International Journal on Advanced Science Engineering Information Technology

Vol.11 (2021) No. 6
ISSN: 2088-5334

Deep Learning-based Video Summarization

Myoungchan Seo [a,*], YoungJin Suh [b], Kyuman Jeong [a]

[a] *Department of Multimedia Engineering, Daegu University, 201, Daegudae-ro, Gyeongsan, South Korea*
[b] *School of Computer Science and Engineering, Soongsil University, 369, Sangdo-Ro, Seoul, South Korea*
*Corresponding author: *smc4001@gmail.com*

***Abstract*— With the development of communication technology, many different kinds of media transmission have become popular. Among various media, video is the most popular media these days. However, users need to spend much time watching the whole video content. Due to the characteristics of video media, many users tend to playback video content quickly or even stop watching in the middle. Some websites provide summary images by capturing only important frames of video content, which is called a video summary. Users can shorten the viewing time by only watching the summary results. In particular, it is highly useful because content such as news articles or speeches can be delivered and utilized quickly. Since video summarization is a labor-intensive task, there is an increasing demand for research on automation techniques. In this paper, an automated process to solve the temporary problem of existing video summary techniques is proposed. The proposed method improves the existing video summarization methods that have been performed manually through human labor by developing artificial intelligence technology that can effectively perform content delivery using video summary automation. In the preprocessing process, the information transfer unit is partitioned using optical flow. In the following process, CNN (Convolutional Neural Network) is used as an in-depth learning method for feature extraction. The results show the efficiency of the proposed algorithm, and some future work will be given in the end.**

***Keywords*— Deep learning; video summarization; scene extraction; convolutional neural network; optical flow.**

Manuscript received 11 Aug. 2020; revised 26 Sep. 2021; accepted 12 Oct. 2021. Date of publication 31 Dec. 2021.

I. INTRODUCTION

With the development of mass media and communication technology, video transmission technology, which has a faster information transmission function than a static method such as text or image, is actively used. However, video-type media has the disadvantage that users have to spend as much time as the length of the video to understand the content. Users who watched the video summarized and reprocessed the video on their blog or website to solve this problem. Users can shorten the viewing time by only watching the summary results. As a condition for providing a video summary, the unit of semantic information contained in the video should be divided, and in this process, a duplicate or repeated process can be excluded to shorten the viewing time. Also, it should be possible to accurately place additional information such as captions and images so that the time to receive the contents through the summarized text or image is no longer than the time to watch the video. In this paper, we propose a method to improve the existing video summarization methods that have been performed manually through human labor by developing artificial intelligence technology that can effectively perform content delivery using video summary automation.

A. Related Work: Camera Scene Extraction

Before summarizing the video, it is necessary to consider the event unit that occurs during video playback. This paper assumes that a semantic unit, which is created to convey a story, such as a movie or animation, is a change in camera scenes. Tejero-de-Pablos *et al.* [4] used the data by uniformly dividing frames into regular time units to summarize the motion of UGSVs (User-generated sports videos). It uses the point that adjacent frames overlap and is a method of dividing by simply arranging frames in units of time.

Chu *et al.* [2] reduce the number of overlapped frames through several preprocessing stages to divide the video and induce features to extract features for similar scene segmentation. First, the amount of change between two consecutive frames in the RGB and HSV color spaces is measured, and a boundary point in which the total amount of change is greater than 75% is used as a boundary for classifying the frames.

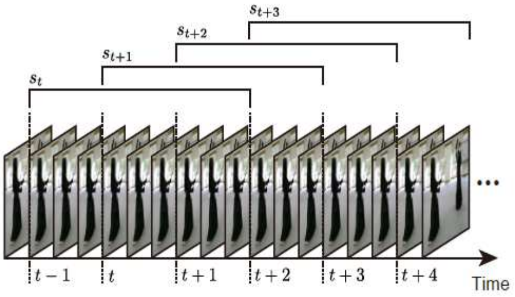

Fig. 1 Dividing video data (Excerpted from [4])

After that, if there are less than ten frames (defined as shots in Chu *et al.* [2]) divided by these boundaries, they are merged and reconstructed into a single frame group, and each shot can have up to 150 frames so that the frame groups can be evenly divided. The characteristics of the frames are extracted, which is subject to the above process, then the interaction characteristics between successive frames are extracted. Finally, after K-means clustering is induced to exclude duplicate frames using samples of each feature.

