Feature Selection and Multimodal Fusion for Estimating Emotions Evoked by Movie Clips

Yasemin Timar
Computer Eng. Dept., Bogazici University
Istanbul, Turkey
yasemin.timar@boun.edu.tr

Nihan Karslioglu
Computer Eng. Dept., Bogazici University
Istanbul, Turkey
nihan.karslioglu@boun.edu.tr

Heysem Kaya
Computer Eng. Dept., Namik Kemal University
Corlu, Tekirdag, Turkey
hkaya@nku.edu.tr

Albert Ali Salah
Computer Eng. Dept., Bogazici University
Istanbul, Turkey
Future Value Creation Research Center, Nagoya University
Nagoya, Japan
salah@boun.edu.tr

ABSTRACT

Perceptual understanding of media content has many applications, including content-based retrieval, marketing, content optimization, psychological assessment, and affect-based learning. In this paper, we model audio visual features extracted from videos via machine learning approaches to estimate the affective responses of the viewers. We use the LIRIS-ACCEDE dataset and the MediaEval 2017 Challenge setting to evaluate the proposed methods. This dataset is composed of movies of professional or amateur origin, annotated with viewers' arousal, valence, and fear scores. We extract a number of audio features, such as Mel-frequency Cepstral Coefficients, and visual features, such as dense SIFT, hue-saturation histogram, VGG16 FC6. We reduce feature dimensionality with PCA, summarize them via Fisher vector encoding and further apply a feature selection stage prior to classification with Extreme Learning Machine classifiers. In a post-processing stage, the predictions of individual models are fused. Moreover, some statistical feature summarization methods are applied to these features as well as facial action units in the first approach. For fear problem, weighted ELM classifier is applied. We contrast this approach with SVM and Random Forest regressors. On these classifiers, different fusion and smoothing strategies are assessed. We demonstrate the benefit of feature selection and multimodal fusion on estimating affective responses to movie segments.

CCS CONCEPTS

• **Information systems → Content analysis and feature selection**; **Sentiment analysis**; **Multimedia and multimodal retrieval**;

KEYWORDS

Affective computing, multimodal interaction, emotion estimation, audio-visual features, movie analysis, face analysis, extreme learning machine

ACM Reference Format:
Yasemin Timar, Nihan Karslioglu, Heysem Kaya, and Albert Ali Salah. 2018. Feature Selection and Multimodal Fusion for Estimating Emotions Evoked by Movie Clips. In *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR'2018).* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.475/123_4

ICMR'2018, June 2018, Yokohama, Japan

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Media content including text, images, audio, and videos in various platforms have been growing at an amazing speed with the recent mobile technologies. Research on multimedia summarization, annotation, indexing and retrieval, suggestion and event detection are subsequently prioritized [13]. Similar to multimedia studies, affective video content analysis focuses on perceptual understanding of emotions of viewers to effectively be applied on emotion based media content delivery, summarization, and protection of younger viewers from harmful content. Emotional engagement of the viewer is driven by audio visual media content with powerful presentation and composition techniques. Methods of computer vision, techniques of machine learning and cognitive research are executed to explore and measure the perceptual affect of the multimedia content. In our study, we have implemented various pipelines for the prediction of affective content of video clips denoted by valence, arousal and fear scores in the MediaEval LIRIS-ACCEDE dataset [5]. We used well-known regression models and a classification model on the audio-visual domain for this purpose. The feature sets extracted available in the dataset have also been used to form a baseline system to understand the properties and relations of the most important features for prediction.

This work explains details of the two pipelines we have implemented while developing emotion estimators. Our first pipeline 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. As audio features, we computed Mel-frequency Cepstral Coefficients and we used three

types of visual features in addition to these audio features; The Hue Saturation Histogram, Dense SIFT and VGG16 FC6 feature. In this approach, Extreme Learning Machines (ELM) and score fusions were applied for both arousal and valence prediction tasks. We have previously obtained very good results with Extreme Learning Machine classifiers on features extracted by deep neural networks, and won two multimodal ChaLearn Challenges at ICPR'16 and CVPR'17. In the first approach, experiments on statistical methods for feature summarization are also performed for facial action units with the features used in this approach. For fear problem, weighted ELM [46] is used as classifier. Our second approach uses audio and visual features without any dimensionality reduction, and adds low level scene features and facial features for emotion estimation. Early fusion of the visual features are fed to a Random Forest classifier and to Support Vector Regressors, whose hyper parameters are explored with grid search. 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. We have used some part of this work in the MediaEval 2017 "Emotional Impact of Movies Task" challenge while participating as team BOUN-NKU [24].

2 RELATED WORK

Affective analysis studies focus on the stimuli that produces certain emotions on the audience or the viewer. There exists many aspects to process the affective properties of multimedia content. Affective computing framework has been a popular subject since Picard has defined the first frameworks for computers to gain the ability to recognize, understand, even to have and express emotions [32].

