2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS) | 978-1-6654-9463-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIS54925.2022.9882448

Film Analysis from the Perspective of Cinemeitrics Based on OpenCV and Deep Learning

1st Jiangnan Sun
School of Computer Science and Cybersecurity
Communication University Of China
Beijing, China
e-mail: 826802411@qq.com

2nd Chunfang Li
School of Computer Science and Cybersecurity
Communication University Of China
Beijing, China
e-mail: LCF@cuc.edu.cn

3rd Ruihan Tang
School of Computer Science and Cybersecurity
Communication University Of China
Beijing, China
email: 2322667344@qq.com

***Abstract*—As a new method of film research, Cinemeitrics uses a systematic and digital way to measure and analyze film style, It focuses more on the measurability of objects and the accuracy of results. As an artistic medium, film has become a popular recording tool for creating historical documents. It is more and more actively getting rid of the shackles of text archives and playing an increasingly important role in recording the times, writing history. So this paper takes the historical film ”The Founding of a Republic” as an example, uses image processing tools such as OpenCV to quantitatively analyze the film. The main work includes: shot segmentation, key frame extraction, subtitles extraction, object recognition and face tracking, providing data can be used to analyze film's style, rhythm and emotion.**

***Index Terms*—Cinemeitrics, OpenCV, image processing, Computer Vision**

I. Introduction

Roberto. Rossini, the Italian film master who made ”Rome without defense”, once said that film should become one of the means of writing history. Perhaps, it is more valuable than other means[1].As an artistic medium, film has become a popular recording tool for creating historical documents. It is more and more actively getting rid of the shackles of text archives and playing an increasingly important role in recording the times, writing history[2].

Recent years, driven by the flourish of Computer Vision and the emergence of data analysis tools, Cinemeitrics' research is spring up in the the U.S. and Europe, different from traditional film research, it measures and analyzes films in a systematic and digital way, which can provide a more objective and accurate film analysis mode[3].

International Cinemeitrics has sprung up, targeting at mainly European and American films, research on Chinese films is comparatively fewer. Chinese Cinemeitrics, at the initial stage, is mainly focusing on box office,the audience and cinema, or external commercial big data, Lacking of quantitative analysis of the film itself. And the existing film measurement tools and algorithms still have some shortcomings, even can not meet the basic needs of film research[4].

Film research scholar usually analyzes films from the following three measurable aspects[5]:

1.lens length distribution, which always used to analyze film style;

2.Editing rate, which is always used to present the rhythm. Rhythm is a high-level semantic structure, that can present the tone or emotion of films.

3.scene classification, which shows the shooting range of the lens, has a very strong visual directivity and guiding function.

So in this paper, we present some research work based on image process technology in the three topics mentioned above, the main tasks of this paper are as follows:

- Shot segmentation. In this part, we used histogram algorithm and face tracking to judge whether the shot has been switched.
- Subtitles extraction. Key technologies include image similarity calculation and Baidu API.
- Scene object recognition. Firstly,using frame difference algorithm to extract key frames of the film, and then the objects in the key frames are recognized by vgg-16 network.

In addition, in order to more comprehensively analyze the film style, we calculated the average frame length, the number of long shots, and the longest shot duration of the film, and use some tools like WordCloud to visualize the experimental results.

II. Relative Tools

The relative techniques of this article based on python, a language which is currently very good at image processing, consist of the OpenCV library and Image similarity calculation algorithm for video frame extraction, image segmentation and other applications, Baidu general scene text recognition API for subtitle recognition, and Vgg-16 network for object recognition. In order to make the analysis more intuitive, this experiment visualize the results using the WordCloud library.

A. OpenCV

OpenCV, consisting of a series of C functions and a small amount of C++ classes, is an open-source computer vision library released from Intel that realizes many general algorithms in image processing and computer vision. Opencv has more than 300 cross-platform medium and high-level API functions based on C. Thus, it is widely used in the field of image recognition at present.