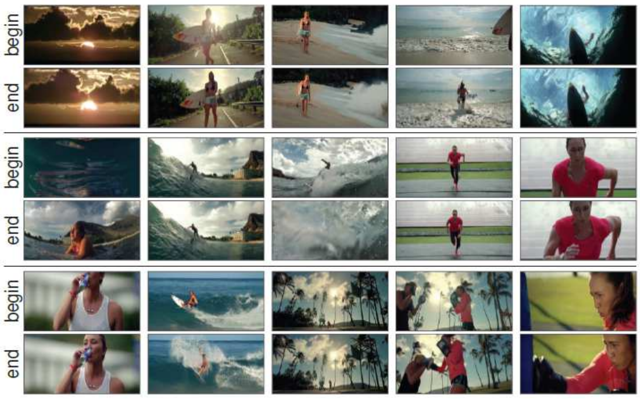

Fig. 2 Video division method from the keyword (Excerpted from Chu *et al.* [2])

Zhai *et al.* [6] use MCMC (Markov Chain Monte Carlo) to randomly calculate the boundary position of the important concept represented by the video through two model parameters from the probability distribution. They proposed a method to measure scene transition based on probability while repeating the process of continuously merging and dividing frames within the divided boundary and the MCMC process of obtaining the boundary position. Using MCMC, a process in which a scene is switched through iterative updating of a single image can be processed stochastic. However, iterating is not suitable for quickly dividing various videos and using them as data.

B. *Related Work: Key Frame Extraction*

Existing research methods for video segmentation, such as Chu *et al.* [2], can be classified into a method using a difference between pixels, edges between successive frame histograms, and a moving vector. These studies require thresholds to segment the scene. Ko and Rhee [5] combined the existing color histogram method with the test to separately calculate the weights according to the contrast level (NTSC standard) in the RGB color space. In this paper, the difference values are emphasized to respond to specific value changes. A variation test technique is proposed, using an average and a standard deviation and automatically setting a threshold. Histogram, test, average, and standard deviation are considered in the process of selecting the threshold value for classifying the boundary, and static transitions such as fade-in/out are performed because the experiments are assumed with the radical transition of video.

Chu *et al.* [2] proposed an MBF (Maximal Biclique Finding) algorithm that extracts key scenes for video summary using common keywords. First, a group of frames divided through a preprocessing process is used, and features occurring in a single frame are extracted. Afterward, a feature sample is obtained by using K-means clustering for interactive features on successive frames to exclude duplicate scenes. Finally, they suggested how to extract the core scene suitable for a common keyword by using the similarity of scenes that often occur in each video. The MBF (Maximal Biclique Finding) method can effectively extract highlight scenes from videos of the same keyword using similar patterns between images. However, the unique pattern scene appearing in each image is ignored, so the temporal flow for content delivery is not considered. For this reason, it is challenging to apply it to video summarization technology, whose primary purpose is content delivery.

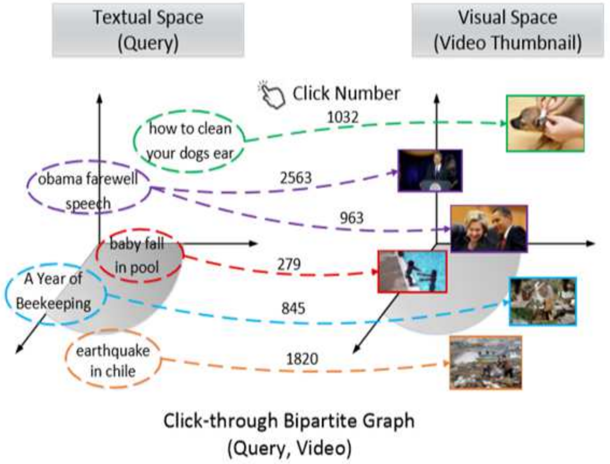

Fig. 3 Creation of association graph between search terms and video (Excerpted from Yuan *et al.* [3])

In the existing video summary methods, subjective opinions are reflected by human hands. Yuan *et al.* [3] proposed a Deep Side Semantic Embedding (DSSE) algorithm that induces automatic classification of videos similar to human subjectivity. The thumbnail selected by the information provider, the number of times the user clicked the video, and the text are used as learning data to generate semantic information on the video side.

Fig. 3 is a graph linking the distribution of video thumbnails and clicks retrieved by sentence or text and their associated data. The method of training the DSSE model uses various semantic information to summarize the video. It has the advantage of using additional information such as tags, captions, and comments that the video represents to extract the core scene. Furthermore, while it has similar advantages to ground truth, it requires a lot of information and time to summarize the entire contents of a single video.

