TCNJ-CS @ MediaEval 2017 Emotional Impact of Movie Task

Sejong Yoon
The College of New Jersey, U.S.A.
yoons@tcnj.edu

ABSTRACT

This paper presents our approaches for the MediaEval Emotional Impact of Movies Task. We employed features from image frames and audio signal. We use support vector regression for the learning and prediction. In addition, we introduce a new feature using exponential decay of the initially predicted emotion labels. The motivation behind this is to computationally model *lingering* effect. Experimental results and future direction are also discussed.

1 INTRODUCTION

MediaEval 2017 Emotional Impact of Movie Task [2] consist of two subtasks. One of them is valence/arousal prediction that predicts a score of *expected* level of two emotional state, valence and arousal for each consecutive ten seconds segments. Both valence (most negative to most positive) and arousal (least active to most active) are defined as a continuous scale within range of $[-1, 1]$. The other is the the fear prediction that makes binary prediction for each of the ten second-segments, whether they are *likely to induce* fear or not. Fear is defined as binary integer $[0, 1]$ where 1 indicates that the segment will induce fear. In the following, we describe the method we used in our prediction system.

2 APPROACH

First, we describe multimodal features we employed. Next, we introduce our new feature based on prediction label to model lingering effect of induced emotions. Lastly, we describe our hierarchical regression framework for emotion prediction.

2.1 Visual and Audio Features

We employed all standard set of visual and audio features provided by the MediaEval task organizers. For image frame-based features, we used 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, fc6 layer of VGG16 network [5]. All features were extracted frame-by-frame, where one frame was extracted per second. The features, except VGG16, were computed using LIRE library. VGG16 features were extracted using the MATLAB Neural Network toolbox.

For auditory features, we employed the audio features provided. In the provided description, there should be 1,582 features which result from a base of 34 low-level descriptors, with 34 corresponding differential coefficients, and 21 functionals applied to each of these 68 contours, thus 1,428 features ($21 \times (34 + 34)$) in total. Out of remaining 154 features, 152 features were computed by applying 19 additional functionals to the 4 pitch-based low-level descriptors and their 4 differential coefficient contours ($19 \times (4 + 4)$). Last two features are additional statistical features, the number of pitch onsets and the total duration of the input. These features were computed every ten seconds segments sliding over the whole movie with a shift of 5 seconds. All these features were computed by openSmile toolbox [3].

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

2.2 Lingering Feature

In addition to the provided features, we introduced additional feature, using the ground truth labels of emotional levels. The motivation behind this new feature, is to computationally model the gradually amplifying or decaying emotional flow, what is typically referred as *lingering* emotion. Traditional and even state-of-the-art affect prediction systems focus on predicting induced emotions as a *spike* noise detection model, regardless of whether they modeled the temporal aspect of affect or not. On the other hand, lingering emotions do not directly induced by the stimuli, rather, they are generated from the emotional change, i.e. response, already existing a priori. In short, we argue that what is called *climax* of a movie is not only a consequence of short segment stimuli, but also amplified (or degraded) by the emotional state change itself across the segments. A similar idea was utilized to predict media interestingness [6], but there was no explicit consideration of amplifying / decaying effect. Here, we consider the change of emotion directly.

To model this lingering emotion, we use emotion level label values. For each segment $t = 1..T$ where T denotes the total number of segments, assume that we are given emotion level label y_t. So, at each time segment t, we have

$$x_1, x_2, \cdots, x_t, \tag{1}$$

$$y_1, y_2, \cdots, y_t, \tag{2}$$

where x_t denotes the vectorized visual / audio features and y_t denotes either the ground truth or predicted emotion level (it could be valence, arousal, or fear label). Then, we can define the lingering feature l_t as an exponential decay function of labels as

$$l_{(t-w)} = y_{(t-w)} \tag{3}$$

$$l_{(t-w+1)} = (1 - \alpha) \cdot l_{(t-w)} + \alpha \cdot y_{(t-w+1)} \tag{4}$$

$$\vdots$$

$$l_s = (1 - \alpha) \cdot l_{s-1} + \alpha \cdot y_s \tag{5}$$

where $s = (t - w), ..., t$, and w denotes the lingering window size. Parameter α is the decay factor. Intuitively, we take weighted accumulated emotions over time, and consider it to model the lingering effect. In training phase, we can utilize the ground truth emotion. In testing phase, we can devise a two-step, hierarchical regression model to obtain the emotion level feature values. We will describe this model in the next section.

There are considerations why we think this can be a reasonable model for the lingering effect. First, with the exponential decay function, we can consider both smoothness and also decaying of emotional change over time. Second, one can view this as simplified version of traditional temporal models, e.g., Hidden Markov Models (HMM), where we fix the transition probability. If we can obtain large number of emotion labels, one may try to learn HMM-based features instead, as in [6].

