Hindawi
Complexity
Volume 2019, Article ID 3581419, 10 pages
https://doi.org/10.1155/2019/3581419

Research Article

Personalized Movie Summarization Using Deep CNN-Assisted Facial Expression Recognition

Ijaz Ul Haq,[1] **Amin Ullah**,[1] **Khan Muhammad**,[2] **Mi Young Lee**,[1] **and Sung Wook Baik**[1]

[1]*Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Republic of Korea*
[2]*Department of Software, Sejong University, Seoul 143-747, Republic of Korea*

Correspondence should be addressed to Sung Wook Baik; sbaik@sejong.ac.kr

Received 16 November 2018; Revised 22 January 2019; Accepted 2 April 2019; Published 5 May 2019

Guest Editor: Li Zhang

Personalized movie summarization is demand of the current era due to an exponential growth in movies production. The employed methods for movies summarization fail to satisfy the user's requirements due to the subjective nature of movies data. Therefore, in this paper, we present a user-preference based movie summarization scheme. First, we segmented movie into shots using a novel entropy-based shots segmentation mechanism. Next, temporal saliency of shots is computed, resulting in highly salient shots in which character faces are detected. The resultant shots are then forward propagated to our trained deep CNN model for facial expression recognition (FER) to analyze the emotional state of the characters. The final summary is generated based on user-preferred emotional moments from the seven emotions, i.e., afraid, angry, disgust, happy, neutral, sad, and surprise. The subjective evaluation over five Hollywood movies proves the effectiveness of our proposed scheme in terms of user satisfaction. Furthermore, the objective evaluation verifies the superiority of the proposed scheme over state-of-the-art movie summarization methods.

1. Introduction

The video data is exponentially increasing over the Internet and personal storage devices including social networks, surveillance, and movies data due to advances and easy access to capturing technologies. Movies data specifically has become one of the most entertaining sources for viewers. However, browsing a movie in enormous collections and searching for a desired scene within a complete movie is a tedious and time-consuming task. Movie summarization (MS) techniques have tried to tackle this problem by producing a short video sequence from the movie, which contains the most important events or scenes. Hence, the viewers may have an idea about the context and the semantics of the movie by watching only the important scenes.

In recent years, many MS techniques have been presented by researchers that can be broadly categorized into automatic MS techniques [1–10] and user-preference based MS techniques [11–15]. In automatic MS techniques, there is no direct preference from the users to generate a summary. These techniques rely on multiple clues such as scripts, subtitles, and movies structure in combination with visual and aural features. For instance, Ngo et al. [1] utilized concept-expansion trees to construct a relational graph for characterizing the semantic concepts of documentary videos. You et al. [2] proposed a summarization method, where four perceptive models are fused according to different cues including motion, contrast, statistical rhythm, and special scenes using a linear combination to generate the summary. In contrast, Weng et al. [3] analyzed the movie from the perspective of relationships between the characters of a movie rather than audio-visual features. They constructed a role-network and then identified the leading roles. The role-network is then used to segment the movie into several substories. Similarly, Salamin et al. [4] used an audio segmentation method along with maximum a posteriori probability (MAP) approach to determine main characters in a movie, which can be used for indexing and retrieval of specific character's shots as well as for summary generation. Sang and Xu [5] used face clustering to get main characters and,

based on their appearance, they generated movie summary which contained only main characters. Evangelopoulos et al. [6] proposed a multimodal saliency based summarization scheme for movies. They extracted features from three different modalities and integrated them to form a multimodal saliency curve. They used spatiotemporal saliency model, an AM-FM speech model, and Part of Speech (POS) tagging to extract features from visual, aural, and textual module, respectively. Aparicio et al. [8] summarized movies and documentaries of different genres by analyzing six different text summarization algorithms on movie scripts and subtitles. The key contribution of their work is selecting a method that best fits among the six techniques for a movie or documentary of a particular genre. Another text based movie summarization scheme presented by Hesham et al. [9] generated a short summary as a trailer using subtitles of the movies. Hang Do et al. [10] summarized movies based on developing characters network. The relationships between characters are based on their appearance, which is used to segment the full-length movie into scenes. Finally, the storyline of the movie is generated as a summary by measuring the social strength of each character in the social network.