Affective video content analysis and multimedia studies on content indexing and retrieval involve describing, storing, and organizing multimedia information and assisting users to conveniently and quickly look up multimedia resources [28]. In general, five main procedures are defined in multimedia indexing and retrieval: structural analysis, feature extraction, data mining, classification and annotation, query and retrieval [21]. Structural analysis aims to segment a video into several semantic structural elements, including shot boundary detection, key frame extraction, and scene segmentation. A number of works have investigated quality assessment of images and videos [3, 18, 30], as well as of paintings and photographs [2, 7, 16, 29, 41].

The Affective Impact of Movies Task was a part of the Medieval Benchmarking Initiative in 2015, and "Emotional Impact of Movies Task" is a part of the Medieval 2016 and 2017 challenges [17, 34]. The overall use case scenario of the task is defined as to design a video search system that uses automatic tools to help users find videos that fit their particular mood, age, or preferences. Audio descriptors and visual features are extracted from the videos with several tools. Support Vector Regression and Random Forest models, Least Square Boosting, Moving Average Smoothing, CNN and LSTM networks have been studied in the challenges by the participants. Dense SIFT, improved trajectories (IDT), histogram of gradients (HOG), histogram of optical flow (HOF) and motion boundary histograms (MBH) feature descriptors with Fisher Vectors, which are successful in activity recognition, are also applied in emotion estimation to model the motion in the movie clips to estimate the response of the viewer [27, 42]. Mel-frequency Cepstral Coefficients (MFCC) are popular choices as audio features used by the teams.

Recent studies on affective movie analysis include cinematographic features with machine learning and pattern recognitions techniques executed on specific fiction movies [6, 10–12]. Canini et al. argue that since the distance between the camera and the subject greatly affects the narrative power of the shot, investigating the characteristics of the shot type would contain indirect information about the distance. In their work on classification of cinematographic shot types [12], 2D scene geometric composition, frame color intensity properties, motion distribution, spectral amplitude and shot content are considered for classifying shots, using C4.5 Decision Trees [33] and Support Vector Machines (SVM) [14]. The features defined for the scene/shot are (1) color intensity distribution on local regions in the frames, (2) motion activity maps [43] for motion distribution, (3) the geometry the scene, through the measurement of the angular aperture of perspective lines found by Hough transform, (4) detecting subjects in the frame by their faces and face dimensions with the well-known Viola-Jones algorithm [38] and (5) the spectral amplitude of the scene and its decay.

According to Smith [35], if a film structure aims to elicit an emotion, the priority of movie makers should be to create a mood. Although mood is not as strong as emotion, it is long-lasting. Being in a mood stimulates our emotion system again and again. As an example, a fearful mood alerts our fear system and drives us to perceive dark shadows and scary objects. This mood orients the viewer towards a particular frightening feeling in the future, and continues until fear emerges [35].

Filmic cues providing emotional information can help to elicit mood or strong emotions. Some of them are facial expressions, dialogue, vocal expression and tone, costume, sound, music, lighting, mise-en-scene, set design, editing, and camera usage [36]. Using a single filmic cue is not enough to arouse a mood. Emotion systems of audiences are very diverse and a single cue can be received by some viewers, but other viewers may miss it. For this reason, redundant filmic cues are used in films and that helps to increase chances of reaching the targeted emotion for many viewers [36]. Accordingly in this work, we have investigated multiple cues, and their fusion for estimating the affective content for movies.

3 DATASET

The main goal for Baveye et al., who proposed the LIRIS-ACCEDE dataset, is to produce ground truth data for training and benchmarking computational models for machine-based emotion understanding [5]. This dataset, including the video clips, valence/arousal and fear annotations, features and protocols, is publicly available at: http://liris-accede.ec-lyon.fr/. It includes 30 movies, from which consecutive ten seconds-segments sliding over the whole movie with a shift of 5 seconds are considered and provided with annotations. The audio features available in the dataset have been extracted using the openSmile toolbox with "emobase2010.conf" configuration [19]. 1582 features are computed, which result from a base of 34 low-level descriptors (LLD) with 34 corresponding delta coefficients appended, and 21 functionals applied to each of these 68 LLD contours (1428 features). In addition, 19 functionals are applied to the 4 pitch-based LLD and their four delta coefficient contours

(152 features). Finally the number of pitch onsets (pseudo syllables) and the total duration of the input are appended (2 features).

The visual features extracted for each of the 30 movies, one frame per second. For each of the frames, several general purpose visual features are provided using the LIRE library (http://www.lire-project.net/). Convolutional neural network (CNN) features (VGG16 FC6 layer output) that have been extracted using the Matlab Neural Networks toolbox. The visual features are the following: Auto Color Correlogram, Color and Edge Directivity Descriptor, Color Layout, Edge Histogram, Fuzzy Color and Texture Histogram, Gabor, Joint descriptor joining CEDD and FCTH in one histogram, Scalable Color, Tamura, Local Binary Patterns, VGG16 FC6 layer, respectively.

