continually fuse them to get more accurate and clear thematic units.

In order to better display the thumbnails, we also designed a visualization system. Fig. 3 shows its main interface, in which news timeline, keyword clouds and news topic cluster are presented. The conducted user survey showed that our method could help users better understand the news content compared with other methods.

The main contributions are as follows.

1) A method of automatically generating thumbnail based on user query is proposed, which can reflect the video content and help user understand the video.

2) A video content structuring method based on fusion of multi-cues is proposed, which can divide videos into many topic units.

## II. Related Work

Video content structuring mainly refers to shot boundary detection. Shot boundary detection is the process of dividing video into shots by comparing the differences of video frame sequences. Janwe [1] detected the histogram distance between frames based on the histogram and judged the shot boundary when the distance was greater than the threshold. Lu [2] further adopted Singular Value Decomposition (SVD) [3] to form a frame feature matrix through the histogram,

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No.2019A1515011953). Jinyu Li and Shujin Lin contribute equally to this article.

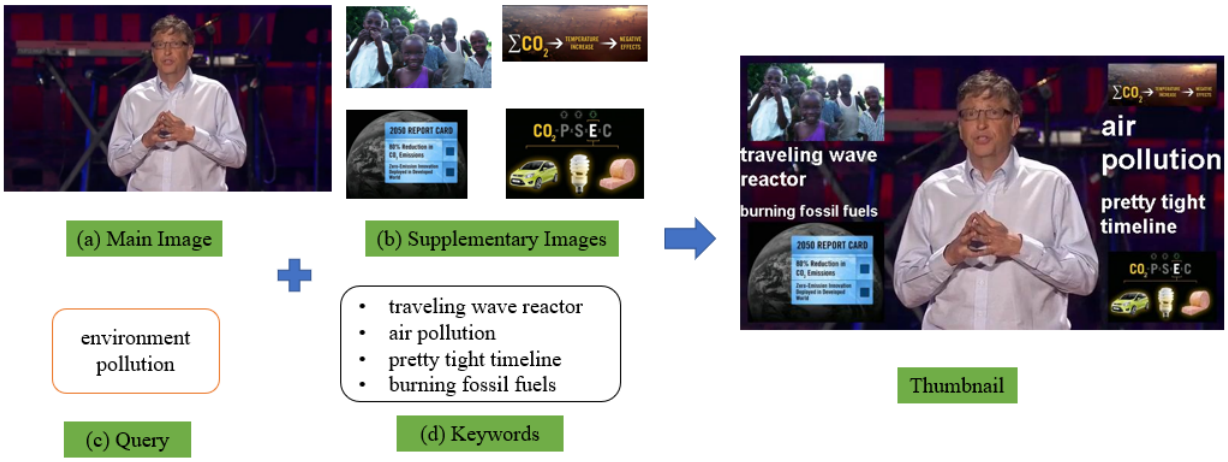



Fig. 1. Thumbnail generation process

which reduced the feature dimension and the amount of calculation. Besides, Heng [4] proposed a boundary-based method. If the boundary between frames changes greatly, it can be considered that shot transfer has occurred. However, histogram and boundary features are too sensitive to illumination and camera motion speed, and the motion-based method solves this shortcoming [5]. Recently, deep learning has become more and more influential. Many researchers tried to use deep convolutional network for shot detection, such as [6] and [7].

A lot of research has been done in order to choose or generate the ideal thumbnail to reduce user stress. Zhang [8] defined 12 features to measure the quality of video thumbnails from three aspects: information richness, user attention and aesthetic value. Song [9] selected attractive thumbnails by evaluating visual quality and aesthetic value, and performed cluster analysis to determine the correlation between thumbnails and video content. But the above methods do not consider the user query. The other text-based methods take it into account. Liu [10] proposed a visual-semantic embedding model, mapping query words and key frames into the common vector space, obtaining their correlation by comparing the similarity of vectors, and selecting the one with the highest correlation as the thumbnail. The above methods are all dedicated to selecting a video frame as a thumbnail. Since the information contained in a picture is very limited and often cannot present enough information to the user, people start to create thumbnails with richer content. Zhao [11] extracted salient visual and textual metadata from videos into magazine-cover-like thumbnail. [12] analyzed people's expressions in the YouTube video frames and selected the most expressive one, then inserted the summary title into the non-salient area of the image to synthesize the thumbnail. Inspired by above methods, we propose a method to generate thumbnail by fusing the phrases and images based on user query. Compared with peer methods, the use of non-salient area of the picture is maximized, thus the generated thumbnail contains rich visual and textual information, which allows users to quickly understand the video content.

