Multi-Modality based Affective Video Summarization for Game Players

Sehar Shahzad Farooq[1] [0000-0002-2571-9121], Abdullah Aziz[2] [0000-0003-4684-4682], Hammad Mukhtar[3], Mustansar Fiaz [1] [0000-0003-2289-2284], Ki Yeol Baek [1][0000-0001-7597-1926], Naram Choi, Sang Bin Yun [1][0000-0002-2940-3204], Kyung Joong Kim[3], and Soon Ki Jung[1] [0000-0003-0239-6785]

[1]School of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea
[2]Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden, 97187
[3]Dept. of Computer Science, National University of Computer and Emerging Sciences, Lahore, Pakistan
[4]Institute of Integrated Technology, Gwangju Institute of Science and Technology, Gwangju, South Korea

Abstract. Games has been considered as a benchmark for practicing computational models to analyze players interest as well as its involvement in the game. Though several aspects of game related research are carried out in different fields of research including development of game contents, avatar's control in games, artificial intelligent competitions, analysis of games using professional gamer's feedback, and advancements in different traditional and deep learning based computational models. However, affective video summarization of gamer's behavior and experience are also important to develop innovative features, in-game attractions, synthesizing experience and player's engagement in the game. Since it is difficult to review huge number of videos of experienced players for the affective analysis, this study is designed to generate video summarization for game players using multi-modal data analysis. Bedside's physiological and peripheral data analysis, summary of recorded videos of gamers is also generated using attention model-based framework. The analysis of the results has shown effective performance of proposed method.

Keywords: Video summarization, Affective analysis, Multi-modal data, and Game Player Modeling

1 Introduction

Games has been widely used as a source of entertainment for every age of groups [1]. Among them, mostly young and child group of people have shown more interest in playing games to keep themselves engage in achieving artificially developed challenges and achievements in the game [2]. Beside this, on the other side, game developers have put their visionary and imaginary ideas to develop such an advanced and challenging

games that it gives a hard time to players to achieve the goals [3, 4]. People invest time and money to cope up with the latest trends and techniques developed to deal the situations in the games [2]. Several video games have been developed in which the players have to control an avatar in the game and participate in a group of people who plays the game at the same time over the internet [5]. This not only gives the players an environment to express their personal behavior and emotions in the game but also help to develop themselves from experience of the other persons playing the game [6]. In recent times, these video games are been recorded and a huge volume of video data is generated [7, 8]. This data is then visualized to extract efficient and affective features. Based on these extractions, a feedback is shared to game developers and game industries to introduce new innovative features in the game [9, 10]. Along with these, this also help to develop in-game attractions to synthesize player's experience and escalate player's engagement in the game [11].

Previously computing was mainly focused to influence on text or numeric data, but as this digital system advanced, it come up with the introduction of several types of data including videos, audios, and images [2]. These heterogeneous types of data provide a platform for the development of huge numbers of applications. Some of the applications includes multimedia surveillance, content generation and analysis, dummy videos in medical experimental studies, advertisements, and games [12, 13]. To effectively manipulate a huge amount of such databases, it is the need of the hour to have a system to fulfil the requirements. It has been seen that traditional methods for data analysis and data management have deficient observations when requested for indexes and labeling. Video summarization techniques fulfil such deficiencies by effectively and efficiently generating and identifying pertinent contents [14, 15].

