A MULTIMODAL MIXTURE-OF-EXPERTS MODEL FOR DYNAMIC EMOTION PREDICTION IN MOVIES

Ankit Goyal[1,2], *Naveen Kumar*[1], *Tanaya Guha*[1,2], *Shrikanth S. Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA
[2] Indian Institute of Technology Kanpur, Kanpur, India
{ankgoyal,tanaya}@iitk.ac.in, komathnk@usc.edu, shri@sipi.usc.edu

ABSTRACT

This paper addresses the problem of continuous emotion prediction in movies from multimodal cues. The rich emotion content in movies is inherently multimodal, where emotion is evoked through both audio (music, speech) and video modalities. To capture such affective information, we put forth a set of audio and video features that includes several novel features such as, Video Compressibility and Histogram of Facial Area (HFA). We propose a Mixture of Experts (MoE)-based fusion model that dynamically combines information from the audio and video modalities for predicting the dynamic emotion evoked in movies. A learning module is presented for the MoE model based on hard expectation-maximization (EM) algorithm. Experiments on a database of popular movies demonstrate that our MoE-based fusion method outperforms popular fusion strategies (e.g. early and late fusion) in the context of dynamic emotion prediction.

Index Terms— Emotion prediction, Multimodal, Fusion, Mixture of Experts

1. INTRODUCTION

Autonomous emotion recognition systems find their place in numerous applications. Computers with the ability to recognize the emotion evoked by media contents can be used to build better human assistive systems [1]. A software capable of recognizing the continuous dynamic emotion evoked by videos can be used in building more personalized video recommendation systems. Emotion recognition systems are very helpful in autonomous video summarization and key event detection tasks [2, 3]. Moreover the emotion profile of a movie, i.e. the continuous dynamic emotion evoked by a movie can be used as hidden layer in predicting outcomes like success and gross income of a movie.

Related Works: Affective analysis in music has been an actively researched area [4] and several well performing systems for predicting emotion in music exist. As compared to emotion prediction in music, emotion prediction in movies is a much more challenging and complex task. In movies, there is a complex interplay between audio and video modalities that determines the perceived emotion. This interaction between modalities is highly dynamic in nature, in the sense that the relative contribution of modalities in emotion prediction may change during the course of the movie. For example let us consider a movie scene which begins with an accompanying musical soundtrack, but the music fades away as the scene proceeds. In such a scene, musical cues would be initially important in setting up the mood, but as we proceed, the visual cues might contribute more to the perceived emotion. Therefore a multimodal framework, which can dynamically captures the interaction between the modalities is necessary for determining the evoked emotion in movies.

Much of the research done in the field of emotion prediction from audio-visual content has focused on accomplishing specific tasks. Chen et al. in [2] were trying to detect violent scenes in a movie. In [5], Nepal et al. focused on automatically detecting goal segments in basketball videos. Recent works have tried to determine categorical emotions in media content. Jiang et al. in [6] and Kang et al. in [7] proposed systems for predicting categorical emotion labels for videos. Some researchers have narrowed their attention on affective analysis of movies from specific genres. Xu et al. analyzed horror and comedy videos by doing event detection using audio cues [8].

A system capable of determining the emotion evoked by a video continuously over time can be very useful in all the above tasks. So in this work, we try to determine emotion evoked by a video by predicting continuous scale and time arousal-valence curves and by validating them against human annotated values. We put forth a list of audio and video features that can be used for the task and also propose several new video features like Video Compressibility and Histogram of Face Area (HFA). We explore different fusion models and show how the complementary information present in audio and video modalities can be exploited to predict emotion ratings. Finally we propose a Mixture of Expert (MoE)-based fusion model that jointly learns optimal fusion weights for the audio and video modalities in a data-driven fashion. We present a learning algorithm based on hard Expectation Maximization (EM) for the MoE-based fusion model.