## III. Video Content Structuring Method

Video is a kind of comprehensive media composed of visual channel and audio channel, which cannot be interpreted intuitively. Therefore, it is necessary to carry out video content structuring, that is, to decompose video content into a series of sub-units and establish semantic relations among them. This chapter proposes a video content structuring method based on multi-channels fusion. Firstly, topic boundaries are extracted from the visual and audio channels respectively, and then the results of the two channels are fused to segment the video into multiple independent and coherent topic units to build a structured index.

### A. Visual channel semantic boundary detection

In general, the structure of the video in vision can be expressed as "frame - shot - scene - video". The scene is composed of a series of shots. And frame contents under the same shot are similar and vary less. If we only consider from the vision perspective, scene boundary can be simply considered as the boundary of video semantic unit.

In this section, we adopt the method proposed by Son [13] to extract video scenes. Different scenes generally correspond to different video semantic themes. However, in certain types of videos (such as news videos and documentaries), the content is relatively compact and the scene switchs with very high frequency, in which case several scenes constitute a semantic unit and scene boundaries can no longer represent theme boundaries, which we will continue to deal with in Part C.

### B. Audio channel semantic boundary detection

The main useful information contained in audio channel is the transformed text, so we need to preprocess audio information. First, we adopt speech recognition toolkit to convert voice into text. Then filter text out stop words and redundancy information such as timestamp. At last, we use [14] to normalize words into a unified form, such as 'playing' into 'play'.

After getting refined text, we leverage the pre-trained LDA topic model [15] to divide the text information into 40 topics. So each word can be represented as a 40-dimensional vector. Next, we leverage cosine similarity to represent the semantic

similarity between the text block and its context, as shown in (1),

$$Cor(c) = \frac{\sum_{t=1}^{40} \omega_{t,c}\omega_{t,cb}}{\sqrt{\sum_{t=1}^{40} {\omega_{t,c}}^2}\sqrt{\sum_{t=1}^{40} {\omega_{t,cb}}^2}} + \frac{\sum_{t=1}^{40} \omega_{t,c}\omega_{t,cf}}{\sqrt{\sum_{t=1}^{40} {\omega_{t,c}}^2}\sqrt{\sum_{t=1}^{40} {\omega_{t,cf}}^2}} \quad (1)$$

Where $c$ represents the current text block. $cf, cb$ represents the blocks of text before and after $c$ respectively. $\omega_{t,x}$ is the value of the $t$-th dimension characteristic of the text block $x$. Learning from [16], we define $depthScore$ as the difference in score between a text block and its context, reflecting the intensity of semantic changes on both sides of a text block. It can be considered as a theme boundary when large enough, which can be calculated as,

$$depthScore(c) = \frac{1}{2}(Cor(c_l) + Cor(c_r) - 2Cor(c)) \quad (2)$$

Where $c_l$,$c_r$ represents the maximum point of correlation score on both sides of text block $c$. For the convenience of subsequent calculations, we set the number of audio boundaries to be 1.5 times as large as the one of visual boundaries.

### C. Multiple channels fusion

After obtaining the boundaries of visual and audio theme units, we try to fuse the cues of the two channels to more accurately divide the boundaries of video semantic units.

At first, we need to determine the number of video themes. Here, we mainly consider the following points, (a) Different theme units of a video are distributed in different scenes, so the number of video scenes is a factor. (b) The longer the video, the more thematic units it may contain, so the length of the video should be considered. (c) There is usually a large gap between the semantic similarity of the boundary in the text unit, so the larger the $depthscore$ is, the more likely it is to be the topic boundary. To sum up, we calculate it as,

$$TopicCount = \alpha * n + \beta * t + \gamma * count \quad (3)$$

Where $n$ represents the number of video scenes. $t$ is the number of minutes in video duration. $\alpha$, $\beta$, $\gamma$ is variable parameter. $count$ indicates the number of text blocks whose $depthScore$ exceeds the threshold $\rho$. The threshold is defined as,

$$\rho = \frac{1}{m}\sum_{i=1}^{m} ds_i - 0.5\mu \quad (4)$$

Where $\mu$ represents the standard deviation of the text block's $depthScore$, and $ds_i$ represents the $depthScore$ of the $i$-th text block, $1 \le i \le m$. $m$ represents the number of text blocks.

