DEEP FEATURES FOR MULTIMODAL EMOTION CLASSIFICATION

Shriman Narayan Tiwari, Ngoc Q. K. Duong, Frédéric Lefebvre, Claire-Helène Demarty, Benoit Huet, Louis Chevallier

DEEP FEATURES FOR MULTIMODAL EMOTION CLASSIFICATION

Shriman Narayan Tiwari[1⋆] *, Ngoc Q. K. Duong*[2]*, Frédéric Lefebvre*[2]
Claire-Hélène Demarty[2]*, Benoit Huet*[1]*, Louis Chevallier*[2]
[1] EURECOM, Sophia Antipolis, France
[2] Technicolor Research and Innovation, Cesson Sévigné, France

ABSTRACT

Understanding human emotion when perceiving audio-visual content is an exciting and important research avenue. Thus, there have been emerging attempts to predict the emotion elicited by video clips or movies recently. While most existing approaches focus either on single modality, *i.e.,* only audio or visual data is exploited, or build on a multimodal scheme with late fusion, we propose a multimodal framework with early fusion scheme and target an emotion classification task. Our proposed mechanism presents the advantages of handling (1) the variation in video length, (2) the imbalance of audio and visual feature sizes, and (3) the middle-level fusion of audio and visual information such that a higher level feature representation can be learned jointly from the two modalities for classification. We evaluate the performance of the proposed approach on the international benchmark, *i.e.,* the MediaEval 2015 Affective Impact of Movies[1] task, and show that it outperforms most state-of-the-art systems on arousal accuracy while using a much smaller feature size.

***Index Terms*—** Multimodal-based emotion classification, affective computing, deep learning, video segmentation, feature aggregation.

1 Introduction

Endowing computers with human like emotion perception techniques would be a technological leap towards creating a real context aware machine. It offers a wide range of applications *e.g.,* in health care system to make visually impaired people more autonomous, in parental control systems to detect violent scenes, or in content recommendation. Thus, understanding emotion elicited by the scene is a promising research avenue and within the past decade there has been an active ongoing research towards this trend [1, 2, 3, 4].

Earlier approaches exploited Hidden Markov Models (HMM) for affective content analysis [5]. But such models allow transition only between the neutral state and other affective states, which made it restrictive. Other probabilistic models [1, 2] were also investigated for affective video content representation. Zhang *et al.* proposed a clustering based approach with arousal and valence features in [6], and then a support vector regression based model using user profile and multimedia features for analyzing music video in [7]. At the same time, a latent Dirichlet allocation model using k-means clustering was also presented in [8]. Recently with the success of deep learning in various applications, it has also been investigated for the considered task [3, 9, 10]. Some state-of-the-art approaches leverage classic image and audio features such as histogram of gradient (HOG), color intensity, scale-invariant feature transform (SIFT), dense trajectories, mel-frequency cepstral coefficients (MFCC) [10, 11], etc., together with features extracted from convolutional neural networks (CNN) pre-trained on public datasets such as ImageNet [11, 12, 13, 14] for the final emotion classification. In [13] features extracted from more dedicated CNN were also used, such as stacked optical flows trained on the UCF-101 dataset[2] for emotion classification in movies. All these works either use late fusion of classifiers [11, 14] or some mid-level fusion scheme, in which aggregated features are used as input to a final classification step [13, 15]. For the later, SVM is the prefered classifier, only [16] attempted to use neural networks as the final classifier, but reported low performance.

Compared to what can be found in the literature, we propose here to combine the extracted audio and visual features at a middle stage so that a higher level feature representation can be learned jointly from the two modalities by a multilayer perception (MLP) for classification. While classic audio features are used, visual features solely consist of features from pre-trained CNN. However since the extracted low-level audio and visual features usually have different size, our further contribution is to provide a mechanism for balancing their contribution in an aggregated feature. We will also compare our results with a baseline SVM-based classifier.

