A Method of Making Video Trailer Based on Artificial Intelligence

Li Na¹, Li Lanyou¹,²

¹ Network and Information Center, Nanjing Institute of Technology, Nanjing, Jiangsu, China
² College of Mechanical and Power Engineering, Nanjing Tech University, NanJing, Jiangsu, China

*Corresponding author’s e-mail: leenamail@163.com

Abstract. Based on the flownet2 optical flow algorithm, cascade CNN face detection and facenet face recognition algorithm, the method of automatically generating video trailers by using the sum of the optical flow vector sum per second and the absolute value of the audio loudness amplitude per second and the character duration information is proposed. It can effectively edit the visual essence clip, the video essence clip corresponding to the audio, the main character clip and the story outline to generate a trailer.

1. Introduction
In the current video streaming service, the service provider will provide a 1~2 minute trailer to allow users to quickly view the video content and attract users to view the entire content or order. Content providers offer well-prepared trailers for most videos, but for videos without trailers, professional video producers need to make hand-made trailers. This paper provides a trailer production method based on artificial intelligence visual analysis, which can automatically generate trailers based on video content.

2. Definition of Trailer
Trailers usually contain the following key elements. First, the wonderful scenes in the trailer. Second, the intense sound in the trailer. Third, the main character segment includes the front, middle, and rear segments of the main character.

Therefore, automatically generating trailers needs to solve the following problems. First of all, visual essence segment recognition refers to visual analysis of video and analysis of more exciting video segments. Secondly, video sound recognition refers to recognizing the sound in the video, and can recognize the segment of the video with relatively high sound intensity. Finally, protagonist recognition refers to visual analysis of the video through face recognition[1-2]. The character can be analyzed and the characters are classified. The main characters are judged according to the duration of the character's appearance, such as the first main character, the second main character, and the third main character.

This paper provides solutions to the above problems.
3. Visual essence recognition

The more exciting video images in the video, such as martial arts scenes, braking scenes, explosions and other visual impact scenes, are more dramatic movements of people or objects in the picture. In this paper, the optical flow method\(^3\) is used to identify these image change points in the video. The optical flow method is an important method for moving image analysis. The optical flow is a motion vector field that describes a pixel in a three-dimensional space that is represented by a moving object reflected in a two-dimensional image. The optical flow contains information about the target motion and expresses the changes in the image. It refers to a two-dimensional instantaneous velocity composed of all the pixels in the image, where the two-dimensional velocity vector is the projection of the three-dimensional velocity vector of the visible point in the object on the imaging surface.

The current optical flow algorithms are Faneback\(^4\), flownet\(^5\), and flownet 2\(^6\). This article is based on the flownet2 optical flow algorithm. The video change point is obtained by calculating the motion vector sum of each optical flow field. Let the pixel position of the previous frame be \(p_1(x_1,y_1)\), and the displacement of the current frame is \(p_2(x_2,y_2)\), then the optical flow vector of the pixel is

\[
v = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}
\]

the sum of the optical flow vectors of the current frame \(O_f\) is

\[
O_f = \sum_{i=1}^{a} v_i
\]

the optical flow vector sum per second \(O_s\) is

\[
O_s = \sum_{f=1}^{framerate} O_f
\]

Among them, framerate is the number of frames per second of video. Analyze the sum \(O_s\) of the optical flow vectors per second of the video, sorted in reverse order, and select the video essence segment.

4. Video sound essence clip recognition

As a kind of wave, the amplitude is an important attribute describing the wave, and the amplitude affects the size of the sound. In order to obtain the sound impact point in the video, this paper converts the audio in the video into Wav format, and calculates the sum of the absolute values of the amplitudes of each sample point in each second.

\[
A_s = \sum_{i=1}^{sampleRate} |a_i|
\]

\(sampleRate\) is the audio sample rate, \(a_i\) is the amplitude of the sample point, and \(A_s\) is the sum of the absolute values of the amplitude of the sample points per second. By sorting the \(A_s\) per second, a segment with a strong video sound is obtained.