As we examine the annotations of arousal, valence and fear of 30 movies clips, we can clearly see that most of the fear annotations exists when there is a increase in arousal and decrease in valence. Still we know that fear is located in negative valence, positive arousal in the Circumplex of Affect. On the other hand, some of the fear annotations are contrary to this assumption (e.g. the movie "Norm or Full Service"). It is important to observe the trend in the valence arousal scores, which gives the most import clue about fear. However, fear annotations are labeled by different annotators whose emotional responses are very subjective and different. Subsequently, it is challenging to produce emotion estimators that will consistently and accurately predict the viewer response.

4 SYSTEM DEVELOPMENT

We contrast two different, but complementary approaches to the problem of affect estimation from movies. We describe them in two separate subsections here.

4.1 First approach: Audio-Visual Features with Dimensionality Reduction

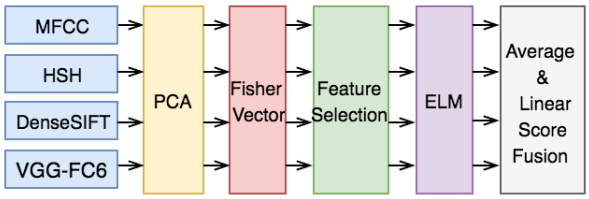


Figure 1: Audiovisual pipeline in the first approach.

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. 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 [8] were extracted at scales {4,6,8}, at 5 pixel intervals and once for every 30 frames of video. Finally, we used the VGG16 FC6 feature available in the dataset, which is extracted from a deep neural network trained for image recognition. We normalized the features with signed square root and L2 normalization. After reducing their dimensionality by 50% via PCA, we encoded them with Fisher vectors (FV) [31], which measures how much the features deviate from a background probability model, in this case a mixture of Gaussians. In other words, FV encoding quantifies the amount of parameter shift needed to best fit the new coming data in the probability model. The number of GMM components were selected as 32 for DenseSIFT and MFCC, and a single Gaussian was used for HSH and VGG16 FC6.

A ranking based feature selection approach was applied using Random Sample versus Labels Canonical Correlation Analysis Filter (SLCCA-Rand) method [25]. 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 [22]. Grid 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.

Other experiment applied for arousal and valence problem is normalizing features with min-max, then summarizing features with their mean, geometric mean, standard variation, maximum and minimum. PCA is used to reduce the dimension of summarization of features with 95% explained variance. Min-max normalization is again applied to output of PCA before applying ELM as similiarly as above experiment. Facial action units extracted using OpenFace [4] are applied to this experiment as a different feature. Another similar approach is applied to the problem of fear with the same features. Weighted ELM[46] is used as a classifier to solve the problem of imbalanced number of classes in fear problem. Results obtained for these problems are shown in Table 1 and 2 for test set.

4.1.1 ELM and Weighted ELM Learners. The "extreme" learning paradigm is based on two fast stages that are also meant to overcome over-fitting. The first is random generation of the hidden node output matrix $\mathbf{H} \in \mathbb{R}^{N \times h}$, where N and h denote the number of instances and the hidden neurons, respectively. This is done via random generation of the first layer weights and the bias vector. While the first layer weights are unsupervised and not tuned for the task at hand, the second layer weights $\beta \in \mathbb{R}^{h \times L}$ are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$, where $\mathbf{T} \in \mathbb{R}^{N \times L}$ is the label matrix and L is the number of classes. $\mathbf{T}$ is a real valued vector $\in \mathbb{R}^{N \times 1}$ in case of regression. In the case of L-class classification, $\mathbf{T}$ is represented in one vs. all coding:

$$\mathbf{T}_{t,l} = \begin{cases} +1 & \text{if } y^t = l, \\ -1 & \text{if } y^t \neq l. \end{cases} \quad (1)$$

The extreme learning rule is generalized to use any kernel $\mathbf{K}$ with a regularization parameter C, without generating $\mathbf{H}$ [22], relating ELM to LSSVM [37]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}, \tag{2}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. Weighted ELM [46] introduces a diagonal weight matrix $\mathbf{W_{t,t}} = \mathbf{1}/(\mathbf{N_l})$, where $y^t = l$ and N_l represents the number of training set instances having class label l:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{WK})^{-1}\mathbf{WT}. \tag{3}$$

Since total weight is equal for each class, models optimize average recall, counter-balancing the under-represented classes. The weighted KELM version is particularly employed in the Fear task.

4.2 Second Approach: Audio-Visual Feature Performances in Emotion Estimation

Our second approach used audio and visual features presented by the organizers, as well as extracted face and low level features without any dimensionality reduction. Dimensions are 1.582 for audio, and 1.271 for visual features. Additionally, we have extracted low level features and face geometric features, which have different pipelines before fusion. 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. Additionally, deep features, which are the FC6 layer from the VGG19 network out of frames sampled with 1 second apart.

