ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.016

Automatic Gaming Highlights Generation Using Facial Expression Analysis of Game Streamer

Kazuma Kii [+], Eiji Kamioka [1] Chanh Minh Tran [1] and Phan Xuan Tan [1]

[1] Graduate School of Engineering and Science, Shibaura Institute of Technology, Japan

Abstract. With the spread of short video platforms, there has been a growing demand to generate short highlights from very long video live streams especially gaming live streams, which has become a burden for video editors. Therefore, it is necessary to introduce an automatic highlights generation method to reduce the burden on editors, while providing viewers with an efficient viewing experience. Existing works on gaming highlights generation relied on viewer comments, and thus cannot analyze videos with few comments. Moreover, since the comments are view-driven information, highlights generated based on comments may not correctly reflect the intention of the streamer. This study focuses on information obtained from the video itself. Specifically, the emotional responses of game streamers based on their facial expressions are utilized for automatically extracting highlight scenes from long gaming live streams.

Keywords: Video Summarization, Facial Expression, Emotion Analysis.

1. Introduction

Short video platforms such as YouTube Shorts [1] or TikTok [2] are among the major means of entertainment nowadays. It has been reported that the average number of views on YouTube Shorts per day has exceeded 70 billion [3]. Due to such high demand, the need to create short highlights from long videos is also increasing, especially for gaming live streams. However, manually creating a short highlight video requires manually selecting and extracting interesting scenes from the long live stream. This has become a heavy burden for the video editors. Therefore, it is necessary to come up with an automation method for extracting the interesting scenes from a gaming live stream to generate gaming highlights. Existing works on extracting highlight scenes from gaming live streams often utilize the viewers' comments. Matsuda et al [4] created a highlight video using a method that suggests scenes with the highest number of comments as important scenes. However, this method is prone to induce false positives due to an increase in the number of comments caused by a single user posting a series of comments.

This study proposes to utilize the information independent of elements outside the video obtained during the gaming live stream. Specifically, by analyzing emotional data obtained from facial expression analysis of game players in professionally created highlight videos, a model that has learned the distribution of emotional data characteristic of highlight videos is created, and highlight videos are automatically created.

2. Related Works

There have been several existing studies on the automatic generation of gaming highlights by analyzing the number of viewer comments at a certain time in live gaming videos. In 2023, Matsuda et al. [4] conducted a study on the automatic generation of highlight videos by analyzing viewer comments on YouTube. In this study, the top 30 to 60 scenes with the highest "number" of comments were selected as important scenes and used to generate highlight videos. However, this method has a problem of false detection of scenes that should not be highlighted, such as when a particular viewer repeatedly throws comments or when the distributor asks viewers for their opinions. In addition, Aoki et al. [5] created a system for detecting chorus in music content by analyzing the number of comments on Nico Nico Video [6], and applied it to videos in various genres, such as live game videos, to generate highlight videos. However,

[+] Corresponding author.
E-mail address: af21058@shibaura-it.ac.jp.

since Nico Nico Video has an upper limit to the number of comments, and the oldest comments are deleted first, it is not appropriate to propose highlight videos based on comment analysis. Furthermore, the method that relies on comments also has the problem of not being able to generate highlight videos when the number of comments is small.

As a study that tackled the extraction of highlight scenes from live game videos from a different perspective, Yoshida et al. [7] introduced a method for generating cut-out videos for games, focusing on their acoustic characteristics. In this study, the top 10 scenes with the highest sound pressure compared to the maximum sound pressure level of the entire original video are proposed as highlight scenes. Unlike methods that analyze information that is not metadata of the video itself, such as comments and postings, this research provides highlight videos by analyzing sound pressure data contained in the metadata of the video itself. Since this proposal is not a method for analyzing comments, it is a step further than research that analyzes comments, since it does not require the addition of a function for analyzing the sentiment of words specific to broadcasts, and it does not allow for analysis when the number of viewers declines. However, false positives are common because the distance between the distributor and the microphone varies greatly, and the maximum sound pressure level varies greatly from distributor to distributor.

3. Proposed Approach

In this study, the information extracted from the video live stream itself is considered. An automatic generation of highlight gaming videos is proposed using facial expression analysis of the video game streamer. Since most video game streamers often display their own faces in their live stream videos, their facial expressions can be extracted and analyzed to create emotional data. Such a method is independent of whether the viewers comment on the video. This section presents the details of how the proposed system identifies highlight scenes and automatically creates highlight videos based on the estimated emotional data.