Due to diverse nature of movies and contradiction between user's preferences, the generated summary using automatic MS techniques may be felicitous for one user, but it may be infelicitous to others. Such MS schemes do not have strength to generate a summary that can fulfill the diverse subjective requirements of users. Therefore, user's preferred shots selection for movie summary is still a challenging problem, which is well addressed by user-preference based MS techniques. For example, Li et al. [11] suggested that substories from movies can be detected using short- and long-term audio-visual temporal features analysis. The length of generated summary is controlled by user input. Similarly, Ellouze et al. [12] used audio-visual features for personalized movie summary, where user can choose contents and type of various shots along with the duration of the summary. However, the comparison of user-preferred contents and movie contents is performed at feature level rather than semantic level. Peng et al. [13] utilized the emotion and attention of a viewer to generate summary according to user's mood by analyzing his facial expression, eye movement, blink, and head motion while watching a video. A movie summarization scheme based on user generated data is presented by Sun et al. [15]. They used real-time comments given by audience on the timestamp of a movie. The contents of the comments show the concepts of the ongoing scene, while the number of the comments indicates excitement level of the audience. Concluding the MS literature, movies are richer sources, providing semantics of complex ideas through audio-visual data. Therefore, segmenting a movie based on semantic level features gives best baseline for MS. Human emotions are one of the semantic level information, which can be extracted from video contents. FER has various applications in different domains such as medical [21], content recommendation [22], surveillance [23], safety [24], and robotics [25]. Similarly, in movies data, emotions of characters are the prominent element that directly gets audience's attention, which can be exploited to generate a meaningful summary.

Recently, a lot of research has been done for human emotion recognition using facial expression analysis. A comprehensive survey on FER is presented by Corneanu et al. [26] for RGB, 3D, thermal, and multimodal schemes. Traditional facial feature extraction schemes such as Gabor wavelet transform [27], optical flow [28], local binary patterns [29], and model-based methods [30] have many limitations like high computation and low performance due to diverse environments, i.e., light changes, pose, and clutter background. Moreover, these schemes are restricted to frontal faces and uniform skin colors. Recently, deep learning technologies [31] have shown tremendous results in the field of computer vision compared to traditional approaches. For instance, Kim et al. [32] proposed a hierarchical committee of deep CNNs by combining the decisions of multiple models trained on public FER databases. A feature redundancy-reduced (FRR-CNN) is proposed by Xie et al. [33] for FER to generate less redundant features and compact representation of the image. Uddin et al. [34] extracted local directional position patterns from depth video data and fed them into a deep belief network (DBN) for FER. Inspired from CNN approaches, in this paper, we proposed a user-preference based MS scheme, based on FER to determine the emotional state of the characters using CNN model. The user should choose the kind of emotional states that he prefers to be part of the final summary. The main contributions of this work are summarized as follows:

(1) Emotions of characters in a movie are the prominent elements that directly get audience's attention. Therefore, users always show interest in certain kind of emotional scenes in a movie. In this paper, we present a framework for generating summary based on user's preferred emotional scenes via an input query during summary generation.

(2) The hierarchical structure of a movie assists generating a sensible summary in which shots segmentation plays an important role. Therefore, we propose an entropy-based shots segmentation mechanism, which segments shots based on visual information. This strategy helps categorize shots into informative and noninformative.

(3) Deep learning approaches need huge data for better learning of parameters, whereas existing FER CNN models are not trained on such huge data. Therefore, for precise training of FER first we trained the model from the scratch over VGG face dataset to learn the structure of face and then we fine-tuned this model for FER on KEDF dataset. Our strategy gives prominent results over state-of-the-art techniques.

The rest of paper is arranged as follows: Section 2 discusses the proposed methodology for movie summarization in detail. Experimental results and discussion are presented in Section 3, followed by the conclusion and future work in Section 4.