With a large interest of the research community, the Affective Impact of Movies task[3] in the MediaEval challenge 2015 [17] has become a state-of-the-art benchmark which attracted more than ten research groups to evaluate their systems on the same dataset. The underlying emotion conveyed

⋆ This work was performed while the first author was doing an internship at Technicolor Research

[1] http://www.multimediaeval.org/mediaeval2015/affectiveimpact2015/

[2] http://crcv.ucf.edu/data/UCF101.php

[3] The large variance of audio-visual content and different shooting styles makes the affective content analysis a very challenging task.

in short videos of the dataset is characterized by two well-known metrics *valence* and *arousal* [18, 19], where valence indicates the emotional value associated with a stimulus and arousal is a state of heightened activity in both our mind and body that makes us more alert. The Affective Impact of Movies task and its 2015 results will serve as a baseline to compare the performance of our system.

The rest of the paper is organized as follows. Section 2 presents the general workflow of the proposed approach with detailed description for each step. We conduct experiment to compare the performance of the proposed method with state-of-the-art systems using the MediaEval benchmark in Section 3. Finally, we conclude in Section 4 with remarks about some future perspectives.

2 Proposed approach

The general workflow of the proposed approach is presented in Figure 1. This framework contains three major steps as: 1) video segmentation and low-level modal-specific feature extraction, 2) feature dimensionality control and aggregation, and 3) classification and voting. Each of these steps will be described in detail in following subsections.

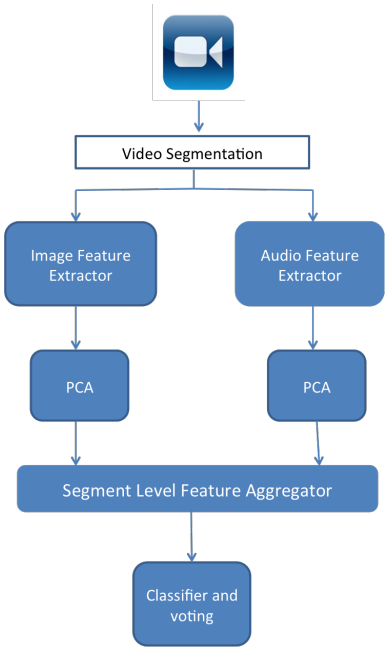

Fig. 1. General work flow of the proposed approach

2.1 Video segmentation and low-level feature extraction

Given an input video, we propose first to segment it into short segments of length t seconds, typically $t = 2$, each and process each segment independently. Thus, with an input video of length T seconds, the number of segments N is given by:

$$N = \left\lfloor \frac{T}{t} \right\rfloor \quad (1)$$

where, $\lfloor\rfloor$ is the greatest integer function. This simple segmentation scheme brings three benefits in our framework as follow. First, it allows us to capture local information presented in certain parts of the video sequence. Second, it allows to handle input videos with different lengths, which is a typical issue in most existing databases, since the extracted audio-visual feature also depends on the signal length. Third, when assigning label of the input video in training data set for all corresponding segmented short segments, it actually increases the number of examples, *i.e.,* by N times, so that we can expect to learn a better classifier.

For each t-second length video segment, low-level features are extracted for the audio and visual modalities as follow. Concerning the visual modality, we select several key frames per segment, ranging from one to ten in our experiments, and compute a visual feature vector for each key frame. For this purpose, we extract the intermediate layers of the CNNs pre-trained by the ImageNet dataset given a key frame as an input image. As these CNNs are trained to classify objects on huge dataset, we can hope that the extracted features from some intermediate layers also provide discriminative visual characteristics that will be relevant for our task. Two well-known CNN networks,*i.e.,* Overfeat proposed in [12] and Visual Geometry Group (VGG)'s network described in [20], are considered in our experiments. From Overfeat [12], features are extracted from layers 18 and 21 of the small network, and layers 21 and 24 from the large network. Similarly with VGG network, different features are extracted for small and large architecture. Overall, visual feature vectors of different length ranging from 128 to 4096 are computed for testing our framework with different feature sizes. Concerning audio feature, audio track corresponding to the video segment is devided into 40 ms frames with 50 % overlap. Then the standard Mel-frequency cepstral coefficients (MFCC) is computed for each frame. In our experiment, we compute 20 MFCCs, their first derivatives Δ, and second derivatives $\Delta\Delta$, and concatenate them together. This results in a 57 length vector as audio feature for each 40 ms frames.

