




A Video Summarization Method for Movie Trailer-Genre Classification Based on Emotion Analysis

Wan En Ng¹, Muhammad Syafiq Mohd Pozi²(✉) , Mohd Hasbullah Omar² ,
Norliza Katuk² , and Abdul Rafiez Abdul Raziff³

¹ Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Kuala Lumpur 50603, Malaysia

² School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
syafiq.pozi@uum.edu.my

³ Kulliyah of Information and Communication Technology, International Islamic University Malaysia, 50728 Kuala Lumpur, Gombak, Malaysia

Abstract. We live in an information world where visual data undergo exponential growth within a very short time window. With diverging content diversity, we simply have no capacity to keep track of those data. While short video platforms (such as TikTok™ or YouTube Shorts™) can help users view relevant videos within the shortest time possible, those videos might have misleading information, primarily if it is derived from long videos. Here, we analyzed several short videos (in terms of movie trailers) from YouTube and established a correlation between one movie trailer and the classified movie genre based on the emotion found in the trailer. This paper contributes to (1) an efficient framework to process the movie trailer and (2) a correlation analysis between the movie trailer and movie genre. We found that every movie genre can be represented by two unique emotions.

Keywords: movie analytic · face detection · emotion recognition · video summarization

1 Introduction

The emergence of social video platforms such as TikTok™ and YouTube Shorts™ have enabled people to consume several videos within a short time window [1]. This benefits both viewer and content creator, as only important and concise information is relayed from creator to viewer. However, some of these videos are redundant, misleading, clickbait or simply misrepresentations of reality [2]. As a result, many resources, such as computing capacity and (most importantly) people's time, have been wasted just to process and consume these short videos.

The misrepresentation issue can be largely observed in many movie trailers. A movie trailer (which is a very short video clip made of multiple scenes derived from the promoted movie) is usually composed when a movie producer needs to promote a new

movie to the public. However, in recent times, many have complained that some of the produced movie trailers are misrepresenting the promoted movie [3]. For example, one of the issues is that some movie trailers' scenes are unavailable or probably edited out from the promoted movie [4]. This false representation is why some people spend their money and 1 - 2 h of their life watching that movie. There is a case where a fan threatened legal action over a particular when the final movie did not include scenes that were promoted in the movie trailers¹.

As the number of movies released to the public increases over time, many producers want to ensure their movies are watched by many people worldwide. This will also increase the number of movie trailers released to the public. Hence, to facilitate consumers in visualizing many movie trailers without having to spend minutes watching the trailer, we introduce a method to summarize the movie trailers into metadata and perform data analytics in determining the relationship between the movie genre and emotion recognition, that occur in the movie trailer.

This paper is organized as follows. Section 2 overviews the previous works on video summarization from the year 1997 to the year 2022. Section 3 explains the used in this work. Section 4 describes the results of the conducted experimentation on several movie trailers. Finally, Sect. 5 presents the conclusion and future perspectives related to the main findings.

2 Related Works

The exponential growth of video data has been overwhelming that it requires some sort of summarization so users can consume many videos within a short period of time [5]. Even though video summarization has recently engaged growing attention in computer vision communities, the research on video summarization can be dated back to 1997 [6]. Here, we categorized the changes in video summarization corpus into three categories, image features statistical analysis, graph-based analysis, and semantic analysis.

2.1 Image Features Statistical Analysis

From 1997 to 2005, video summarization is mainly focused on statistical analysis of the extracted image features. Hence, many efforts are focused on building quality image features from a given video, based on low-level image analysis [7]. For example, a novel technique based on SVD (Singular Value Decomposition) which derives the refined feature space for clustering visual similar frames and a metric from measuring the amount of visual content available in each frame based on the degree of visual changes [8]. Some have proposed an automatic construction of video summarization based on similarities of images in the video which will be used to select key frames of video segments [9]. Later, a more robust two-stage clustering based video summarization technique has also been proposed to improve the relevancy of the generated video summary, whereby, the first stage provided shot separation and determined the periods of stable content with good keyframe candidates while the second stage selected the final set of key-frames [10]. A

¹ <https://bit.ly/3VTwcWS>.

much more sophisticated low-level image analysis to select representative frames based on the result of analysis of the events has also been proposed as a method to improve the relevancy of the generated video summary [11].