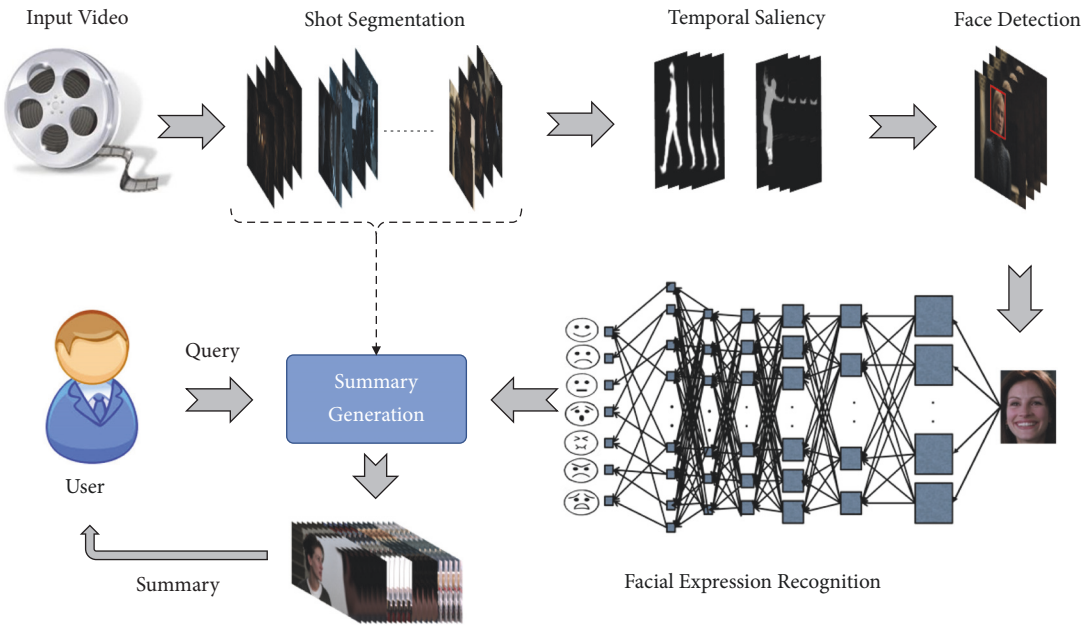

Figure 1: Overall framework of the proposed movie summarization scheme.

2. Proposed Methodology

Our proposed MS scheme is four folded: (1) entropy-based shots segmentation, (2) saliency extraction and face detection, (3) FER based on deep CNN model, and (4) summary generation. All the steps are discussed in subsequent sections in detail. The proposed system can generate summaries based on user-preference for any genre of movie. The overall framework of the proposed scheme is given in Figure 1.

2.1. Entropy-Based Shots Segmentation. Movies are also called structured videos because they include hierarchical structure of scenes and shots. A single shot is an uninterrupted segment of a movie that consists of sequential frames with static or continuous camera motion, while a scene consists of one or more shots of the same place or activity captured from different angles [1]. This structure assists in MS at initial stage by segmenting the full-length movie into shots and scenes. Shots segmentation is a key step for any summarization technique, especially, when dealing with entertainment videos. Recently, numerous domain specific shots segmentation techniques have been proposed such as color histogram based [15], deep feature based [35], person appearance based [31], and spare coding based methods [36]. In this paper, we proposed entropy-based shots segmentation technique, which analyzes frame sequences and selects the frame with sharp change in visual contents. Entropy E_N for a single frame can be calculated using (1) and (2).

$$p_i = \frac{n_i}{N_c} \tag{1}$$

$$E_N = -\sum_{i=1}^{N_c} p_i \log\left(p_i\right) \tag{2}$$

Herein, p_i is the probability of pixel, $q_i \in N_C$, N_C is the number of pixels in the neighborhood of pixel q_i, and n_i is the number of pixels having same intensities. Entropy of a frame represents the amount of information and semantics of the visual contents. Thus, it helps categorize the shots into informative and noninformative. Furthermore, the generated summary represents the informative shots only by excluding the shots with no or less information.

2.2. Saliency Extraction and Face Detection. Generally, saliency of an image is used to extract the foreground information, which can also be used for predicting the amount of information in it [37, 38]. To select the most informative shots, we calculated the average saliency score of a single shot using a saliency optimization technique [39]. Firstly, saliency map of a frame is obtained and the sum of all nonzero pixels is divided by the total number of image pixels. Secondly, average saliency score for a single shot is calculated by dividing the total sum of individual frame by total number of frames in a shot. Finally, the average saliency score is compared with a predefined threshold to select the most salient shots. In this way, all the nonsalient or noninformative shots are discarded, and the salient shots become the part of generated summary. The salient shots are further analyzed to detect characters' faces. For face detection, a multitask cascaded network [40] is used with additional constraint of size. The size constraint for a face is applied due to variation found in scale and poses to remove unwanted faces. Therefore, we selected only those faces, which are 15% of the frame size because main characters are filmed focused and closed. Also, FER is not working perfectly for small-sized faces [41]. Figure 2 represents some sample movie frames with detected faces and their corresponding saliency maps.

(a)

(b)

Figure 2: Sample frames from test movies: (a) highly salient shots with detected actor's faces and saliency maps and (b) low salient shots with detected actor's faces and saliency maps.