978-1-6654-9463-2/22/$31.00 ©2022 IEEE
ICIS2022, June 26-28, 2022, Zhuhai, China

B. Histogram algorithm

The shot segmentation part of this paper applies Histogram algorithm.

Histogram algorithm is to collect the histogram data of the source image and the image to be screened, normalize the histogram of each collected image, and then calculate the histogram data using the Babbitt coefficient algorithm. Finally, the image similarity value is obtained. Its value range is between [0, 1]. 0 indicates extremely different, and 1 indicates extremely similar (the same). The algorithm steps can be roughly divided into two steps, generating their own histogram data according to the pixel data of the source image and the candidate image. Step 2: calculate the similarity value by using the histogram output in step 1 and the Bhattacharyya coefficient algorithm.Bhattacharyya coefficient is an approximate measure of the amount of overlap between two statistical samples and can be used to determine the relative proximity of the two samples considered. The formula is shown in formula (1).

$$Bhattacharyya = \sum_{i=1}^{n} \sqrt{\Sigma a_i . \Sigma b_i} \tag{1}$$

C. VGG-16

VGG Net is a deep convolutional neural network developed by researchers from the Visual Geometry Group of Oxford University and Google DeepMind who explore the relationship between the depth and performance of a convolutional neural network by repeatedly stacking 3*3 small convolution kernel and 2*2 max pooling layer. VGGNet successfully construct into a convolutional neural network with a depth of 16 to 19 layers. Compared with the state-of-the-art network structure, its error rate is greatly reduced. The meaning of 16 in the VGG-16 network is: there are 16 layers with parameters, including about a total of 138 million parameters[6].

D. MTCNN

Mtcnn(multi task convolutional neural network)is a multi task neural network model for face detection proposed by Shenzhen Research Institute of Chinese Academy of Sciences in 2016. It combines face region detection with face key point detection, and adopts the idea of candidate frame and classifier for fast and efficient face detection. The three cascaded networks are P-Net for quickly generating candidate windows, R-Net for high-precision candidate window filtering and selection, and O-Net for generating final bounding box and face key points[7].

E. Facenet

Facenet (a unified embedding for face recognition and clustering) directly turns the input image into a feature vector in European space. The European distance between the two feature vectors can be used to measure the similarity between the two. It can be used in face verification, recognition and clustering tasks. The main idea of facenet is: calculate the European distance of the feature vectors directly learned by CNN of the two inputs images, the smaller it is the greater the possibility that it's the same person in both images. Once we have this face image feature extraction model, face verification becomes a problem of comparing the similarity of two images with the specified threshold; Face recognition becomes a KNN classification problem of feature vector set; Face clustering can be completed by K-means clustering of face feature sets. Fig. 1. is a simple example in the paper "facenet: a unified embedding for face recognition and clustering"[8]. The numbers in the figure represent the European distance between the two adjacent images' feature vectors. It can be seen that the intra class distance of the images is obviously less than the inter class distance, and the threshold is about 1.1.

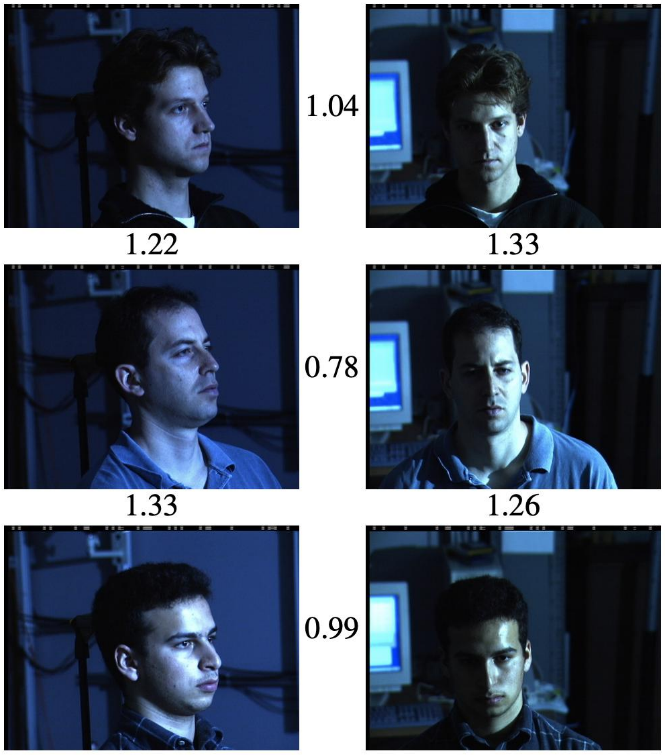

Fig. 1. A simple example for Facenet.

