Visual and Audio Aware Bi-Modal Video Emotion Recognition

Siqi Xiang (sqxiang@buaa.edu.cn)
Wenge Rong (w.rong@buaa.edu.cn)
Zhang Xiong (xiongz@buaa.edu.cn)
School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Min Gao (gaomin@cqu.edu.cn)
Qingyu Xiong (xiong03@cqu.edu.cn)
School of Software Engineering, Chonqing University, Chongqing 401331, China

Abstract

With rapid increase in the size of videos online, analysis and prediction of affective impact that video content will have on viewers has attracted much attention in the community. To solve this challenge several different kinds of information about video clips are exploited. Traditional methods normally focused on single modality, either audio or visual. Later on some researchers tried to establish multi-modal schemes and spend a lot of time choosing and extracting features by different fusion strategy. In this research, we proposed an end-to-end model which can automatically extract features and target an emotional classification task by integrating audio and visual features together and also adding the temporal characteristics of the video. The experimental study on commonly used MediaEval 2015 Affective Impact of Movies has shown this method's potential and it is expected that this work could provide some insight for future video emotion recognition from feature fusion perspective.

Keywords: videos; multi-modal scheme; modal fusion; end-to-end; temporal characteristics

Introduction

To better understand and analyse people's emotion response during watching videos, it is essential to study the cognitive determinants beneath the video presentation. Currently, content based approaches are the main trend for video emotion analysis, and a lot of models have been proposed to help identify the emotions evoked by videos (Hanjalic, 2006), among which affective analysis based on video visual contents have been studied for several years. Several approaches which employed different machine learning models such as Bayesian network (Soleymani, Kierkels, Chanel, & Pun, 2009), Hidden Markov Models (Kang, 2003) have been proposed and proven applicable to tackle with this challenge.

Though visual content based video emotion analysis has proven applicable in real applications, there still exists challenges since even the same scene could cause different emotions (Choe, Chun, Noh, Lee, & Zhang, 2013). Recently audio related features have also proven its effectiveness in emotion analysis (Cui, Jin, Zhang, Luo, & Tian, 2010). For example, Xu et al. tried to use audio emotional events (AEE) such as laughing, horror sounds and other features to detect horror and comedy movies (Xu, Chia, & Jin, 2005).

While previous studies focused on video or audio features alone in detecting video emotion have proven their ease in implementation, to further improve the classification performance, some researchers indicate the possibility by combining visual features with audio features to form a hybrid feature that can carry information from two different modalities (domains) at the same time. Such methods can be roughly divided into two categories in terms of the way the features are combined, i.e., later fusion of classifiers (Yi, Wang, Zhang, & Yu, 2015), and early fusion scheme, in which features are concatenated into a final classifier (Dai et al., 2015; P., Hayrapetyan, Tapaswi, & Stiefelhagen, 2015).

In this research, we employed the idea of modal fusion and then proposed an end-to-end framework to integrate the visual and audio features for video emotion analysis. Recently with the development of deep learning techniques, a lot of advanced methods have been proposed for feature extraction. In this research, we used convolutional neural network (CNN) to extract video emotion related features as CNN has proven its success in learning intermediate representations from low-level features (Acar, Hopfgartner, & Albayrak, 2014). Afterwards, taking into account the temporal characteristics of video, we further use Long Short Term Memory (LSTM) model (Hochreiter & Schmidhuber, 1997) to integrate the extracted temporal features since it performs well on tasks that require integration of state information over time. Finally a multi-layer perceptron (MLP) is employed to classify the final video emotions.

To confirm the validity of the proposed method, we implement it in the Affective Impact of Movies Task 3 in the MediaEval challenge 2015 (Sjöberg et al., 2015). The task has now become a state-of-the-art benchmark which attracted a large number of research teams to test their models on this data set. The experimental study result against different bench experiments on this dataset shows the proposed method's potential in detecting video's emotion.