2.3. Facial Expression Recognition Using Transfer Learning. Training a deep CNN model requires a huge amount of data for learning its parameters from the scratch. However, transfer learning becomes a key concept in deep learning since it effectively deals with the problems having small datasets [42, 43]. Recently, CNN has beaten human error on image classification, when trained on a dataset with millions of data samples. However, some tasks such as FER are still facing the lack of data. Therefore, to tackle this problem, we used the concept of transfer learning using ResNet [20] CNN model for FER. ResNet CNN model has many versions including ResNet-34, ResNet-50, ResNet-101, and ResNet-152 layers networks. We have utilized ResNet-50 to balance the accuracy and time complexity of the system. Originally, it is trained on 224×224 images of ImageNet [44] dataset, which contains millions of samples for 1000 categories. We did not achieve good results when using weights of the pretrained ResNet-50 model for fine-tuning it on KDEF FER dataset [45]. The reason is that we have only face images that represent emotion of a human with very little changes on face in KDEF dataset and the model, which is pretrained on general categories data, is not effective for it. For this purpose, we introduced two step learning procedure where the first step includes training a CNN model from the scratch for face identification and second is transfer learning of the same model for FER. For face identification, we used large-scale VGG face [46] dataset to train weights of ResNet-50 for face data. The ResNet is primarily inspired by the structure of VGG [47] model, where both models use small size kernels for convolutional features extraction. The small-sized filters help learn all kind of tiny patterns in data, which are very common in FER [48]. ResNet utilized multiple consecutive branches of convolutional layers stacked on each other and performed down sampling with stride of 2. The network is ended with a global average pooling layer and one fully connected layer, presenting the number of classes for classification. The architecture of 50 weighted layers is given in Figure 3. The implementation details about transfer learning are discussed in the experimental section.

2.4. Summary Generation. The shots with high saliency and faces are forward propagated to the proposed trained CNN model for FER. In our experiments, we observed that FER from a single frame of a shot is not effective in representing the emotional state of the entire shot. Also, there is a possibility that a shot may contain multiple faces. Hence, the maximum detected emotion is selected as emotional state for the entire shot. Finally, the summary is generated according to the user's query specifying emotional state from the predefined seven classes of emotions. The flow of summary generation is visualized in Figure 4.

3. Experimental Evaluation

In this section, we have discussed the experimental evaluation of the proposed MS scheme. We performed two sets

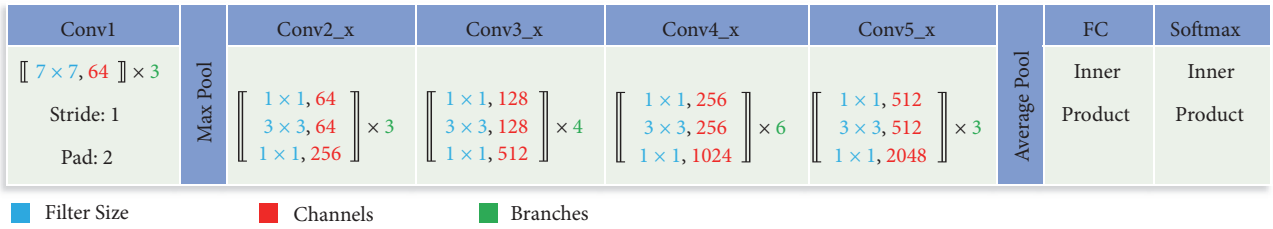

Figure 3: Architecture of 50-layer residual CNN model, which is changed for 128×128 face images. Conv3_1, Conv4_1, and Conv5_1 are down sampled using max pooling with stride of 2.

