Specific Designed Facial Expression Recognition System for Interactive Film Applications

Rui Qin[1], Jingxin Liu[1], Hongying Meng[1], and Tong Chen[2]

[1]Brunel University London, London, UB8 3PH, UK
[2]Southwest University, Chongqing, 400715, China
Hongying.Meng@brunel.ac.uk

Abstract. Emotion Artificial Intelligence (AI) is a novel technology for advanced human machine interaction in real-world applications. In interactive films, variant contents can be adjusted and displayed in the film based on audiences' interaction such as voice, hand gesture or body movement, etc. In this paper, a specific emotion detection system is designed and implemented that can detect emotion continuously through the audience's facial expression and give the feedback to the film immediately. Then the film can change its contents accordingly. In this system, A pre-trained convolutional neural network is used for feature extraction from video frames and then the emotion is predicted by a support vector regression model. The environmental noise is reduced in the pre-processing stage and the final prediction is smoothed in the post-processing. A database is recorded for this particular scene and the proposed system is trained on it. The experimental results demonstrate the effectiveness of the system and the built interactive film "RIOT" has been exhibited on several occasions with good performance.

Keywords: Facial Expression Recognition, Interactive Film, Convolutional Neural Network, Support Vector Regression

1 Introduction

Interactive film, also been called interactive film game, is a kind of video games presented in a cinematic method usually with full-motion video captured by actors or animated object [21]. The early equipment of playing these video games concludes a video player connected with a computer and the software with a visual interface displaying different buttons of choices in check points. With the correct buttons pressed, the film will go to the next chapter; With the wrong buttons pressed, the film will stop playing usually with a 'game over' shown in the screen.

Emotion AI is a subset of artificial intelligence that measures, understands, simulates, and reacts to human emotions. It's also known as affective computing, or artificial emotional intelligence. With the development of emotion AI technology, there exist some innovations in interactive film. One of the most important aims in interactive film is to enhance the empathy of audiences such as 'Dark

Sides' [10]. The control system is added with speech or emotion recognition [19] [18].

The first emotion recognition system used for interactive film is proposed by Nakatsu et al. [19] in 1999 when the emotion of human speech was detected and transferred to a computer agent that played a character role in the interactive film system. There is limited progress since then. However, the emotional feedback for interactive film is so important and it has been identified as a future development direction of interactive film making area [3].

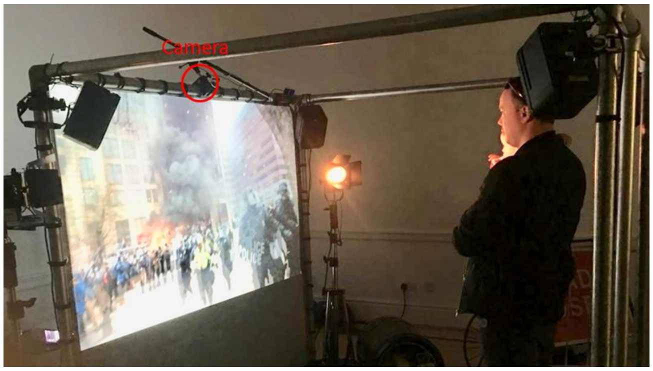

Fig. 1. Interactive film system. The audience's facial expression is captured from a web camera and the film will be run to different directions based on the audience's emotion.

The RIOT prototype [20] is a interactive film to display a riot situation as shown in Fig. 1 It is used as a training tool for the users to strength their mental toughness gradually in a disastrous situation. The system includes an interactive film with a set of stereo sounds. In order to avoid mental crash of the users when they see terrible scenarios, the emotion of the users should be carefully monitored and the film should response and adapt quickly if there is clear anger or fear emotion detected from the users. A simplified emotion recognition system is needed for this particular application.

2 Related works

Facial expression recognition (FER) is still a challenging task as it requires high speed of emotion detection under the constrains such as limitation of computing power, high data rate of the video data, etc.