2.2 Feature dimensionality control and aggregation

As presented in Section 2.1, the final audio and visual features extracted per video segment have very different sizes. Indeed, we have an audio feature vector of size 57 for each 40 ms audio frame (thus a feature size of 5700 for 2s of video) and several hundreds or thousands as visual feature vector

for each key frame image. Thus when aggregating the two modalities, the classifier may learn mostly from one of them, as their contribution is very imbalanced.

To overcome this issue, we propose to perform an additional step of feature dimensionality control before aggregating the two modalities. This is done by performing principle component analysis (PCA) on the input audio/visual feature matrices where the feature output, as the number of principle components, can be well-controlled. As a result, PCA will help setting the audio and visual features to equivalent sizes, so that the two modalities contribute equally to the task. Note that, in addition to controlling the respective modality dimensions as a main target, PCA also helps in eliminating redundancy in the extracted features. As a practical example, key frames extracted from some static video segments are very similar. Thus using visual feature vectors computed from all key frames may not be necessary.

Once the low-level audio and visual features are balanced, they are concatenated to create a single joint feature vector. In the next step, this balanced and aggregated feature will serve as input to our joint-modality classifier.

2.3 Classification and voting strategy

In order to learn a deeper feature for classification, the concatenated low-level audio-visual feature vectors are served as inputs to a fully connected neural network, known as multi-layer perceptron (MLP). We expect that MLP will learn a higher level feature representation jointly from the two modalities to improve the classification performance. In the implementation, an exhaustive grid-search is launched to find the optimal parameters of the MLP. More about the MLP will be discussed in Section 3.

The output layer of the MLP, which serves as a classifier for each video segment, offers probability values from the softmax function for each of the label classes. In order to obtain a final emotion label for the whole input video, which contains N segments, we then perform majority voting from the results obtained by classifying each individual segments as

$$l = \arg\max_{l_m} \sum_{k=1}^{N} S(l_m|k), \tag{2}$$

where, l is the emotion label given to the video and $S(l_m|k)$ is the score obtained from the classifier when observing label l_m given the k^{th} segment of the video. For the SVM classifier the score is the log-probability values associated with each possible outcome class for a video segment, whereas for MLP, the score is the class probability for each output class. This voting strategy can be understood as choosing the label which maximizes the total score over all classes for all video segments as the final label for the input video.

3 Experiment

After describing the dataset and the evaluation metric in Section 3.1, we present in Section 3.2 the result obtained by the proposed approach. This result is compared with an SVM-based baseline framework as well as with state-of-the-art systems which participated in the MediaEval 2015 campaign.

3.1 Dataset and evaluation metric

For easier and fair comparison with existing methods, we evaluated the performance of the proposed approach on the benchmarked dataset provided in the MediaEval 2015 Affective Impact of Movies task, which were designed for detecting the emotional impact of movies [17]. This dataset, is an extension of the LIRIS-ACCEDE dataset [21], consisting of 10,900 short video clips extracted from 199 Creative Commons-licensed movies of various genres. It is independently split into development set containing 6144 video shots from 100 movies, and test set containing 4756 video shots from remaining 99 movies. Each of these shots are between 8 and 12 seconds long and start and end with either a cut or a smooth transition. The dataset has discrete labels for both valence and arousal. As such valence is labeled as *negative, neutral and positive*, while for arousal the classes are *passive, neutral and active*.

The induced affect detection is officially evaluated using the *global accuracy* metric [17], calculated separately for valence and arousal dimensions, according to:

$$A_i = \frac{C_i}{N_i} \tag{3}$$

where, $i \in \{\text{Valence, Arousal}\}$, C_i, is the total number of correct classifications in all three classes and N_i, is the total number of test examples.

