# CNN Features for Emotional Impact of Movies Task

Yun Yi[1,2], Hanli Wang[2,*], Qinyu Li[2,3]
[1]Key Laboratory of Jiangxi Province for Numerical Simulation and Emulation Techniques, Gannan Normal University, Ganzhou 341000, P. R. China
[2]Department of Computer Science and Technology, Tongji University, Shanghai 201804, P. R. China
[3]Department of Computer Science, Lanzhou City University, Lanzhou 730070, P. R. China

## ABSTRACT

A framework is proposed to predict the emotional impact of movies by using the audio, action, object and scene features. First, four state-of-the-art features are extracted from four pre-trained convolutional neural networks to depict video contents, and an early fusion strategy is used to combine vectors of these features. Then, the linear support vector regression or linear support vector machine is employed to separately learn affective models or fear models, and the strategy of cross-validation is utilized to select training parameters. Finally, the Gaussian blur function is used to smooth scores of video segments. The experiments show that the combination of these features obtains promising results.

## 1 INTRODUCTION

The 2018 emotional impact of movies task consists of two subtasks, including the valence-arousal prediction and the fear prediction. A brief introduction about this challenge has been given in [1]. This paper mainly introduces the proposed framework and discusses the experimental results.

The selection of features is crucial to emotional analysis. Intuitively, the audio, action, object and scene features can influence emotions. Therefore, vectors of four state-of-the-art features are calculated in this framework. Then, the affective models or fear models are learned by using linear support vector regression (SVR) or linear support vector machine (SVM) [2]. Finally, the function of Gaussian blur is utilized to smooth scores of temporal segments.

## 2 FRAMEWORK

Figure 1 shows the key components of the proposed framework, and the highlights of our framework are introduced below.

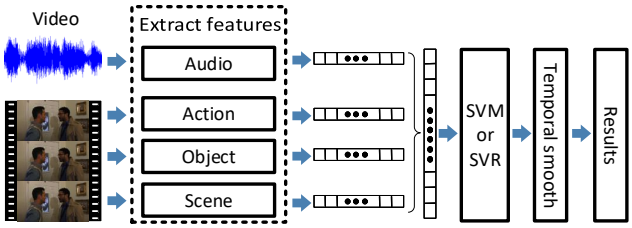


**Figure 1: An overview of the proposed framework.**

### 2.1 Features

To depict a video, four features are separately extracted from four pre-trained Convolutional Neural Networks (CNNs), including audio, action, object and scene features.

*2.1.1 Audio Feature.* The audio signals are important information that describes emotions. VGGish [4] is a famous audio feature extractor, so it is used to calculate the vectors of audio feature. First, the audio files are extracted from videos. Then, the pre-trained model[1] provided by [4] is utilized to calculate the feature vectors of audio files. Therefore, the audio signals are converted into semantically meaningful high-level 128-dimensional feature vectors by VGGish. In conclusion, for the audio feature, a video is described as a sequence of 128-dimensional vectors.

*2.1.2 Action Feature.* The actions in the video can influence viewer's emotions. The two-stream Convolutional Networks (ConvNet) [6] is a well-known framework for video-based action recognition, and includes the spatial ConvNet and the temporal ConvNet. The temporal segment network [8] builds the model of long-range temporal structure to improve this framework, and Inception-v3 [7] is the basic network architecture of the two ConvNets. The pre-trained models provided by [8] are utilized to calculate the vectors from the 'top_cls_global_pool' layer. As a result, a frame is described by two 1024-dimensional vectors. By connecting the two vectors of a frame, a video is depicted as a sequence of 2048-dimensional vectors.

*2.1.3 Object Feature.* The objects in the video may affect emotions of the viewer. The Squeeze-and-Excitation Network (SENet) [5] is the state-of-the-art model for object

*Hanli Wang is the corresponding author, E-mail: hanliwang@tongji.edu.cn.

This work was supported in part by the National Natural Science Foundation of China under Grants 61622115 and Grant 61472281, Shanghai Engineering Research Center of Industrial Vision Perception & Intelligent Computing (17DZ2251600), and IBM Shared University Research Awards Program.


*MediaEval'18, 29-31 October 2018, Sophia Antipolis, France*

[1]https://github.com/tensorflow/models/tree/master/research/audioset

classification. We utilize the pre-trained SENet model[2] to calculate the vectors from the 'pool5/7 × 7_s1' layer. Therefore, the dimension of object features is 2048.

*2.1.4 Scene Feature.* The scenes of the video affect the emotions of the audience. The Places365 dataset is a large dataset for scene classification [9]. We utilize the pre-trained ResNet-50 [3] model[3] to calculate the vectors from the 'avg-pool' layer. So a frame is depicted by a 2048-dimensional vector.