Related Work

In the content-based video research, many researchers have used a lot of models to identify the emotions triggered by the video. Hanjalic argued the possibility to classify films according to their emotions and proposed the concept of "expectation of emotion", which is defined as one or a group of emotions that a filmmaker wishes to use to communicate with a certain culture or a particular audience through the film (Hanjalic, 2006). Through this concept, he proposed the video content information and its underlying characteristics to predict emotion. Later on, Soleymani et al. proposed a Bayesian framework to detect scene affect and the arousal

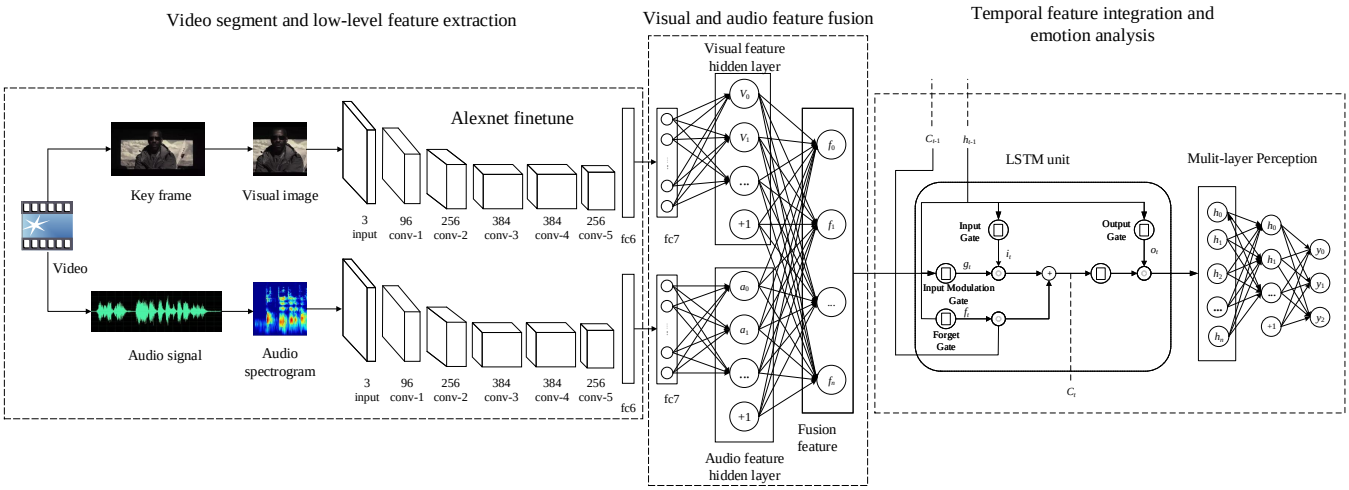

Figure 1: Visual and Audio Aware Video Emotion Analysis Framework

and valence values with content features were used to classify video emotions into 3 classes, i.e., calm, excited positive and excited negative (Soleymani et al., 2009). Similarly Arifin and Cheung established a framework based on the hierarchical coupling of dynamic Bayesian networks to establish the dependencies of the pleasure-activity-dominance emotion model (Arifin & Cheung, 2008).

There are also various studies on video affective characterization using audio features, e.g., rhythm, tempo, mel-frequency cepstral coefficients (MFCC), pitch, zero crossing rate. For example, three feature sets, i.e., intensity, timbre and rhythm were extracted from audio to classify video emotion using Gaussian mixture models (Lu, Liu, & Zhang, 2006). Similarly, Xu et al. tried to use audio emotional events (AEE), including laughing, horror sounds and other features to detect horror and comedy movies (Xu et al., 2005).

In fact, there is a complex interaction between the audio and visual contents to determine the perceived mood. As such the video emotion analysis has begun to use feature fusion method to classify emotion into different classes (Yi et al., 2015). Similarly, Trigeorgis et al. selected the low level descriptors with the traditional adaboost as a classifier (Trigeorgis et al., 2015). Wang and Cheong derived the characteristics of multimodality by probabilistic inference based on two SVM models (Wang & Cheong, 2006), where one SVM model is designed to process audio data and extracts the corresponding advanced audio information, while another SVM model is used to classify the captured video segments.