II. Material and Method

The final goal of this paper is to generate a video summary output with a cartoon layout given the input video. An image evaluation algorithm using deep learning is used as the underlying technology to achieve this goal. This research aims to develop 1) Camera Scene Extraction technology and 2) Image Aesthetic Evaluation technology.

In order to divide the camera scene, which is a video summary learning unit, a criterion for recognizing changes in the scene is required. It uses optical flow that recognizes the movement of objects and removes some frames to exclude cases that are not classified by the fade in/out effect. The group of divided images is trained in the artificial intelligence model. Then, a video summary is generated by inducing the user to summarize as many frames as desired.

A. Camera Scene Extraction

In this paper, it is used to classify the camera scene by estimating the change of the scene through the optical flow. It is difficult to apply only the face model like Cohn *et al.* [7] because the video contains various objects and humans. Therefore, after finding all the feature points that can be obtained from a single image, it is necessary to check the information that can be obtained in the scene transition by calculating the difference between the two frames using the average vector value between before and after frames.

If there is no change in the camera scene, the average amount of change in the optical flow approaches zero. Also, it can be seen that the value of the graph changes rapidly in the section where the camera scene changes. If a scene is divided using a random value of 0.5 on one side of a movie trailer, the result will be like as shown in Fig. 6.

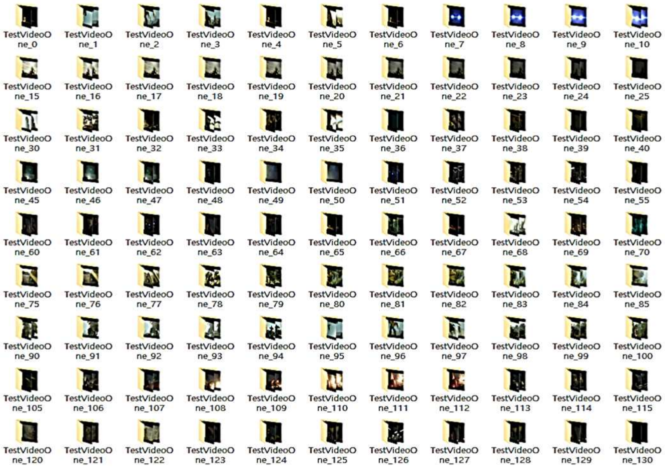

Fig. 4 The result of scene separation in the movie "BumbleBee" (Running time: 2 min 22 sec, separated into 130 folders)

There are some errors in classifying where the brightness of an image is changed due to the fade in/out phenomenon in a fixed scene even though most camera scenes are normally well divided.