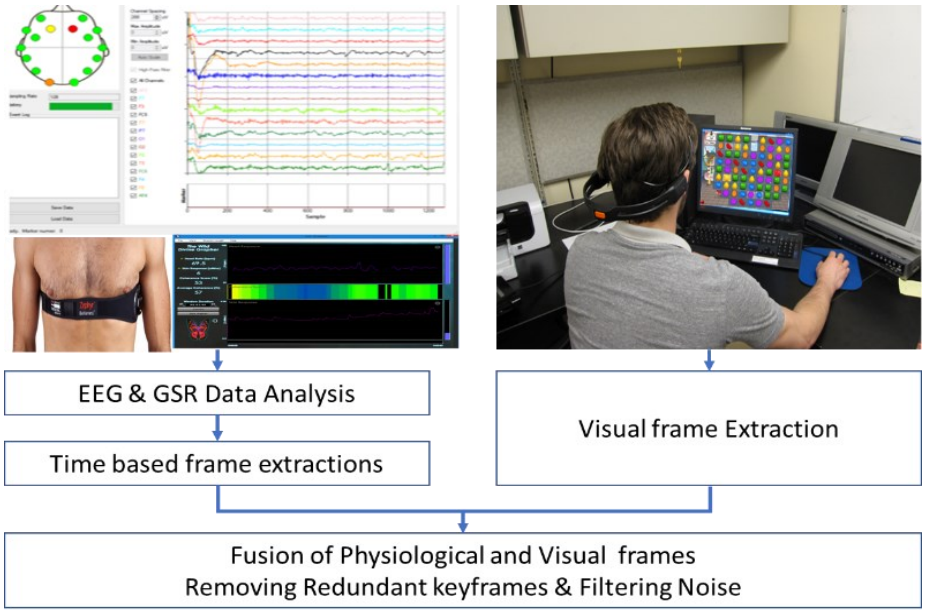

Fig. 1. Framework of proposed multi-modality based affective video summarization

Summary of affective videos can be generated in several forms, but most popular methodology is the combination of video skims, static storyboards, and affective content based keyframes [16, 17]. The final goal of video summarization is to develop highlights of the whole video of a particular event in humanly manners. The purpose of such highlights or short videos is to annotate contents and index long videos with-in database. It helps not only to save storage but also access time to find desired contents quickly. Although, such short videos can be generated manually but due to the long videos and limited manpower, it becomes impossible. A vital scenario is to develop an automated system to reduce manual processing.
Emotions play a crucial role for game players to develop their interest in the game [18]. These emotions are based on the physiological fluctuations caused by an object, a situation in the game as well as the surrounding environment [1, 19]. Emotions can be elicited in many situations within the games as well as the feelings of the players during the game [20, 21]. Such emotions represent behavior and experience of the gamers. Game player modeling refers to the descriptions of the players based on the framework of the data derived from the interaction of the player within the game as well as the association of the human player during the game [22]. In simple words what does a player do in the game is known as its behavior and how does a player feel during the game is its experience. Both behavior and experience of the gamers can be visualized in the game video analysis. To do this it is necessary to find out the most prominent and important events being happened in the videos. To figure it out, a framework is developed to extract smaller video skims with highly affective activities independently and combining them to a short summary. For the affective content analysis an attention-based summary in terms of time frames is generated from the analysis of physiological and peripheral data of the game players. Attention is a helpful mental procedure in cognition and permits humans to interrelate with the outer world in a more concentrated and specialized manner. The authors in [23] proposed the first attention-model based video summarization framework, which decomposes an original video sequence into the primary elements of its basic channels. Next, a set of features related to visual, aural, and linguistic attention is extracted to generate a comprehensive attention curve, which is used as an importance ranking, or to index the video content. On the other hand, a content-based summary is extracted from the recorded video of the players playing the game. These summaries then merge for affective video summarization. A framework of the proposed method is depicted in figure 1.

2 Proposed Methodology

2.1 Visual Frame extraction

Due to the limited resources and time complexities, it is impracticable sometimes to analyze the long videos of games. However, parsing videos [24] is a suitable option in which we divide the lengthy videos into several chunks. In these chunks there are also many shots that have a considerable temporal component. The first step in video summarization is the detection of shots. These depends upon the changes in a scene or an activity from the previous one. These boundaries define a hierarchy to make short

videos. A transition between these boundaries is analyzed to break the large videos in to small video chunks. Among these short chunks of videos, there are abrupt as well as gradual transitions [25]. An example of these transitions includes a totally different scene and a fade scene. The scenes are most importantly available in multiplayer video games where one person while performing its tasks is also dependent upon the other's performances. His progress is widely dependent upon his partners in a real time strategy game. Based on abrupt transition and multi-scene transitions, several methods have been developed in recent years. Though many considered low-level features to identify these transitions from successive frame including frame's intensity, edge detection, entropy and color histogram [25, 26]. In this research we utilized the approach developed in [27] to compute summaries based on the parsed videos. It is done by considering histogram difference of consecutive frames as can be seen in figure 2.

