# BOUN-NKU in MediaEval 2017 Emotional Impact of Movies Task

Nihan Karslioglu[1], Yasemin Timar[1], Albert Ali Salah[1],
Heysem Kaya[2]
[1]Boğaziçi University, İstanbul, Turkey, {nihan.karslioglu, yasemin.timar, salah}@boun.edu.tr
[2]Namık Kemal University, Tekirdağ, Turkey, hkaya@nku.edu.tr

## ABSTRACT

In this paper, we present our approach for the Emotional Impact of Movies task of Mediaeval 2017 Challenge, involving multimodal fusion for predicting arousal and valence for movie clips. In our system, we have two pipelines. In the first one, we extracted audio/visual features, and used a combination of PCA, Fisher vector encoding, feature selection, and extreme learning machine classifiers. In the second one, we focused on the classifiers, rather than on feature selection.

## 1 INTRODUCTION

The challenge we tackle in this paper is the prediction of affective content of video clips, denoted by valence and arousal scores. We used well-known regression models on the audio-visual domain for this purpose. The feature sets extracted by the organizers have been used to form a baseline system to understand the properties and relations of the most important features for this task. The description of the task is provided in [1].

One of the proposed tasks is the prediction of "fear", which is represented by a binary value in the ground truth. However, the sections denoted with fear are rare (only 5%); and this requires classifiers capable of dealing with class imbalance (e.g. Gradient Boosting Classifier). We have not worked on this part of the challenge.

The Emotional Impact of Movies task has been included in the MediaEval challenges since 2015. Various approaches have been studied for the problem in terms of features and regression models in recent years [2]. Audio features, visual descriptors and deep learning based features have been popular among the participants of the 2016 challenge [3].

## 2 FIRST APPROACH

Our first pipeline, given in Fig.1, extracts a number of features, reduces their dimension with PCA, summarizes them with Fisher vector encoding, and further applies a feature selection stage prior to classification.

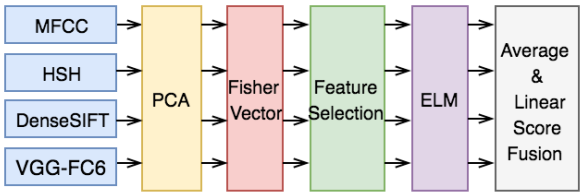


**Figure 1: First system for valence-arousal prediction.**

As audio features, we computed *Mel-frequency Cepstral Coefficients (MFCC 0-12),* from 32ms windows (with 50% overlap). First and second derivatives were added, resulting in a 39-dimensional feature vector.

We used three types of visual features in addition to these audio features. The *Hue Saturation Histogram (HSH)* feature is a 1023-dimensional histogram of color pixels, in 33 hue and 31 saturation levels. They were sampled from one frame per second, and frames were resized to 240x320. For the *Dense SIFT* feature, the frames were further resized to 120x160, and Dense SIFT features [4] were extracted at scales {4,6,8}, at 7 pixel intervals and once for every 30 frames of video. Finally, we used the *VGG FC6* feature provided by the organizers, extracted from a deep neural network trained for image recognition**.**

After reducing the dimensionality of the features by 50% via PCA, we encoded them with Fisher vectors (FV) [5], which measures how much the features deviate from a background probability model, in this case a mixture of Gaussians. The number of clusters were selected as 32 for Dense SIFT and MFCC, and a single Gaussian was used for HSH and VGG-FC6. We normalized the feature vectors with signed square root and L2 normalization.

A ranking based feature selection approach was applied using Random Sample versus Labels Canonical Correlation Analysis Filter (SLCCA-Rand) method [6]. The main idea is to apply CCA between features and target labels, then sort the absolute value of the projection weights to get a ranking. Features that sum up to 99% of the total weight for each modality are selected in this approach.

For regression, Extreme Learning Machines (ELM) were applied for both arousal and valence prediction tasks [7]. Grid


*MediaEval'17*, 13-15 September 2017, Dublin, Ireland

search is applied to find the best parameters of ELM. Regularization coefficient was searched from the range of [0.01,1000] with exponential steps. Radial basis function (RBF) and linear kernels were tested. The RBF kernel scale parameter is optimized in the range of [0.01,1000], also with exponential steps. Pearson Correlation Coefficient (PCC) is taken as performance measure, and optimized over 5-fold cross validation on the development partition. Results in Table 1 are obtained on the test set, for which the ground truth was sequestered.

## 3 SECOND APPROACH

Our second approach used audio and visual features presented by the organizers, without any dimensionality reduction. Dimensionalities are 1.582 for audio, and 1.271 for visual features, respectively. Early fusion of the visual features (except FC6) are fed to Random Forest and support vector regressors (SVR). Hyper parameters are explored with grid search. For SVR, the cost and gamma parameters range from 0.001 to 100. For Random Forests, the number of trees range from 100 to 1000, and the maximum number of features per tree from 3 to 20. Five train and test folds (balanced according to duration and fear labels) are defined to ensure that each movie appears in either in the train set or the test set. The best regressors were chosen via grid search, and tested on each fold to evaluate the performance on a subset of the development set. According to MSE and PCC scores on each fold, the regressors are trained with the best group. The audio and visual subsystem scores are fused with simple averaging, and the scores for a given movie are smoothed with Holt-Winters exponentially weighted moving average method [8]. The pipeline is visually presented in Figure 2.