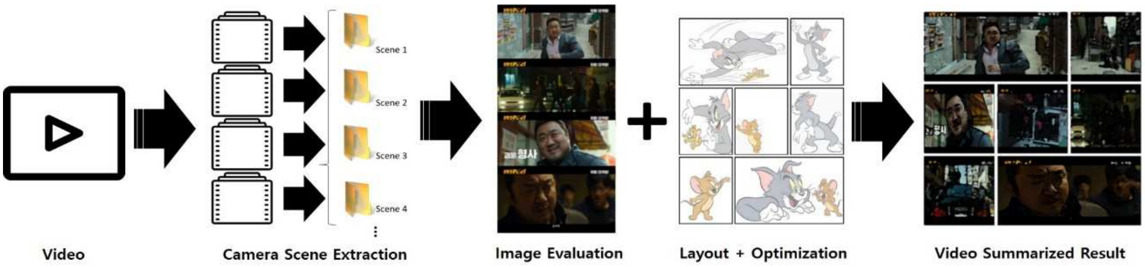

Fig. 5 System overview

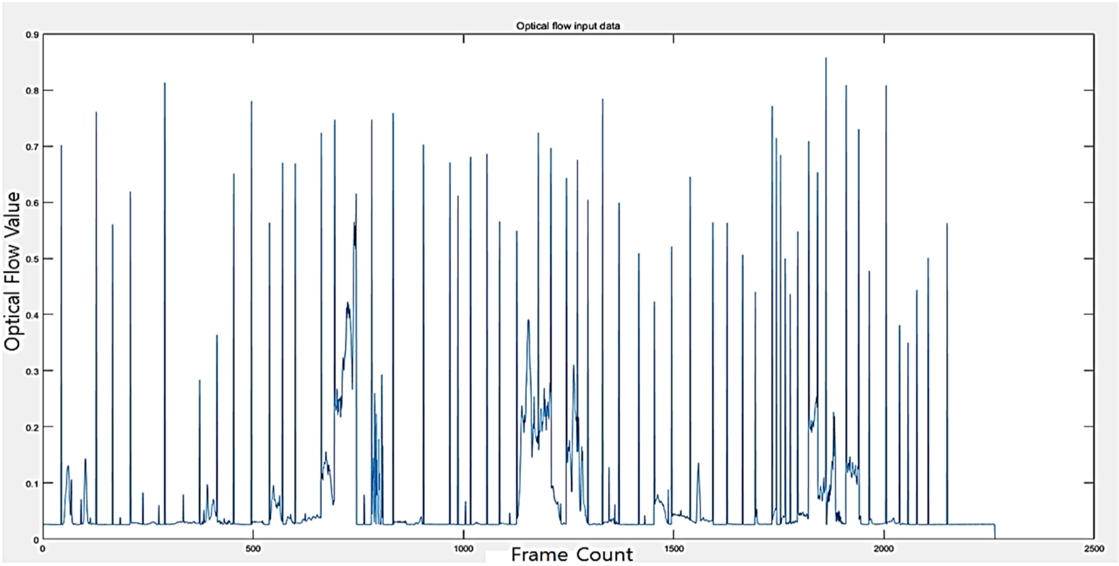

Fig. 6 Variation graph of optical flow average value in the movie trailer "The Spy Gone North."

Fig. 7 Result with wrong scene extraction because of fade in/out

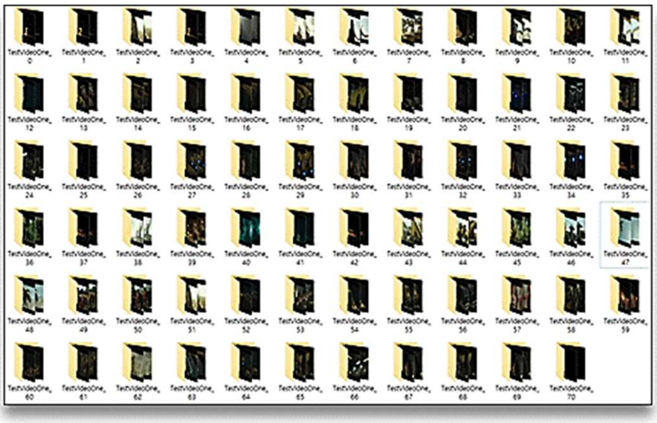

Fig. 8 The result with fade in/out effect consideration in the movie trailer "BumbleBee" (130 folders reduce to 71 folders)

It is necessary to exclude some frames to minimize camera scenes that are misidentified due to fade in/out. According to Raymond Fielding [10] and Apple Motion [9], when expressing fade in/out in 1 ~ 2 seconds of video (24 ~ 48 frames), minimum 6 ~ 12 frames or maximum 24 ~ 48 frames are used. The camera scene was newly divided according to this consideration, except that less than 12 frames were stored in one folder.

As shown in Fig. 8, it is confirmed that the unsorted scenes are stored in a single folder, and it can be seen that the camera scene changes are more precisely separated than conventional methods.

B. *Image Aesthetic Evaluation*

The word embedding method of Ren *et al.* [8], is suitable for application to a system that analyzes information contained in an image. However, it is difficult to extract the scene to be used for the video summary, which is the purpose of this study. In extracting summary scenes, it is not possible to specify objects or select sentences or characters. Therefore, when the data set was created to extract the scene, the subject of the individual was reflected in the selection process after watching the movie trailer among the image files where the camera scene segmentation was completed. Data set creation was performed by labeling the weight of the scene judged to be the most important in one camera scene as 2, 1 for normal or general scenes, and 0 for unnecessary scenes.

When creating a data set, a significantly darker scene or overlaps two scenes compared to a bright one is assumed to have a low cost. A scene that contains characters, such as a movie title or an introduction, is also assumed to have a low score. We can conclude that it is difficult to summarize the contents of the video in these scenes.