2.3 Hierarchical Regression Framework

To combine features, we utilized standard multiple kernel learning approach [1, 4]. We first compute kernels of each feature, and build combined kernel using either addition or multiplications. We use multiplication within same modality, e.g., combining Color Correlogram kernel and Edge Histogram kernel, and use addition between different modalities, i.e., combining combined visual kernel and audio kernel. The lingering feature is considered as another modality than visual and auditory. In summary, our combined kernel was computed as

$$K_{vis} = K_{acc} \cdot K_{cedd} \cdots K_{fc6} \tag{6}$$

$$K_{all} = K_{vis} + K_{aud} + K_{lin} \tag{7}$$

where each $K_{\cdot}$ denotes the kernel computed using the features. We used Radial Basis Function (RBF) kernel with median of training data as the hyperparameter.

Once the combined kernel is computed, we can use it as feature vectors. For the regression model, we used linear Support Vector Regression (SVR). We used MATLAB's `fitrsvm` function for this. One important aspect of our approach is that we use emotion prediction labels to compute the lingering features. Since we do not have ground truth labels for testset, we design a two-step, hierarchical regression framework. In this framework, we need to train two SVR models in the training phase. One model (Model A) is trained with the kernel computed using training data, but the kernel is only combines visual and auditory features. The other model (Model B) is trained with the kernel computed using all modalities. In the testing phase, we first perform an initial emotion prediction on the test data using Model A. Then, we compute the lingering feature using the predicted affect labels. Note that this is computationally *not* expensive since the lingering feature itself is easy to compute and the labels of all training data is only 1 dimensional vector. Finally, we perform final emotion prediction on the test data using Model B. We applied regression framework for all subtasks. For the fear subtask, we first rescaled the output into $[0, 1]$ range and thresholded at 0.75.

3 RESULTS AND ANALYSIS

For the measure, we used Mean Squared Error (MSE) and Person's correlation coefficient (ρ) for the valence and arousal subtasks, and accuracy, precision, recall, and F_1 score for the fear subtask. We used $\alpha = 0.5$ for all experiments.

In the Devset, shown in Table 1, one can see that there is no significant benefit in using lingering feature in this case. We used 50-50 split to obtain the result but the readers should take this result with a grain of salt (particularly, accuracy of fear) since we did not make strict data split based on the information which frame belongs to which video, to obtain this result.

Table 1: Result of All Subtasks in Devset

Subtask	Measure	w/o Linger	w/ Linger
Valence	MSE	0.13106	0.09893
	ρ	0.12826	0.20826
Arousal	MSE	0.08817	0.08415
	ρ	0.08390	0.09001
Fear	Accuracy	0.96581	0.96543

Table 2: Result of All Subtasks in Testset

Subtask	Measure	w/o Linger (run 1)	w/ Linger (run 2)
Valence	MSE	0.20276	0.04640
	ρ	0.19748	0.00583
Arousal	MSE	0.12304	0.11335
	ρ	0.11340	0.21485
Fear	Accuracy	0.77862	0.72956
	Precision	0.22474	0.25530
	Recall	0.09922	0.19224
	F_1	0.10113	0.17399

In the Testset, shown in Table 2, the official results are more interesting. It is obvious that the lingering feature does not help (actually hinders) the valence prediction. On the other hand, for the arousal and fear, lingering feature seems to make positive contribution to the prediction although the overall MSE and accuracy sacrificed a little. It is also notable that the similar tendency could be observed in the Devset in Table 1. One intuitive explain here would be following: what we are modeling with lingering feature, is how the prior, recent emotional change might affect or induce the new emotion. In case of arousal (either active or passive to the stimuli) or fear (feeling horror, anxiety or not), this happens often. On the other hand, valence (positive or negative) is rather difficult to capture with a fixed window size of the linger feature. Moreover, changing from positive to negative emotional state, or vice versa, requires more contextual (or semantic) information of stimuli to understand why that change has happened.

4 DISCUSSION AND OUTLOOK

In this paper, we introduced a new feature modeling lingering effect and presented a hierarchical regression framework to predict emotions. We found promising applications of the new features in arousal and fear prediction, with limitations in valence prediction. In the future, it would be interesting to investigate how one can more robustly capture this lingering effect with in-depth understanding of the feature's impact on the valence prediction.

ACKNOWLEDGMENTS

This work was supported in part by The College of New Jersey under Support Of Scholarly Activity (SOSA) 2017-2019 grant.

REFERENCES

[1] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. 2004. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML) (ICML '04)*. ACM, New York, NY, USA, 6–. https://doi.org/10.1145/1015330.1015424

[2] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, and Mats Sjöberg. 2017. The MediaEval 2017 Emotional Impact of Movies Task. In *MediaEval 2017 Workshop*.

[3] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 835–838. https://doi.org/10.1145/2502081.2502224

[4] Mehmet Gönen and Ethem Alpaydin. 2011. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research (JMLR)* 12 (July 2011), 2211–2268.

[5] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[6] Sejong Yoon and Vladimir Pavlovic. 2014. Sentiment Flow for Video Interestingness Prediction. In *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia (HuEvent '14)*. ACM, New York, NY, USA, 29–34. https://doi.org/10.1145/2660505.2660513