However, since these framework extracts basic features, they lack the ability to use raw inputs to automatically learn mid-level representations. With the development of deep learning techniques, some deep learning based approaches are also proposed in the literature. For example, Kahou et al. used a deep convolution neural network to analyse facial expressions within a frame and used a deep belief net to capture audio feature (Kahou et al., 2016). Levi and Hassner also used convolution neural network to capture visual features to classify video into seven emotions (Levi & Hassner, 2015).

Proposed Approach

The overall pipeline of the proposed visual and audio aware emotion analysis framework is depicted as Fig. 1, where the whole process is divided into three steps: 1) video segment and low level feature extraction; 2) bi-modal visual and audio feature fusion; and 3) temporal feature integration and emotion classification.

Video segment and low level feature extraction

To analyze video emotion, it is necessary to firstly divide a video into short videos with a length of t seconds. In this study we set $t = 1$ so that a video of length T will have T slices. This segmentation has two benefits. First, since the length of each video is different, this segmentation gives us better access to the visual and audio features. Second, Because of the temporal characteristics of the video, cutting the video into the same segments can be used for subsequent recurrent neural networks.

For each segment, we need to extract its visual and audio features separately. As to the visual features, we extract the k key frames for each segment. Due to the strong correlation among frames within a second, we select $k = 1$. The key frame is defined as the frame with the closest RGB histogram to the mean RGB histogram of the whole video clip using the Manhattan distance (Zhu, Jiang, Peng, & Zhong, 2016). Assume that a video clip V contains n frames, the RGB histogram of i-th frame is defined as $h(i)$. The Manhattan distance D between two frames i and j is calculated as follows:

$$D(i,j) = |h(i) - h(j)| \quad (1)$$

and the key frame will be:

$$\arg\min_i D(i, \frac{1}{n}\sum_{j=1}^{n} h(j)) \tag{2}$$

After getting the key frame, it will be resized to $256 * 256$ pixel, as suggested in (Krizhevsky, Sutskever, & Hinton, 2012) as input for fine-tuning. The concept of fine-tuning is to use a model pre-trained on a large dataset, replacing its last layers, and fine-tune the weights on new task using back-propagation. In this study, AlexNet (Krizhevsky et al., 2012) is employed. AlexNet consists of five convolution layers and three fully connected layers. Here we select the fc7 layer of AlexNet which has 4096 neurons as our visual features.

As to the audio features, the traditional methods for audio emotion analysis need to select proper audio features, e.g., MFCC, energies, flatness, and etc. But they often have to conduct a lot of repeat tests to choose the best features. In order to take full advantage of the depth convolution neural network model in extracting data features, the original features of the data should be kept as much as possible in order to avoid losing information. In this research, we process the audio to spectrogram (Barker & Virtanen, 2016), which is a visual representation of the spectrum of frequencies in a sound. We set the window function to 40ms and the hop size to 20ms to generate a spectrogram every second using short-time Fourier transform with a Hamming window (Allen, 1977). The resulting image is resized to $256 * 256$ pixels, here we also use the method of AlexNet finetune to extract the fc7 layer as a feature of the spectrogram.

Bi-modal visual and audio feature fusion

From the last step we obtained visual and audio features for video emotion analysis. However, the length of features of both visual and audio is long and there maybe many redundancy in the features. It will be helpful if we can combine the two types of features and then reduce the overall dimension.