2.2 Graph Based Analysis

From 2005 to 2015, video summarization is slowly transformed into graph-based analysis. Chong-Wah Ngo *et. al* have proposed a unified approach for video summarization based on video structure analysis and video highlights, consisting of two major components: scene modelling and high-light detection [12]. The scene modelling consists of a normalized cut algorithm and temporal graph analysis meanwhile the highlight detection focuses on the motion attention modeling [12]. In 2006, Yuxin Peng and Chong-Wah Ngo proposed an algorithm for similarity measure of video clips based on bipartite graph matching algorithms which are Maximum Matching (MM) and Optimal Matching (OM) [13]. Maximum matching can filter the irrelevant video clips meanwhile Optimal Matching is able to rank the similarity of the clips based on visual and granularity factors [13]. D. Besiris *et. al* have proposed an automatic video summarization technique based on graph theory methodology and the dominant sets clustering algorithm [14]. Graph clustering and mining for multi-video summarization were proposed at the year of 2010 which the separated shot from visual data and extracted keywords from transcripts are structured into a complex graph and perform clustering while the hidden topics in the keyframes and keywords will be mined from clustered complex graph to maximize the coverage of summary over the original video [15]. A video summarization based on the Segments Summary Graphs is proposed, which is the coherency analysis of segmented video frames as represented by region adjacency graphs [16].

2.3 Semantic Analysis

Nowadays, many authors approach video summarization tasks within semantic analysis methodology, using every bit of video data and metadata. For example, in 2015, Yale Song *et. al* proposed TVSum. This unsupervised video summarization framework uses video title to find visually important scenes [17]. In 2016, Aidean Sharghi *et. al* proposed the Sequential and Hierarchical Determinantal Point Process (SH-DPP) for query-focused extractive video summarization, in which, given a user query and a long video sequence, the algorithm will generate a summary by selecting key shots from the video [18]. Later, Mohaiminul Al Bahian *et. al* proposed a convolutional neural network (CNN)-based architecture to mimic the frame-level shot for user-oriented video summarization [19]. Kaiyang Zhou *et. al* proposed a deep summarization network (DSN) that predicts a probability for each video frame which indicates how likely a frame is selected, selects the frames based on probability distributions and generates a video summary [20]. Lebron Casas *et. al* Proposed two models which are Video Summarization Long Short-Term Memory (vsLSTM) and Determinantal Point Process Long Short-Term Memory (dppLSTM) deep network that enable model frame relevance and similarity and additionally incorporate attention mechanism to model user interest [21]. LSTM generate a summary from the video by extracting the most relevant segments and vsLSTM contains the bidirectional chains of LSTM units [21]. The dppLSTM combines vsLSTM

with Determinantal Point Process (DPP) to model pair-wise repulsiveness within video frames [21]. In 2020, Wencheng Zhu *et al.* proposed a *Detect-to-Summarize network (DSNet) framework* containing anchor-based and anchor-free counterparts [22]. The anchor-based method generates temporal interest proposals to determine and localize the representative contents of video sequences while the anchor-free method eliminates the predefined temporal proposals and directly predicts the importance scores and segment locations [22]. Later, UN Yoon *et. al* proposed an unsupervised video summarization method with piecewise linear interpolation (In-terp-SUM) that improve summarization performance and generate a natural sequence of keyframes by predicting importance scores of each frame utilizing the interpolation method [23]. Figure 1 summarizes how video summarization research domain evolves over time.

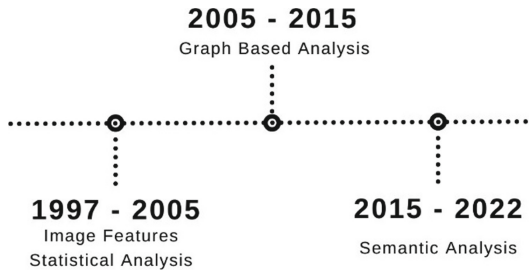


Fig. 1. A temporal summary of video summarization literature from 1997 to 2022.

