Visual and audio analysis of movies video for emotion detection @ Emotional Impact of Movies task MediaEval 2018

Elissavet Batziou[1], Emmanouil Michail[1], Konstantinos Avgerinakis[1], Stefanos Vrochidis[1], Ioannis Patras[2], Ioannis Kompatsiaris[1]
[1]Information Technologies Institute, Centre for Research and Technology Hellas
[2]Queen Mary University of London
batziou.el@iti.gr,michem@iti.gr,koafgeri@iti.gr
stefanos@iti.gr,i.patras@qmul.ac.uk,ikom@iti.gr

ABSTRACT

This work reports the methodology that CERTH-ITI team developed so as to recognize the emotional impact that movies have to its viewers in terms of valence/arousal and fear. More Specifically, deep convolutional neural newtworks and several machine learning techniques are utilized to extract visual features and classify them based on the predicted model, while audio features are also taken into account in the fear scenario, leading to highly accurate recognition rates.

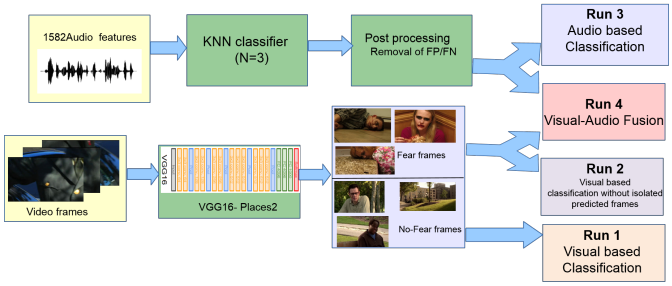

Figure 1: Block diagram of our approach for fear recognition

1 INTRODUCTION

Emotion based content have a large number of applications, including emotion-based personalized content delivery[2], video indexing[7], summarization[5] and protection of children from potentially harmful video content. Another intriguing trend that appears to get a lot of attention lately is style transferring and more specifically recognizing the emotion from some painting or some specific section from a movie and transferring its affect to the viewer as a style to a novel creation.

Emotional Impact of Movies Task is a challenge of MediaEval 2018 that comprises of two subtasks: (a) Valence/Arousal prediction and (b) Fear prediction from movies. The Task provides a great amount of movies video, their visual and audio features and also their annotations[1].Both subtasks ask from the participants to leverage any available technology, in order to determine when and whether fear scenes occur and to estimate a valence-arousal score for each video frame in the provided test data [3].

In this work, CERTH-ITI introduces its algorithms for valence/arousal and fear recognition subtasks, which include the deployment of deep learning and other classification schemes to recognize the desired outcome. More specifically, a 3-layer neural network(NN) and a simple linear regression model are deployed, with and without PCA, so as to predict the correct emotion in the valence-arousal subtask, while a pre-trained $VGG16$ model [6] is combined with a K Nearest Neighbors (KNN)- classification scheme, so as to leverage the visual and audio attributes respectively and identify the correct boundary video frames in the fear subtask.

Copyright held by the owner/author(s).
MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

2 APPROACH

2.1 Valence-Arousal Subtask

In the valence-arousal recognition subtask, keyframe extraction is initially applied so as to extract one video frame per second and correlate them with the annotations that were provided from MediaEval emotion organizers, who has also used the same time interval to record human extracted groundtruth data. The provided visual features are then concatenated into one vector representation so as to have a common and fixed representation scheme throughout different video samples.

The first recognition approach that was deployed concerns the valence/arousal estimation by adopting a linear regression model. Linear regression try to minimize the residual sum of squares between the groundtruth and predicted responses by using linear approximation (Run 3). PCA is also deployed on our final visual features vectors so as to reduce their dimensionality and keep only the most discriminant principal components (in our case the first 2000) to represent all features (Run 4).

A Neural Network (NN) framework has also been deployed so as to fulfil the valence/arousal recognition substask. For that purposes, a 3-hidden layer NN with ReLU activation function and Adam optimizer with learning $rate = 0.001$ was deployed. The size of each hidden layer is 64, 32 and 32 respectively. We use batch size equal to 10 and 10 epochs. The size of the training set is 2/3 of the development set and the remaining 1/3 for validation set. The input of the NN is the set of vectors of concatenated visual features(Run 3). PCA has also been used in order to downsample the concatenated highly dimensional size (5367) in the golden section of 2000 principal components(Run 4).

2.2 Fear Subtask

For the fear recognition subtask, we initially keyframe extraction every one second, as we perform in valance subtask. The frames annotated as "fear" were significantly less than the "no-fear" class and, therefore, in order to balance our dataset we used data augmentation techniques. Firstly, we downloaded from Flickr about 10,000 images with tag "fear" and we also download emotion images [1] and kept those which are annotated as "fear". In order to further increase the number of fear frames, we additionally use data augmentation techniques on the provided annotated frames. We randomly rotate and translate pictures vertically or horizontally and we randomly apply shearing transformations, randomly zooming inside pictures, flipping half of the images horizontally and filling in newly created pixels which can appear after a rotation or a width/height shift. Finally, we reduce the set of no-fear frames. After these, we had about 23,000 "fear" and 30,000 tagged as "no fear" images to train our model. We used transfer learning to gain information from a large scale dataset and also trained our model in a very realistic and efficient time. The architecture that we chose to represent our features is the *VGG*16 pre-trained on Places2 dataset [8] because the majority of the movies have places as background and so we assume that it would be helpful. We use Nadam optimizer with learning rate 0.0001. The batch size is 32 and the number of epochs 50. Finally, we set a threshold of 0.4 on their probability (Run 1). In a different approach, we used the same architecture without isolated predicted frames (Run2).