Then we fuse the two channels. Since visual information is more intuitive than voice information, the theme boundary detected by visual cues should be given priority. In addition, if the detected boundary in two channels is similar, we have reason to believe this boundary is very likely to be the real boundary of the video.

We map the detected boundary to the coordinates. Assuming the boundary of the visual channel is $\{x_1, x_2, x_3, \ldots, x_a\}$, and the boundary of the audio channel is $\{y_1, y_2, y_3, \ldots, y_b\}$. From above, we conclude $b = 1.5 * a$. Assuming $n$ to be the number of topics. However, the magnitude of $n$ and $a$ is uncertain. We define the objective function as

$$\begin{cases} \min \sum^{n} |x_i - y_j|, & a \ge n \\ \min \sum^{a} |x_i - y_j| - \theta \sum_{i=1}^{n-a} depthScore(c_i), & a < n \end{cases} \quad (5)$$

If $n \le a$, we directly select $n$ mappings with minimum distance$(x_i, y_i)$; If $a < n$, $(n-a)$ points with the maximum $depthScore$ should be selected from the audio boundary, that is $\{c_1, c_2, c_3, ..., c_{n-a}\}$. For each specific mapping $(x_i, y_i)$, the boundary position can be calculated as,

$$e = \begin{cases} y_j + |x_i - y_j|\phi, & \text{if } x_i \ge y_j \\ y_j - |x_i - y_j|\phi, & \text{if } x_i < y_j \end{cases} \quad (6)$$

Where $\phi, \theta$ is coefficient.

## IV. Automatic thumbnail generation

The main purpose of this chapter is to introduce a method of news video thumbnail generation called NewsThumbnail. Firstly, we extract video keyframes and key phrases. Then we filter out the content not relevant to user query. Finally a content layout method is proposed to arrange images and text to get thumbnail.

### A. Extract keyframes and key phrases

In the last chapter we divide the video into different theme units, each unit includes a series of frames and text. But the graphic material may have high redundancy and be of low quality. So we still need to further refine it.

Extracting keyframes is divided into two steps, first filtering out blurry and dim frames, and then removing redundant frames. To remove blurred images, it is necessary to evaluate the image quality of the video frame. In this chapter, [17] is used to determine the quality of the frame, which is a common way for calculating image blurring in microscopy. Next, we perform shot segmentation on the video frame to remove redundant frames. In order to select the most representative frame within each theme unit, we adopt k-means clustering on it based on the color histogram distance, and select the center of each cluster as the representative frame of each theme unit. Because the semantics that a single word can express is very limited, key phrases can show more information than keywords. Thus we leverage [18] to extract key phrases in text content.


Fig. 2. The process of extracting regions of interest


Fig. 3. The main interface of NewsThumbnail

### B. Query correlation measurement method

To make the thumbnail better meet the user query, it is necessary to select more relevant graphic materials as the thumbnail content. Frames and text cannot directly correspond. Because news video frames themselves contain rich text, we use the speech recognition text from between this to the next to match the query.

Firstly, we leverage GloVe [19] to turn each word in the phrase into a word vector representation, and use the cosine similarity to represent the distance between words $w_1$ and $w_2$ as $Sim(w_1, w_2)$. Suppose $p_i = \{{w_1}^i, {w_2}^i, {w_3}^i, \ldots, {w_n}^i\}$ and $p_j = \{{w_1}^j, {w_2}^j, {w_3}^j, \ldots, {w_m}^j\}$ are two phrases, the similarity between phrases is calculated from the similarity between the words that make up the phrase as follows,

$$Relation(p_i, p_j) = [\sum_{a=1}^{n} \max_{1 \le b \le m} Sim(w_a^i, w_b^j) + \sum_{b=1}^{m} \max_{1 \le a \le n} Sim(w_b^j, w_a^i)]/(n+m) \tag{7}$$

### C. Video thumbnail content layout method

After preparing the graphic materials, we perform saliency region extraction on the image, and then insert the saliency regions of several supplementary images and key phrases into the non-salient regions of a main image. The saliency region extraction process is shown in Fig. 2

The main image is the video frame that best reflects the video content and is the background of the thumbnail. When selecting the main image, we believe that the correlation between the theme unit and the user query, and the duration are two very important indicators. Because thumbnails are adapted to the user query, the correlation with the user query must be considered. Besides, the theme units with longer duration are likely to be more representative and should be preferred, the specific calculation is,

$$S(c_i) = \eta \frac{t(c_i)}{\sum_{j=1}^{N} t(c_j)} + (1-\eta) Relation(query, c_i) \tag{8}$$

Where $C = \{c_1, c_2, c_3, \ldots, c_n\}$ is a series of thematic units. $t(c_i)$ indicates the duration of the theme unit. $query$ indicates the user query. $Relation(query, c_i)$ represents the correlation between user query and the graphic content of the theme unit, which is defined in (7). $\eta$ is the coefficient.