3 Methodology

Figure 2 shows the overview of our methodology. In the first step, we collect several movie trailers from YouTube™, and group it based on the genre of the movie. Then, each video is splitted into several chunks, so that every chunk can be processed in parallel. Each chunk is processed to achieve the following tasks:

1. Face detection task, where we are going to identify whether there is at least one person in each chunk.
2. Emotion detection task, where we will identify the detected person's emotion.

Finally, a video summary will be produced by analyzing the distribution of detected emotion in each movie genre.

3.1 Data Collection

We used the movie trailers, downloaded from YouTube™ as our main data. Each movie trailer is based on 5 movie genres: Comedy, Action, Fantasy, Horror and Romance. A total of 25 trailers have been selected with five trailers for each movie genre. The duration of each movie trailer is about 3 to 4 min. Table 1 shows the list of movies that have been selected for each movie genre.

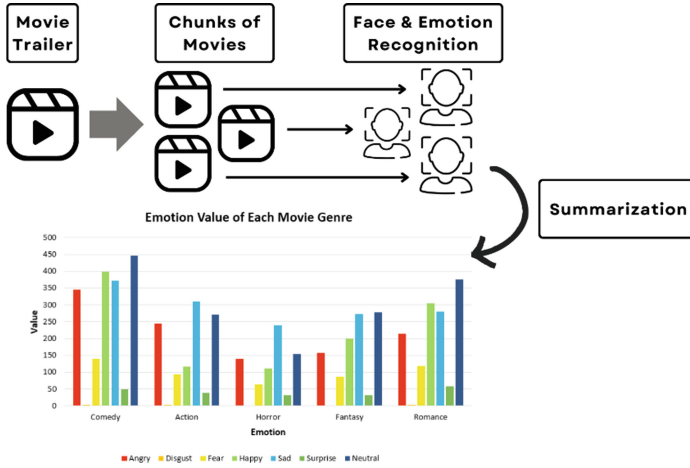


Fig. 2. Overview of our methodology used in this research.

Table 1. Table captions should be placed above the tables.

Movie Genre	Movie Title
Comedy	Ghostbusters, 2016
	Once Upon a Time in Hollywood, 2019
	Tag, 2018
	The Intern, 2015
	The Bucket List, 2007
Action	The Tomorrow War, 2021
	The Finest Hours, 2016
	Shooter, 2007
	Security, 2017
	Wonder Woman, 2017
Fantasy	Fantastic Beasts: The Secrets of Dumbledore, 2022
	Dolittle, 2020
	The Lord of the Rings: The Two Towers, 2002
	Miss Peregrine’s Home for Peculiar Children, 2016
	A Monster Calls, 2016
Horror	No One Gets Out Alive, 2021
	The Conjuring, 2013
	Don’t Breathe 2, 2021
	The Silence, 2019
	The Invisible Man, 2020

(continued)

Table 1. (continued)

Movie Genre	Movie Title
Romance	Little Women, 2019 Passengers, 2016 Crazy Rich Asians, 2018 Beauty and the Beast, 2017 Titanic, 1997

3.2 Data Processing

Each movie trailer is around 3 to 4 min. Even though the duration is shorter compared to the full movie, we still chunk the video into several chunks to benefit from the parallel processing facility provided by Python programming language. Prior to face detection and emotion recognition tasks, the movie trailer is chunked into multiple smaller chunks, in which each chunk only consists of 60 s short video data. These chunks are then processed in parallel for face detection and emotion recognition tasks.