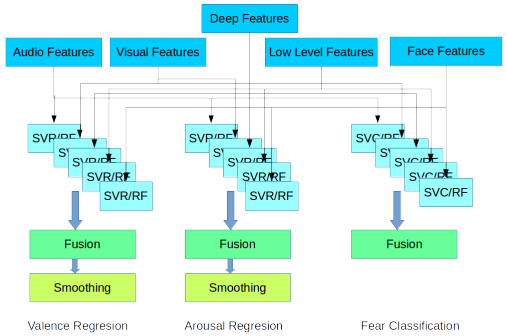


Figure 2: Valence Arousal Regression and Fear Classification.

4.2.1 Faces of Characters. The most important features in a scene is the human presence which is related to main characters in the movie. The physical properties seen on the screen of the characters are perceived by the viewer to understand the dynamics of the scene in general sense. Face of the character is the most important clue, while the viewer has intrinsic knowledge with human body proportions with the surrounding. The Viola-Jones face detection algorithm has been used in many application since its first published [38]. On the other hand, with the advances in convolutional neural networks (CNN), new approaches trained with millions of face images are giving successful results. One of them is the dlib [26] state of the art face recognition system, built with deep learning, which was reported to have an accuracy of 99.38% on the Labeled Faces in the Wild benchmark. We have utilized the face geometrics to derive scene features such as the shot scale, as well as for the basic emotional impact of the character. A closer shot of the character signifies stronger affect in the scene, while the relative dimensions of the faces in a group can suggest relations between the characters.

4.2.2 Low Level Features of Scenes. Various color spaces are studied in computer vision research. However hue-saturation-value HSV space is very popular in the domain of human visual perception modeling. In general terms, *hue* refers to the color itself, *saturation* represents the boldness of the color, and *value* gives the brightness level. As an example pastel blue would be less saturated than a very bold blue. Saturated colors are more preferred by humans over non-saturated colors [9]. According to recent studies, since 1940s when the color film became the standard, saturation in films has been steadily increasing. Hue, which tends to be the more easily identifiable color dimension, also plays a significant part in our narrative understanding. The average hue of the whole movie may not be particularly useful, but the average hue of a shot can be very important [9].

Luminance is a measurement of how much light is present in an image or a series of images. Luminance is controlled not only during shooting but also in post production by manipulating the contrast and exposure of the film. Well-known concepts of "low-key" and "high-key" in the photography are also used in movies. The histogram of the key frames image can be used as the features of a shot. However, as Brunick et al. state, most films are composed of slighter luminance changes. So a luminance indicator of the movie can be estimated by first calculating the median luminance for each frame of the film and then averaging across the entire film. We propose the same approach to estimate the luminance of each shot by transforming the key frame to gray scale and calculating the median value.

Within a movie, two basic on-screen activities are defined; (1) *motion* refers to the actions of an agent in front of the camera, and (2) *camera movement* refers to results of change in camera position or lens length, like pans, tilts and zooms. However, people generally do not consciously realize the distinction between these two activities when processing visual information [9]. Cutting et al. proposed to combine both motion end movement into a *visual activity* descriptor [15]. We propose to measure the visual activity by utilizing background models and background subtraction. One of the commonly used approaches for this is applying probability density functions and estimators for the variance of each pixel in the scene model. In the early works, one Gaussian model is used per pixel [40], but later more complex models, like Gaussian mixture models, improved the background subtraction [20, 44]. The method we have used is the Gaussian Mixture-based Background/Foreground Segmentation Algorithm, implementing the improved Gaussian mixture model background subtraction proposed by Zivkovic et al. [44, 45]. In this model, the scene has a history of frames which affect the background model. And there is a threshold on the squared Mahalanobis distance between the scene

pixel and the model to decide whether a pixel is well described by the background model. We have set the history to 500 (frames). We assume the background model generated at the end of each shot includes all the visual activity occurred during the shot. We calculate the mean intensity to summarize this information to be used as a low level feature of the scene.

4.2.3 Regressors. Support Vector Regression models are widely used in affective content analysis. SVR models construct a hyperplane by mapping vectors from an input space into a high dimensional feature space such that they fall within a specified distance of the hyperplane. C parameter value is critical; when it is too large, there exists a high penalty for non-separable points and many support vectors will overfit. When it is too small, we may have underfitting [1]. The behavior of the model is also sensitive to the γ parameter. If γ is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. We use grid search to optimize these parameters.

The second approach we use is the Random Forest, which is a widely used ensemble learning method in affective computing. A random forest fits a number of decision trees on various subsamples of input data and uses averaging to improve the predictive accuracy and to control over-fitting.