## 2.2 Emotional Prediction

To combine vectors of these features, we utilize the early fusion strategy because of its simplicity and efficiency. As shown in Fig. 1, we directly connect vectors of these features for each sample.

For different subtasks, the linear SVR and the linear SVM are used to learn the emotional models, separately. The number of positive samples is less than that of the negative samples in the fear subtask. To solve this problem, we weight positive and negative samples in an inverse manner. The regularization parameter $C$ is set by the strategy of cross-validation. The LIBLINEAR toolbox[4] is used to implement the L2-regularized L2-loss SVM and SVR.

After obtaining the scores of video segments, we use the function of Gaussian blur to smooth these scores. Let the score vector of a video be $V$. Then, the Gaussian blur function is defined as

$$\text{Gaussianblur}(V) = V \otimes K,$$

where $\otimes$ is the convolution operator, $K$ is the specified Gaussian kernel. In experiments, we set the size of Gaussian kernel to 11 for the valence-arousal subtask and 5 for the fear subtask.

# 3 RESULT AND DISCUSSION

In order to evaluate the aforementioned features described in Section 2.1, the features provided by the task organizers are selected as the baseline features. As required in the task, we submit five runs for each of the two subtasks. Table 1 shows the features used in these runs.

**Table 1: Features used in five runs.**

| Runs | Features |
|---|---|
| Run 1 | features provided by the task organizers |
| Run 2 | audio and scene features |
| Run 3 | audio, scene and object features |
| Run 4 | audio, scene and action features |
| Run 5 | audio, scene, action and object features |

For the sake of fair comparison, the five runs utilize the same framework except the features used. Regarding the learning algorithm, SVR is employed in the valence-arousal subtask, and SVM is used in the fear subtask. The Mean Square Error (MSE) and Pearson Correlation Coefficient (PCC) are reported for the valence-arousal subtask, and the Intersection over Union (IoU) of time intervals is considered as the evaluation metric for the fear subtask [1]. The results are given in Table 2 and Table 3

**Table 2: Results of the valence-arousal subtask.**

| Runs | Valence | | Arousal | |
|---|---|---|---|---|
| | MSE | PCC | MSE | PCC |
| Run 1 | 0.09142 | 0.27518 | 0.14634 | 0.11571 |
| Run 2 | **0.09038** | **0.30084** | **0.13598** | 0.15546 |
| Run 3 | 0.09163 | 0.26326 | 0.14056 | 0.14310 |
| Run 4 | 0.09105 | 0.25668 | 0.13624 | **0.17486** |
| Run 5 | 0.09243 | 0.24679 | 0.13950 | 0.15226 |

**Table 3: Results of the fear subtask.**

| Runs | IoU of time intervals |
|---|---|
| Run 1 | 0.14360 |
| Run 2 | 0.12900 |
| Run 3 | 0.13067 |
| Run 4 | **0.15750** |
| Run 5 | 0.14969 |

As shown in Table 2, Run 2 obtains the best result in the valence-arousal subtask. This suggests that the combination of audio feature and scene feature is sufficient to predict valence-arousal values. In the fear subtask, Run 4 achieves the top performance as shown in Table 3. This demonstrates that the combination of audio, scene and action features is enough to describe fear, and that the method using more features does not necessarily lead to better experimental results. By comparing the results of Run 2 and Run 3 in Table 2 and Table 3, the usage of the object feature improves the performance in the fear subtask, but it decreases the performance in the valence-arousal subtasks. This may be due to the reason that some objects can cause people's fears, such as blood, guns, etc. In Table 3, Run 4 obtains better performances than Run 3. This partly demonstrates that actions are more likely to cause fear than objects.

# 4 CONCLUSION

In this work, we propose a framework to predict the emotional impact of movies. Vectors of four features are calculated by using four pre-trained convolutional neural networks. The affective models or fear models are separately learned by using SVR or SVM, and the function of Gaussian blur is utilized to smooth the temporal scores. Experimental results show that the combination of audio feature and scene feature is enough in the valence-arousal subtask, and that additional action feature improve the performance in the fear subtask.

[2] https://github.com/hujie-frank/SENet
[3] https://github.com/CSAILVision/places365
[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/multicore-liblinear

## REFERENCES


[1] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, Zhongzhe Xiao, and Mats Sjöberg. 2018. The MediaEval 2018 emotional impact of movies task. In *MediaEval 2018 Workshop*.

[2] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[4] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *ICASSP*. 131–135.

[5] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*. 7132–7141.

[6] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 568–576.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*. 2818–2826.

[8] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*. 20–36.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Transactions Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452 – 1464.