

AUTH-SGP in MediaEval 2016 Emotional Impact of Movies Task

Timoleon Anastasia

Hadjileontiadis Leontios

School of Electrical Computer & Engineering, Aristotle University of Thessaloniki, Greece
{timoanas,leontios}@auth.gr

ABSTRACT

This paper presents all the aspects expected for the MediaEval Workshop. The tested and adopted solutions are well described and the interest of using a set of features versus another one is discussed. The conclusion follows state-of-the-art findings and allows bringing new inputs in the understanding of emotion prediction.

1. INTRODUCTION

Recent years videos have been the main medium for many people to interact with each other and share information. So, there is a further need to evaluate the quality of this interaction in terms of emotions, not only to analyze the video-content. To serve this purpose, video affective content analysis has gained interest among researchers[12]. Many audio-visual video features can be found useful to depict emotion. For example, imagine a film where the background is full of warm colors. This can induce the viewers to have positive emotions, namely emotions with high valence values. Motion is another important film element that can control a video's emotion. Films with large motion intensity can cause stronger emotions, where the arousal score is higher. This task aims exactly at predicting the emotional feedback of the users while watching different genres of films[6].

2. SYSTEM DESCRIPTION

2.1 Feature Extraction

The key points of our system can be summarized to the followings: first we extract multi-modal features that can successfully represent emotion. These can be either local features, from specific patches of the video frames or from overlapping time windows of sound signals, or directly global features from the entire image[9]. In the first case, a feature encoding technique must be applied, in order to convert these local features to global. We examined the Bag-Of-Words and Fisher Vector approaches[9]. Finally, the extracted features are regressed and/or combined in order to predict the emotion scores.

2.1.1 Development-data feature

These features were provided by the organizers of the task. A great variety of features were given, including diverse fea-

tures from the audio signals of the movies, features regarding the scene cuts and much more. These features were used almost directly, the only preprocessing step included the normalization of them, by subtracting the mean value and dividing with the standard deviation of each column.

2.1.2 Improved Dense Trajectories (IDT)

This kind of features provide information about the movement of the videos and are calculated in different spatial and temporal scales[11]. They are extensively used to classify human actions. We resized the original videos to 320x240. Then, several descriptors were calculated for each trajectory (length of 15 frames), including Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram along x and y axes (MBHx and MBHy). The total number of descriptors for each trajectory is 426 (30+96+108+96+96)[1].

For the conversion of the local features into global, the Fisher Vector approach was used. A Gaussian Mixture Model (GMM) was employed to construct a codebook with k words for each descriptor ($k = 64$). A total of 2500000 points were sampled from the descriptors of the development-train set to train the GMM. The features of each descriptor are then individually projected via PCA to the half of their dimensions, resulting in 213 dimensions for each trajectory, and encoded using the Fisher Kernel method. The power and L2-normalization schemes were applied to each descriptor and to the resulting vectors, which hopefully can improve the performance of the system. Finally, an entire video can be described by a vector of 27264 features (=2 [mean value, standard deviation due to the Gaussian model]*213 [features]*64[codebook size]).

2.1.3 Deep Learning Feature

Deep learning is a modern sub-field of computer vision and machine learning, which uses artificial neural networks combined with the principles of convolution in images, to describe pictures using more abstract and high-level features. We used the famous BVLC Caffe deep learning framework, and treated the BVLC Reference CaffeNet pre-trained model as a feature extractor[8]. In particular, this network contains 5 convolutional layers, 2 fully connected layers and a soft-max classifier. We extracted features from the last fully-connected layer which outputs 4096 neurons.

The input frames were the keyframes from the 10-seconds videos, size 256x256[5]. Instead of averaging the results over the 10 random-crops that the network produces for each image, the 4096 output activations of each one of the 10 crops were kept, resulting in 10x4096 feature representations for

each video. Then, the classic Bag-of-Works concept was used to encode these features. The size of the codebook was 8, and the BOWKMeansTrainer class from OpenCV[7] was used to find the clusters. Each video was finally represented by a 8-bin normalized histogram of the frequency of appearance of each codeword. These features were added to the development-data features to explore, whether the performance is actually improved with their presentation.

2.1.4 Dense SIFT Feature

The SIFT descriptor was used on the re-scaled videos. One common approach when we are dealing with videos is, to densely compute SIFT features along neighbors of pixels in images, with specific stride step (counted in pixels) and specific frame step. In our approach, the neighborhood size is 10×10 , and a new SIFT descriptor is calculated every 5 pixels and every 5 frames[10]. After the extraction of the dense SIFT feature, PCA is applied to reduce the dimension of the descriptor from 128 to 64. Finally, the fisher vector is applied, in a similar manner to the IDT approach.

2.1.5 Hue Saturation Histogram (HSH)

As mentioned above, different colors can depict different genres of emotions. We converted the frames from RGB to Hue Saturation Value (HSV) space and then computed a two-dimensional histogram keeping only the hue and saturation channels. The number of hue bins were 15, while the number of saturation bins were 16. A HSH was calculated every 5 frames, exactly like the Dense SIFT descriptor. Finally, PCA and fisher vector approaches were applied.