F. Baidu general scene text recognition

Baidu General Scene Text Recognition, a multi-scene, multi-language, high-precision full-image text detection and recognition service, developed by the Baidu Smart Cloud team, can recognize Chinese, English, Japanese, Korean and more than 20 languages, ranking first in a number of ICDAR indicators in the world. The model is specially optimized for blur, tilt, flip and other conditions on the images with strong robustness, and supports more than 20000 large fonts, with an overall recognition accuracy rate of 99%.

III. Prepare Your Paper Before Styling

The experiment of this paper is mainly divided into the following parts: movie subtitle extraction, shot segmentation and key frame object recognition. Based on the extracted subtitles and recognized objects, the theme of the film can be clearly displayed. According to the results of shot segmentation, the average frame length, the number of long shots and the longest shot duration are calculated, which is of great significance for the analysis of film style.

A. Subtitles extraction

a) subtitle located: In order to carry out subtitle recognition, we must first locate the subtitle. The purpose of subtitle location is to prevent other words in the video from interfering with the experimental results, and to reduce the amount of calculation to speed up the process. Location includes upper and lower area location and left and right area location. Generally speaking, the upper and lower areas of subtitles in a video generally keep in their position. Therefore, read the first 50 frames of subtitles in the video, obtain the upper and lower areas of subtitles of 50 pictures, and average the upper and lower boundary positions respectively, that is, the upper and lower areas of subtitles of the video. Since the length of each sentence in the subtitle is different, the left and right positions of the subtitle are constantly changing, so it is necessary to analyze and locate each frame's left and right caption positioning:

- In order to reduce the amount of calculation, segment the places in the picture we may use.
- Blur the image to remove the noise and preserve the edge details at the same time. The median filter technology is selected here. Median filter is a typical nonlinear filtering technology. The basic idea is to replace the gray value of the pixel with the median of the gray value in the neighborhood of that pixel.
- Detect the edge of the image, and then binarize the image.
- Convert the picture data into a matrix, sum each line, and traverse the sum of each line. The two lines with the largest difference from the values of adjacent lines are the upper and lower boundary positions of subtitles.

The process of left and right subtitle positioning is basically the same as that of up and down subtitle positioning. The difference is that after converting the picture data into a matrix in the fourth step, the matrix needs to be transposed, and then the subsequent operation is carried out to obtain the left and right boundary positions of subtitles. The process of subtitles' location is as Fig.2.

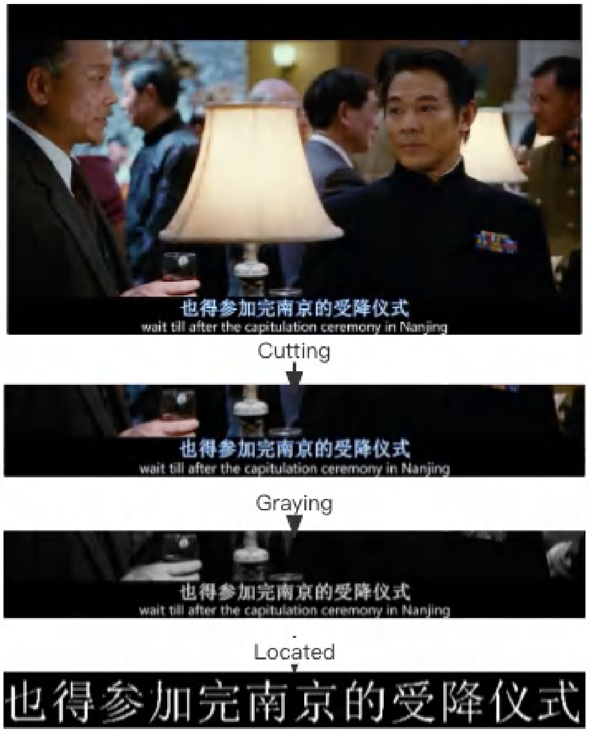
也得参加完南京的受降仪式
wait till after the capitulation ceremony in Nanjing
Cutting
也得参加完南京的受降仪式
wait till after the capitulation ceremony in Nanjing
Graying
也得参加完南京的受降仪式
wait till after the capitulation ceremony in Nanjing
Located
也得参加完南京的受降仪式