A simple flow of the proposed system is shown in Figure 1.

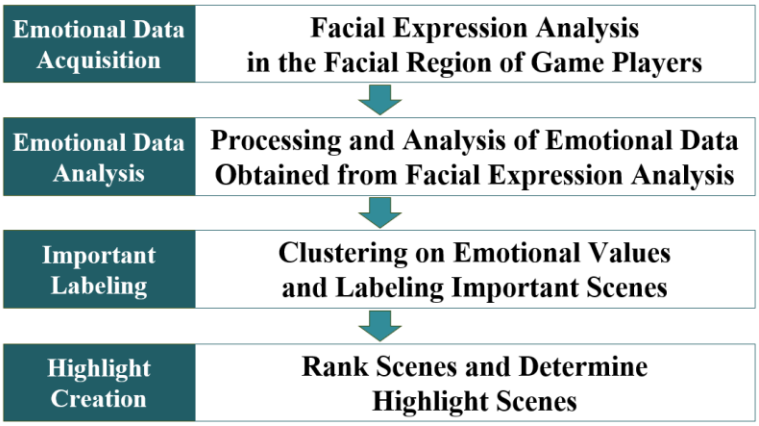

Fig. 1: Flow of Proposed Method

In this study, OpenCV [8], dlib [9], and HSEmotionRecognizer [10][11] are used for facial expression analysis. HSEmotionRecognizer is a model specialized for facial emotion recognition and is a lightweight neural network model built on the EfficientNet [12] deep learning model. The proposed system uses an 8-class model, which analyses the face region and classifies the facial expression based on 8 emotions: anger, contempt, disgust, fear, happiness, neutral, sadness, or surprise. This work adopts the enet_b0_8_best_vgaf model, which has the highest prediction accuracy for the 8-class model. The proposed system calculates each emotion frame-by-frame and averages them every 30 seconds to create emotional data points for analysis of highlight scenes.

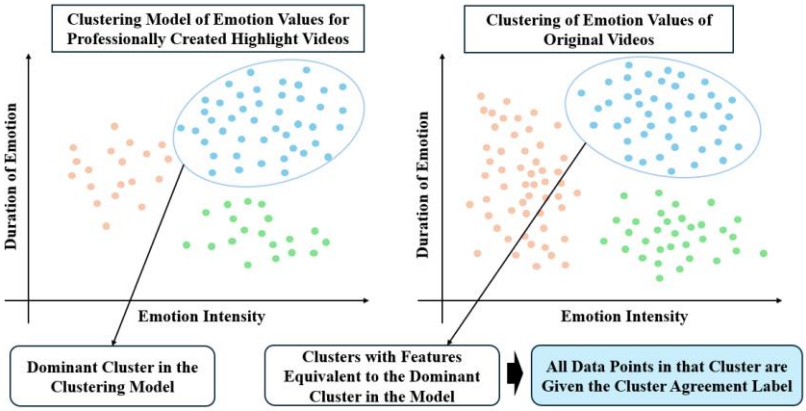

Fig.2: Cluster Agreement Labeling

Next, the optimal number of clusters for clustering the 8 emotional data points is determined. In determining the optimal number of clusters, the number of clusters is determined for each of the eight emotions based on the silhouette score, which evaluates the within-cluster variance and between-cluster variance. This number of clusters is used to build a model that learns the distribution of emotional values characteristic of professionally created highlight videos. This clustering model is then used to perform clustering on the emotional values obtained from the facial expression analysis of the original video. Based on the clustering results, important labeling of scenes is performed. As Figure 2, the data points of clusters with features equivalent to the dominant clusters of the emotional value clustering model for professionally created highlight videos are assigned to the Cluster agreement label. Next, assign a Strong Emotion Label to the top 10% with the highest emotional values as Figure 3.