2.1.6 Audio Feature

We used the Mel Frequency Cepstral Coefficients (MFCC) as representative audio feature [3]. Each video can be described by three different types of MFCCs. The first type is the short-term descriptor, where the input audio signal is divided into overlapping windows of size 32ms (and overlap 50%) and then a cepstral representation is computed for each one of them. The other two types of descriptors are the mean and standard deviation of the above-mentioned features, resulting in a 39-dimensional (3×13) vector. Finally, PCA dimension reduction and encoding with fisher vector were employed.

2.2 Regression

As far as regression is concerned, the Support Vector Regression (SVR)[4] is employed in this project. For each task, a grid search cross-validation scheme was used, in order to determine the best hyper-parameters C , γ and the type of kernel for each model. We investigated radial basis function and linear kernels, while C and γ were in the range $[0.01, 10]$ and $[0.001, 1]$ respectively. The objective function to be maximized was the Pearson-Correlation Coefficient between predicted and real output values. The cross-validation scheme we followed was simple k-fold validation with $k=5$. The distribution of different types of movie genres in each set (train and validation) was not taken into account, although it is a good alternative future direction.

3. RESULTS AND DISCUSSION

1st sub-task. We submitted a total of 5 runs for the first sub-task only. The first run was only with the presence of the already-extracted features from the development-data. The

second run was the combination of the above features with the deep-learning ones. The features were concatenated horizontally and then regressed. The third run includes only the features from the improved dense trajectories. The fourth run contains only the HSHs, MFCCs, DSIFT as well as IDT features. The fifth run mixes the features from the two previous runs. Due to the large size of the feature-space, for the last run, a linear late fusion strategy was implemented and the scores of the two regressors were combined linearly[2].

Table 1 displays the name of each run, whether it was an external or a required run, and the Pearson-Correlation coefficient for valence and arousal models separately, both in development- and release test-set. Some cells of the matrix do not provide scores for the release test-set, because these runs were executed after the corresponding deadline. It should be pointed out, that some videos had too little movement and no IDT features could be extracted. So, models of Runs 3,4 and 5 were trained, validated and evaluated in a slightly smaller set of videos (9786 instead of the total 9800 movie-segments).

Table 1: **AUTH-SGP Results**, Pearson-Correlation Coefficient on Development Test-set (Dev-Test) and Release Test-set (Rel-Test)

Run	Arousal		Valence	
	Dev-Test	Rel-Test	Dev-Test	Rel-Test
Run1	0.308	0.247	0.264	0.076
Run2_ext	0.303	0.265	0.290	0.11
Run3	0.264	-	0.192	-
Run4	0.244	-	0.209	-
Run5	0.307	-	0.247	-

2nd sub-task. It is worth mentioning also, that an attempt was made for the second sub-task. A deep-learning model was trained from scratch for the two variables (valence, arousal) separately. Because there were difficulties with the converge of these models and the results were not encouraging, we decided not to publish them.

4. CONCLUSIONS

Comparing Run1 and Run2, we can conclude that, deep learning features do actually improve the performance of the system. From Run3 and Run4 we can notice, that IDT features (Run3), which represent motion, are more important for the arousal prediction (emotion intensity), while HSH features in Run4, which symbolize color, better affect the performance of the valence model (positive-negative emotions). These conclusions are confirmed also from our findings in bibliography[12]. Finally, combining the features from Run3 and Run4 leads to a satisfying improvement of both models.

5. REFERENCES

- [1] Activity Recognition in Videos using UCF101 dataset. https://github.com/anenbergb/CS221_Project.
- [2] Finding optimized weights when combining classifiers. <https://www.kaggle.com/c/otto-group-product-classification-challenge/forums/t/13868/ensemble-weights/75870#post75870>.
- [3] pyAudioAnalysis: A Python library for audio feature extraction, classification, segmentation and applications. <https://github.com/tyiannak/pyAudioAnalysis>.

- [4] Scikit-learn: Machine learning in Python.
<http://scikit-learn.org/stable/>.
- [5] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Deep Learning vs. Kernel Methods: Performance for Emotion Prediction in Videos. In *2015 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.
- [6] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Sjöberg and Christel Chamaret. The MediaEval 2016 Emotional Impact of Movies Task. In *Proc. of the MediaEval 2016 Workshop*, Hilversum, Netherlands, Oct. 20-21 2016.
- [7] Itseez. Open source computer vision library.
<https://github.com/itseez/opencv>, 2015.
- [8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14*, pages 1725–1732, Washington, DC, USA, 2014. IEEE Computer Society.
- [9] D. Paschalidou and A. Delopoulos. Event detection on video data with topic modeling algorithms. Master’s thesis, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Nov. 2015.
- [10] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM ’10, pages 1469–1472, New York, NY, USA, 2010. ACM.
- [11] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.
- [12] S. Wang and Q. Ji. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, 6(4):410–430, Oct. 2015.