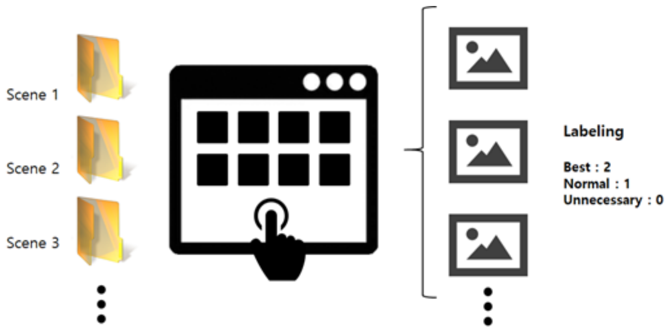

Fig. 9 Generation of training set/test set

The ratio of the training set and the test set is approximately 3.45: 1, and 84,967 images from the 29 movie trailers are divided into 65902(training set): 19065(test set) and labeled according to the same criteria. The data set creation results are shown in Table 1 below.

TABLE I
LABELLING RESULT OF 84,967 FRAMES IN 29 MOVIE TRAILERS

	Training Set	**Test Set**
Ratio	3.45	1
Frame Count	65902	19065
Best	2138	1181
Normal	47464	12267
Unnecessary	16300	5617

The convolutional layer[11] for finding the most suitable frame for summarizing the video is organized, as shown in Fig. 10. These layers utilize the data set generated by the division and labeling of the camera scene unit.

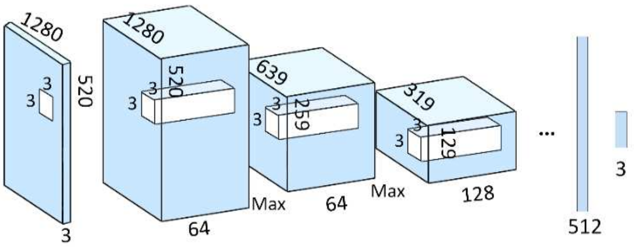

Fig. 10 CNN structure for scene extraction

The image files with 1280X520 size, in which the letterbox area for adding additional information such as subtitles or movie titles is removed, are used as an input value in the HD resolution (1280X720) released by movie companies. In this research, 3X520X1280 color information is needed since the RGB color model is used. A 3X3 sized mask is used to reduce width and height by half the size of the mask, excluding an outer edge of the mask. The process of decreasing the size of masks must be repeated until the result is obtained by extracting the probability of belonging to Best, Normal, and Unnecessary, respectively.

In order to clearly distinguish the result value, the scenes that are not the most influential in the video summary can be excluded by adding the average probability and the best probability, then subtracting the unnecessary value through the above process.

The best-rated frames are aggregated for each camera scene, and the scores of each frame are sorted in descending order. Finally, dynamic programming is used to prevent the phenomenon of a high-numbered specific section. The desired number of images can be obtained by comparing the scores of each frame to obtain the final result.

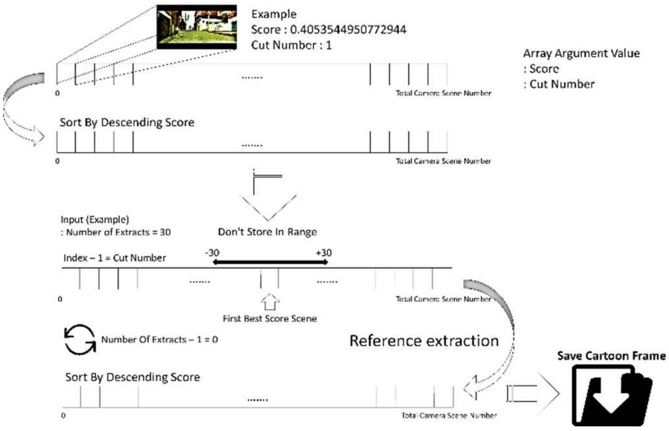

Fig. 11 Dynamic programming method for distributed scene extraction

First, image information, which the model has evaluated, is stored in an array with the number of camera scenes and is sorted in descending order from its highest score to its lowest score. Then, using the same sized array from the sorted array, the order of the frames with the highest score is stored as the index of the newly declared array and placed in the left and right index ranges as many as the number of frames for extraction with high scores. If it is out of range, it is possible to evenly extract the desired number of frames throughout the image by repeating the process until the number of frames becomes 0 to store a new frame.

III. Results and Discussion

In order to create a summarized video, a labeling data set of 84,967 images from 29 movie trailers are used to train the network. The CNN model structure repeats the convolution and max-pooling process, as shown in Table 2.