2. THE DATASET AND THE EXPERIMENTAL SETUP

In the current work we have used the dataset described in [9]. The dataset consists of 12 video clips, each from a different movie and around 30 min long. The movies in the dataset are academy award-winning and are from different genres. For each video clip, there are two curves, one for intended/evoked valence and the other for arousal. These curves vary in range from -1 to 1 and are sampled at a rate of 25 Hz. On doing a frequency response analysis of these curves, we found that more than 99% of their energy was contained in frequencies less than 0.2 Hz. This implies that the emotions vary slowy over time and it is sufficient to sample arousal and valence ratings every 5 sec interval. So we split all movies into non-overlapping 5 sec samples, giving us around 3800 samples with one valence and arousal rating associated with each sample. For each of these samples, we separate the audio and video channels and extract audio and video features from each, as described in Sec.3.

3. FEATURE DESIGN

Various features have been proposed in the literature for the emotion recognition task. In addition to this list of audio and video features we propose two novel features: Video Compressibility and Histogram of Facial Area (HFA). The features used in the current work are described as follows.

3.1. Audio features

Mel Frequency Spectral Coefficients (MFCC) and Chroma: MFCC and Chroma [10] features have been widely used in emotion recognition tasks. We extract MFCC features for each 25 ms window with 10 ms shift and the Chroma features for each 200 ms window with a shift of 50 ms. We also compute Delta MFCC and Delta Chroma features, which are time derivatives of the MFCC and Chroma features respectively. Finally for each sample, we compute statistics (mean,min,max) of the previously mentioned features.

Audio Compressibility: This audio feature was introduced in [11], where it had been shown to be highly correlated human annotated arousal and valence ratings. For an audio clip, the Audio Compressibility feature is defined as the ratio of the size of losslessly compressed audio clip to raw audio clip. We compress the raw audio using the FLAC lossless codec [12] using ffmpeg [13] and include the ratio of the size of the compressed audio clip to original audio clip as a feature.

Harmonicity: The presence of music in a media content helps in triggering emotional response in the viewers [14]. To capture this information, we use Harmonicity feature which was introduced in [15] as a measure of presence of music in an audio clip. We firstly divide a sample audio clip into 50 ms mini clips and extract pitch in each of these 50 ms clip using the aubio tool [16]. Harmonicity for that sample audio clip is then taken as the ratio of number of 50 ms clips that have a pitch to the total number of 50 ms clips.

3.2. Video features

Shot Frequency: Cuts in the video have been widely used by cinematographers to set the pace of action [17]. In order to capture this information, we follow an approach similar to that in [18]. We detect shots, i.e. the sequence of frames recorded continuously by a camera, in the sample video clip using ffmpeg [13] and count the total number of shots present.

Histogram of Optical Flow (HOF) : Motion or activity in a scene affects the emotional response of viewers [19]. For capturing this information we use the HOF feature [20]. First, we extract the Lukas-Kanade Optical Flow [21] for all frames except the ones near a shot boundary. Frames near a shot boundary were excluded because they would exhibit a spurious high optical flow value because of discontinuity. Corresponding to each frame for which optical flow is calculated we construct a 8 bin histogram as follows. For each optical flow vector $[x, y]$ in a frame, we calculate its angle with the positive x axis i.e equal to $\tan^{-1}(\frac{x}{y})$ and find the bin in which it will lie, using the fact that the i_{th} bin represents angles $\in [\frac{(i-1)\pi}{4}, \frac{(i)\pi}{4}]$. Then its contribution to that bin is taken proportional to its L2 norm as$\sqrt{x^2 + y^2}$. We have opted for an 8 bin histogram because it is robust and sufficient for the task. For each sample video clip we then compute statistics of the HOF features across its frames.

3d Hue Saturation Value (HSV) Histogram: Color has a significant influence on the human affective system [22, 23]. This information is captured using the 3d HSV feature. First we convert the frames from RGB to HSV color space. Then for each frame we construct a 3d Histogram as follows. We quantize each of the hue, saturation and value into 4 equal sized intervals. So a pixel has 4 choices for hue, 4 for saturation and 4 for value, and therefore it can lie in any of the $4 \times 4 \times 4$ (64) bins. Finally for each sample video clip, we compute statistics from the 3d HSV Histogram features across all the frames in it.