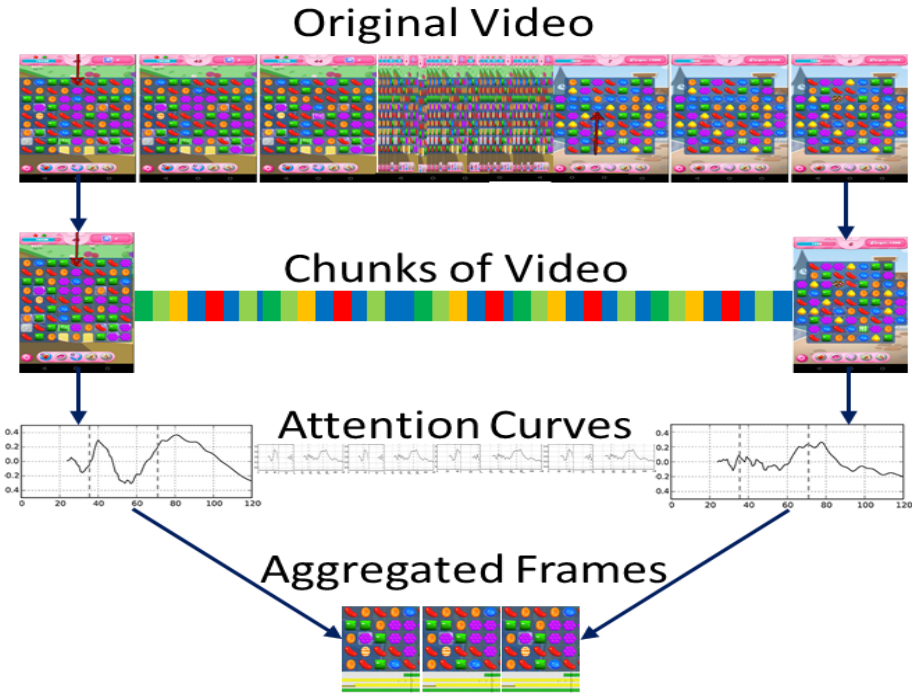

Fig. 2. Video summarization using attention curve modal

2.2 EEG and GSR based frame extraction

Electroencephalogram (EEG) reveal an electrical activity of the brain through electrodes that are placed at specific locations of the human scalp [28]. It generates waves of different heights and frequencies. The height displays the strength and frequency displays number of cycles. These signals of different amplitude can be classified into several categories. These classifications represent different psychophysiological situations. Among them, beta-frequency band ranging from 12~30 Hz represent attentiveness. The up and down of human emotional arousal represents its attention or awakens/

alertness and hence beta-band can be used to find the attention of the game players. Galvanic Skin Response (GSR) on the other hand reflect changes in our skin glands known as sweat glands and represent intensity of emotional state [29]. This is mostly dependent on the surrounding environment and basically related to the experience. Thus, EEG represents the in-human emotional activities and GSR represents the gamer's association while playing the game.

It is been evidenced by the neurobiologists that the attentiveness is regulated by arousal level known as the reactivity state in human. It can be panic, anger or excitement state. This is because of the beta-band in EEG signals. Hence, the attention features are measured by extracting the power spectral densities (PSDs) of beta-band of EEG Data [30]. Similarly experience based intensity of emotional state is also extracted using GSR's data. During pre-processing of raw data, the attention curve was normalized between 0 and 1. Careful synchronization of EEG data and player's video frames was the key element in this study which could give negative results if altered. So, at each particular time frame, a video frame is synchronized, and a specific frame is extracted from the video shots. The value close to 1 is considered as a high attention and values close to zero as lower or less attention.

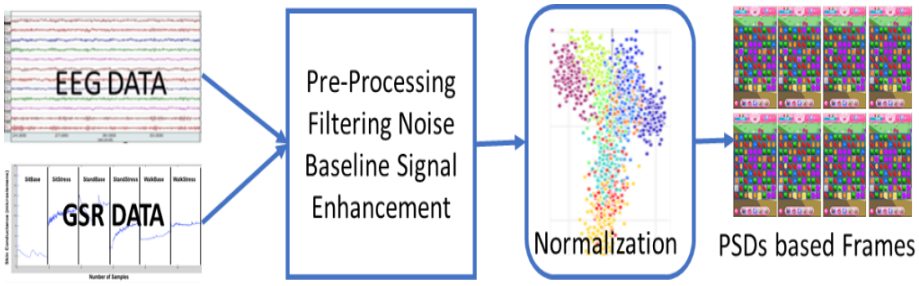

Fig. 3. EEG and GSR data-based frame extraction process

2.3 Inter-modality attention fusion

The visual frames synchronized with the same time frame of high attention curves of EEG and GSR data modal and visual frames extracted from the attention curve model of the parsing videos are than combined linearly to obtain the final video frames. This inter-modality fusion combines the strength of above individual modals and generate a non-redundant final summary. To reduce the duplicate scenes, we removed redundant frames.