Fig. 2. The process of subtitles' location.

b) Subtitle processing: After the subtitle location is completed, the next important problem is how to extract the continuous subtitles without repetition. Because a caption will appear in multiple frames, if the duplicate caption is not removed, the running speed will be greatly reduced in the subsequent recognition process. Therefore, we need to eliminate the duplicate caption.

Here, use formula (2) to calculate the average percentage of the sum of square errors of each pixel between two images, If the similarity degree of two pictures is greater than a certain value and both subtitles exist, we can think that subtitle has been switched.

$$e = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - b_{ij})^2 \times 100\% \tag{2}$$

We selected four frames to show the process of calculating the threshold. As shown in the Fig. 3, *im0* is the frame with empty subtitles, in *im1* and *im2*, subtitles are the same, and in *im3* subtitles have changed. First, read the picture and convert it into a three bit array. Then, according to the above formula, it is found that when $e > 1$ subtitles are switched at that time.

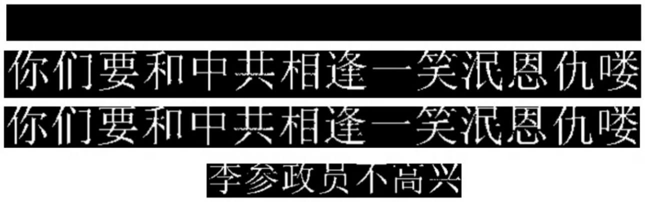

Fig. 3. *im0* to *im3*.

One complete subtitle file includes the content and time of that subtitle. When subtitle switching occurs each time, obtain the time of the current frame as the subtitle time.The video is processed according to the above process. When the subtitle is switched, the subtitle picture obtained from the subtitle positioning processing of the frame is stored in the corresponding folder with the name of the frame for the next subtitle recognition processing.

However, in the subtitle location, due to the change of subtitle background and other factors, the location of the same subtitle is not necessarily the same. This step has certain limitations in judging whether the subtitle is switched. It is necessary to conduct similarity analysis again in subtitle recognition.

c) Subtitle recognition: The first tool selected in the subtitle recognition part is Tesseract. After experimental test, it is found that the accuracy of Chinese recognition by Tesseract is too low. Finally, we selected baidu general scene character recognition API. In the recognition processing,we compared the similarity between the identified sentence and the previous sentence. If the text similarity is greater than 0.55, it is considered that the subtitles of the two sentences are the same and will not be written into the subtitle file; If the text similarity is less than 0.55, it is considered as different subtitles, which are written into the subtitle file, and the subtitle time is written into the file at the same time.

We choose to calculate the cosine similarity of two sentences to calculate the text similarity (formula (3)). The principle of cosine similarity is to take the cosine value between the angle between two vectors in a vector space as a measure of the difference between two individuals. The cosine value is close to 1 and the angle tends to 0, indicating that the more similar the two vectors are, and when the cosine value is close to 0 and the angle tends to 90 degrees, indicating that the more dissimilar the two vectors are.

$$cos\theta = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i^2)} \times \sqrt{\sum_{i=1}^{n}(y_i^2)}} = \frac{a \cdot b}{\|a\| \times \|b\|} \qquad (3)$$

Finally, we draw the word cloud(Fig. 5) with wordcloud according to the obtained subtitle file. From the high-frequency words ”China”, ”democracy”, ”Communist Party of China” and ”Kuomintang” in the Fig. 4, we can clearly see the type of the film and the events described: The civil war(1946-1949) in China.