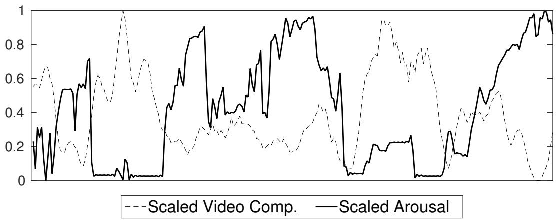


Fig. 1: Plot showing the variation in scaled video compressibility and scaled arousal value for a sample movie

Video Compressibility: Along the lines of audio compressibility, we define a video compressibility feature to capture aspects of motion and change in a video. Most video compression algorithms tend to exploit redundancy in video information over time by using motion and change predictions. As we expect them to be correlated with the perceived emotion ratings, we use video compressibility as a compact feature to combine the effects of such factors over a clip. To calculate video compressibility for a sample video clip, we first compress the raw video with the lossless huffyuv [24] codec using ffmpeg [13] and then calculate the ratio of the size of the compressed video to the original raw video. We have found that the video vompressibility feature has a correlation -0.25 with human annotated arousal values. The p-value for the correlation is 0, asserting that the correlation is significant. We have plotted the variation in scaled arousal values and scaled video compressibility for a movie sample in Fig.1, where we can clearly observe how video compressibility and arousal values vary inversely.

Histogram of Facial Area (HFA): Face closeups have been frequently employed by cinematographers to attract the attention of the audience and evoke aroused emotions [25]. We attempt to extract this information using the HFA feature. We begin by carrying out face detection in all the frames using a Deep Convolutional Network based face detector [26]. Of all the faces detected in a frame, the ones with the largest area is taken as the primary face. For a sample video clip, we detect the primary faces in all the frames. All the frames containing a face are binned according to the primary face area to construct a histogram. We construct a 3 bin histogram, with the bins representing small, medium and large sized faces. Fig.2 shows the formation of HFA for a sample video clip.

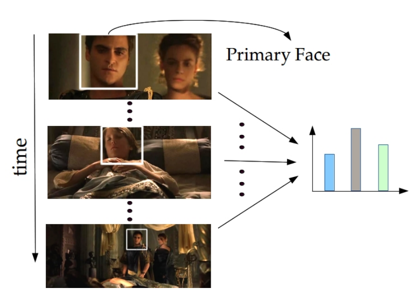


Fig. 2: Schematic representation showing the formation of Histogram of Face Area (HFA) for a sample video

4. SYSTEMS FOR EMOTION PREDICTION

As described in Sec.2, it is sufficient to predict valence and arousal ratings for a movie at an interval of 5 sec. To accomplish this task we split each movie into non-overlapping 5 sec sample clips and extract the above mentioned audio and video features from them. We learn different regression models which try to predict the arousal and valence ratings corresponding to each sample from the extracted features. We perform a leave one movie out cross validation. For all the experiments, we learn independent models for arousal and valence using the sample clips from training movies. These models are then used to predict arousal and valence ratings for every 5 sec sample on the held out test movie. To incorporate the temporal context information, we apply a temporal Gaussian smoothing on the predicted values. This ensures that the predicted value for each clip is consistent with its neighbors. The length of the smoothing window in case of arousal is represented as l_{ar} and in case of valence as l_{vl}. For each model, l_{ar} and l_{vl} are chosen through a grid search on the cross validation sets so as to maximize the performance of that model.

4.1. Audio Only, Video Only and Early Fusion

In the audio only and video only model we try to predict the arousal and valence values using only the audio features or video features. We learn a simple linear regression model to predict the arousal and valence values from the features of each sample clip. We also tried other regression models like Support Vector Regression (SVR) and Gaussian Process Regression (GPR) but there was not much improvement in the prediction so we focused on simple linear regression. In the early fusion model we simply concatenate the audio and video feature vectors and learn a linear regression model using the fused feature vector.

4.2. Late Fusion Model

In the case of the late fusion model, we learn two independent models, one from only the video features and other from only the audio features. Then we try to fuse the predictions from the two models to give the final prediction.