3.3 Face Detection Task

A pretrained face detection model is used to identify faces from the video. It is based on Facial Emotion Recognition [24] and has been implemented as an open-source Python library² for emotion analysis of images and videos data which will categorize each of the faces based on the emotion into angry, disgust, fear, happy, sad, surprise and neutral. By default, Facial Emotion Recognition (FER) uses OpenCV's HaarCascade classifier to detect faces in image. However, to increase the face detection accuracy, FER implemented MTCNN network 3 in their face detection model. MTCNN is a Multi-task Cascaded Convolutional Networks which is a framework that consists of three stages of convolutional networks: Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net), that are able to recognize face based on the eye, nose and lip's location in an image frame [25, 26]. Figure 3 illustrates MTCNN network architecture, and how it relates to our dataset.

Stage 1: Proposal Network (P-Net).

This stage is a fully convolutional network which is known as Proposal Network (P-Net) which obtains the candidate windows and the bounding box regression vectors [25]. The bounding box regression vectors will be used to calibrate the candidate. Next, Non-Maximum Suppression (NMS) is employed to merge the highly overlapped candidates.

Stage 2: Refine Network (R-Net).

The candidates are fed to Refine Network (R-Net) to further reject a huge number of false candidates, calibrate the candidate with bounding box regression and merge the highly overlapped candidates with NMS.

Stage 3: Output Network (O-Net).

² <https://pypi.org/project/facial-emotion-recognition/>.

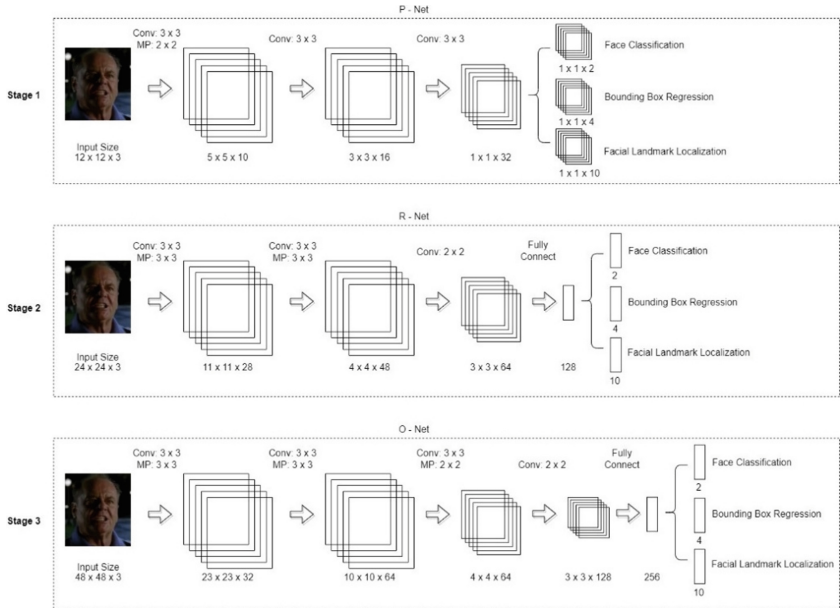


Fig. 3. An illustration of MTCNN network architecture. “Conv” refers to convolution while “MP” refers to max pooling.

Output Network (O-Net) is similar to the R-Net but it describes the detected face in more detail. At the end, the network will output the five facial landmarks’ positions as shown in Fig. 4.

When the MTCNN parameter is set to ‘True’, the FER model will use MTCNN network to detect faces meanwhile if it set to ‘False’, OpenCV HaarCascade classifier will be used to detect the faces. The model will run analysis on each frame of the video chunk and create a rectangle box around the face on every image which the emotion values next to it. It will publish a processed video which will have a rectangle box around the detected faces with live emotion values. Figure 5 shows the example of a detected face with live emotion values of comedy movie title with the name of The Bucket List.



Fig. 4. The detected faces with facial landmark.