Let $x_a \in R^D$ denotes audio features and $x_v \in R^D$ denotes visual features, where $D \in R$ is the dimension of audio and visual features, the joint representation of features by fusion modal can be written as:

$$x_f = \alpha_a g(x_a; w_a) + \alpha_v g(x_v; w_v) \tag{3}$$

where $g(.)$ denotes the hidden layer of both audio and visual channel. α_a defines the weights of audio features and α_v defines the weights of visual features at the same time. The hidden layer of audio features is:

$$g(x_a; w_a) = \theta((w_a, x_a) + b_a) \tag{4}$$

where θ denotes the activation function (rectified linear units (Zeiler et al., 2013), sigmoid etc.) of the audio hidden layer. Similarly the hidden layer of visual feature is:

$$g(x_v; w_v) = \theta((w_v, x_v) + b_v) \tag{5}$$

Temporal feature integration and emotion analysis

Though previous steps we have obtained fused features from visual and audio perspective, there is still a challenge about how to predict corresponding emotion status. Furthermore, in previous step the features are about a single frame, taking into account the temporal characteristics of video, it is necessary to study how these features can be used over time. In this research we will use the LSTM model to fuse sequence features together.

Recurrent Neural Networks (RNNs) are powerful networks and it can model input sequences of different lengths, because the parameters of the network can be shared over different parts (Mikolov, Karafiát, Burget, Cernocký, & Khudanpur, 2010). RNNs are often trained by Back-Propagation Through Time (BPTT) algorithm, but the main problem with the BPTT is that the gradients tend to vanish or explode which was resulted by propagating the gradients down through layers. Therefore it is difficult to learn efficient long-term dependencies. To overcome this limitation, the Long-Short-Term-Memory (LSTM) (Hochreiter & Schmidhuber, 1997) units have been created to capture long-term dependencies. LSTMs have the ability to remove or add information to the cell state through a well-designed structure called a "gate". It is believed that the LSTMs can model the temporal aspect of induced emotions in our task. Various units have been proposed in the community to constitute a LSTM. In this research, we employed the LSTM units described in (Zaremba & Sutskever, 2014). The LSTM unit of time step t consists of three sigmoidal gates, i.e., input gate i_t, output gate o_t, forgetting gate f_t. The most important part of the LSTM unit is a linear self-loop state cell c_t. The memory cell unit c_t is a sum of two terms: the previous memory cell unit c_{t-1} which is modulated by f_t, and g_t, a function of the current input and previous hidden state, modulated by the input gate it. h_t denotes the hidden layer's output at step t. We can update our hidden layer for time step t as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{8}$$

$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{9}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{10}$$

$$h_t = o_t \odot tanh(c_t) \tag{11}$$

where x_t is the current fusion feature, h_{t-1} is the previous hidden layer vector. $x \odot y$ denotes the element-wise product of vectors x and y. In addition, W_{xi}, W_{xf}, W_{xo}, W_{xc}, W_{hi}, W_{hf}, W_{ho}, W_{hc} are weights for the gates, and b_i, b_f, b_o, b_c are biases for the gates. σ is the nonlinear methods (*e.g.*, sigmoid or tanh).

The output of the last time step of LSTM unit will be the input of the fully connected neural network, also known as multi-layer perception (MLP).The hidden layers and parameters of MLP will discuss in experiment. The prediction layer

will have 3 units y_l $(l = 0, 1, 2)$ and the class probability is calculated by taking the softmax as below:

$$y_l : p(y_l = c) = \frac{exp(y_l, c)}{\sum_{c' \in C} exp(y_l, c')} \quad (12)$$

where C denotes the three emotion states. Finally the label with the max probability will be the expected label.

Experimental Study

Dataset

In order to fairly verify the performance of our proposed method, we implement it on the dataset provided by MediaEval 2015 Afective Impact of Movies task (Sjöberg et al., 2015), which consists of 10,900 short video clips extracted from 199 Creative Commons-licensed movies of various genres. It is an extension of the LIRIS-ACCEDE dataset (Baveye, Dellandréa, Chamaret, & Chen, 2015), which originally contains 9,800 excerpts extracted from 160 movies.The MediaEval 2015 task added 1,100 video clips additionally from 39 movies. The dataset is divided into training set and test set. The training set consists of 6,144 videos extracted from 100 movies while the test set includes 4,756 videos extracted from the remaining 99 movies. These videos last from 8 to 12 seconds and start and end with a cut or fade. The ground truth for each of 10,900 video clips consists of discrete labels for arousal (calm-neutral-active) and valence (negative-neutral-positive).