In our approach, we use early fusion of visual features (except VGG FC6), which are fed to Random Forest classifiers and support vector regressors (SVR). CNN (i.e. VGG16 FC6) features are fed to separate regressors. Hyper parameters are explored with grid search. In our pipeline evaluation with grid search, for SVR, standard scaling is used and the C and γ 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 validation set, but not in both. 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 [39]. The pipeline is visually presented in Fig. 2.

5 RESULTS AND DISCUSSIONS

To quantify the performance of the two pipelines, mean-squared error (MSE) and Pearson's r are computed. The MSE for regression models is widely used to quantify the difference between estimated values and the true values estimated. It measures the amount by which the estimated values differ from the ground truth and assesses the quality of the regression in terms of its variation and degree of bias, while the Pearson product-moment correlation coefficient (or Pearson's r) is a measure of the linear correlation between estimated and true values. The measurements on the test set provided by the organizers are used to evaluate the pipelines separately. The results of the first approach are presented in Table 1. The first run is the average scores of MFCC, HSH, Dense SIFT and VGG FC6 subsystems. 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. Results using statistical feature summarization methods of action units, Dense SIFT, HSH, FC6 and MFCC are also in Table 1.

For the arousal task, action units have the best PCC results in the first approach. While Dense SIFT, FC6, HSH and MFCC have very low results for PCC according to Table 1, their average score fusion is promising for the arousal task. For the valence task, dense SIFT seems to work best alone. Score fusion of MFCC, HSH, Dense SIFT, FC6 provide better results.

The Tables 3 and 4 present the results of the test conducted with the second approach. They are obtained on the sequestered test set of Mediaeval 2017 Emotion Task challenge. Bar charts in Figs. 5 and 6 visualize the results in the tables.

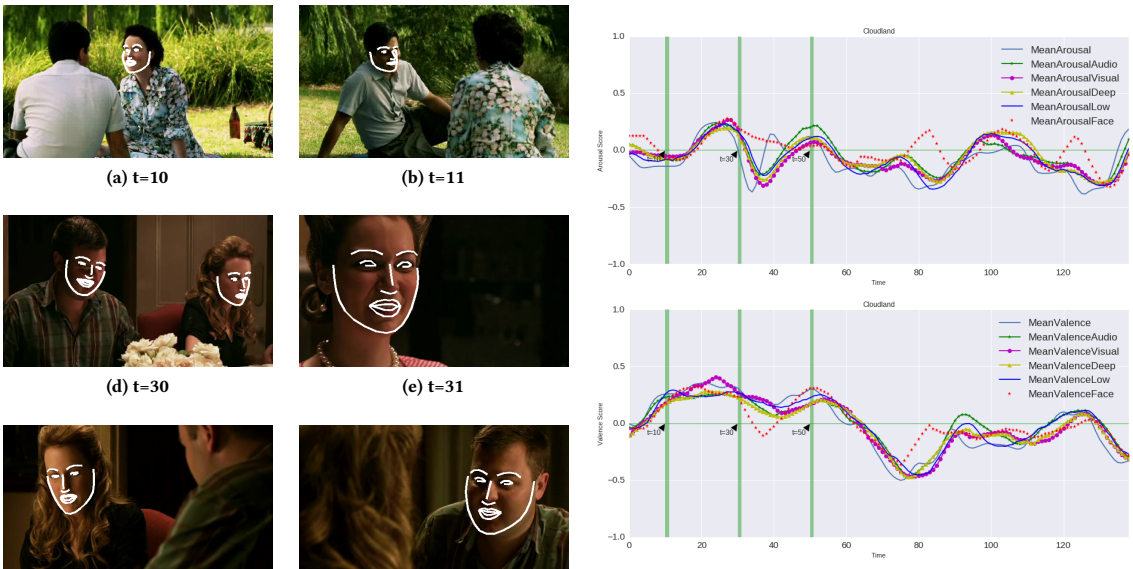
Regarding the valence predictions in Table 3, the visual and VGG FC6 features are more successful, while fused values produce an average MSE and Pearson's r. Still, low level features are not quite successful as the visual features. Regarding the arousal results in Table 4, we see that the fused results are better than individual visual and audio features. On the other hand, smoothing is making the results of low level features and audio features inaccurate.

Face geometrics are not quite successful compared to the other features. One of the reasons is that the face is not being recognized in most of the shots. For example, in the movie "Island" there is a human in most of the frames, but since the character is shot from the back, we can not see her face most of the time. As a result, human detection fails in those frames and this creates an inaccuracy.

Table 2 shows the performance of individual features for the fear classification task. This is a very challenging task, with a very unbalanced training and test set. The F1 scores are low on the overall, with MFCC achieving the best F1 with a small margin.