B. Object Recognition

Identify the objects in the video, summarize the characteristics of the objects in the film and the characteristics of the scene, can strengthen the understanding of the film and television content. There is usually not only one content of a shot, but different scenes, characters and change of time will appear with the movement of the shot. Therefore, for the understanding and generalization of a shot, we can extract the frames containing key contents.Video object recognition is divided into two steps. Firstly, the key frame should be extracted, and then the objects in the key frame should be recognized.

a) key frame extraction: This paper selects the method based on inter frame difference to extract key frames.

Firstly we process the differential calculation frame by frame. Preprocess the two images before and after respectively, grayscale and Gaussian filter the image, then calculate the absolute value difference value, binarize the calculated result, and then add all the values in the matrix as the difference value. This method compresses the whole matrix into one dimension. Although it is a little simple and rough and cause some loss of the information, it doesn’t bother because we only want to measure the gap between frames. The difference value we obtained is discrete, but the peak value selected after smoothing is more representative (some highly disturbing burrs will be removed), we perform exponential smoothing on the difference value. If the difference between the front and back frames is large, we can think that the lens has switched. We use the sliding window algorithm, the window length is 25, and select the maximum value in the window to obtain the frame with large difference as the key frame. Finally, according to the key frame index obtained in the previous step, the corresponding frame is obtained from the video as the key frame of the video.

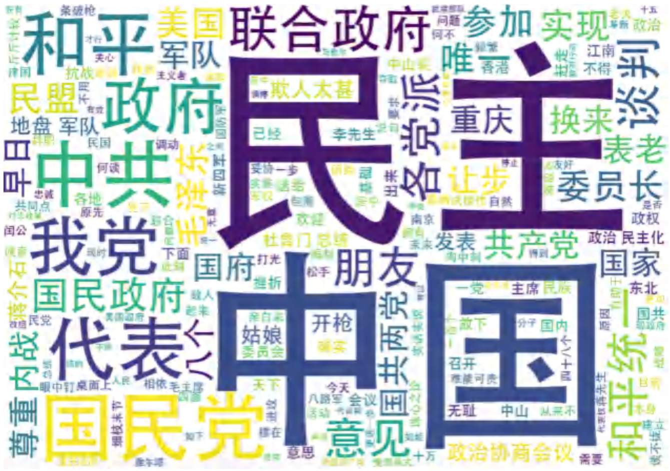

Fig. 4. Word cloud of subtitles.

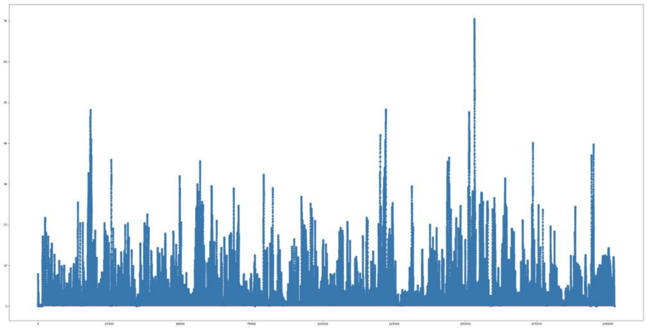

Fig. 5. List of smoothed difference values.

b) key frame object recognition: Then the vgg-16 network is used to recognize the key frame obtained in the previous step, where input the key frame picture, and output the object name using the pre trained model. Firstly, the vgg-16 network structure (forward propagation) is reproduced and encapsulated in a vgg16 class. We use the trained vgg16 so it is not necessary to compute the parameter of NPY model for training. As mentioned above, the vgg-16 network test image must be an RGB image with a fixed size of 224 * 224. Therefore, it is necessary to preprocess the test image which include resize the image to the size of 224 * 224 and normalize the image pixels. Then input the picture, conduct classification test, and output the object name. All key frames in the video are recognized. Fig. 6 shows the visualization of the recognition results.