Evaluation Metrics & Baseline

In order to evaluate the affective detection task, the official and complete method is global precision (Sjöberg et al., 2015), which is the proportion of the number of correctly assigned videos in the total video samples and is defined as:

$$Precision = N_c / N_t \quad (13)$$

where N_c is the number of videos which are assigned to the correct class, and N_t is the total number of test videos. In this research, we only compare the results obtained for the arousal classification. This is because compared to arousal, valence is not sensitive in the dataset. As such comparing the results of the arousal classification is a commonly adopted choice (Sjöberg et al., 2015).

To evaluate applicability of the model fusion approach, in this research we compared it against the proposed approach in predicting arousal values using only the image features or audio features. Furthermore, we also compared the proposed approach against early fusion and later fusion methods, respectively. In the early fusion model we simply concatenate the audio and video features together, while in later fusion schema, we firstly trained two MLP classifiers to represent the two modalities separately. Their predictions are denoted as p_a and p_t and the overall output emotion class can be assigned by

$$p = \alpha p_a + (1 - \alpha) p_t \quad (14)$$

where α indicates the relevant importance between audio and visual features. In this research we set $\alpha = 0.56$, as indicated in (Goyal, Kumar, Guha, & Narayanan, 2016).

Afterwards we also compare our results against state-of-art systems in the MediaEval 2015 challenge. These systems include: later fusion models with manually selected features (Yi et al., 2015; Chakraborty et al., 2015), early fusion models with manually selected features (P. et al., 2015; Trigeorgis et al., 2015), later fusion models with automatically selected features (Tiwari et al., 2016), early fusion models with automatically selected features (Dai et al., 2015; Seddati et al., 2015).

Experiment Settings

We tested the different feature dimensions and found that the final result did not change much in the range of 250 to 1000. We decided to use feature size of 512 for both visual pathway and audio pathway. Therefore the fusion feature as the final LSTM model input has 512 dimensions. LSTM model can handle different video length, the longest video is 18 seconds that is 18 time steps. The system is trained end-to-end to predict the videos emotion class at each time step. It is found that the most significant parameter is the number of LSTM hidden layers. We compared LSTM networks with 64, 128, 256, and 512 hidden units, separately. Finally, we found that 256 hidden units can be selected to achieve the best results, as shown in Fig. 2. Afterwards we selected MLP as our classifier in which rectified linear units were used as nonlinear functions and stochastic gradient descent with minibatches was used for parameter updates (Zeiler et al., 2013). Also we used categorical cross-entropy loss function to get the best results. The hidden layer uses dropout to prevent overfitting, and the factor is set 0.5. The number of hidden layer of MLP and the units' number can also affect the model results, and ultimately we chose one hidden layer with 64 hidden units.

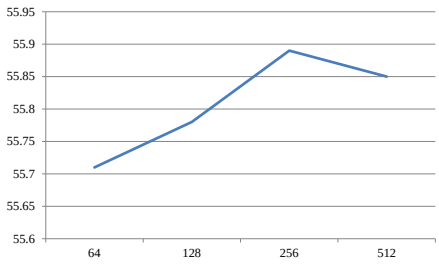

Figure 2: Arousal Accuracy with Different Number of Hidden Layer Units

Result and Analysis

Table 1 presents the proposed method's performance using different feature space and fusion strategy. It is observed that the performance of all feature fusion strategies are better than using only single feature. It is because that video images are the main cause of people's emotions, but audio

can complement the lack of information in video images. It is further found that our proposed model fusion method is better than both the simple early fusion and late fusion, affirming the effectiveness of multi-modal emotion classification. This maybe because early fusion leads to the sparsity of input vectors and late fusion has little consideration for visual and audio's correlation (Williams et al., 2009).