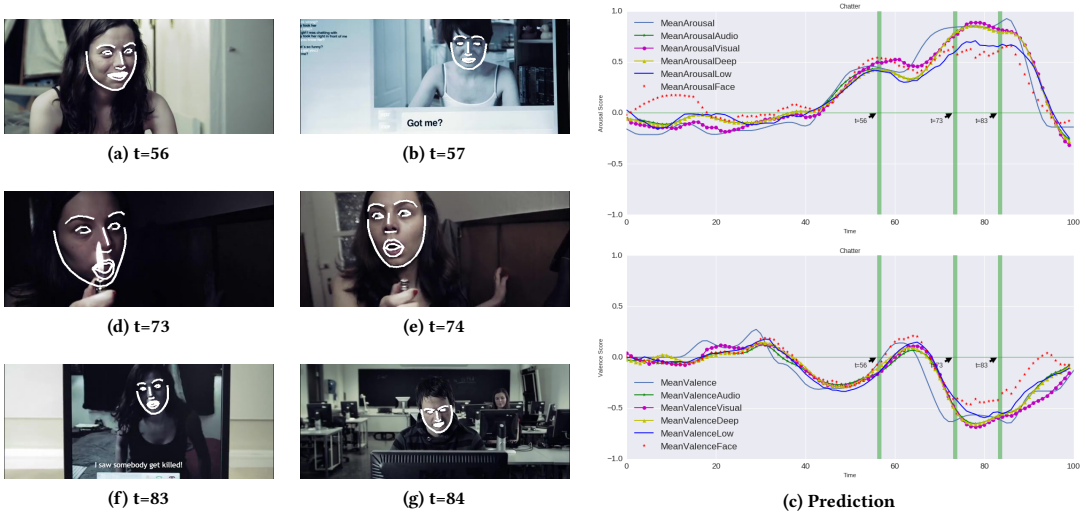
To understand the effects of various features, we have visualized the prediction scores. We have displayed the emotion predictions for two movies in Figs. 4 and 3, namely "Cloudland" and "Chatter" respectively. We have selected sample scenes with human presence and annotated the corresponding approximate time in the prediction curve. As one can see, visual features and CNN features with SVR regression have the lowest error. Face features changes more rapidly compared to the low level features such as color, lighting, and motion of the scene. This suggests that there should be a more elaborate approach to detect humans in the scene.

The valence and arousal predictions of our models are very similar to the results of the Mediaeval 2017 Emotion Impact challenge. Among the models of the participants, the system with the best PCC for valence prediction used VGG features with SVR regressors, while the best MSE scores were produced with fusion of custom audio features, VGG and visual features fed to 2-layer LSTM networks [23]. We have produced the best MSE scores in arousal with audio visual regression and late fusion. On the other hand, we have obtained the best PCC scores using VGG features with SVR. The variance of MSE scores in the challenge is low, despite the fact that PCC scores vary noticeably, mostly because the cost functions



(a) t=10 (b) t=11 (c) Prediction (d) t=30 (e) t=31 (f) t=50 (g) t=51

Figure 3: "Cloudland" faces in dual-frames t=10, 30, 50



(a) t=56 (b) t=57 (c) Prediction (d) t=73 (e) t=74 (f) t=83 (g) t=84

Figure 4: "Chatter" faces in dual-frames t=56, 73, 83

of the regressors favor minimizing the error over maximizing the correlation.

6 CONCLUSIONS

In this work, we have investigated various approaches to predict viewer emotions while watching movies. In the first approach, MFCC and various visual features, different feature summarization methods, score fusions and single feature runs are performed for arousal and valence task. From these experiments, single representation can sometimes work better as action units but it can be also said that it doesn't give a good result for defining emotion of a movie but fusion of features can have better results. This conclusion may have confirmed that filmic emotional clues when used alone will not elicit expected emotion by movie makers on viewers but the expected emotion will be achieved when used together. As a future work, experiments will be performed with feature based fusions and using different filmic clues including emotion information to be able to get better results.

As it is shown in our second approach, grid search enables models with a high performance with traditional regressors, without

Table 1: Results of the first approach

Run	Arousal MSE	Arousal PCC	Valence MSE	Valence PCC
avg: MFCC, HSH, Dense SIFT and FC6	**0.123**	**0.129**	**0.186**	0.026
weighted comb.: MFCC, HSH, Dense SIFT and FC6	0.143	0.099	0.225	**0.046**
selective avg: MFCC and FC6 (valence) + HSH (arousal)	**0.123**	0.105	**0.189**	0.039
linear selective comb.: MFCC, FC6, Dense SIFT (valence),HSH (arousal)	0.143	0.099	0.225	**0.046**
Stats + PCA (Statistical Feature Summarization with PCA): Action Units	0.165	**0.230**	0.208	-0.280
Stats + PCA: Dense SIFT	0.255	0.052	0.247	-0.028
Stats + PCA: FC6	0.176	0.026	0.209	**0.055**
Stats + PCA: HSH	0.173	-0.075	0.228	-0.005
Stats + PCA: MFCC	0.166	-0.006	0.221	-0.010

Table 2: Fear task classification results.