3 Experiments and results

The proposed method is evaluated based on the videos recorded during play time of the gamers in laboratory environment. The details are explained in the following section.

3.1 Data collection

Devices

To access the potential response of the proposed methodology, the input data is generated by the subjects (game players) during their game play. For this purpose, we have set up an experimental environment where the EEG and GSR data is collected for the players during their game play. Along with it, we also set up a recording medium for the recording of the videos of the game screen during game play. This is done in such a way that each subject is wearing an EMOTIV EPOC headset[1] on its head with its electrodes at specific places of the scalp. The Bio Harness[2] device is worn around the chest of each game player. Both devices are controlled by a separate person to turn on and off the recordings of the game players. A mobile device is used to let the game players play the game. The game selected for this experiment is Candy Crush Saga[3]. It is freely available on the android platform. The recording capabilities of the Samsung Galaxy S7 mobile is used to record the screen of the game players during the game play. Windows 10 pro (64-Bit) operating system with Processor Intel® Core™ i7-4790 CPU@3.60Ghz (8 CPUs), RAM 32 Gigabyte and LG ULTRAWIDE (HDMI) MONITOR is used to monitor EEG and GSR signals and accurate video recordings for better synchronization of time at every device.

Data Types

Three types of data are collected in a way that the EEG signals of 14 channels is recorded by Emotive device with its real-time display on the monitor for every channel to work fine. The GSR signal display is connected to a computer and is manually recorded (start and end time). The video is recorded by itself from the mobile device. Before the experiment, each step of the experiment is explained to the subjects and upon their fully understanding, the experiment was conducted. The purpose of the experimental setup was to carefully synchronize the EEG and GSR data and video on the same time frame so that it become easy to extract the specific key frames from the videos based on the attention curve model response.

Game Levels and Subjects

Two Candy Crush Sage game levels were chosen for each player to play and its neuronal responses were recorded These levels are number 8and 343. Each subject is requested to play calmly and without extra stress to avoid depression-based alertness. After each level played by the subject, it is requested to each player to produce a summary of the game play by its own experience to maintain the ground truths for comparative analysis. A total of 10 subjects participated in this experiment and 10 videos were recorded. The subjects chosen for this experiment are the university

[1] https://www.emotiv.com

[2] https://www.biopac.com/product-category/research/telemetry-and-data-logging/bioharness/#:~:text=BioHarness%20with%20AcqKnowledge%20software%20is,respiration%2C%20posture%2C%20and%20acceleration.

[3] https://king.com/game/candycrush

Fig. 4. Comparison of affective video contents

students at graduate school. There was different type of players categorized as novice, intermediate and expert players. It took around 8~10 minutes for each player to play two levels whereas a 5-minute gap is given between two levels. In this rest period, it is requested to manually define a summary. To synchronize the EEG and GSR data with the visual frames of the video, EEGLAB's toolbox was used. It acquires EEG data wirelessly from the EMOTIV device and stores it. MATLAB platform was used to extract the features from EEG Data as well as from the Videos.

3.2 Case study: extracting keyframes from a single video

To demonstrate the effectiveness of the proposed model, one subject's data is presented in detail. It will help the reader to easily and correctly understand the proposed method and results acquired through it. The subject is a graduate student of one of the universities in South Korea, and is young, healthy, and intermediate game player. The video recorded for this subject consists of 2490 frames in which he plays two levels (easy and hard). The subject easily achieves the target in an easy level whereas it fails to complete hard level. However, the ground truth of the subject reveals that he felt emotional stimuli during the game play.