In the work of [13], [17], [8] and [7], images are treated individually for the classification and the dynamic facial expression information is lost. In [1], face detection and extraction are used and the best performance is achieved among all other methods. However, a cloud is needed for this system that is not useful while the cloud or Internet is not available. A fast prediction speed is achieved

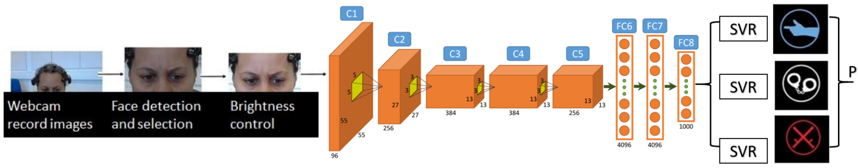

Fig. 2. Main block diagram of the proposed FER system. In pre-processing, face is detected and brightness is adjusted. Then CNN is used for feature extraction and SVR for regression on each emotion. After that, post-processing is carried out to smooth the emotion predictions of multiple frames continuously.

in [23], but the performance is limited due to the algorithm simplicity restricted by hardware resource.

The existing deep learning FER system like [14] and [15] cost too much calculation sources and can not achieves the short time response requirement of interactive film. The existing short time response FER system like [23], the accuracy is too low. Another existing FER system in [2] is good for using on mobile phones with only one face in screen and good condition of light. But with lots of audience in exhibition of interactive film and an extremely change of light for film, the system will not adapted.

In RIOT film system, a 3D sound with 8 speakers is used in the platform, that takes lots of computing power of the computer already. So the computing resource for FER system is very limited. In addition, the computing speed requirement is very strict as the response has to be fast enough.

3 Specific designed FER system

The interactive film is built on a Java platform that coordinates the film segments to be shown in a particular order, synchronises the 3D audio and emotion predictions. On the platform, the FER system is run all the time to output the emotions on the faces in real time. The FER system works as soon as the camera is turned on. There are essential 4 check points in the interactive film where the emotion prediction will be used for showing different segments.

In this system, three types of emotions detected are angry, fear and calm as they are the main response while watching the film. The scale of each emotion is produced as real numbers, so it is a regression problem. There are essential 4 check points in the interactive film and at each checking point, the overall emotion scale is calculated and the dominate emotion is reported.

3.1 System overview

The main components of the proposed FER system is shown in Fig. 2. There are 4 stages. Firstly, human face is detected and the brightness of the environment is adjusted. Secondly, a CNN network is used for feature extraction and three SVRs are used for regression on each emotion. Finally, post-processing is carried out to produce emotion predictions by smooth the predicted emotion from multiple frames.

3.2 Pre-Processing

Face detection Firstly, Continues Adaptive Mean Shift (CAMSHIFT) [4] face detection method is used to detect the faces from the camera. It is a classic face detection algorithm that tracks faces in video by using a probability distribution of the faces.

Face selection In general, there are multiple faces in front of the camera. For this interactive film system, the users stand in a circle, and the camera usually captures several faces. The main user usually stands close to the screen, the face extracted will be the largest. Thus, a simple algorithm is developed for face selection. Only the largest face image is chosen for emotion detection.

Brightness adaptation Unlike the system built in the lab, real application systems need to work under all lighting conditions. So a brightness adaptation method is used to adjust the illumination. The basic idea is to average the brightness in the faces of subjects. Based on the average light, some faces that are too dark will be brightened, and some faces that are too bright will be dimmed. Without brightness control system, the facial recognition system may lose the face when it is too dark or too bright. There are lots of brightness control system such as bi-histogram equalisation introduced by Chen et al. [5]. But considering the limitation of calculation sources, a simple average brightness control method is utilised.

After changing from colour space to gray-scale on each frame, a video dataset can be represented as $I^{v}_{ijf}(1 \leq i \leq W, 1 \leq j \leq H, 1 \leq f \leq F, 1 \leq v \leq N)$

where W and H are the width and height of a frame, F is the number of frames in a video and N is the total number of videos.