5. Character recognition in video

In order to obtain the main character segment in the video, the video is visually analyzed, the face is detected\(^7\), and the faces are clustered to count the duration of each character. The commonly used face detection algorithms are cascade\(^8-9\) and cascade CNN\(^10-12\). This article uses cascade CNN face detection algorithm. Facenet\(^13\) face coding and recognition methods are used to encode and cluster faces. The face detection and clustering algorithm flow is shown in Figure 1.
Figure 1. Face detection and face clustering.

(1) Get a frame of video.
(2) Perform face detection on the frame video screen to detect whether there is a face.
(3) If there is a face, obtain the face position and crop the face photo, and perform face coding. If the code is accurate, it is considered to be a recognizable face.
(4) Compare the face photos that have been cropped. If the comparison is successful, classify the face into the known face information and record the time position information of the face.
(5) If the comparison is unsuccessful, save it as a new face.

The face set Role is obtained by face detection and face clustering. Each face cluster is a character R, and the duration of each character (if a character appears multiple times within one second, only once), for example, Role={R₁, R₂, ..., Rᵢ, ..., Rₙ}.

By calculating the duration of each character and sorting them in reverse order, the three characters C₁, C₂, and C₃ with the longest duration are obtained.

6. Trailer generation combination algorithm

The data for generating the trailer is obtained by visual essence recognition, audio essence recognition, and character recognition. The trailer is selected according to the following rules.

(1) Divide the source video into multiple segments. Each segment is called a segment. When the video is greater than or equal to 60 minutes, every 10 minutes is a segment. When the video is less than 60 minutes, 15% of the video duration is a segment (For example, 30 minutes of video, every 4.5 minutes is a segment). The number of segments selected in a segment is no more than 3.

(2) The visual essence segment has the highest priority, and the selected video essence segment number is 80% of the video segment number. Each video essence segment is 3 seconds, and the interval between two adjacent essence segments is not less than 2 minutes. The essence of the fragment is called the featured slice.

(3) The audio essence segment is the second priority. The selected video essence segment is 80% of the video segmentation number, and each video essence segment is 3 seconds. The interval between the two adjacent essence segments is not less than 2 minutes.
(4) According to the face analysis, the characters are sorted according to the duration of the character to make a preferred selection.

1) The essence of each character is 5 seconds long. If it is less than 5 seconds, make up 5 seconds.
2) Each segment is divided into 3 parts, each part is called part1, part2, part3, and the part of the blank part is preferentially filled.
3) If the segment already has 3 feature slices, the segment is no longer selected as the essence segment.
4) For character segments that match the part of the segment, the best segment with the most continuous segments in the middle is preferred.

7. Numerical simulation experiment
Select the first episode of the first season of "Sherlock" as the source of the experiment. The duration of the film source is 1 hour 27 minutes 36 seconds. The frame rate is 23.98. The resolution is 1024*576. The audio sample rate is 44.1 kHz. Through the optical flow detection, calculate the sum of the optical flow vectors per second of the video, as shown in Figure 2.

![Figure 2. Video optical flow vector sum.](image)

Select the 5 highest points of the sum of the optical flow vectors, and select 3 seconds for each point. The results of the visual essence segment selection are shown in Table 1.

| Fragment serial number | Starting time | End Time |
|------------------------|---------------|----------|
| 1                      | 0:0:14        | 0:0:16   |
| 2                      | 0:12:26       | 0:12:28  |
| 3                      | 0:17:19       | 0:17:21  |
| 4                      | 0:53:33       | 0:53:35  |
| 5                      | 1:12:54       | 1:12:56  |

By separating the video and encoding it into the Wav format, the sampling rate is 44.1 kHz, and the sum of the absolute values of the amplitudes of the sampling points per second is as shown in Figure 3.

![Figure 3. The absolute amplitude of the audio amplitude.](image)

5 audio segment highlight points, each segment is selected for 3 seconds. The results of the audio essence segment selection are shown in Table 2.

| Fragment serial number | Starting time | End Time |
|------------------------|---------------|----------|
| 1                      | 0:04:09       | 0:04:11  |
| 2                      | 0:33:21       | 0:33:23  |
| 3                      | 0:35:07       | 0:35:09  |
| 4                      | 0:56:06       | 0:56:08  |
| 5                      | 1:08:36       | 1:08:38  |
Using face detection and face clustering, the three characters C1, C2, and C3 with the longest sorting time are shown in Figure 4, respectively.