Table 1. F-measure comparison of proposed method and STIMO

Video No.	STIMO	proposed
1	0.51	**0.60**
2	0.55	**0.70**
3	0.65	0.55
4	0.47	**0.48**
5	0.60	0.58
6	0.66	0.64
7	0.72	0.67
8	0.74	**0.77**
9	0.50	0.46
10	0.48	**0.48**

For the underlying video recording, the video frames extracted from the video are presented. The attention features are measured by extracting the power spectral densities (PSDs) of beta-band of EEG and GSR Data. Based on that the relevant frames synchronized with the same time frame are extracted and displayed in Fig.4 (a). For video frame extractions, we used histogram difference of consecutive frames. However, since the background and overall game display remains the same, we have multiplied the difference several times to measure major changes in the video frames. Hence the frames extracted using this method are shown in fig.4 (b). Fig.4 (c) merges the video frames extracted using video frame extraction method and EEG and GSR based frame extractions into a combined summary. To reduce the duplicate scenes, we removed redundant

frames. Also, to keep the most interesting frames we combined the strengthened frames and removed the frames with weak features.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{F} - \text{measure} = 2 * \frac{\text{recall*precision}}{\text{recall+precision}} \tag{3}$$

3.3 Ground truth comparison-based summary evaluation

To access the effectiveness of the proposed video summarization, we used evaluation matrices for ground truth comparison-based summary. In this research, we requested each subject to evaluate the summary of the video for the ground truth. Though it is a difficult task for each subject, but we evaluate based on their experience. Along with it, we also get help from the expert group of multimedia team members. The f-measure, recall and precision matrices are usually used to evaluate the effectiveness of the models. The ratio of relevant keyframes chosen and the total frames either relevant or irrelevant is known as precision. Whereas recall is ratio of chosen keyframes to the total number of keyframes available in ground truth summary. The average of precision and recall is known as the f-measure. Higher the f-measure represents higher precision and higher recall values. In this research we find out the f-measure after evaluating precision and recall from the given formulas in equation 1, 2 and 3 respectively. According to the equations, TP represents true positive frames chosen and FP are false positive frames. To compare the proposed method with state-of-the-art techniques, the dataset is modified and facilitated to the previous known methods used in [17].

4 Conclusions and Future Work

In this research work, multi-modal based affective video summarization for game players is proposed. This research is based on a combination of extracting affective key frames for game players using two different models. The attention features are measured by extracting the power spectral densities (PSDs) of beta-band of EEG and GSR data. Based on that the relevant frames synchronized with the same time frame are extracted and combined to generate a short summary of the whole video. We also used histogram differences of consecutive frames from the whole video for video frame extractions. At last, we combined these two-attention based short summaries linearly into a single summary and removed redundant frames and less strengthened frames. The game player's attention is modeled based on several sensory perceptions. i.e., GSR signals, EEG, or neurological signals. In EEG attention model, we preferred beta-band frequency of the neuronal signals of the game player. It is found out that EEG based attention model reveal the emotional attachment of the game player within the game in terms of interest and focus. The f-measure of the proposed method is not ignorable.

Comparing with previous method STIMO, though the proposed model does not perform well for every video, but the results are considerable. Since most of the background of the Candy Crush Saga game is similar in every scene. Hence, it becomes difficult to compare the change in each key-frame. Though it is a primarily study, in the future, it can be improved using other videos of the game players and results shall be comparable.

Acknowledgement

This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394).
This work was also supported by Global University Project (GUP) grant funded by the GIST in 2020.

References

[1] S. S. Farooq and K.-J. Kim, "Game Player Modeling," in *Encyclopedia of Computer Graphics and Games*, N. Lee, Ed., ed Cham: Springer International Publishing, 2015, pp. 1-5.

[2] D. Hooshyar, M. Yousefi, and H. Lim, "Data-Driven Approaches to Game Player Modeling: A Systematic Literature Review," *ACM Computing Surveys (CSUR),* vol. 50, p. 90, 2018.

[3] C. Bateman and R. Boon, *21st Century Game Design (Game Development Series)*: Charles River Media, Inc., 2005.

[4] R. Lamb, L. Annetta, D. Hoston, M. Shapiro, and B. Matthews, "Examining human behavior in video games: The development of a computational model to measure aggression," *Social neuroscience,* vol. 13, pp. 301-317, 2018.

[5] J.-L. Hsieh and C.-T. Sun, "Building a player strategy model by analyzing replays of real-time strategy games," in *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 2008, pp. 3106-3111.