Feature	Accuracy	Precision	Recall	F1 Score
MFCC	0.539	0.161	0.340	0.192
Action Units	0.505	0.150	0.351	0.187
Dense SIFT	0.590	0.152	0.255	0.166
HSH	0.707	0.121	0.093	0.090
FC6	0.836	0.017	0.014	0.015

Table 3: Valence pipeline evaluations on challenge test set

Features	Regressors	Smoothing	MSE	PCC
Visual	SVR	yes	0.181	0.107
FC6	SVR	yes	0.183	0.339
FC6	SVR	no	0.185	0.207
Visual	SVR	no	0.185	0.061
avg Audio-Visual	SVR	yes	0.188	0.090
Face	RF	yes	0.206	-0.058
Face	RF	no	0.213	0.010
Low Level	RF	no	0.240	-0.061
Low Level	RF	yes	0.242	-0.223
Audio	SVR	yes	0.243	-0.025
Audio	SVR	no	0.245	0.013

Table 4: Arousal pipeline evaluations on challenge test set

Features	Regressors	Smoothing	MSE	PCC
avg Audio-Visual	SVR	yes	0.113	0.219
FC6	SVR	yes	0.126	0.150
FC6	SVR	no	0.128	0.073
Low Level	RF	yes	0.132	-0.027
Face	RF	yes	0.136	0.031
Visual	SVR	yes	0.136	0.122
Low Level	RF	no	0.139	-0.004
Visual	SVR	no	0.140	0.070
Face	RF	no	0.142	-0.017
Audio	SVR	no	0.147	0.018
Audio	SVR	no	0.152	-0.044

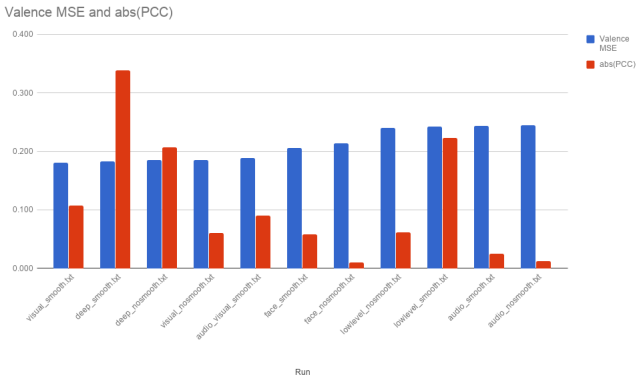


Figure 5: Valence results

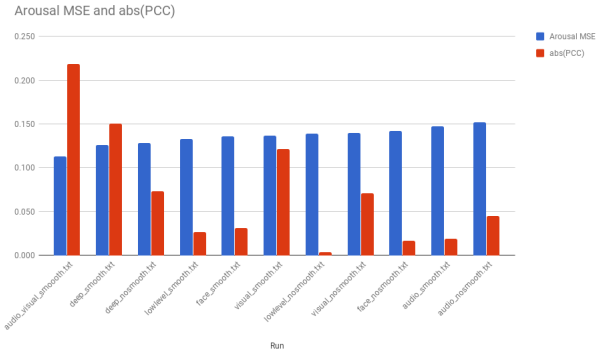


Figure 6: Arousal results

any further need for feature selection. Additional visual cinematographic features and face features produce promising results, indicating room for improvement through stylistic features. Post-processing is also very helpful; the positive effect of smoothing is clearly observed in the arousal results. In fact, the best performance scores for valence are produced by CNN features with SVR models and arousal prediction is generally better than valence prediction, which is consistent with the literature and MediaEval Emotional

Impact challenge results. The high subjectivity of emotions is the main challenge, for which we will continue to develop more accurate emotion estimators for movie viewers. In our future works, we plan to analyze the features of speech, music and sound effects separately in the audio domain. Each modality has a different effect on the viewer, which we hope to observe in the dataset.

REFERENCES

[1] Ethem Alpaydin. 2014. *Introduction to machine learning.* MIT Press.

[2] Seyed Ali Amirshahi, Michael Koch, Joachim Denzler, and Christoph Redies. 2012. PHOG analysis of self-similarity in aesthetic images. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 82911J–82911J.

[3] Seyed Ali Amirshahi and Mohamed-Chaker Larabi. 2011. Spatial-temporal video quality metric based on an estimation of QoE. In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*. IEEE, 84–89.

[4] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 6. IEEE, 1–6.

[5] 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.

[6] Sergio Benini, Luca Canini, and Riccardo Leonardi. 2010. Estimating cinematographic scene depth in movie shots. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 855–860.

[7] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the international conference on Multimedia*. ACM, 271–280.

[8] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 1–8.