TABLE II
CNN Structure for Video Summary

Processing	Width	Height	Depth	Mask Size
Input	1280	520	3	3
Convolution	1280	520	64	3
Max Pooling	639	259	64	3
Convolution	639	259	64	3
Max Pooling	319	129	64	3
Convolution	319	129	128	3
Max Pooling	159	64	128	3
Convolution	159	64	128	3
Max Pooling	79	31	128	3
Convolution	79	31	256	3
Max Pooling	39	15	256	3
Convolution	39	7	256	3
Max Pooling	19	7	512	3
Convolution	19	3	512	3
Max Pooling	9	3	512	3
Dense	**512**	-		
Result	**3**	-		

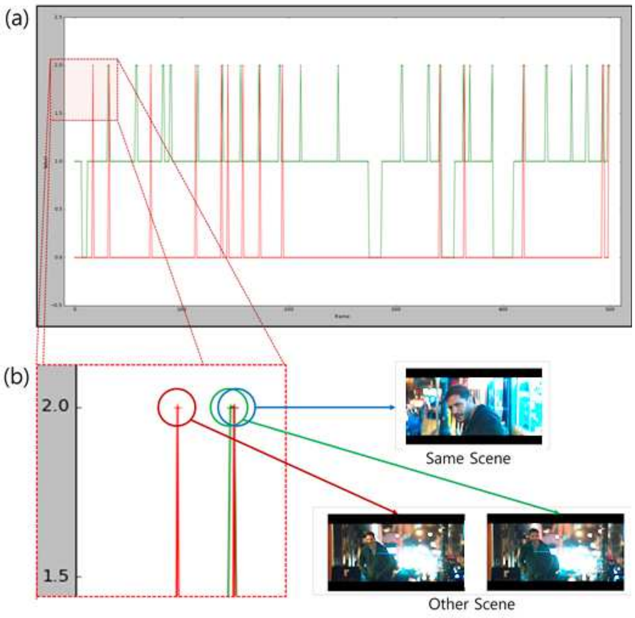

Fig. 12 (a) Scene extraction graph (Green: ground truth, Red: CNN) (b) Error point (green & red), same scene extraction (blue)

The system used in the experiment was a multi-connection of AMD RYZEN 1800X (Boost Clock 4.0GHz) CPU and two GTX 1080TI GPUs and was performed using the deep learning framework CNTK (Computational Network Toolkit) released by Microsoft under the Windows operating system.

In this experiment, each movie trailer must be summarized in 30 frames. Fifteen scenes were extracted from the randomly selected 500 frames to confirm the effectiveness of the proposed algorithm. The comparison to the ground truth images of the corresponding camera scene is shown in Fig. 12.

Table 3 shows the ratio of the difference between the average value of each pixel deviation by converting the extracted image into grayscale and the scene extracted using Ground Truth and CNN.

TABLE III
Gray Scale Average Deviation and Scene Match Ratio

Scene	**1**	**2**	**3**	**4**	**5**
Gray Scale AVG Difference Value	12.6239	0	10.0665	8.2391	2.0821
Scene	**6**	**7**	**8**	**9**	**10**
Gray Scale AVG Difference Value	19.0071	21.0586	8.4469	15.7346	0
Scene	**11**	**12**	**13**	**14**	**15**
Gray Scale AVG Difference Value	19.1798	0	0	11.8698	0
Match Rate			**26%**		

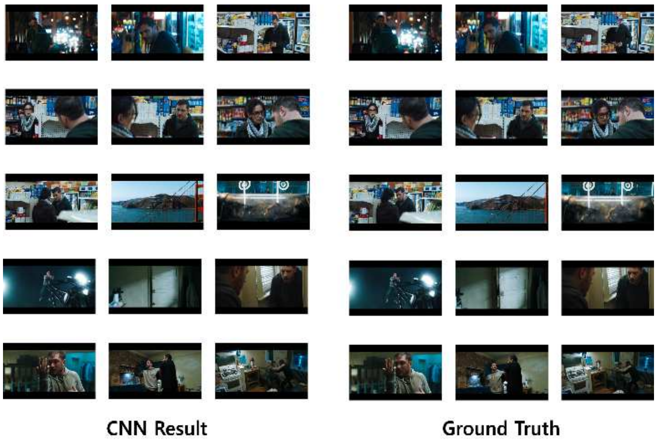

Fig. 13 Comparison of scene extraction (CNN vs. ground truth)

Among the 611 camera scenes of the 19,065 frames in the entire test data set, 135 images extracted using CNN compared to Ground Truth have the same image ratio of approximately 22%. Fifteen randomly extracted video summary scenes are shown in Fig. 13. Fig. 14, Fig. 15, and Fig. 16 are generated to summarize the video in 30 images through a completed learning model from a segmented camera scene using optical flow. This results from generating a summary video cartoon by adjusting the cartoon layout ratio of two neighboring images using the probability that the frame belongs to the best and cutting and arranging it to fit the size.