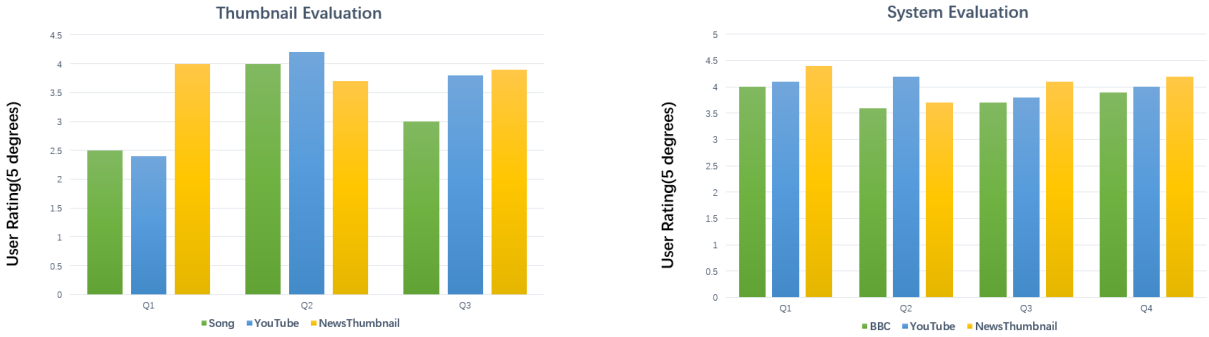
In the process of image embedding, we mainly consider the following factors. (a) Too many embedded images will reduce the readability of thumbnails, so the number of embedded images should not exceed 4, and the embedded positions should be arranged at the four corners; (b) In order to leave enough space for text content, the sum of the salient regions of the embedded images should not exceed 60% of the entire thumbnail area; (c) To ensure the quality of the thumbnail, the idea of the backtracking method is used. The optimization objective function is shown as,

$$\max_{c_i \in C} \sum_{i=1}^{n} S(c_i) \tag{9}$$

Pruning function during search is,

$$\sum_{i=1}^{n} \sum_{c_i \in C_1} S(c_i) + \max \sum_{j=1}^{4-n} \sum_{c_j \in (C - C_1)} S(c_j) \le \sum_{k=1}^{N} \sum_{c_k \in C_k} S(c_k) \tag{10}$$

Where $C_1$ represents the currently embedded semantic unit. $C_k$ represents the other searched embedding result. The semantic units in them are arranged in descending order according to the weights calculated in (8). The first expression to the left of the inequality sign represents the sum of embedded image weights. The second expression represents the maximum possible weight value in the unembedded images. The whole left side represents the maximum possible weight under the current situation. The right side represents




(a) (b)

Fig. 4. The result of user study

the maximum value of the weights of other feasible solutions that have been searched. During the search process, if the maximum possible weight of the search subtree is smaller than the weight of other feasible solutions, then this path does not need to be searched, and the pruning strategy is implemented.

In the process of phrases embedding, we mainly consider the following factors. (a) We stipulate that the sum of the remaining area occupied by the phrase does not exceed 80% of it; (b) To highlight the importance of phrases in thumbnails, here we preset 10 fonts with decreasing sizes, and the phrases with high relevance should also be with larger fonts in thumbnails; (c) Phrases with high relevance should be inserted first. According to the length of the phrases, the phrases can be arranged in various forms. Due to its complexity, we try to insert the phrases with high correlation to smallest rectangle with the largest font. The process of generating thumbnails is shown in Fig. 1.

## V. SYSTEM OVERVIEW AND USER STUDY

To better present the results of NewsThumbnail, we designed a visualization system. After the user enters the query words, the system will cluster the search results by time and content. In the middle is a timeline and a word cloud generated by the video content summary. The left and right sides show videos from official and personal institutions respectively. The cover of the video is generated by our method, and can be enlarged with suspension. The main interface is shown in Fig. 3. We can click to see the detailed video information, as shown in Fig. 5. The upper part is the video list under the category, and the lower part is the video player part, showing the detailed information of the video such as title, abstract, release time, etc.