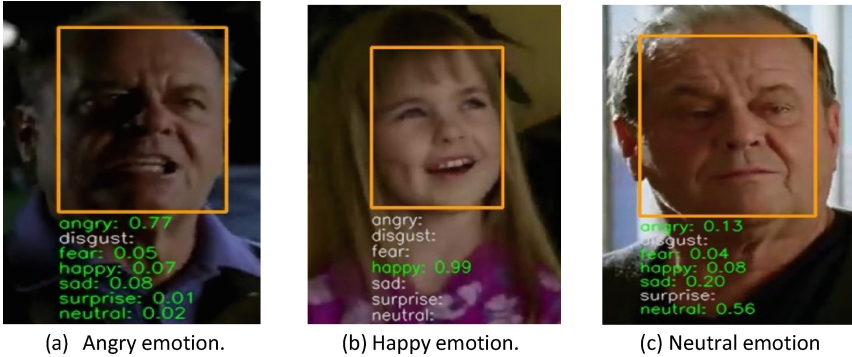


Fig. 5. Each detected face will have 7 emotions quantified with probability of that emotion to be true. The emotions are angry, disgust, fear, happy, sad, surprise and neutral.

3.4 Emotion Recognition Task and Computing Emotion Value

The FER implementation also has a function that can be used to perform emotion detection tasks. The function can be used to classify detected faces into seven emotion categories: angry, disgust, fear, happy, sad, surprise and neutral, as shown in Fig. 5. Hence, for each detected face, there will be seven emotion values associated with the face. Each emotion value is a probability for that emotion to be true in that specific face instance (detected face in one frame).

Figure 6 tracks the emotion value and score of each of the emotions of the movie named “The Bucket List”. In Fig. 6, every emotion in one frame is a probability value of all possible emotions, that sum up to 1, such as in Eq. 1:

$$frame = \sum_{j=1}^k p(y_j) \tag{1}$$

where each frame represents the probability of every emotion to be true in relative to all available emotions: $y_j \in \{\text{angry, disgust, fear, happy, sad, surprise, neutral}\}$.

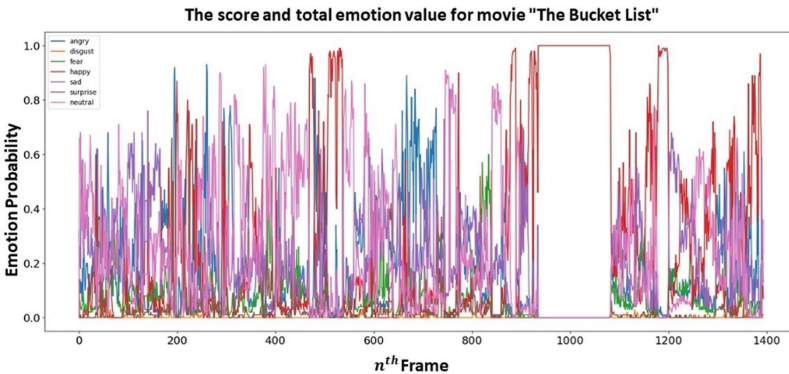


Fig. 6. The score and total emotion value for movie trailer “The Bucket List”.

In the movie “The Bucket List”, we can see that happy is the dominant emotion in that movie, especially at around 1000th frames, and disgust is the uncommon emotion in that movie.

4 Results and Analysis

Next, we compute the cumulative emotion value $e_{i,j}$ of video i , with emotion j , such as defined in Eq. 2:

$$e_{i,j} = \sum_{m=1}^n (p(y_j) f(\text{face})) \quad (2)$$

where $y_j \in \{\text{angry, disgust, fear, happy, sad, surprise, neutral}\}$ and $f(\text{face})$ is a function that describes the emotion of detected face, for every m face detected in the video up to n faces.

Figure 7 shows the cumulative emotion value of each movie genre, in which each genre is represented by 5 movie trailers. Note that a -second clip has 30 frames, hence why Fig. 7 showing values greater than 1.