3.2 Algorithm setup and results

We tested the approach with different audio-visual feature sizes obtained using dimensionality reduction, ranging from 128 to several thousands. It was found that, the classification performance was not sensitive to the feature size ranging form 250 to 1000. Therefore we decided to use smaller feature sizes of the order of 500 as input for both the baseline SVM and the MLP classifiers to jointly learn the feature representation. Note that, with a view of adding temporal information, experiments were also performed with several key frames ranging from one to ten per video segment but, it turns out that, generally due to high correlation among the frames, equal performance was achieved by using only a few number of key frames (2 or 3 frames per segment).

In the MLP architecture, rectified linear units [26] were used as nonlinear functions and stochastic gradient descent with minibatches was used for parameter updates. The best

Approach	Feature type	Classifier	Feature Length	Arousal Accuracy (%)
Yi et al. [11]	IDT + SIFT + MFCC + HSH + CNN	Linear SVM	$>$ 4,000	55.93
Lam et al. [14]	HOG + MBH +SIFT + MFCC + VDFULL + CNN	SVM	$\sim$ 40960	55.90
Trigeorgis et al. [22]	Low Level Descriptors + CNN	AdaBoost	$\sim$ 1000	55.72
Seddati et al. [16]	OFM + CNN trained on HMDB-51[23]	Two layer fully connected neural network	$\sim$ 20,000	52.44
Marin et al. [15]	GIST + IDT + CNN	linear SVM	$>$ 100,000	51.90
Chakraborty et al. [24]	Low Level Descriptors	Artificial Neural Network	$\sim$ 1000	48.95
Dai et al. [13]	LSTM trained on UCF-101	SVM with linear and χ^2 kernels	$>$ 10,000	48.7
Mironică et al. [25]	Fisher kernel + CNN	non-linear SVM	$\sim$10,000	44.36
Proposed approach with SVM baseline	Intermediate CNN features + MFCC	SVM with RBF kernel	$\sim$ 500	***55.14***
Proposed approach	Intermediate CNN features + MFCC	MLP	$\sim$ 500	**55.85**

Table 1. Comparison with state-of-the-art approaches participated in the MediaEval 2015: Affective Impact of Movies Task

results were obtained using categorical cross-entropy loss function. The hidden layers had a dropout factor of 0.5, which helps in avoiding the over-fitting the model. The number of hidden layers, the number of neurons per layer and the learning rate were varied and optimized by the development data. Finally, with the considered dataset, we obtain the following parameters of the MLP: 2 hidden layers with 100 hidden units in hidden layer one and 10 units in hidden layer two.

We compare the proposed approach, *i.e.,* where MLP is used for jointly learning a higher level feature representation from the balanced visual and audio modality, with the baseline setup, *i.e.,* where the balanced audio-visual feature is used directly as input to the SVM classifier (named "Proposed approach with SVM baseline" in Table 1). We also compare our results with state-of-the-art systems in the MediaEval 2015 challenge. The results are shown in Table 1, where the *feature type* column summarizes the features exploited by each approach. The *classifier* column shows the final classifier used by each approach and *feature length* column presents the final multimodal feature size required by each algorithm. As example, in [11] CNN visual features are used along with dense SIFT features, MFCCs, hue-saturation histogram (HSH) and improved dense trajectory (IDT), where as [25] uses Fisher kernels with CNN visual features. Also, some submissions used external datasets to train the models like HMDB-51 is used in [16] and UCF-101 is used in [13]. Note that we propose here solely the results obtained for arousal classification. For valence, there was no significant improvement compared to the proposed methods in MediaEval 2015. Nevertheless none of these systems did improve much above the random baseline (+7% for the best system), compared to what was achieved for Arousal (+22%). Thus, it was a relevant choice to exhibit the results for arousal classification only.