Table 1: Comparison of accuracy by different fusion models

Approaches	Arousal Accuracy(%)
Visual features only	55.51
Audio featues only	55.14
Early fusion	55.71
Later fusion	55.70
Modal fusion	**55.89**

Table 2 is the experimental result of the propose method against most recently revealed results. The result demonstrates the feasibility and superiority of end-to-end training for video emotion classification. It is found that one system's result (Yi et al., 2015) is slightly higher (less than 0.1%) than the proposed one. However, its features are selected manually, which is time-consuming, not universal and not portable. What's worse, their feature dimension is also long. End-to-end training has better transfer learning properties and the training process is convenient. Using a well-trained model for another similar problem only needs a simple refinement. It is also observed from the table that the method proposed in (Tiwari et al., 2016) has the similar feature size to ours, while the proposed model outperforms their final arousal accuracy. This may because their feature fusion approach is rough and does not consider the temporal characteristics. This demonstrates that temporal features could play a role in video emotion analysis to a certain extent. As for the other methods, our result can outweigh them which shows that modal fusion has a great advantage compared with simple early fusion and later fusion. Fusing visual and audio feature in a mid-level is a potential strategy since visual and audio information in video have a certain interaction. It can also inferred that CNN has good performance in visual and audio feature extraction.

Conclusion and Future Work

Video emotion recognition is an important challenge as detecting affective attitudes is an important research field in cognitive science. It is argued that visual and audio information are both important in detecting video emotion. Therefore in this paper we used a deep learning architecture to fuse visual and audio modalities for video affective classification. This end-to-end framework has the advantages of simple training and convenient transplantation and demonstrates that modal fusion with small size of features can compare against most state-of-art results obtained by participants of the MediaEval 2015 Affective Impact of Movies task. Furthermore, it would be interesting to study if it is feasible to include information from other domains/modalities, e.g., abstract words (Siakaluk, Knol, & Pexman, 2014), which deserve future study in the future work.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (No. 61332018), the National Department Public Benefit Research Foundation of China (No. 201510209), and the Basic and Advanced Research Projects in Chongqing (No. cstc2015jcyjA40049).

References

Acar, E., Hopfgartner, F., & Albayrak, S. (2014). Understanding affective content of music videos through learned representations. In *Proceedings of the 20th Anniversary International Conference on MultiMedia Modeling, Part I* (pp. 303–314).

Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, *25*(3), 235-238.

Arifin, S., & Cheung, P. Y. K. (2008). Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *IEEE Transactions on Multimedia*, *10*(7), 1325–1341.

Barker, T., & Virtanen, T. (2016). Blind separation of audio mixtures through nonnegative tensor factorization of modulation spectrograms. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *24*(12), 2377–2389.

Baveye, Y., Dellandréa, E., Chamaret, C., & Chen, L. (2015). LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, *6*(1), 43–55.

Chakraborty, R., Maurya, A. K., Pandharipande, M., Hassan, E., Ghosh, H., & Kopparapu, S. K. (2015). TCS-ILAB - mediaeval 2015: Affective impact of movies and violent scene detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Choe, W., Chun, H., Noh, J., Lee, S., & Zhang, B. (2013). Estimating multiple evoked emotions from videos. In *Proceedings of 35th Annual Meeting of the Cognitive Science Society* (pp. 2046–2051).

Cui, Y., Jin, J. S., Zhang, S., Luo, S., & Tian, Q. (2010). Music video affective understanding using feature importance analysis. In *Proceedings of the 9th ACM International Conference on Image and Video Retrieval* (pp. 213–219).

Dai, Q., Zhao, R., Wu, Z., Wang, X., Gu, Z., Wu, W., & Jiang, Y. (2015). Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Goyal, A., Kumar, N., Guha, T., & Narayanan, S. S. (2016). A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 2822–2826).