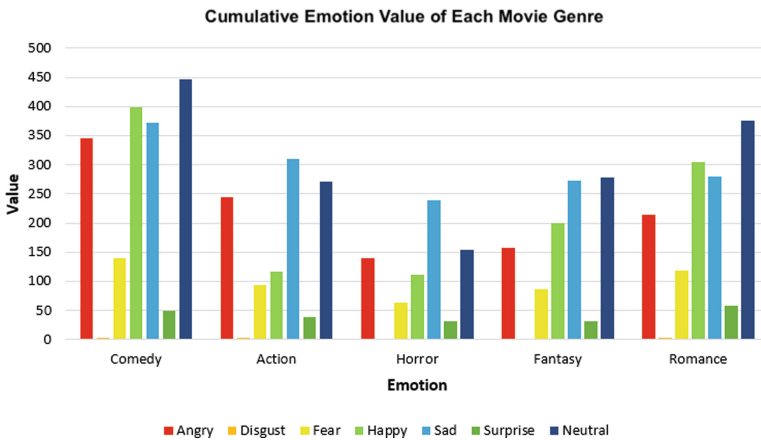


Fig. 7. Cumulative emotion value of each movie genre.

Based on Fig. 7, we found that:

1. We could not associate a movie genre with just one single emotion.
2. This experiment suggests each movie genre could be represented by the top three highest cumulative emotion value, such as:
 - (a) *Comedy* genre movies can be represented by *Neutral*, *Happy*, and *Sad* emotions.
 - (b) *Action* genre movies can be represented by *Sad*, *Neutral*, and *Angry* emotions.
 - (c) *Horror* genre movies can be represented by *Sad*, *Neutral*, and *Angry* emotions.
 - (d) *Fantasy* genre movies can be represented by *Neutral*, *Sad*, and *Happy* emotions.
 - (e) *Romance* genre movies can be represented by *Neutral*, *Happy*, and *Sad* emotions.

3. Some of the movie trailers that we collected could overlap with other genres. For example, the “*Crazy Rich Asians, 2018*” movie can also be considered a Romance and Comedy genre movie, instead of a Romance-only movie. Hence, we found that *Neutral* and *Happy* are mostly associated with *Romance* and *Comedy* movie genres.
4. For *Action* and *Horror* movies, three common emotions associated with those genres are *Sad*, *Neutral* and *Angry*.
5. Based on our dataset, emotion *Disgust* is not a popular emotion to be portrayed in a movie.

5 Conclusion and Future Works

In this paper, we run an experiment to investigate the relationship between short videos that are derived from long videos, based on the description of the long videos. We done this by collecting 25 movie trailers, classified based on 5 movie genres. Next, we establish a relationship between emotion found in the movie trailers and with its movie genres, respectively. We found that a single emotion cannot represent each movie genre. In our case, at least 2 emotions are required to represent a movie genre.

In the future, we plan to collect much more movie trailers and build a more accurate emotion detection tool to verify the relationship between emotions detected in each movie trailer and classified movie genre, which has been established in paper.

Acknowledgments. The authors thank the Ministry of Higher Education Malaysia for funding this study under the Fundamental Research Grant Scheme (Ref: FRGS/1/2019/ICT02/UUM/02/2, UUM S/O Code: 14358), and Research and Innovation Management Centre, Universiti Utara Malaysia for the administration of this study.

References

1. Cuesta-Valiño, P., Gutiérrez-Rodríguez, P., Durán-Álamo, P.: Why do people return to video platforms? millennials and centennials on TikTok. *Media Commun.* **10**(1), 198–207 (2022)
2. Gothankar, R., Troia, F.D., Stamp, M.: In: Stamp, M., Aaron Visaggio, C., Mercaldo, F., Di Troia, F. (eds.) Clickbait Detection for YouTube Videos, pp. 261–284. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97087-1_11
3. Korsgaard, M.B.: Fake trailers as imaginary paratexts. *MedieKultur: J. Media Commun. Res.* **36**(68), 107–125 (2020)
4. Garcia, R., Watson, W.: Fake it while you make it: When do fantasy and science fiction movie trailers become deceptive advertising? In *A Stranger Field. Studies of Art, Audiovisuals and New Technologies in Fantasy, SciFi and Horror Genres.*, 122
5. En, N.W., Mohd Pozi, M.S., Jatowt, A.: A face recognition module for video content analysis in malaysian parliament sessions. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, JCDL 2020*, pp. 533–534. Association for Computing Machinery, New York (2020). <https://doi.org/10.1145/3383583.3398628>
6. DeMenthon, D., Kobla, V., Doermann, D.: Video summarization by curve simplification. In: *Proceedings of the Sixth ACM International Conference on Multimedia*, pp. 211–218 (1998)
7. Zhang, H.J., Wu, J., Zhong, D., Smoliar, S.W.: An integrated system for content-based video retrieval and browsing. *Pattern Recogn.* **30**(4), 643–658 (1997)