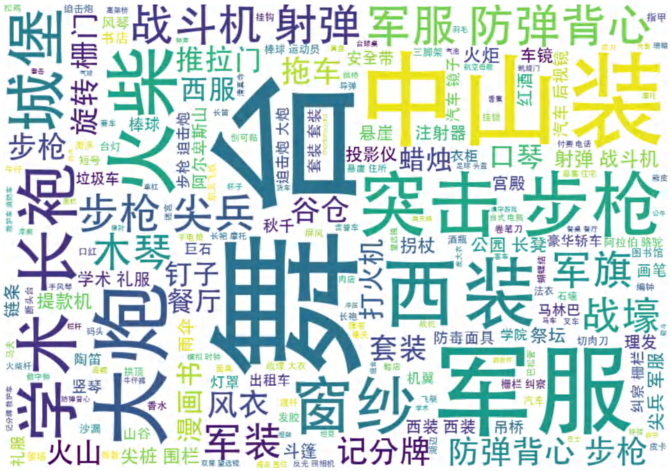

Fig. 6. Visualization of the recognition results.

The objects in the film can highly show the theme of the film. It can be seen from the Fig. 6 that the theme of the film is related to the characteristics of the times; The word ”stage” appears many times because the film contains three important events - Chongqing negotiation, Chiang Kai Shek’s election as president of the Republic of China and The establishment of the people’s Republic of China, as shown as 7; From the frequency of the weapons related words, we can suggest that there are a large number of war related scenes in the film. According to the results of words extraction in the previous step, it is not difficult to deduce the story line of the film.

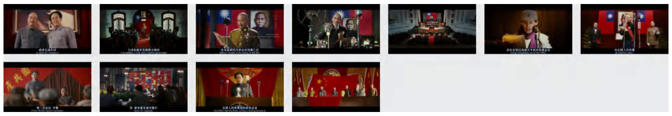

Fig. 7. Scenes including the stage appear in the film.

C. shot segmentation

We intercepted some frames of two adjacent shots in the film to show the changes of frames when the shots are switched, as shown in Fig. 8, it’s easy to find that the switching of two shots is often accompanied by the change of scene and people, so we use this feature to segment shots.

this paper selects the image similarity calculation algorithm to segment the shot. First, the film is divided into frames. If the similarity of two adjacent frames is less than a certain value, it is considered that the shot has been switched. After the comparative test, we selected Histogram algorithm with threshold of 0.7 to compute the similarly. In order to better show the effect of lens segmentation, we intercepted the five minute segment of the climax of the film and showed the results as shown in Fig. 9.

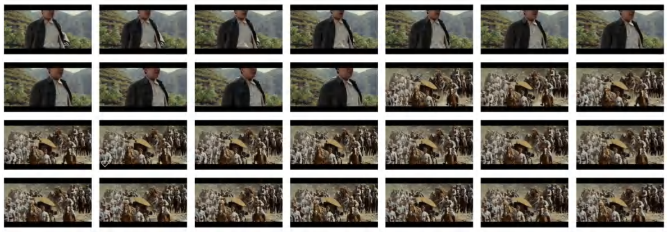

Fig. 8. Shot switching.

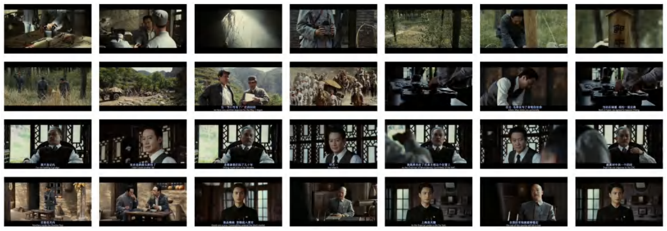

Fig. 9. Preliminary shot segmentation results.