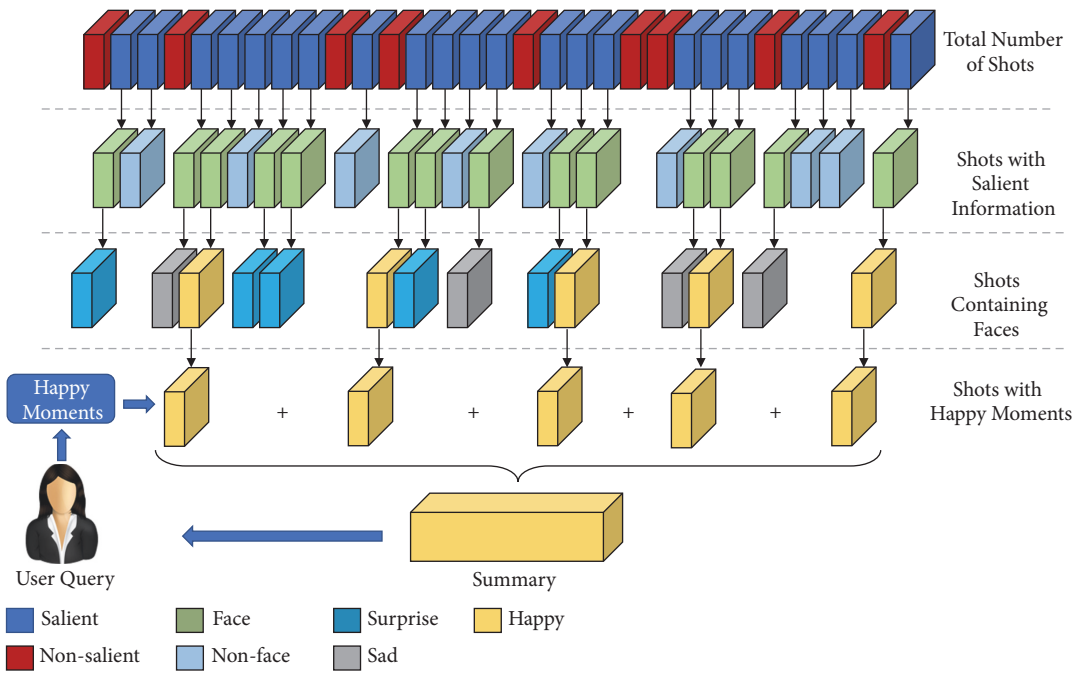

Figure 4: Flow diagram of summary generation based on user-preferences.

of experiments: (1) an evaluation of the trained model on KDEF dataset and its comparison with other models and (2) a subjective evaluation on five Hollywood movies of different genres. The experiments are performed using a deep learning framework known as Caffe [49] that is installed over Ubuntu16.04 operating system and equipped with NVIDIA TITAN X GPU having 12 GB dedicated memory running over a hardware of Intel™ Core i5 CPU with 64 GB RAM.

3.1. Datasets. We have used two datasets: VGG face [46] to train ResNet CNN model from scratch and KDEF dataset [45] for transfer learning. VGG face dataset contains 2.6 M images of 2.6 K celebrities around the world. KDEF dataset contains 4900 images of 70 subject including 35 males and 35 females. This dataset contains 7 classes, i.e., afraid, angry, disgust, happy, neutral, sad, and surprise. For each class, image samples are taken from five different angles in two sessions. This dataset best fits for our problem because, in movies, characters' faces are also found in a variety of poses. Figure 5 represents some sample images from each class of KDEF dataset. For subjective evaluation, five Hollywood movies are used. The detailed description of test movies is given in Table 1.

3.2. Objective Evaluation of Facial Expression Recognition. The ResNet CNN model is first trained on VGG face [46] dataset having 2597 classes. We resized all the images to 128×128 face regions and each pixel of the image is subtracted from the mean image to normalize the intensities. The full VGG face dataset is filtered out and some images were discarded in training. The dataset used in our experiments contains 0.42 million training and 0.14 million validation images. The original ResNet-50 is trained on 224×224 images; therefore, if we use pretrained ResNet-50, then its kernel size, stride, and padding information is not fitting for 128×128 image classification problem. In our method, the weights of ResNet for 128×128 face images are initialized from the scratch. The detailed description can be seen in the original ResNet article [20]. The model is trained for 50 epochs with 64 batch size and the learning rate is initialized with 0.01, which is decreased after every 10 epochs by the factor of learning rate ratio 10. The reason behind decreasing the

Figure 5: Sample images from each class of KDEF dataset.

Table 1: Description of the test movies data.

Movie ID	Movie title	Genre	Length (min)	Number of shots	Number of scenes
TLH	The Lake House	Fantasy/Drama	105	865	76
MBBN	My Blueberry Nights	Romance/Drama	90	1009	73
YHGM	You've Got Mail	Comedy/Romance	119	1049	62
SA	Salt	Action/Adventure	100	2339	78
NH	Notting Hill	Comedy/Romance	124	1623	42

Table 2: Transfer learning results of different CNN model for KDEF dataset.

CNN models	Overall accuracy (%)
MobileNet [16]	40.65
SqueezNet [17]	45.37
AlexNet [18]	46.81
GoogleNet [19]	52.63
ResNet-50 (224×224) [20]	64.83
ResNet-50 (128×128)	*93.65*

learning rate is to prevent the model from overfitting problem during the training. We achieved precise results on VGG face dataset using ResNet-50 layers network, where the accuracy after 50 epochs reached 96.82% and the loss decreased up to 7×10^{-5}. Results of different CNN models fine-tuned on KDEF datasets are given in Table 2. It is clear from Table 2 that all the pretrained models have achieved very less accuracy when fine-tuned on the original weights. The reason of low accuracy is that these models are trained on general categories dataset and we need a model whose parameters are previously trained on face data. Therefore, first we trained ResNet-50 using large-scale VGG face dataset to learn face structure and then fine-tuned the trained model on KDEF dataset for FER.