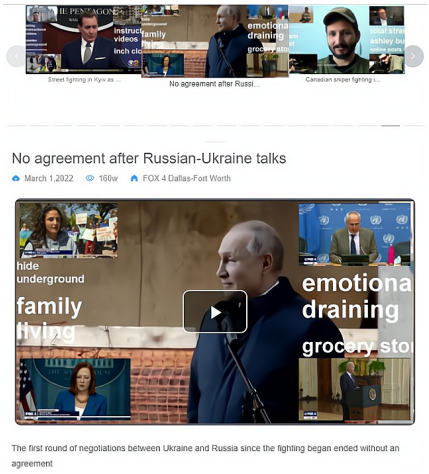



Fig. 5. The detailed information about video

Besides, we designed a user survey to verify the effectiveness of NewsThumbnail and the performance of our visual system. In our research, we hired 30 volunteers (15 males and 15 females) to participate, all between the ages of 18 and 40, and all have the experience of retrieving English news on video sites. Before starting the test, we designed a warm-up session for the volunteers to understand the function of the system and learn how to use it.

### A. Thumbnail evaluation

We compared NewsThumbnail with Song's method [9] and video thumbnails provided by YouTube from 3 aspects: richness(Q1), beauty(Q2) and attractiveness(Q3). Each volunteer is asked to watch 3 different videos on YouTube carefully and then evaluate thumbnails generated by each method with a rating from 1 to 5. The result is shown in Fig. 4(a). Compared with traditional thumbnail, our method contains richer information and is more attractive. However, due to the embedding of graphic materials, the readability may be slightly inferior. The comparison of the thumbnails is shown in Fig. 6 and the question is set as follows:

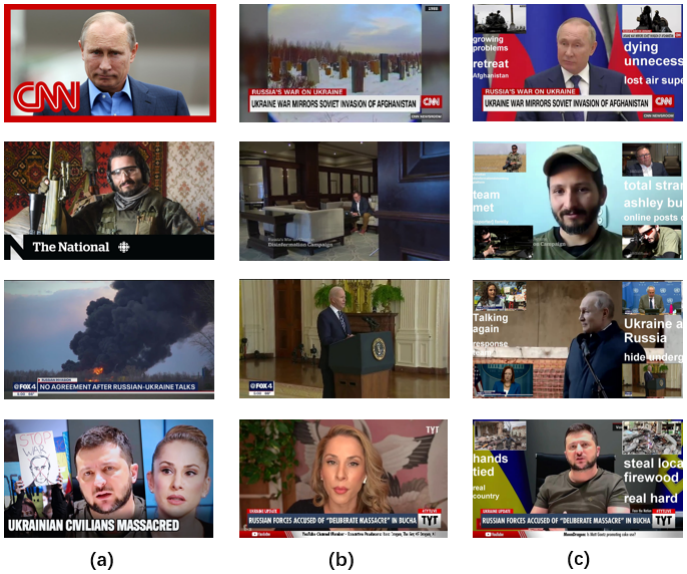
Q1: Do you think the thumbnail contains rich information?
Q2: Do you think the thumbnail is beautiful?
Q3: Does the thumbnail grab your attention?

### B. System evaluation

To verify the effectiveness of our method in practical applications. We compare our visual system with YouTube and BBC. BBC is a news broadcaster established in the

TABLE I
STATISTICAL SIGNIFICANCE TESTS OF THUMBNAIL EVALUATION

| | | *Song* | *YouTube* | *NewsThumbnail* |
|---|---|---|---|---|
| Q1 | mean | 2.53 | 2.41 | **4.01** |
| | stddev | 0.30 | 0.24 | 0.28 |
| Q2 | mean | 4.00 | **4.17** | 3.71 |
| | stddev | 0.31 | 0.19 | 0.15 |
| Q3 | mean | 2.97 | 3.77 | **3.93** |
| | stddev | 0.32 | 0.12 | 0.19 |




(a) (b) (c)

Fig. 6. (a) is the thumbnail generated by YouTube; (b) is the thumbnail generated by Song's method; (c) is the thumbnail generated by our method

United Kingdom and the largest news media in the world. YouTube is one of the largest comprehensive video retrieval systems and allows users to upload their own videos. We selected the parts of the two systems that are most relevant to our work for a comparative study. This paper refers to the methods proposed by Crabb [20] and Yadav [21] for system performance evaluation, which are commonly used in the literature to evaluate the efficiency and effectiveness of video systems. The question is set as follows:

Q1: Do you think the function of the system can help you understand the video content?