Let $\boldsymbol{y^{(v)}}$ be the prediction from the video features and $\boldsymbol{y^{(a)}}$ be the prediction from the audio features, then final prediction $\boldsymbol{y^{(pre)}}$ is given by Eqn.1. Please note the value of α remains the same for all samples across all the cross validation folds and its value is chosen so as to maximize the correlation between the actual and predicted values. We further analyze the performance of the late fusion model with changing α in Sec.6 using Fig.4.

$$\boldsymbol{y^{(pre)}} = \alpha \boldsymbol{y^{(v)}} + (1-\alpha)\boldsymbol{y^{(a)}} \ , \alpha \in [0,1] \tag{1}$$

4.3. Proposed Mixture of Experts (MoE)-based Fusion Model

In the MoE-based model we have two experts, one that uses only the audio features and the other that uses only the video features. Along with the two experts we also learn a gating function, which determines the contribution from each expert in the final prediction. The final prediction for the MoE-based model is very similar to the Late Fusion model except for the fact that here we do not have a fixed α. The value of α depends on the audio and video features of the current sample. So a MoE-based model can be thought of as comprising of two independent learners and a gating function, where the gating function decides the contribution of each learner as shown schematically in Fig.3.

Let $y_1, y_2, .., y_n$ be our target label for each sample, $\boldsymbol{x_1^{(a)}}, \boldsymbol{x_2^{(a)}}, .., \boldsymbol{x_n^{(a)}}$ be the audio features for each sample, $\boldsymbol{x_1^{(v)}}, \boldsymbol{x_2^{(v)}}, ..., \boldsymbol{x_n^{(v)}}$

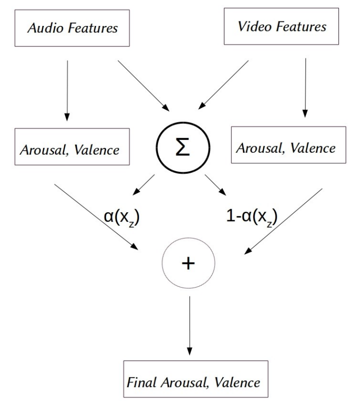


Fig. 3: Schematic representation of Proposed Mixture of Experts (MoE)-based Fusion Model

be the video features for each sample and $\boldsymbol{x_1^{(z)}}, \boldsymbol{x_2^{(z)}}, ..., \boldsymbol{x_n^{(z)}}$ be the features for determining α in each sample. In general one can choose any feature set for $x_i^{(z)}$. In our case we first concatenated all the audio and video features and then did a Principal Component Analysis (PCA) [27] to reduce their dimension. The principal components explaining 90% of the variance were retained in order to construct $x_i^{(z)}$. The predicted label for the i_{th} sample, $y_i^{(pre)}$ is given by Eqn.2.

$$\begin{aligned} y_i^{(pre)} &= \alpha_i y_i^{(v)} + (1-\alpha_i) y_i^{(a)} \\ where\ y_i^{(a)} &= \boldsymbol{\omega_a^\intercal x_i^{(a)}},\ y_i^{(v)} = \boldsymbol{\omega_v^\intercal x_i^{(v)}}, \\ \alpha_i &= \frac{1}{1+e^{-\boldsymbol{\omega_z^\intercal x_i^{(z)}}}} \end{aligned} \tag{2}$$

In order to learn the parameters of the model, we follow an algorithm similar to hard expectation maximization as described next. The loss function for the algorithm, $L(\boldsymbol{\omega_a}, \boldsymbol{\omega_v}, \boldsymbol{\omega_z})$ is given by Eqn.3.

$$\begin{aligned} L(\boldsymbol{\omega_a}, \boldsymbol{\omega_v}, \boldsymbol{\omega_z}) &= \sum_{i=1}^{n} \left\{ y_i - y_i^{(pred)} \right\}^2 \\ &= \sum_{i=1}^{n} \left\{ y_i - \alpha_i y_i^{(v)} - (1-\alpha_i) y_i^{(a)} \right\}^2 \\ = \sum_{i=1}^{n} &\left\{ y_i - \frac{\boldsymbol{\omega_v^\intercal x_i^{(v)}}}{1+e^{-\boldsymbol{\omega_z^\intercal x_i^{(z)}}}} - \left(1 - \frac{1}{1+e^{-\boldsymbol{\omega_z^\intercal x_i^{(z)}}}}\right) \boldsymbol{\omega_a^\intercal x_i^{(a)}} \right\}^2 \end{aligned} \tag{3}$$