After training ResNet with face recognition dataset, we claim that its weights can now learn face features and its structure effectively. Therefore, we have used the parameters of the trained model for transfer learning of FER using KDEF dataset. For fine-tuning process, all images of the KDEF dataset are resized to 128×128 face regions and each pixel of image is subtracted from mean image to normalize the intensities. In transfer learning process, we have initialized the learning rate from 0.001 and decreased it after each 10 epochs by the factor of learning rate ratio 10. The model is fine-tuned for 30 epochs, achieving 92.08% validation accuracy with loss of 0.192 at final epoch. Confusion matrix and overall accuracy for the test set of KDEF dataset are given in Table 3. All categories are not much confused with each other, i.e., afraid, angry, and sad classes achieved per-class accuracy under 90% while the rest of all classes have accuracy above 90%. The results for this dataset are very convincing, making our trained model capable of FER in the heterogamous movie data. The KDEF dataset has various categories of face poses and viewpoint variations, which help to easily analyze the character's facial expressions in the movie.

3.3. Subjective Evaluation of Generated Summary. One of the challenging steps in movie summarization is the evaluation of the generated summary due to the lack of standards. Generally, there are two types of assessment used in video summarization literature that can be categorized into intrinsic and extrinsic techniques. In intrinsic evaluation, the generated summary is directly analyzed from its contents. For instance, fluency in generated summary, coverage of the main theme of original video, and similarity with referenced summary generated by movies expert are checked. In extrinsic evaluation, the performance is evaluated as information retrieval problem using a multichoice questionnaire. The excellence of summary is then measured by the increase in quiz scores. In this paper, we followed the second technique because it generates summary based on user's query to select emotional shots of a specific class. In our experiments, a total of ten subjects participated in the subjective evaluation, in which six students are selected from graduate and four from undergraduate program having age in the range from 20 to 25 years. All the participants were instructed to watch the selected movies before the evaluation and asked to rate the following three questions between 1 and 10 after watching the summary generated by their desired query. Table 4 represents the statistics of all the detected emotions in the informative shots of the test movies.

Table 3: Confusion matrix and overall accuracy for the test set of KDEF dataset.

	Afraid	Angry	Disgust	Happy	Neutral	Sad	Surprise	Per-class accuracy (%)
Afraid	*51*	1	1	1	0	1	7	82.26
Angry	1	*62*	0	0	1	6	0	88.57
Disgust	0	1	*73*	0	0	2	0	96.05
Happy	0	0	2	*71*	0	0	0	98.03
Neutral	1	1	0	0	*70*	1	2	93.33
Sad	3	1	1	1	3	*64*	0	87.67
Surprise	1	0	0	0	0	0	*43*	97.73
				Overall accuracy				**93.65**

Table 4: Statistics of all the detected emotions in the informative shots of the test movies.

Movie ID	Number of shots detected in each emotion category						
	Afraid	Angry	Disgust	Happy	Neutral	Sad	Surprise
TLH	12	17	0	24	113	87	0
MBBN	19	13	1	18	211	48	6
YHGM	9	16	2	46	186	38	4
SA	13	20	0	4	102	72	2
NH	128	101	3	157	389	203	12