Q2: Are you satisfied with the user interface of the system?

Q3: Do you think the system is easy to operate?

Q4: What is your overall satisfaction assessment of the system?

Each volunteer will watch 3 short news video using each system and then answer the question with a rating from 1 to 5. The result is shown in Fig. 4(b). It shows that NewsThumbnail achieved the highest scores for understanding video content (Q1), ease of operation (Q3), and overall system satisfaction (Q4). Compared with other systems, the thumbnails generated by NewsThumbnail contain richer graphic content, which can help users better understand the video content. At the same time, NewsThumbnail is also the simplest in function, so the operation will be relatively simple. The statistical significance tests is shown in TABLE I. The $stddev$(standard deviation) of all do not exceed 0.5. Thus it can be proved that there is no malicious grading in the experiment.

## VI. CONCLUSION

This paper proposes a method for structuring video content based on multi-channels fusion, which combines the results of semantic boundary detection on both visual and audio channels. In addition, on the basis of above method, an automatic news video thumbnail generation method is proposed to combine the graphic content related to user query into a rich thumbnail. At the same time, this paper also involves a visualization system to show the convenience of thumbnails in news video search, which has achieved high user satisfaction in user study.

## REFERENCES


[1] Nitin J Janwe and Kishor K Bhoyar. Video shot boundary detection based on jnd color histogram. In *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, pages 476–480. IEEE, 2013.

[2] Zhe-Ming Lu and Yong Shi. Fast video shot boundary detection based on svd and pattern matching. *IEEE Transactions on Image processing*, 22(12):5136–5145, 2013.

[3] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear algebra*, pages 134–151. Springer, 1971.

[4] Wei Jyh Heng and King N Ngan. An object-based shot boundary detection using edge tracing and tracking. *Journal of Visual Communication and Image Representation*, 12(3):217–239, 2001.

[5] Patrick Bouthemy, Marc Gelgon, and Fabrice Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE transactions on circuits and systems for video technology*, 9(7):1030–1044, 1999.

[6] Lifang Wu, Shuai Zhang, Meng Jian, Zhe Lu, and Dong Wang. Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks. *IEEE Access*, 7:77268–77276, 2019.

[7] Rui Liang, Qingxin Zhu, Honglei Wei, and Shujiao Liao. A video shot boundary detection approach based on cnn feature. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 489–494. IEEE, 2017.

[8] Boyan Zhang, Zhiyong Wang, Dacheng Tao, Xian-Sheng Hua, and David Dagan Feng. Automatic preview frame selection for online videos. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2015.

[9] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 659–668, 2016.

[10] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3707–3715, 2015.

[11] Baoquan Zhao, Shujin Lin, Xin Qi, Zhiquan Zhang, Xiaonan Luo, and Ruomei Wang. Automatic generation of visual-textual web video thumbnail. In *SIGGRAPH Asia 2017 Posters*, pages 1–2. 2017.

[12] Akari Shimono, Yuki Kakui, and Toshihiko Yamasaki. Automatic youtube-thumbnail generation and its evaluation. In *Proceedings of the 2020 Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia*, pages 25–30, 2020.

[13] Jeong-Woo Son, Sang-Yun Lee, So-Young Park, and Sun-Joong Kim. Video scene segmentation based on multiview shot representation. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 381–383. IEEE, 2016.

[14] Martin F Porter. An algorithm for suffix stripping. *Program*, 1980.

[15] Xuan-Hieu Phan and Cam-Tu Nguyen. Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference, 2006.

[16] Martin Riedl and Chris Biemann. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, 2012.

[17] Yu Sun, Stefan Duthaler, and Bradley J Nelson. Autofocusing in computer microscopy: selecting the optimal focus algorithm. *Microscopy research and technique*, 65(3):139–149, 2004.

[18] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20, 2010.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[20] Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J Hughes. Online news videos: the ux of subtitle position. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers & accessibility*, pages 215–222, 2015.

[21] Kuldeep Yadav, Ankit Gandhi, Arijit Biswas, Kundan Shrivastava, Saurabh Srivastava, and Om Deshmukh. Vizig: Anchor points based non-linear navigation and summarization in educational videos. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 407–418, 2016.