The task of the learning algorithms is to estimate parameters $\boldsymbol{\omega_a}, \boldsymbol{\omega_v}$ and $\boldsymbol{\omega_z}$ that minimize the loss function. We adopt a co-ordinate descent approach and subdivide the algorithm into two steps corresponding to the individual learners and gating function respectively. We repeat these steps iteratively till convergence.

STEP I : In this step we fix $\boldsymbol{\omega_z}$ and try to minimize the loss function by estimating optimal values for $\boldsymbol{\omega_a}$ and $\boldsymbol{\omega_v}$. Since $\alpha_i, \forall i$ depends only on $\boldsymbol{\omega_z}$ and $\boldsymbol{x_i^{(z)}}$, they are also fixed in this step.

$$\begin{aligned} &\underset{\boldsymbol{\omega_a}, \boldsymbol{\omega_v}}{\text{minimize}} \sum_{i=1}^{n} \left\{ y_i - \alpha_i y_i^{(v)} - (1-\alpha_i) y_i^{(a)} \right\}^2 \\ &\underset{\boldsymbol{\omega_a}, \boldsymbol{\omega_v}}{\text{minimize}} \sum_{i=1}^{n} \left\{ y_i - \boldsymbol{\omega_v^\intercal} \alpha_i \boldsymbol{x_i^{(v)}} - \boldsymbol{\omega_a^\intercal} (1-\alpha_i) \boldsymbol{x_i^{(a)}} \right\}^2 \\ &\underset{\boldsymbol{\omega_a}, \boldsymbol{\omega_v}}{\text{minimize}} \sum_{i=1}^{n} \left\{ y_i - \begin{bmatrix} \boldsymbol{\omega_v} \\ \boldsymbol{\omega_a} \end{bmatrix}^\intercal \begin{bmatrix} \alpha_i \boldsymbol{x_i^{(v)}} \\ (1-\alpha_i) \boldsymbol{x_i^{(a)}} \end{bmatrix} \right\}^2 \end{aligned} \tag{4}$$

From Eqn.4, it is clear that $\boldsymbol{\omega_a}$ and $\boldsymbol{\omega_v}$ can be found by solving the linear regression problem which tries to predict $y_i \forall i$ using $[\alpha_i \boldsymbol{x_i^{(v)}}\ (1-\alpha_i)\boldsymbol{x_i^{(a)}}]^\intercal$ as the feature vector.

STEP II : In this step we try to minimize the loss function by changing $\boldsymbol{\omega_z}$ keeping $\boldsymbol{\omega_a}$ and $\boldsymbol{\omega_v}$ fixed. This is achieved by following a gradient descent formulation where the gradient is given by Eqn.5.

$$\frac{\partial C}{\partial \boldsymbol{\omega_z}} = 2\sum_{i=1}^{n} \Big\{ [y_i - \alpha_i y_i^{(v)} - (1-\alpha_i)\, y_i^{(a)}] \alpha_i (1-\alpha_i) \left(y_i^{(v)} - y_i^{(a)} \right) \boldsymbol{x_i^{(z)}} \Big\} \quad (5)$$