## 4 RESULTS AND ANALYSIS

We submitted five runs for the valence/arousal prediction task. The first run is the average scores of MFCC, HSH, Dense SIFT and VGG-FC6 subsystems, and obtains our lowest MSE on the valence task.

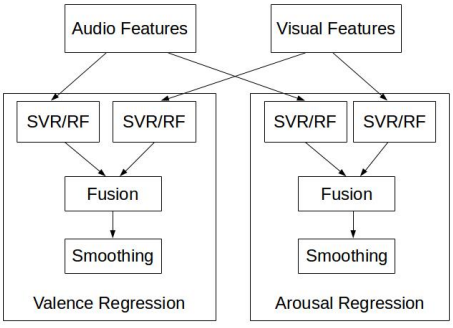


**Figure 2: Pipeline without dimensionality reduction.**

The second run is a linear weighted combination of the predictions used in the first run. In the third run, while an average of MFCC and FC6 are computed for valence, the average of MFCC, HSH and FC6 are computed for arousal. In the fourth run, linear combination scores of MFCC, Dense SIFT and FC6 are computed for valence, and linear combination scores of MFCC, HSH and FC6 are computed for arousal. For the fifth run, the regression pipelines are selected after grid search, resulting in four separate SVRs with RBF kernels (with best scoring hyper parameters from cross-validation). AV scores of test-set data are fused and smoothed to generate the run outputs.

**Table 1: Arousal/Valence Prediction Results (MSE)**

| Run | Approach | Arousal | | Valence | |
|---|---|---|---|---|---|
| | | MSE | PCC | MSE | PCC |
| 1 | 1, simple avg. | 0.1231 | 0.1289 | **0.1859** | 0.0263 |
| 2 | 1, weighted comb. | 0.1433 | 0.0986 | 0.2249 | 0.0464 |
| 3 | 1, selective avg. | 0.1237 | 0.1046 | 0.1889 | 0.0386 |
| 4 | 1, linear selective comb. | 0.1434 | 0.0990 | 0.2251 | 0.0460 |
| 5 | 2, smoothed | **0.1126** | **0.2186** | 0.1881 | **0.0904** |

When we compare Run1 and Run2 from Table 1, we can say that combining all features from the first approach with simple average fusion method is better than combining them with weighted fusion technique for arousal task but this situation is opposite for valence in terms of PCC. Comparing Run1 with Run3 and Run4, Dense Sift is important for better arousal prediction in PCC metric. Run1 shows that fusing all features from the first approach with simple averaging method gives the best MSE result for valence. The best results are obtained in Run5 for arousal prediction for two metrics. PCC result of Run5 for valence is also the best result between the other runs. We also observe that prediction of arousal is more accurate compared to valence.

The computation power of our computer is limited in terms of time and memory. Therefore we plan to choose more components for higher explained variance for PCA and more clusters for GMM for the first 4 runs in our future works. In addition to this, we plan to employ CCA to extract arousal and valence correlates as mid-level features, so as to optimize PCC and MSE measures simultaneously.

## ACKNOWLEDGMENTS

This work is supported by Bogazici University Project BAP 16A01P4 and by the BAGEP Award of the Science Academy.

## REFERENCES

[1] E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, M. Sjöberg. "The MediaEval 2017 Emotional Impact of Movies Task," *Proc. of the MediaEval 2017 Workshop*, Dublin, Ireland, Sept. 13-15, 2017.

[2] Y. Baveye, E. Dellandrea, C. Chamaret, L. Chen, “Deep Learning vs. Kernel Methods: Performance for Emotion Prediction in Videos,” In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.

[3] E. Dellandrea, L. Chen, Y. Baveye, M. Sjöberg, C. Chamaret, “The Mediaeval 2016 Emotional Impact of Movies Task”, In *MediaEval 2016 Workshop*, 2016.

[4] A. Bosch, A. Zisserman and X. Munoz, “Image classification using random forests and ferns,” In *IEEE 11th International Conference on Computer Vision, (ICCV 2007),* 2007.

[5] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” In *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[6] H. Kaya, T. Özkaptan, A.A. Salah and F. Gürgen, “Random discriminative projection based feature selection with application to conflict recognition,” *IEEE Signal Processing Letters, 22(6)*, pp. 671-675, 2015.

[7] G.B. Huang, H. Zhou, X. Ding and R. Zhang, “Extreme learning machine for regression and multiclass classification”. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(2), pp. 513-529, 2012.

[8] P. R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science, 6(3), pp. 324–342, 1960. doi:10.1287/mnsc.6.3.324*.