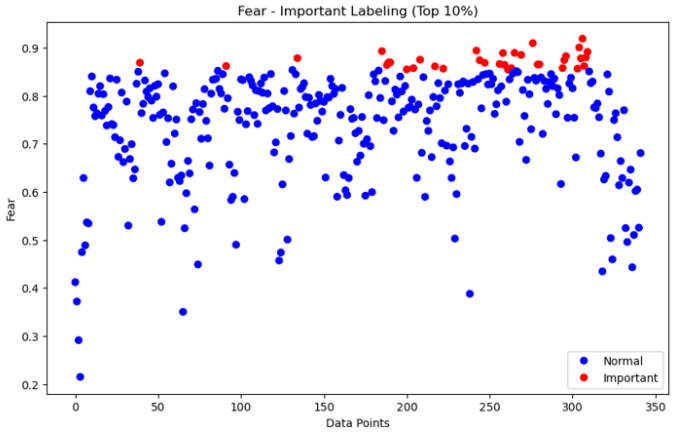

Fig. 3: Strong Emotion Labeling

Finally, these important labels are used to rank the scenes. For cluster agreement labels, the more labels there are at the same time, the higher the importance is rated as Figure 4. A strong emotion label is a label that is weighed up to the genre of the game. For example, in the case of a horror game, a scene with a high Surprise emotional value is rated higher as a scene with high importance. Preliminary research has shown that this is likely to be a highlight scene when the game player's emotions shift from Fear to Surprise, To avoid the false detection of Surprise, scenes in which both Fear and Surprise have a strong emotion label, and frames in which Surprise has a strong emotion label and Fear have a strong emotion label in the previous time interval are considered as important scenes like Figure 5. This operation eliminates sudden false increases in Surprise's emotional values due to false positives in facial expression analysis. By combining these scenes from the top, the highlight video is output.

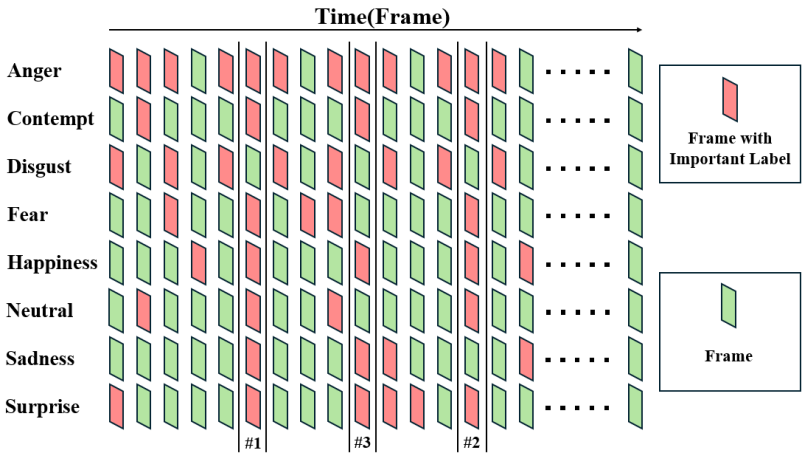

Fig. 4: Determination of Important Scene by Cluster Agreement Label

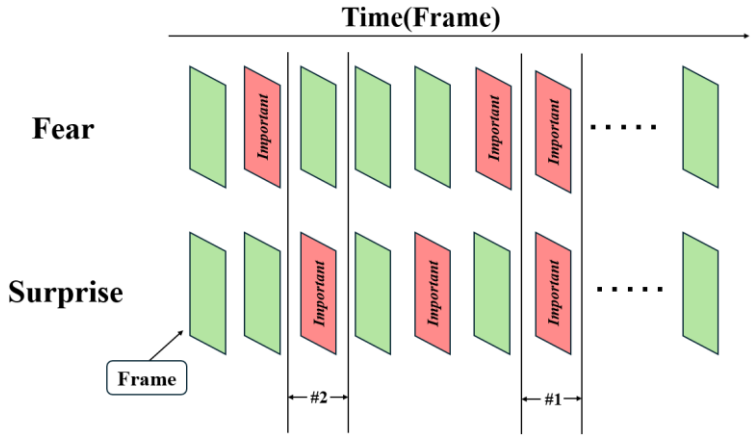

Fig. 5: Determination of Important Scene by Strong Emotion Label

4. Evaluation

An experiment was conducted to confirm the effectiveness of the proposed method for automatically generating highlight videos using facial expression analysis of video game streamers. In this experiment, 10 highlight videos created by the proposed method (system-created videos) were evaluated in two ways. Firstly, the system-created videos were evaluated whether they included the same scenes as those created by professional editor (streamer-created videos). To ensure the generalizability of the proposed model, the evaluation was conducted using a model that excluded professionally created videos from the training data. Secondly, a subjective human evaluation was conducted to compare the system-created videos with streamer-created videos based on the following criteria: whether the video was uncomfortable, whether the flow of the game could be understood, and whether the video was interesting. The results of these evaluations are summarized in Table 1 and Table 2.