5. EVALUATION

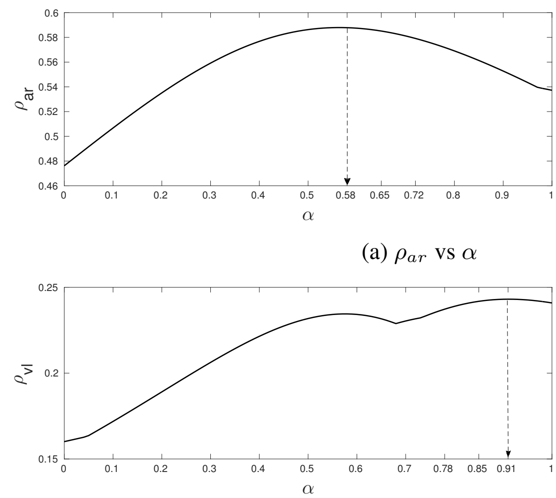
As briefly mentioned in Sec.4, we perform a leave one movie out cross validation to test the performance of different models. For each model we compute the mean absolute correlation between the predicted label and ground truth label for all movies. ρ_{ar} refers to the mean of absolute correlation between predicted arousal and ground truth arousal and similarly ρ_{vl} refers to the mean absolute correlation between predicted valence and ground truth valence. For arousal, the correlations in all cases is positive so mean absolute correlation would be same as mean correlation. But for the case of valence, we do sometimes get negative correlation values. This can be attributed to the fact that unlike arousal, valence involves higher cognitive thinking and similar features elicit opposite valence. For example a fighting scene can evoke very different valence response depending on whether the hero or the villain is dominating or a laughing scene takes opposite sign on the valence scale depending on whether it is the hero or villain who is laughing. The models proposed are unable to capture this aspect of valence and sometimes automatically give inverted prediction for valence, resulting in significant negative correlation with the actual valence value.

Out of the 12 movies in the dataset, 2 are animated movies. Since the video features for animation movies would be very different from usual movies, we have excluded these 2 video from the video and fusion models. We have evaluated the audio model twice once with the animation movies and ones without them. The audio model with animation movies is referred as Audio Only$_1$ and the audio model without animation movies is referred as Audio Only$_2$.

6. RESULTS AND OBSERVATIONS

Table 1 shows the ρ_{ar} and ρ_{vl} value for different models. We have considered the Early Fusion Model, i.e. all audio-video features together with linear regressions as our baseline. It can be seen that fusion models perform better than individual audio or video model. This shows audio and visual modalities contain complementary information and their fusion helps in emotion prediction. Also in all the models the prediction for arousal is better than that for valence. This can be attributed to the fact that valence prediction is far more challenging than predicting arousal as it requires higher semantic and cognitive thinking. The proposed MoE-based fusion model which dynamically adjusts the contribution from audio and video modalities outperforms all other models. Overall, considering the complexity of the task the MoE-based model is does a good job in predicting the valence and arousal curves.

As mentioned in Sec.4, we apply a Gaussian window at the end of each model to incorporate the context information. This increases the agreement between the predicted label and its neighbors. We found that $l_{vl} > l_{ar}$ for all the models, which clearly shows that valence requires a longer context information than arousal. Further we investigated how audio and video modalities contribute to the final prediction in the case of late fusion model. The final prediction in the case of late fusion is given by Eqn.1. We have plotted the change in ρ_{ar} and ρ_{vl} with changing value of α in Fig.4. Please note that $\alpha = 0$ corresponds to Video Only Model and $\alpha = 1$ corresponds to Audio Only Model. From the plots we can see that $\alpha = 0.56$ for the best performing arousal system and $\alpha = 0.91$ for the best performing valence system. We can conclude that in our model audio and video contribute almost equally for arousal prediction but for valence prediction audio contributes more.



(a) ρ_{ar} vs α

(b) ρ_{vl} vs α

Fig. 4: Plots showing the variation in ρ_{ar} and ρ_{vl} with α for the Late Fusion Model

7. CONCLUSIONS

In this paper we addressed the problem of tracking time varying continuous emotion ratings using multimodal cues of media content. We suggested a list of audio and video features suitable for the task, including novel features like Video Compressibility and HFA. We compared and analyzed the performance of audio only, video only and fusion models. Further we proposed a MoE-based fusion model which dynamically fuses the information from the audio and video channels and outperforms the other models. We also presented a hard EM based learning algorithm for the MoE-based model. The MoE-based model in general performs well in the emotion recognition task except sometimes for valence when high level semantic information is required. Future research and development of systems that can capture the semantic information in a video can help in improving the emotion prediction models.