The average brightness B^v of video v can be present in Equ. 1.

$$B^v = \frac{\sum_{k=1}^{F}\sum_{i=1}^{W}\sum_{j=1}^{H} I^{v}_{ijk}}{W \times H \times F} \tag{1}$$

So the average brightness of each frame in all the videos B_{avg} can be presented by Equ. 2.

$$B_{avg} = \frac{\sum_{v=1}^{N} B^v}{N} \tag{2}$$

For a new testing video I^{t}_{ijf}, value of the pixels I^{New}_{ijf} is adapted using Equ. 3.

$$I^{New}_{ijf} = \frac{I^{t}_{ijf} \times B_{avg}}{B^t} \tag{3}$$

3.3 CNN based feature extraction

CNN models have been used for human face recognition and FER with very good accuracy. However, many of them are too large on size and too slow on computing. Therefore, instead of using more accurate CNN like VGG19 [9] and ResNet [22], a simple pre-trained AlexNet [12] is chosen as a feature extraction method.

The pre-trained AlexNet has totally 25 layers, including 5 convolution layers, 5 max pooling layers, 3 fully connected layers, 2 dropout layers, 1 softmax layer and 8 Rectified Linear Units. The fine-tuning is made by a dataset includes 20 subjects and each image has been automatically reshaped to 227 × 227 as AlexNet required in MATLAB.

3.4 Post-processing

Human emotion is not changed suddenly within one second, so a slide window with one second length is used to smooth the predictions. Therefor, a majority voting process has been used after all the regression has been made by SVR. Playing the film segments takes anywhere from dozens of seconds to a few minutes. Within this window, all the regression results are averaged together to produce one prediction for each emotion. The highest value of three emotions will be considered as the dominate emotion of the user that is used for the film flow control. There are three regression emotion, 'anger', 'fear' and 'calm' in a video segment length as L seconds and prediction frequency g, which have three SVR scores present as Ra_i, Rf_i and Rc_i respectively, where $i = 1, 2, \ldots, L \times g$. So the final predict label for this video segment can be presented as equation 4.

$$\max(\sum_{i=1}^{L\times g} Ra_i, \sum_{i=1}^{L\times g} Rf_i, \sum_{i=1}^{L\times g} Rc_i) \tag{4}$$

4 Evaluation

4.1 Data collection and emotion annotation

For this particular application, there is no public available database available for training the machine learning system. So a database is recorded and annotated. In this interactive film, the users are asked to watch the film. A camera on the top of the screen is set up to record the their face while watching the film. These are spontaneous emotions on the faces as they are the natural expressions. The 4 segments of the film last from a few seconds to a few minutes. After that, the users are asked to draw their level of three emotions: anger, fear and calm separately during the time they watch the film segments. The level of emotions are between 0 to 1, 0 represents low level of this emotion and 1 represents high level of this emotion. Then, these labels are used as emotion ground-truth for regression.

A database has then been created with three emotion labels: calm, fear and anger. For each emotion, a scale value between [0,1] was recorded by the subjects who completed the recording. The labels are levels of these three emotions, numbered from 0 to 1 representing from no fear/anger/calm to extreme fear/anger/calm.

Table 1. Details of the recorded database.

Number of subject	20
Number of segments per subject	4
Average frame per segment	207
Average time per segment	23 seconds
Type of emotion	3 (fear, anger, calm)
Scale range of emotion label	[0,1]

The detailed information of the database is shown in Table 1. There are total 20 subjects, including all races, cultures and ages from youths to old people, with 12 male and 8 female subjects. Each subject is asked to watch 4 film segments and each video lasts for 12 seconds to 36 seconds unequally. The details can be see in Table 1. In actual testing circumstances, the emotions of testing subjects are detected using the same 4 film segments. These 4 film segments are selected as the most radical and essential in the film. If the subjects keep calm in all 4 film segments, the film will be kept playing with a good ending. If the subjects display a high level of fear or anger, the film will take a different route with a bad ending.

4.2 Experimental results

A 10-fold cross-validation method is used for subject independent evaluation. In each fold, the data from 18 subjects are used for training and the data of rest 2 subjects are used for testing. The first predictions are made according to the predictions from all the frames in one second. Then for each segment, all the prediction values on 3 emotions are added and the biggest one is used to predict the label of that segment. Then the classification accuracy were calculated on the 3 emotion labels.

For comparison purpose, Edge Orientation Histogram (EOH) [6] features are also extracted replacing CNN features in the whole system. EOH has been proved to be an effective traditional feature extraction method on FER, like depression recognition in Meng et al. [16] and Jan et al. [11]. The overall results are shown in Table 2. From this Table, it can be seen that CNN is better than EOH feature. The best classification accuracy reaches 80.3%.