Table 1: Performance Evaluation Results of the Proposed Method

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Accuracy	0.84	0.84	0.83	0.84	0.96	0.87	0.67	0.71	0.78	0.82
Precision	0.40	0.35	0.45	0.40	0.15	0.35	0.65	0.35	0.50	0.40
Recall	0.26	0.35	0.35	0.35	0.14	0.39	0.62	0.39	0.30	0.38
F1 Score	0.32	0.35	0.39	0.37	0.15	0.37	0.63	0.37	0.38	0.39

Table 2: Results of Subjective Evaluation

	Highlight 1			Highlight 2		
	Positive	Neutral	Negative	Positive	Neutral	Negative
Comfortability	4	2	4	4	4	2
Game Flow Understandability	7	1	2	6	3	1
Enjoyment	4	5	1	8	2	0

Table 1 shows that the mean values of the ratings in the confusion matrix were 0.82 for accuracy, 0.40 for precision, 0.35 for recall, and 0.37 for F1 score. In this study, the importance of the F1 score, which is the harmonic mean of the precision, which evaluates how many highlights are included in a highlight video, and the recall, which represents how few highlights are undetected, was considered important. However, the average F1 score for this study was low at 0.37. Next, Table 2 shows that the number of positive opinions exceeded negative opinions in terms of discomfort, ease of understanding the flow of the match, and interest in the video. This result likely indicates that the highlight videos produced by the system may have selected highlight scenes that were not chosen by the professionals. However, 90% of respondents said that professionally produced highlight videos were better than system-generated highlight videos.

5. Conclusion and Future Work

In this study, an automatic highlights generation method for gaming videos was proposed using facial expression analysis of video game streamers. Focusing on horror games, the proposed method loaded several highlight videos created by professionals, clustered the emotional values obtained by analyzing the facial expressions of the game streamers and created a clustering model. It was confirmed that the highlight videos created by the proposed system tended to include scenes with many highlights. The quality of the selected scenes was highly evaluated by the viewers' subjective evaluation. However, it is also clear that there is still room for improvement due to the slightly low scores of the evaluation indexes, such as the average F1 score of only 0.37. Also, it is necessary to further improve the automatic highlight video generation system and increase the accuracy of the system, for example, by adding conditions for assigning important labels, ranking scenes, flexible automatic setting of highlight video length, and adding other information to be evaluated.

6. References

[1] YouTube https://www.youtube.com/

[2] TikTok https://www.tiktok.com/

[3] YouTube Official Blog, https://blog.youtube/inside-youtube/shorts-revenue-sharing-update/

[4] Junya Matsuda, “Study on Automatic Creation of Clipped Video based on User Comments,” Proceedings of the 2023 Information Processing Society of Japan Kansai Branch Conference, 2023.

[5] Hidenori Aoki, Homei Miyashita, “A Trial Video Summarization and Chorus-Section Detecting on Nicovideo,” IPSJ Research Report Music Information Science (MUS), 2008, 2008.50 (2008-MUS-075): 37-42.

[6] Nico Nico Video https://www.nicovideo.jp/

[7] Yoshida Taiki, Ogawa Tomomi, “Highlight Generation Method for Game Commentary Using Acoustic Features,” 2023 IEEE 3rd International Conference on Software Engineering and Artificial Intelligence (SEAI), Xiamen, China, June 16-18, 2023, pp. 266-270.

[8] OpenCV https://opencv.org/

[9] dlib http://dlib.net/

[10] Savchenko Andrey, “Facial Expression Recognition with Adaptive Frame Rate based on Multiple Testing Correction,” Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR, Vol.202, Honolulu Hawai, July 23-29, 2023, pp. 30119-30129.

[11] Andrey V Savchenko, Lyudmila V Savchenko, Ilya Makarov, “Classifying emotions and engagement in online learning based on a single facial expression recognition neural network,” IEEE Transactions on Affective Computing, Vol. 13, Issue 4, 2022, pp. 2132-2143.

[12] Tan Mingxing, Le Quoc, “Efficientnet: Rethinking model scaling for convolutional neural networks,” International conference on machine learning, PMLR, 2019, pp. 6105-6114.