However, for the shots with some characters talking, the scene, characters and light of the previous frame and that of the next frame will change greatly. Therefore, for this kind of shots, this paper selects the face tracking method. In a small time period, if the two faces appear alternately repeatedly, it is considered that it is a dialogue shots. Within the time range of the dialogue, we think the shots does not switch. Firstly, we get all the people’s picture who appearing in the film and mark their names at the same time, then used MTCNN and FaceNet to locate and recognize face in the frames of the last step’s results, when the face appear repeatedly, marked them belonging the same shot.The result of shot segmentation after this step is shown in the Fig. 10.

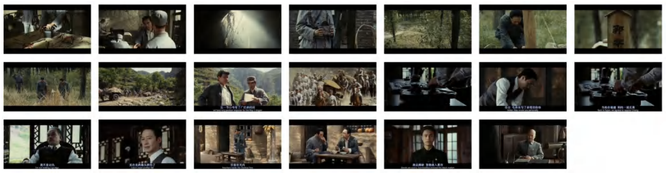

Fig. 10. Shot segmentation results.

Finally, according to the segmentation results, we obtained the average shot length, the number of long shots and the editing rate, as shown in TABLE I. We can see from the data that the film has fewer shots, longer average shot length and more long shots, which is also a feature of historical films: Longer shots provide a lot of information and build the whole story framework to let the audience see the process

TABLE I
SHOTS RELATED DATA

Number of shots	1091
Average shot length	7.286s
Number of full-length shots(>30s)	28
Longest shot	67s

of story development at a slower pace[9]. Different from the decomposition shots assembled by montage in the later editing, long shots are shots that take a long time and show a complete action or event in a unified time and space without later cutting. It is this objective representation, aiming at restrainedly restoring the original essence of life, that keep the shots from the creators' subjective intervention, maintain the continuity and integrity of the subject in time and space as much as possible, and put the viewer from the perspective of a coldly and calmly third party. The longest shot of the film, which describes the dialogue between Chiang Kai Shek and his wife before they teared up the ”October 10th Agreement”, is up to 67 seconds, the key frame of the shot is shown in the figure11. It not only show their mental states and thoughts at that moment which is also the fuse of The civil war in China and pave the way for the subsequent development of the film.

Fig. 11. The key frame of the longest shot.

IV. Conclusion

This paper analyzes the historical film ”The Founding of a Republic” from the perspective of Cinemeitrics. It analyzes the shooting characteristics of the film from three aspects: shots length, editing rate and picture. Combined with historical materials, the story line of the film could be restored through subtitles and object recognition.

However, in film metrology, almost all analysis is based on the shots, so the accuracy of shots segmentation is in the first place. The algorithm of image segmentation is based on the calculation of shot similarity, which makes it difficult to overcome the lens boundary detection in gradual transition, so the accuracy of shots segmentation calculation needs to be improved. Moreover, the subtitles' background of the film selected in this paper is a solid color, so the recognition accuracy is high. But for the film with subtitles above the scenes, the recognition accuracy is relatively low, which still needs to be improved.

Acknowledgment

This work was supported by the National Social Science Art Project of China (18BC034), the Fundamental Research Funds for the Central Universities (CUC210A008), and Key Cultivation Engineering Project of Communication University of China (CUC19ZD007-2).

References

[1] Tao Tao. Waiting History In Image [M].Bei Jing:China Film Press ,2015:1-22.
[2] White H. Historiography and historiophoty[J]. The American historical review, 1988, 93(5): 1193-1199.
[3] Gang Chen. Econometrics and visual path of fermu's film structure [J] Film art, 2020 (04): 45-52
[4] Shizhen Yang. Theory, method and application of Metrology [J] Contemporary film, 2019 (11): 32-38
[5] Daoxin Li. Digital Humanities, filmmaker chronology and new path of film research [J] Film art, 2020 (05): 27-35
[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[7] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE signal processing letters, 2016, 23(10): 1499-1503.
[8] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823
[9] Ding Feng, Luan Lingfei, Richard Anderson. What can filmmakers learn from Ang Lee films—— A case study from the perspective of film perception [J] Contemporary film, 2019 (01): 29-38