[9] KL Brunick, JE Cutting, and JE DeLong. 2013. Low-level features of film: What they are and why we would be lost without them. *Psychocinematics: Exploring cognition at the movies* (2013), 133–148.

[10] Luca Canini, Sergio Benini, and Riccardo Leonardi. 2011. Affective analysis on patterns of shot types in movies. In *Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on*. IEEE, 253–258.

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

[12] Luca Canini, Sergio Benini, and Riccardo Leonardi. 2013. Classifying cinematographic shot types. *Multimedia tools and applications* 62, 1 (2013), 51–73.

[13] Min Chen, Shiwen Mao, and Yunhao Liu. 2014. Big data: a survey. *Mobile Networks and Applications* 19, 2 (2014), 171–209.

[14] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[15] James E Cutting, Kaitlin L Brunick, Jordan E DeLong, Catalina Iricinschi, and Ayse Candan. 2011. Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception* 2, 6 (2011), 569–576.

[16] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV*. Springer, 288–301.

[17] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Viktor Sjöberg, Christel Chamaret, et al. 2016. The mediaeval 2016 emotional impact of movies task. In *MediaEval 2016 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2016 Workshop*.

[18] Ahinet M Eskicioglu. 2000. Quality measurement for monochrome compressed images in the past 25 years. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vol. 6. IEEE, 1907–1910.

[19] Florian Eyben and Björn Schuller. 2015. openSMILE:): the Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records* 6, 4 (2015), 4–13.

[20] Nir Friedman and Stuart Russell. 1997. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 175–181.

[21] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 797–819.

[22] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. 2012. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, 2 (2012), 513–529.

[23] Zitong Jin, Yuqi Yao, Ye Ma, and Mingxing Xu. [n. d.]. THUHCSI in MediaEval 2017 Emotional Impact of Movies Task. ([n. d.]).

[24] Nihan Karslioglu, Yasemin Timar, Albert Ali Salah, and Heysem Kaya. [n. d.]. BOUN-NKU in MediaEval 2017 Emotional Impact of Movies Task. ([n. d.]).

[25] Heysem Kaya, Tuğçe Özkaptan, Albert Ali Salah, and Fikret Gürgen. 2015. Random discriminative projection based feature selection with application to conflict recognition. *IEEE Signal Processing Letters* 22, 6 (2015), 671–675.

[26] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[27] Vu Lam, Sang Phan Le, Duy-Dinh Le, Shin'ichi Satoh, and Duc Anh Duong. 2015. NII-UIT at MediaEval 2015 Affective Impact of Movies Task.. In *MediaEval*.

[28] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 2, 1 (2006), 1–19.

[29] Congcong Li and Tsuhan Chen. 2009. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE Journal of* 3, 2 (2009), 236–252.

[30] Yiwen Luo and Xiaoou Tang. 2008. Photo and video quality evaluation: Focusing on the subject. In *Computer Vision–ECCV*. Springer, 386–399.

[31] Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 1–8.

[32] Rosalind W Picard and Roalind Picard. 1997. *Affective computing*. Vol. 252. MIT press Cambridge.

[33] J Ross Quinlan. 2014. *C4.5: programs for machine learning*. Elsevier.

[34] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. 2015. The MediaEval 2015 Affective Impact of Movies Task.. In *MediaEval*.

[35] Greg M Smith. 1999. Local emotions, global moods, and film structure. *Passionate views: Film, cognition, and emotion* (1999), 103–26.

[36] Greg M Smith. 2003. *Film structure and the emotion system*. Cambridge University Press.

[37] Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9, 3 (1999), 293–300.

[38] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Vision and Pattern Recognition*. IEEE.

[39] Peter R Winters. 1960. Forecasting sales by exponentially weighted moving averages. *Management science* 6, 3 (1960), 324–342.

[40] Christopher Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. 1997. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 7 (1997), 780–785.

[41] Shao-Fu Xue, Qian Lin, Daniel R Tretter, Seungyon Lee, Zygmunt Pizlo, and Jan Allebach. 2012. Investigation of the role of aesthetics in differentiating between photographs taken by amateur and professional photographers. In *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 83020D–83020D.

[42] Yun Yi, Hanli Wang, Bowen Zhang, and Jian Yu. 2015. MIC-TJU in MediaEval 2015 Affective Impact of Movies Task.. In *MediaEval*.

[43] Wei Zeng, Wen Gao, and Debin Zhao. 2002. Video indexing by motion activity maps. In *Int. Conf. on Image Processing*, Vol. 1. IEEE, I–912.

[44] Zoran Zivkovic. 2004. Improved adaptive Gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, Vol. 2. IEEE, 28–31.

[45] Zoran Zivkovic and Ferdinand van der Heijden. 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters* 27, 7 (2006), 773–780.

[46] Weiwei Zong, Guang-Bin Huang, and Yiqiang Chen. 2013. Weighted extreme learning machine for imbalance learning. *Neurocomputing* 101 (2013), 229–242.