Table 2: Comparison with state-of-the-art approaches of MediaEval 2015 Affective Impact of Movies task

Approaches	Fusion method	Feature selection type	Feature length	Arousal Accuracy(%)
(Yi et al., 2015)	Later fusion	Manual	>4,000	55.93
(Trigeorgis et al., 2015)	Early fusion	Manual	1000	55.72
(P. et al., 2015)	Early fusion	Manual	>100,000	51.90
(Chakraborty et al., 2015)	Later fusion	Manual	1,000	48.95
(Tiwari et al., 2016)	Early fusion	Automatic	500	55.85
(Seddati et al., 2015)	Early fusion	Automatic	20,000	52.44
(Dai et al., 2015)	Early fusion	Automatic	10,000	48.70
Proposed approach	Modal fusion	Automatic	512	55.89

Hanjalic, A. (2006). Extracting moods from pictures and sounds: towards truly personalized TV. *IEEE Signal Processing Magazine*, *23*(2), 90-100.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gülçehre, Ç., Michalski, V., Konda, K., ... Bengio, Y. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal of Multimodal User Interfaces*, *10*(2), 99–111.

Kang, H. (2003). Affective content detection using HMMs. In *Proceedings of the 11th ACM International Conference on Multimedia* (pp. 259–262).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of 26th Annual Conference on Neural Information Processing Systems* (pp. 1106–1114).

Levi, G., & Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of 2015 ACM on International Conference on Multimodal Interaction* (pp. 503–510).

Lu, L., Liu, D., & Zhang, H. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(1), 5–18.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of 11th Annual Conference of the International Speech Communication Association* (pp. 1045–1048).

P., M. V., Hayrapetyan, S., Tapaswi, M., & Stiefelhagen, R. (2015). KIT at mediaeval 2015 - evaluating visual cues for affective impact of movies task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Seddati, O., Kulah, E., Pironkov, G., Dupont, S., Mahmoudi, S., & Dutoit, T. (2015). Umons at mediaeval 2015 affective impact of movies task including violent scenes detection. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Siakaluk, P. D., Knol, N., & Pexman, P. M. (2014). Effects of emotional experience for abstract words in the stroop task. *Cognitive Science*, *38*(8), 1698–1717.

Sjöberg, M., Baveye, Y., Wang, H., Quang, V. L., Ionescu, B., Dellandréa, E., ... Chen, L. (2015). The mediaeval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Soleymani, M., Kierkels, J. J. M., Chanel, G., & Pun, T. (2009). A bayesian framework for video affective representation. In *Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction.*

Tiwari, S. N., Duong, N. Q., Lefebvre, F., Demarty, C.-H., Huet, B., & Chevallier, L. (2016). *Deep features for multimodal emotion classification.* Retrieved from `https://hal.inria.fr/hal-01289191`

Trigeorgis, G., Coutinho, E., Ringeval, F., Marchi, E., Zafeiriou, S., & Schuller, B. W. (2015). The ICL-TUM-PASSAU approach for the mediaeval 2015 "affective impact of movies" task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Wang, H. L., & Cheong, L. F. (2006). Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, *16*(6), 689–704.

Williams, S., Oliker, L., Vuduc, R. W., Shalf, J., Yelick, K. A., & Demmel, J. (2009). Optimization of sparse matrix-vector multiplication on emerging multicore platforms. *Parallel Computing*, *35*(3), 178–194.

Xu, M., Chia, L., & Jin, J. S. (2005). Affective content analysis in comedy and horror videos by audio emotional event detection. In *Proceedings of 2005 IEEE International Conference on Multimedia and Expo* (pp. 622–625).

Yi, Y., Wang, H., Zhang, B., & Yu, J. (2015). MIC-TJU in mediaeval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Zaremba, W., & Sutskever, I. (2014). Learning to execute. *CoRR*, *abs/1410.4615*.

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M. Z., Yang, K., Le, Q. V., ... Hinton, G. E. (2013). On rectified linear units for speech processing. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3517–3521).

Zhu, Y., Jiang, Z., Peng, J., & Zhong, S. (2016). Video affective content analysis based on protagonist via convolutional neural network. In *Proceedings of 17th Pacific-Rim Conference on Multimedia, Part I* (pp. 170–180).