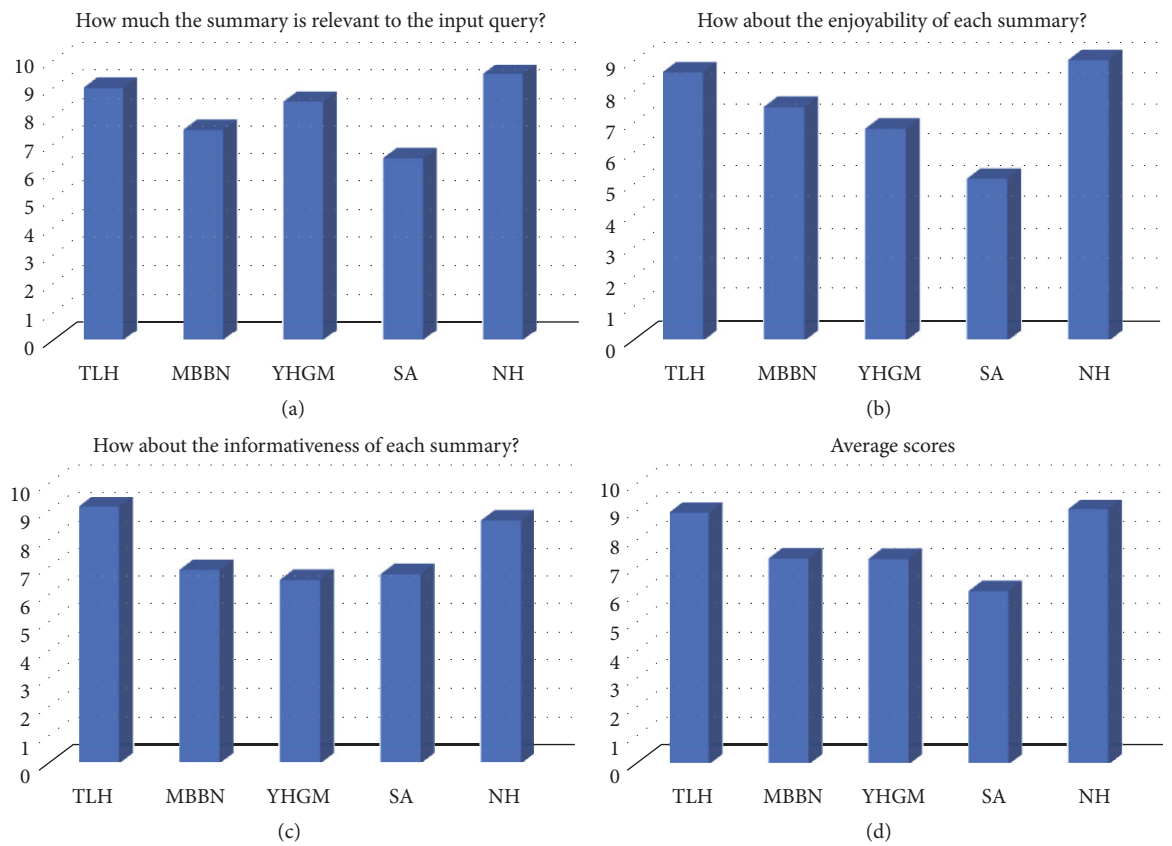

Figure 6: Subjective evaluation of the generated summaries: (a), (b), and (c) represent average scores for Q 1, Q 2, and Q 3, respectively, and (d) shows the average scores of all the three questions.

Figure 7: Sample shots from movie "*Notting Hill*" with emotional states predicted by our proposed scheme.

Q 1: How much the summary is relevant to the input query?

Q 2: How about the enjoyability of each summary?

Q 3: How about the informativeness of each summary?

In Figure 6, the average scores of all the participants for each test movie are calculated and represented in the form of graphs. It is clear from Figure 6 that all the test movies give good results except "*Salt*", which is an action movie. In action movies, the human movements are fast, making the task of FER very challenging due to blur effects. Our proposed scheme achieved best results for movie "*Notting Hill*", which is a romantic movie, containing very rich emotions. Figure 7 represents some shots from movie "*Notting Hill*" with corresponding emotional state of the shots. Concluding the overall evaluation and discussion, we claim that the performance of our proposed scheme is best on movies of genre drama, comedy, romance, and fantasy compared to action and adventures.

4. Conclusion

In this article, we presented a user-preference based movie summarization scheme. First, we segmented the movie into shots using a novel entropy-based shots segmentation mechanism. Secondly, we computed temporal saliency for each shot to discard nonsalient shots. Next, character's faces are detected in the salient shots and fed into a deep CNN model for FER. Finally, the summary is generated according to the user's query of any emotion state from the predefined seven classes. We evaluated our trained model for FER and the overall proposed scheme of movie summarization using objective and subjective analysis. We found that our proposed scheme demonstrates better performance compared to other movies summarization techniques. In future, we aim to conduct experiments on animated movies and fuse both aural and visual features for movies summary generation.

Data Availability

All the datasets and testing movies are publicly available. We provided the citation of each dataset and the movies are downloaded from online databases, i.e., YouTube and Crackle. The python code for analysis and the trained model will be made available to readers upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1D1A1B07043302).

References

[1] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on*

Circuits and Systems for Video Technology, vol. 15, no. 2, pp. 296–304, 2005.

[2] J. You, G. Liu, L. Sun, and H. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 273–285, 2007.

[3] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: Movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[4] H. Salamin, S. Favre, and A. Vinciarelli, "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1373–1380, 2009.

[5] J. Sang and C. Xu, "Character-based movie summarization," in *Proceedings of the 18th ACM International Conference on Multimedia ACM Multimedia 2010, MM'10*, pp. 855–858, October 2010.

[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos et al., "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[7] C.-M. Tsai, L.-W. Kang, C.-W. Lin, and W. Lin, "Scene-based movie summarization via role-community networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1927–1940, 2013.

[8] M. Aparício, P. Figueiredo, F. Raposo, D. Martins De Matos, R. Ribeiro, and L. Marujo, "Summarization of films and documentaries based on subtitles and scripts," *Pattern Recognition Letters*, vol. 73, pp. 7–12, 2016.

[9] M. Hesham, B. Hani, N. Fouad, and E. Amer, "Smart trailer: Automatic generation of movie trailer using only subtitles," in *Proceedings of the 1st International Workshop on Deep and Representation Learning, IWDRL 2018*, pp. 26–30, 2018.

[10] T. T. Do, Q. H. Tran, and Q. D. Tran, "Movie indexing and summarization using social network techniques," *Vietnam Journal of Computer Science*, vol. 5, no. 2, pp. 157–164, 2018.