[6] M. Ahmad, L. Ab Rahim, K. Osman, and N. I. Arshad, "Towards Modelling Effective Educational Games Using Multi-Domain Framework," in *Encyclopedia of Information Science and Technology, Fourth Edition*, ed: IGI Global, 2018, pp. 3337-3347.

[7] C. Bauckhage, A. Drachen, and R. Sifa, "Clustering game behavior data," *IEEE Transactions on Computational Intelligence and AI in Games,* vol. 7, pp. 266-278, 2015.

[8] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*: Springer, 2017.

[9] R. Newbery, J. Lean, J. Moizer, and M. Haddoud, "Entrepreneurial identity formation during the initial entrepreneurial experience: The influence of

simulation feedback and existing identity," *Journal of Business Research,* vol. 85, pp. 51-59, 2018.

[10] M. Ambinder, "Biofeedback in gameplay: How valve measures physiology to enhance gaming experience," in *game developers conference*, 2011.

[11] D. Arsenault, "Video game genre, evolution and innovation," *Eludamos. Journal for Computer Game Culture,* vol. 3, pp. 149-176, 2009.

[12] H. Ekanayake, "CognitiveEmotional User Correction for Multimedia Interactions Using Visual Attention and Psychophysiological Signals," 2009.

[13] S. Mei, M. Ma, S. Wan, J. Hou, Z. Wang, and D. D. Feng, "Patch based Video Summarization with Block Sparse Representation," *IEEE Transactions on Multimedia,* 2020.

[14] N. Ejaz and S. W. Baik, "Video summarization using a network of radial basis functions," *Multimedia Systems,* vol. 18, pp. 483-497, 2012.

[15] Z. Ji, Y. Zhao, Y. Pang, X. Li, and J. Han, "Deep Attentive Video Summarization With Distribution Consistency Learning," *IEEE transactions on neural networks and learning systems,* 2020.

[16] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of visual communication and image representation,* vol. 19, pp. 121-143, 2008.

[17] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," *Multimedia Tools and Applications,* vol. 46, p. 47, 2010.

[18] R. Cowie, C. Pelachaud, and P. Petta, "Emotion-Oriented Systems," *Cognitive Technologies,* pp. 9-30, 2011.

[19] K. Loderer, R. Pekrun, and J. L. Plass, "Emotional foundations of game-based learning," *Handbook of Game-Based Learning,* p. 111, 2020.

[20] G. Du, W. Zhou, C. Li, D. Li, and P. X. Liu, "An Emotion Recognition Method for Game Evaluation Based on Electroencephalogram," *IEEE Transactions on Affective Computing,* 2020.

[21] B. K. Miller, "Guess the Emotion: A Tablet Game to Support Emotion Regulation Skills for Children with Autism," 2020.

[22] S. S. Farooq, J.-W. Baek, and K. Kim, "Interpreting behaviors of mobile game players from in-game data and context logs," in *Computational Intelligence and Games (CIG), 2015 IEEE Conference on*, 2015, pp. 548-549.

[23] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia,* vol. 7, pp. 907-919, 2005.

[24] S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," *IEEE transactions on circuits and systems for video technology,* vol. 11, pp. 522-535, 2001.

[25] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. de Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognition,* vol. 109, p. 107567, 2020.

[26] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang, and D. D. Feng, "Video summarization via block sparse dictionary selection," *Neurocomputing,* vol. 378, pp. 197-209, 2020.

[27] I. Mehmood, M. Sajjad, S. Rho, and S. W. Baik, "Divide-and-conquer based summarization framework for extracting affective video content," *Neurocomputing,* vol. 174, pp. 393-403, 2016.

[28] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *The Scientific World Journal,* vol. 2014, 2014.

[29] M. Val-Calvo, J. R. Álvarez-Sánchez, J. M. Ferrández-Vicente, A. Díaz-Morcillo, and E. Fernández-Jover, "Real-Time Multi-Modal Estimation of Dynamically Evoked Emotions Using EEG, Heart Rate and Galvanic Skin Response," *International Journal of Neural Systems,* vol. 30, pp. 2050013-2050013, 2020.

[30] S. Sanei and J. A. Chambers, *EEG signal processing*: John Wiley & Sons, 2013.