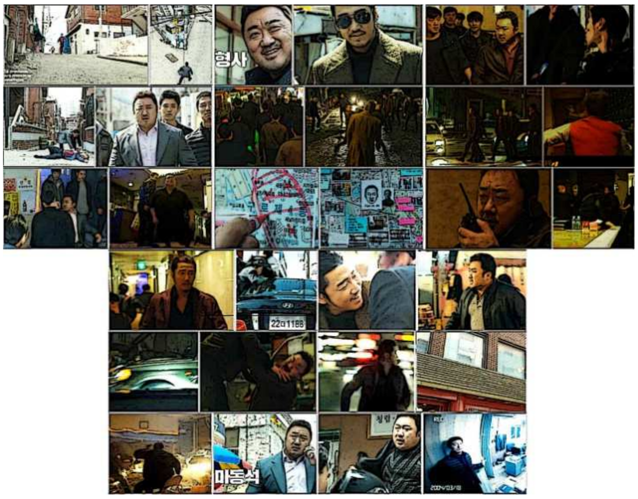

Fig. 14 Video summary result with 30 frames from "The Outlaws."

IV. Conclusion

In this paper, we divide the video's content delivery unit through the camera scene's division using optical flow and extract the scene suitable for delivering the content to the viewer through the deep learning model CNN to summarize the video. Suggestions are given for expressions using styles and layouts. However, in dividing the camera scene, there is a problem that is not entirely classified, as shown in Fig. 17. Continuous model improvement to secure objectivity is needed.

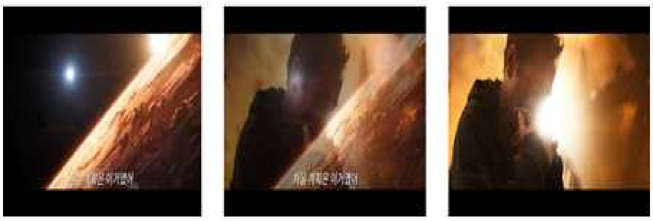

Fig. 15 Camera scene without division because of overlap effect

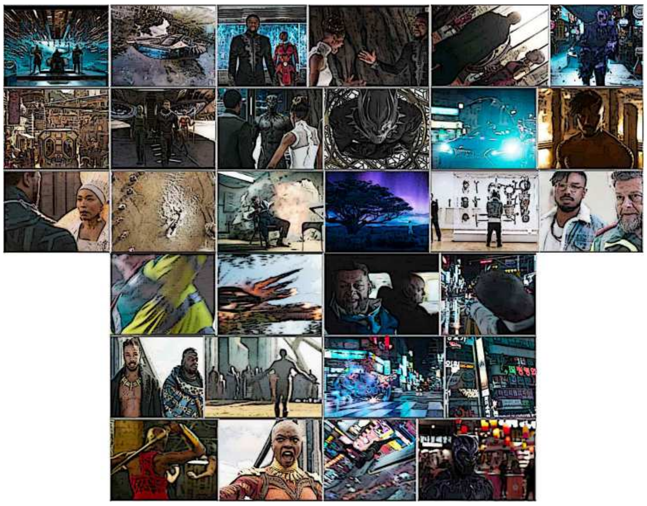

Fig. 16 Video summary result with 30 frames from "Black Panther."

When summarizing video content, it is necessary to use subtitles or captions. In addition, when applying a cartoon style [1], the size of the layout depends on the probability that the scene will be extracted. Various attempts such as an iterative cropping technique [12]-[20] required considering the image's composition.

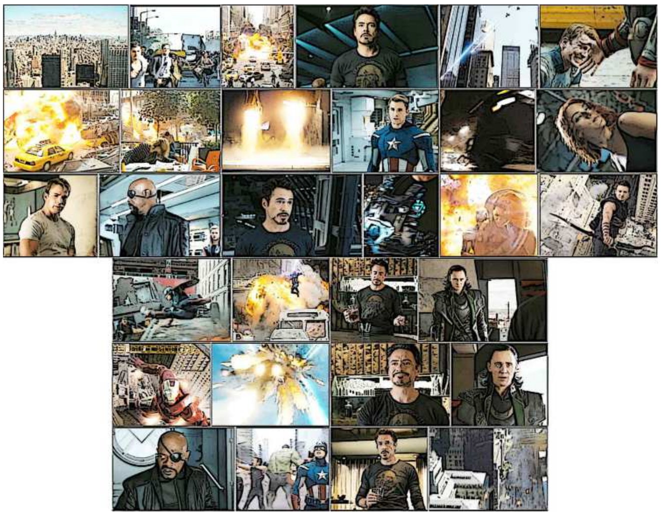

Fig. 17 Video summary result with 30 frames from "Avengers 1."