8. Gong, Y., Liu, X.: Video summarization using singular value decomposition. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No. PR00662), vol. 2, pp. 174–180. IEEE (2000)
9. Yahiaoui, I., Merialdo, B., Huet, B.: Automatic video summarization. In: Proceeding of CBMIR Conference (2001)
10. Farin, D., Effelsberg, W., de With, P.H.: Robust clustering-based video-summarization with integration of domain-knowledge. In: Proceedings of IEEE International Conference on Multimedia and Expo, vol. 1, pp. 89–92. IEEE (2002)
11. Corchs, S., Ciocca, G., Schettini, R.: Video summarization using a neurodynamical model of visual attention. In: IEEE 6th Workshop on Multimedia Signal Processing 2004, pp. 71–74. IEEE (2004)
12. Ngo, C.-W., Ma, Y.-F., Zhang, H.-J.: Video summarization and scene detection by graph modeling. *IEEE Trans. Circuits Syst. Video Technol.* **15**(2), 296–305 (2005)
13. Peng, Y., Ngo, C.-W.: Clip-based similarity measure for query-dependent clip retrieval and video summarization. *IEEE Trans. Circuits Syst. Video Technol.* **16**(5), 612–627 (2006)
14. Besiris, D., Makedonas, A., Economou, G., Fotopoulos, S.: Combining graph connectivity & dominant set clustering for video summarization. *Multimedia Tools Appl.* **44**(2), 161–186 (2009)
15. Shao, J., Jiang, D., Wang, M., Chen, H., Yao, L.: Multi-video summarization using complex graph clustering and mining. *Comput. Sci. Inf. Syst.* **7**(1), 85–98 (2010)
16. Demir, M., Isil Bozma, H.: Video summarization via segments summary graphs. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 19–25 (2015)
17. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5179–5187 (2015)
18. Sharghi, A., Gong, B., Shah, M.: Query-focused extractive video summarization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016, Part VIII*, pp. 3–19. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_1
19. Al Nahian, M., Iftekhhar, A., Islam, M.T., Rahman, S.M., Hatzinakos, D.: Cnn-based prediction of frame-level shot importance for video summarization. In: 2017 International Conference on New Trends in Computing Sciences (ICTCS), pp. 24–29. IEEE (2017)
20. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
21. Lebron Casas, L., Koblents, E.: Video summarization with lstm and deep attention models. In: International Conference on MultiMedia Modeling, pp. 67–79. Springer (2019).
22. Zhu, W., Lu, J., Li, J., Zhou, J.: Dsnet: a flexible detect-to-summarize network for video summarization. *IEEE Trans. Image Process.* **30**, 948–962 (2021). <https://doi.org/10.1109/TIP.2020.3039886>
23. Yoon, U.-N., Hong, M.-D., Jo, G.-S.: Interp-sum: Unsupervised video summarization with piecewise linear interpolation. *Sensors* **21**(13), 4562 (2021)
24. Goodfellow, I.J., et al.: Challenges in Representation Learning: A Report on Three Machine Learning Contests. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) *Neural Information Processing*, pp. 117–124. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42051-1_16

25. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**(10), 1499–1503 (2016)
26. Saini, P., Kumar, K., Kashid, S., et al.: Video summarization using deep learning techniques: a detailed analysis and investigation. *Artif. Intell. Rev.* (2023). <https://doi.org/10.1007/s10462-023-10444-0>