5 Conclusion

In this paper, a specific designed emotion recognition system is proposed and implemented based on CNN, SVR, pre-processing and post-processing methods.

Table 2. Comparison on classification accuracy (%)

Feature	Brightness control	Accuracy	Frames per second
EOH	No	58.01	42
EOH	Yes	61.44	40
CNN	Yes	**80.30**	9

It has been trained and tested on a small database with satisfied performance. It was then integrated into interactive film RIOT prototype and demonstrated in several occasions with good success. Although it was designed under specific requirements of the interactive film for demonstration purpose under various environmental conditions, the proposed emotion detection system can be run as a stand alone software and it has the potential to be applied in many real-world applications.

Acknowledgement

We'd like to thank Karen Palmer, for the design and coordination of the RIOT prototype. We'd also like to thank The National Theatre Immersive Storytelling Studio, The Perception Institute and Crossover Labs for the support.

References

1. Humaid Alshamsi, Veton Kepuska, and Hongying Meng. Automated facial expression recognition app development on smart phones using cloud computing. In *Proceedings of 8th IEEE Annual Conference on Ubiquitous Computing, Electronics and Mobile Communication (UEMCON), 2017*, pages 577–583.
2. Humaid Alshamsi, Hongying Meng, and Maozhen Li. Real time facial expression recognition app development on mobile phones. In *Processing on 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1750–1755.
3. Keith Bound. Future of filmmaking: Artificial intelligent and biofeedback interactive movies. https://www.linkedin.com/pulse/future-filmmaking-artificial-intelligent-biofeedback-keith. Website.
4. Gary R Bradski. Real time face and object tracking as a component of a perceptual user interface. In *Proceedings of Fourth IEEE Workshop on Applications of Computer Vision, 1998. WACV'98.*, pages 214–219.
5. Soong-Der Chen and Abd Rahman Ramli. Minimum mean brightness error bi-histogram equalization in contrast enhancement. *IEEE transactions on Consumer Electronics*, 49(4):1310–1319, 2003.
6. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Processing on IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. (CVPR 2005).*, volume 1, pages 886–893.
7. Pratik Gala, Raj Shah, Vineet Shah, Yash Shah, and Mrs Sarika Rane. Moodyplayer: A music player based on facial expression recognition. *International Research Journal of Engineering and Technology (IRJET)*, 5(4):3703–3707, 2018.

8. Mariana-Iuliana Georgescu, Radu Tudor Ionescu, and Marius Popescu. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836, 2019.
9. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
10. Rebecca Hills-Duty. darksides. https://www.vrfocus.com/2017/10/mindtree-pictures-developing-interactive-vr-movies/. Website.
11. Asim Jan, Hongying Meng, Yona Falinie A Gaus, Fan Zhang, and Saeed Turabzadeh. Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings on the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM, 2014.
12. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
13. Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.
14. Yuanyuan Liu, Xiaohui Yuan, Xi Gong, Zhong Xie, Fang Fang, and Zhongwen Luo. Conditional convolution neural network enhanced random forest for facial expression recognition. *Pattern Recognition*, 84:251–261, 2018.
15. André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.
16. Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30. ACM, 2013.
17. Shervin Minaee and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv preprint arXiv:1902.01019*, 2019.
18. Shigeo Morishima. Real-time talking head driven by voice and its application to communication and entertainment. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*, 1998.
19. Ryohei Nakatsu, Joy Nicholson, and Naoko Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 343–351. ACM, 1999.
20. Karen Palmer. RIOT - KAREN PALMER. http://karenpalmer.uk/portfolio/riot/. Website.
21. Marie-Laure Ryan. *Narrative as virtual reality: Immersion and interactivity in literature and electronic media.* Johns Hopkins University Press, 2001.
22. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, volume 4, page 12, 2017.
23. Saeed Turabzadeh, Hongying Meng, Rafiq M Swash, Matus Pleva, and Jozef Juhar. Facial expression emotion detection for real-time embedded systems. *Technologies*, 6(1):17, 2018.