Acknowledgment

Daegu University Research Grant, 2020 supported this research.

REFERENCES

[1] T. Bhattacharjee, S. Saha, A. Konar, and A. K. Nagar, *Static Video Summarization Using Artificial Bee Colony optimization*, Computational Intelligence (SSCI) 2018 IEEE Symposium Series on, pp. 777-784, 2018.

[2] W.-S. Chu, Y. Song, and A. Jaimes, *Video Co-summarization: Video Summarization by Visual Co-occurrence*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3584-3592, 2015.

[3] Y. Yuan, T. Mei, P. Cui, and W. Zhu, *Video Summarization by Learning Deep Side Semantic Embedding*, IEEE Circuits and Systems Society, 2017.

[4] A. Tejero-de-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, *Summarization of User-Generated Sports Video by Using Deep Action Recognition Features*, IEEE Transactions on Multimedia, vol. 20, no. 8, pp. 2000-2011, 2018.

[5] K.-C. Ko, and Y.-W. Rhee, *Video Segmentation using The Automated Threshold Decision Algorithm*, Journal of the Korean Society of Computer Information, vol. 10, no. 6, pp. 65-75, 2005.

[6] Y. Zhai, and M. Shah, *Video Scene Segmentation using Markov Chain Monte Carlo*, IEEE Transactions on Multimedia, vol. 8, no. 4, pp. 686-697, 2006.

[7] X. Fan, X. Yang, Q. Ye, and Y. Yang, *A Discriminative Dynamic Framework for Facial Expression Recognition in Video Sequences*, Journal of Visual Communication and Image Representation, vol. 56, pp. 182-187, 2018.

[8] M. Ren, R. Kiros, and R. S. Zemel, *Exploring Models and Data for Image Question Answering*, Advances in Neural Information Processing Systems, 2015.

[9] Apple Support, *Motion: Fade In/Fade Out*, World Wide Web https://support.apple.com/kb/PH15957?locale=ko_KR

[10] R. Fielding, *The Technique of Special Effects Cinematography*, Focal Press, pp. 151-152, 1985.

[11] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, *A Closed-form Solution to Photorealistic Image Stylization*, in the proceeding of ECCV 2018, pp. 469-483, 2018.

[12] W. Wang, J. Shen, and H. Ling, *A deep network solution for attention and aesthetics aware photo cropping*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1531–1544, 2019.

[13] P. Lu, H. Zhang, X. Peng, and X. Peng, *Aesthetic guided deep regression network for image cropping*, Signal Processing: Image Communication, vol. 77, pp. 1 – 10, 2019.

[14] Y. Kao, R. He, and K. Huang, *Deep aesthetic quality assessment with semantic information*, IEEE Transactions on Image Processing, vol. 26, no. 3, pp. 1482–1495, 2017.

[15] G. Guo, H. Wang, C. Shen, Y. Yan, and H. M. Liao, *Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression*, IEEE Transactions on Multimedia, vol. 20, no. 8, pp. 2073–2085, 2018.

[16] D. Li, H. Wu, J. Zhang, and K. Huang, *A2-RL: Aesthetics aware reinforcement learning for image cropping*, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8193–8201.

[17] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, *Good view hunting: Learning photo composition from dense view pairs*, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[18] M. B. Islam, W. Lai-Kuan, and W. Chee-Onn, *A survey of aesthetics driven image recomposition*, Multimedia Tools Appl., vol. 76, no. 7, pp.9517–9542, 2017.

[19] H.-J. Lee, K.-S. Hong, H. Kang, and S. Lee, *Photo Aesthetics Analysis via DCNN Feature Encoding*, IEEE Transactions on Multimedia, vol. 19, no. 8, pp. 1921-1932, 2017.

[20] Y. Deng, C. C. Loy, and X. Tang, *Image Aesthetic Assessment: An experimental survey*, IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 80-106, 2017.