Model	ρ_{ar}	ρ_{vl}
Audio Only$_1$	0.56 ± 0.23	0.24 ± 0.15
Audio Only$_2$	0.54 ± 0.23	0.24 ± 0.15
Video Only	0.49 ± 0.18	0.16 ± 0.12
Baseline (Early Fusion)	0.58 ± 0.17	0.22 ± 0.12
Late Fusion	0.59 ± 0.2	0.24 ± 0.14
Proposed MoE	$\mathbf{0.62 \pm 0.16}$	$\mathbf{0.29 \pm 0.16}$

Table 1: Performance of different models in predicting continuous in time and scale arousal-valence curves

8. REFERENCES

[1] Rosalind W Picard, *Affective computing*, vol. 252, MIT press Cambridge, 1997.

[2] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su, "Violence detection in movies," in *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*. IEEE, 2011, pp. 119–124.

[3] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *Multimedia, IEEE Transactions on*, vol. 15, no. 7, pp. 1553–1568, 2013.

[4] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*, 2010, pp. 255–266.

[5] Surya Nepal, Uma Srinivasan, and Graham Reynolds, "Automatic detection of'goal'segments in basketball videos," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 261–269.

[6] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue, "Predicting emotions in user-generated videos," in *Proceedings of The AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[7] Hang-Bong Kang, "Affective content detection using hmms," in *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003, pp. 259–262.

[8] Min Xu, Liang-Tien Chia, and Jesse Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 4–pp.

[9] Nikos Malandrakis, Alexandros Potamianos, Georgios Evangelopoulos, and Athanasia Zlatintsi, "A supervised approach to movie emotion tracking," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2376–2379.

[10] Takuya Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. ICMC*, 1999, vol. 1999, pp. 464–467.

[11] Naveen Kumar, Rahul Gupta, Tanaya Guha, Colin Vaz, Maarten Van Segbroeck, Jangwon Kim, and Shrikanth S Narayanan, "Affective feature design and predicting continuous affective dimensions from music," in *MediaEval Workshop, Barcelona, Spain*, 2014.

[12] Josh Coalson, "Flac-free lossless audio codec," *Internet: http://flac. sourceforge. net*, 2008.

[13] Fabrice Bellard, M Niedermayer, et al., "Ffmpeg," *Availabel from: http://ffmpeg. org*, 2012.

[14] Gordon C Bruner, "Music, mood, and marketing," *the Journal of marketing*, pp. 94–104, 1990.

[15] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis, "Music tracking in audio streams from movies," in *2008 IEEE 10th Workshop on Multimedia Signal Processing*. IEEE, 2008, pp. 950–955.

[16] "aubio," `http://aubio.org/`, Accessed: 2015-06-30.

[17] Brett Adams, Chitra Dorai, and Svetha Venkatesh, "Novel approach to determining tempo and dramatic story sections in motion pictures," in *2000 International Conference on Image Processing*. IEEE, 2000, vol. 2, pp. 283–286.

[18] Tanaya Guha, Naveen Kumar, Shrikanth S Narayanan, and Stacy L Smith, "Computationally deconstructing movie narratives: An informatics approach," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2264–2268.

[19] Benjamin H Detenber, Robert F Simons, and Gary G Bennett Jr, "Roll em!: The effects of picture motion on emotional responses," *Journal of Broadcasting & Electronic Media*, vol. 42, no. 1, pp. 113–127, 1998.

[20] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *Computer Vision–ECCV 2006*, pp. 428–441. Springer, 2006.

[21] Bruce D Lucas, Takeo Kanade, et al., "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, 1981, vol. 81, pp. 674–679.

[22] Alan Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *Signal processing magazine, IEEE*, vol. 23, no. 2, pp. 90–100, 2006.

[23] Patricia Valdez and Albert Mehrabian, "Effects of color on emotions.," *Journal of Experimental Psychology: General*, vol. 123, no. 4, pp. 394, 1994.

[24] Ben Rudiak-Gould, "Huffyuv v2. 1.1 manual," 2004.

[25] Carl R Plantinga and Greg M Smith, *Passionate views: Film, cognition, and emotion*, Johns Hopkins University Press, 1999.

[26] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep convolutional network cascade for facial point detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 3476–3483.

[27] Svante Wold, Kim Esbensen, and Paul Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1, pp. 37–52, 1987.