[11] Y. Li, S.-H. Lee, C.-H. Yeh, and C.-C. J. Kuo, "Techniques for Movie Content Analysis and Skimming Tutorial and overview on video abstraction techniques," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 79–89, 2006.

[12] M. Ellouze, N. Boujemaa, and A. M. Alimi, "IM(S)2: Interactive movie summarization system," *Journal of Visual Communication and Image Representation*, vol. 21, no. 4, pp. 283–294, 2010.

[13] W.-T. Peng, W.-T. Chu, C.-H. Chang et al., "Editing by viewing: Automatic home video summarization by viewing behavior analysis," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 539–550, 2011.

[14] R. Kannan, G. Ghinea, and S. Swaminathan, "What do you wish to see? A summarization system for movies based on user preferences," *Information Processing & Management*, vol. 51, no. 3, pp. 286–305, 2015.

[15] Z. Li, X. Liu, and S. Zhang, "Shot Boundary Detection based on Multilevel Difference of Colour Histograms," in *Proceedings of the 1st International Conference on Multimedia and Image Processing, ICMIP 2016*, pp. 15–22, June 2016.

[16] A. G. Howard, M. Zhu, B. Chen et al., *Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, 2017, https://arxiv.org/abs/1704.04861.

[17] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, *SqueezeNet: Alexnet-Level Accuracy with 50x Fewer Parameters and < 0.5 MB Model Size*, 2016, https://arxiv.org/abs/1602.07360.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, 2012.

[19] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15)*, pp. 1–9, IEEE, June 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 770–778, July 2016.

[21] L. Hazelhoff, J. Han, S. Bambang-Oetomo, and P. H. de With, "Behavioral state detection of newborns based on facial expression analysis," in *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 698–709, 2009.

[22] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 636–647, 2013.

[23] M. Sajjad, M. Nasir, F. U. M. Ullah, K. Muhammad, A. K. Sangaiah, and S. W. Baik, "Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services," *Information Sciences*, 2018.

[24] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," in *Proceedings of the International Workshop on Human-Computer Interaction*, pp. 6–18, 2007.

[25] L. Zhang, M. Jiang, D. Farid, and M. A. Hossain, "Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5160–5168, 2013.

[26] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[27] M. Tkalcic, A. Odic, and A. Kosir, "The impact of weak ground truth and facial expressiveness on affect detection accuracy from time-continuous videos of facial expressions," *Information Sciences*, vol. 249, pp. 13–23, 2013.

[28] C.-K. Hsieh, S.-H. Lai, and Y.-C. Chen, "An optical flow-based approach to robust face recognition under expression variations," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 233–240, 2010.

[29] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: a comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[30] M. A. A. Dewan, E. Granger, G.-L. Marcialis, R. Sabourin, and F. Roli, "Adaptive appearance model tracking for still-to-video face recognition," *Pattern Recognition*, vol. 49, pp. 129–151, 2016.

[31] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. Albuquerque, "Activity recognition using temporal optical flow convolutional features and multi-layer LSTM," *IEEE Transactions on Industrial Electronics*, 2018.

[32] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[33] S. Xie and H. Hu, "Facial expression recognition with FRR-CNN," *IEEE Electronics Letters*, vol. 53, no. 4, pp. 235–237, 2017.

[34] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[35] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," *Pattern Recognition Letters*, 2018.

[36] J. Li, T. Yao, Q. Ling, and T. Mei, "Detecting shot boundary with sparse coding for video summarization," *Neurocomputing*, vol. 266, pp. 66–78, 2017.

[37] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1746–1758, 2017.

[38] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, p. 38, 2018.

[39] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 2814–2821, June 2014.

[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[41] I. U. Haq, K. Muhammad, A. Ullah, and S. W. Baik, "DeepStar: detecting starring characters in movies," *IEEE Access*, vol. 7, pp. 9265–9272, 2019.

[42] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

[43] M. Seera and C. P. Lim, "Transfer learning using the online fuzzy min–max neural network," *Neural Computing and Applications*, vol. 25, no. 2, pp. 469–480, 2014.

[44] J. Deng, W. Dong, R. Socher et al., "ImageNet: a large-scale hierarchical image database," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 248–255, 2009.

[45] M. G. Calvo and D. Lundqvist, "Facial expressions of emotion (KDEF): Identification under different display-duration conditions," *Behavior Research Methods*, vol. 40, no. 1, pp. 109–115, 2008.

[46] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015*, p. 6, 2015.

[47] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2014, https://arxiv.org/abs/1409.1556.

[48] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.

[49] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the ACM Conference on Multimedia (MM '14)*, pp. 675–678, 2014.