Additionally, in order to exploit auditory information, we developed a classification method applied on audio features already extracted from the challenge committee using openSmile toolbox [4]. Audio feature vectors, consisting of 1582 features, extracted from videos every second, were separated into training (80%) and validation set (20%). In order to equalize the size of the two classes in the training set we randomly removed "no-fear" samples. We apply KNN classification method with N=3 on the test set, results were further processed, in order to remove erroneous false negatives (single "no-fear" samples around "fear" areas) and false positives (isolated small "fear" areas consisting of one or two "fear" samples).

Results from visual and audio analysis were submitted both separately, as different runs, and in combination by taking the post probabilities of visual and auditory classifications and setting a threshold of 0.7 on their average probability. The overall block diagram of this approach is depicted in Figure 1.

3 RESULTS AND ANALYSIS

We have submitted 4 runs for valence/arousal prediction and their results are introduced in Table 1. In the experiments two evaluation measures are used: (a) Mean Square Error (MSE) and (b) Pearson Correlation Coefficient (r). We observe that the *NN* approach that we describe in the previous section has the best performance amongst all the others. Furthermore, it is worth mentioning that the linear regression model produces some extremely high scores, probably because the original feature vectors weren't neither discriminative nor adequate enough to create the regression model. However, PCA projection to lower dimensional space, with higher discriminative power show to solve this problem as it reduces the redundant noise and keep the most important features. Moreover, there is a "NaN" score for the Pearson measure in the arousal prediction scores, because we accidentally set the training value stable and so our model predicts the same score for all frames, but this score does not characterize our model, since it does not appear in any other prediction within the valence/arousal prediction sub-task.

[1] http://www.imageemotion.org/

Table 1: CERTH-ITI Predictions

	Valence		Arousal		Fear
Run	MSE	r	MSE	r	IoU
1	396901706.564	0.079	1678218552.19	**0.054**	**0.075**
2	0.139	0.010	0.181	-0.022	0.065
3	**0.117**	**0.098**	**0.138**	nan	0.053
4	0.142	0.067	0.187	-0.029	0.063

We have also submitted 4 runs for fear prediction subtask and their results are also presented in Table 1 and are evaluated in terms of Intersection over Union (IoU). From Table 1 we see that, the best performance for the fear recognition subtask is Run 1, using all predicted scores of the pre-trained VGG16 model. In addition, our intuition to remove isolated predicted frames (Run 2), as they are not associated with any duration, did not perform better than Run 1, hence we miss significant information (video frames that invoke fear).

4 DISCUSSION AND OUTLOOK

In this paper we report the CERTH-ITI team approach to the MediaEval 2018 Challenge "Emotional Impact of Movies" task. The results in valence/arousal prediction subtask shows that according to MSE, the best result obtained in Run 3 for both valence and arousal, while regarding to Pearson Correlation Coefficient, Run 1 has the best performance for arousal and the second best performance for valence. The Pearson correlation is able to measure linear correlations between two or more variables. However, the MSE is obtained by a sum of squared deviations between predicted and ground-truth values, no matter if they are linearly correlated or not.

The results of the fear prediction subtask shows that the inclusion of audio features failed to enhance the classification performance, as expected. This could be due to several reasons, with the prominent one to be the incapability of performing data augmentation on audio features such as in the case of visual analysis. Both the aforementioned reason and the large inequality between the two classes, which led us discard many "no-fear" annotations, in order to balance the training set, resulted on a very limited training set. These drawbacks could be overcome by using classification methods able to handle unbalanced training sets, such as penalized models, or by enriching the training set with external annotated datasets and by exploring more efficient fusion methods, such as performing classification on fused audiovisual features, instead of a posterior combining separate classification results.

ACKNOWLEDGMENTS

This work was funded by the EC-funded project V4Design under the contract number H2020-779962

REFERENCES

[1] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[2] Luca Canini, Sergio Benini, and Riccardo Leonardi. 2013. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology* 23, 4 (2013), 636–647.

[3] 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 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2018 Workshop*.

[4] Florian Eyben, Felix Weninger, Florian Gross, and BjÎŞÂűrn Schuller. 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*. 835–838.

[5] Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Chua Tat-Seng. 2011. Affective video summarization and story board generation using pupillary dilation and eye gaze. In *Multimedia (ISM), 2011 IEEE International Symposium on*. IEEE, 319–326.

[6] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[7] Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao, and Qi Tian. 2010. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia* 12, 6 (2010), 510–522.

[8] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2018), 1452–1464.