![Figure 4. Lead analysis results.](image)

Select a 5 second continuous segment every 10 minutes. If there are 3 segments in 10 minutes, discard and reselect. The result of the character segment selection is shown in Table 3.

Table 3. Character fragment.

| Role  | Starting time | End Time  |
|-------|---------------|-----------|
| role C1 | 0:07:01 | 0:07:05 |
| role C1 | 0:16:06 | 0:16:10 |
| role C1 | 0:21:41 | 0:21:45 |
| role C1 | 0:24:06 | 0:24:10 |
| role C2 | 0:27:30 | 0:27:34 |
| role C1 | 0:41:37 | 0:41:41 |
| role C2 | 0:47:11 | 0:47:15 |
| role C1 | 0:57:51 | 0:57:55 |
| role C1 | 1:02:31 | 1:02:35 |
| role C2 | 1:07:31 | 1:07:35 |
| role C1 | 1:16:23 | 1:16:27 |
| role C1 | 1:19:28 | 1:19:32 |
| role C3 | 1:23:05 | 1:23:09 |

Through the above steps, the method obtains a trailer of 95 seconds duration. For the analysis of the source video playback, the visual essence scenes are all segments with relatively large changes in motion, such as traffic scenes. The audio highlights are clips that are more intense in background music. The roles C1, C2, and C3 through face recognition and clustering are Dr. John Watson, Sherlock Holmes, and DI Lestrade, which are the three most important characters in the video. Dr. John Watson appeared the most, followed by Sherlock Holmes and DI Lestrade again, meeting the expectations of the trailer.

8. Conclusion
This paper obtains the data needed for the trailer through visual optical flow analysis algorithm, audio amplitude analysis, character face detection and clustering algorithm. The method of automatically generating a trailer through the trailer generation algorithm satisfies the requirements of the trailer of the ordinary video, so that the user can quickly understand the main characters, story outlines and essence fragments of the video.

Acknowledgment
This work is supported by the School-Level Research Fund Project of Nanjing Institute of Technology (No.QKJA201607).

References
[1] Liang Luhong, Ai Haizhou, Xu Guangyou, et al.(2002) Review of face detection research. Journal of Computers, 25 (5): 449-458.
[2] Li Yuelong, Yan Yan, Wang Jianming, et al.(2016) Overview of face feature point extraction methods. Journal of Computers, 39(7): 1356-1374.
[3] Barron J L , Fleet D J , Beauchemin S S . (1994)Performance of optical flow techniques. International Journal of Computer Vision, 12(1):43-77.
[4] G. Farneback. (2003) Two-frame motion estimation based on polynomial expansion. In: 13th Scandinavian Conference on Image Analysis (SCIA 2003). Springer-Verlag, 363-370.

[5] Dosovitskiy A, Fischer P, Ilg E, et al. (2015) FlowNet: Learning Optical Flow with Convolutional Networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2758-2766.

[6] Ilg E, Mayer N, Saikia T, et al. (2017) FlowNet2.0: Evolution of Optical Flow Estimation with Deep Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii, USA, 1647-1655.

[7] Yan Yan, Zhang Yujin. (2009) Research progress of face recognition based on video. Journal of Computers, 32 (5): 878-886.

[8] Viola P, Jones M. (2001) Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Kauai, HI, USA, vol. I-I.

[9] Chen Lifeng, Wang Jiabin, Zheng Lixin. (2017) Optimization of Face Detection and Tracking Algorithm and Implementation of Web Side. Computer Applications and Software. 2017(09).

[10] Li Haoxiang, et al. (2015) A convolutional neural network cascade for face detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. pp. 5325-5334.

[11] Qin H, Yan J, Li X, et al. (2016) Joint training of cascaded CNN for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3456-3465.

[12] Jing Chenkai, Song Tao, Zhuang Lei, et al. (2018) Overview of face recognition technology based on deep convolutional neural network. Computer Applications and Software, 35 (1): 223-231.

[13] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering [C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, 2015: 815-823.