As can be seen, the proposed approach with MLP offers higher accuracy than the baseline workflow with SVM classifier. This supports our strong expectation that a higher level feature representation learned jointly from two modalities would improve the performance. It can also be inferred that multimodal feature fusion at middle stage, in which the contribution of each modality is well-balanced, could be a good strategy in a multimodal framework. As compared to the state of the art, the proposed approach outperforms all the proposed approaches but two. Note that the systems [11, 14] with slightly better performance (0.1%) employed much larger feature vectors than ours.

4 Conclusion

In this paper we have presented a multimodal, deep learning-based, framework for emotion classification on the arousal scale. This framework has the advantage of handling the variation in the video length and the imbalance of feature sizes from the different modalities through a mid-level feature aggregation scheme. The proposed approach was compared with those submitted by the participants of the MediaEval 2015 Affective Impact of Movies task. We have demonstrated that even with a small feature size jointly learned from only two features types, CNN-based visual feature and MFCCs, the proposed system outperforms all state-of-the-art systems but two in terms of arousal accuracy. The proposed multimodal framework could be further improved by adding external dataset for training. Also, taking into account the temporal aspect of videos, either by adding temporal features (*e.g.,* motion-based) or by using recurrent neural networks, is worth being tested.

5 References

[1] A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[2] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[3] S. E. Kahou, C. Pal, et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[4] M. Paleari, R. Chellali, and B. Huet, "Features for multimodal emotion recognition: An extensive study," in *Cybernetics and Intelligent Systems (CIS), 2010 IEEE Conference On*. IEEE, 2010, pp. 90–95.

[5] H. B. Kang, "Affective content detection using HMMs," in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 259–262.

[6] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective mtv analysis based on arousal and valence features," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 1369–1372.

[7] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 510–522, 2010.

[8] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 523–535, 2010.

[9] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *MultiMedia Modeling*. Springer, 2014, pp. 303–314.

[10] E. Acar, F. Hopfgartner, and S. Albayrak, "Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos," in *Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on*. IEEE, 2015, pp. 1–6.

[11] Y. Yi, H. Wang, B. Zhang, and J. Yu, "MIC-TJU in MediaEval 2015 affective impact of movies task," 2015.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[13] Q. Dai, R. W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y. G Jiang, "Fudan-huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with deep learning," 2015.

[14] V. Lam, S. Phan, D. D. Le, S. Satoh, and D. A. Duong, "NII-UIT at MediaEval 2015 Affective Impact of Movies Task," 2015.

[15] P Marin Vlastelica, S. Hayrapetyan, M. Tapaswi, and R. Stiefelhagen, "KIT at MediaEval 2015–Evaluating visual cues for affective impact of movies task," 2015.

[16] O. Seddati, E. Kulah, G. Pironkov, S. Dupont, S. Mahmoudi, and T. Dutoit, "UMons at MediaEval 2015 Affective Impact of Movies Task including violent scenes detection," 2015.

[17] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen, "The MediaEval 2015 Affective Impact of Movies Task," in *MediaEval 2015 Workshop*, 2015.

[18] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989.

[19] P. J Lang, M. M. Bradley, and B. N Cuthbert, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, pp. 39–58, 1997.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *Affective Computing, IEEE Transactions on*, vol. 6, no. 1, pp. 43–55, 2015.

[22] G. Trigeorgis, E. Coutinho, F. Ringeval, E. Marchi, S. Zafeiriou, and B. Schuller, "The ICL-TUM-PASSAU approach for the MediaEval 2015 Affective Impact of Movies Task," 2015.

[23] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2556–2563.

[24] R. Chakraborty, A. K. Maurya, M. Pandharipande, E. Hassan, H. Ghosh, and S. K. Kopparapu, "TCS-ILAB-MediaEval 2015: Affective Impact of Movies and Violent Scene Detection," 2015.

[25] I. Mironica, B. Ionescu, M. Sjöberg, M. Schedl, and M. Skowron, "Rfa at MediaEval 2015 Affective Impact of Movies Task: A multimodal approach," .

[26] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton, "On rectified linear units for speech processing," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2013, pp. 3517–3521.