CEUR-WS.org/Vol-2328/1_paper_9.pdf

Partners in Crime: Utilizing Arousal-Valence Relationship for Continuous Prediction of Valence in Movies

Tanmayee Joshi*, Sarath Sivaprasad*, and Niranjan Pedanekar

TCS Research, Tata Consultancy Services Limited,
54B Hadapsar Industrial Estate, Pune 411002, India
{tanmayee.joshi, sarath.s7, n.pedanekar}@tcs.com

Abstract. The arousal-valence model is often used in characterizing human emotions. Arousal is defined as the intensity of emotion, while valence is defined as the polarity of emotion. Continuous prediction of valence in entertainment media such as movies is important for applications such as ad placement and personalized recommendations. While arousal can be effectively predicted using audio-visual information in movies, valence is reported to be more difficult to predict as it also involves understanding the semantics of the movie. In this paper, for improving valence prediction, we utilize the insight from psychology that valence and arousal are interrelated. We use Long Short Term Memory networks (LSTMs) to model the temporal context in movies using standard audio features as input. We incorporate arousal-valence interdependence in two ways: 1. as a joint loss function to optimize the prediction network, and 2. as a geometric constraint simulating the distribution of arousal-valence observed in psychology literature. Using a joint arousal-valence model, we predict continuous valence for a dataset containing Academy Award winning movies. We report a significant improvement over the state-of-the-art results, with an improved Pearson correlation of 0.69 between the annotation and prediction using the joint model, as compared to a baseline prediction of 0.49 using an independent valence model.

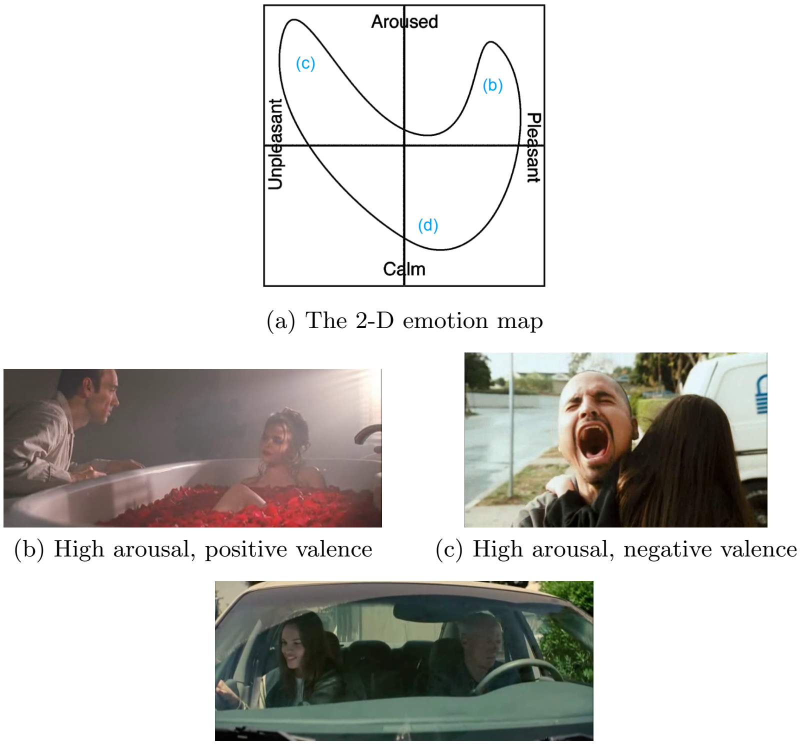
Keywords: Emotion Prediction · Movies · Audio · LSTM.

1 Introduction

Entertainment media such as movies can create a variety of emotions in viewers' minds. These emotions vary in intensity as well as in polarity, and keep on changing continuously with time in media such as movies. A single scene can go from low intensity to high intensity and from positive to negative polarity in a matter of seconds. Such changes are often accompanied by cinematic devices such as variation in music intensity, speech intensity, shot framing, composition and character movements. In addition, static aspects such as scene color tones

* Both authors contributed equally to this work.

and ambient sound also contribute towards setting the polarity of the scene. Prediction and profiling of emotions that movies can generate in viewers finds utility in a variety of affective computing applications. For example, predicted intensity of emotions in a movie can be used to place advertisements. A viewer is likely to pay attention where emotional intensity is low. Similarly, the viewer experience is likely to get adversely affected if one places a happy advertisement after a sad scene. Using such insights, Yadati et al. used motion, cut density and audio energy to predict emotion profile of YouTube videos for optimizing advertisement placement in videos [10]. Additional uses of emotion prediction have been reported for content recommendation [4] and content indexing [12].



(a) The 2-D emotion map

(b) High arousal, positive valence

(c) High arousal, negative valence

(d) Low arousal, neutral valence

Fig. 1: (a) shows the 2-D emotion map as suggested by [6], while (b), (c) and (d) show scenes from the movies *American Beauty*, *Crash* and *Million Dollar Baby*, respectively. They occur in different parts of the 2-D emotion map as shown in (a).

Hanjalic and Xu proposed that emotional content in entertainment media such as movies and videos be modeled as a continuous 2-dimensional space of *arousal* and *valence*, the 2-D emotion map, shown in Fig. 1(a) [6]. *Arousal* is a measure of how intense a perceived emotion is, while *valence* is an indication of whether it is positive or negative or neutral. For example, *excited* is a high arousal and positive valence emotional state, *distressed* is a high arousal and negative valence emotional state, while *relaxed* is a low arousal and neutral valence emotional state. One can find scenes from movies corresponding to such emotional states. For example, a scene from *American Beauty* in Fig. 1(b) shows an *excited* protagonist in a high intensity romantic dream sequence and is located in the top right of the parabolic contour of the 2-D emotion map. Similarly, a high intensity scene from the movie *Crash*, where the character is *distressed* thinking that his daughter is shot, is located on the top left part of this contour. A scene from the movie *Million Dollar Baby* where the protagonist and her coach are taking a *relaxed* car ride is located near the bottom at the centerline.

Continuous prediction of arousal and valence, while important to the aforementioned applications in entertainment, is a challenging task since movies feature a dynamic interplay of audio, visual and textual (semantic) information [5]. Baveye et al. predicted continuous valence and arousal profiles for a dataset of 30 short films using kernel methods and deep learning [1]. Malandrakis et al. predicted continuous valence and arousal profiles using hand-crafted audio-visual features on an annotated dataset of 30 minute clips from 12 Academy Award winning movies [7]. Goyal et al. reported an improvement over these results using a Mixture-of-Experts (MoE) model for fusing the audio and visual model predictions of emotion [5]. Sivaprasad et al. improved the predictions further by using Long Short Term Memory networks (LSTMs) to capture the context of the changing audio visual information for predicting the changes in emotions [8].

A consistent observation across the aforementioned results of continuous emotion prediction has been that the correlation of valence prediction to annotation is worse than that for arousal. This is because valence prediction often requires higher order semantic information about the movie over and above lower order audio visual information [5]. For example, a violent fight scene has a negative connotation, but if the protagonist is winning, it is perceived as a positive scene. Also, a bright visual of a garden may lead to a positive connotation, but the dialogs might indicate a more negative note.

We found that in all aforementioned results for continuous prediction, arousal and valence were modeled separately. Zhang and Zhang suggested that arousal and valence for videos be modeled together [11]. They created a dataset of 200 short videos (5 to 30 seconds) consisting of movies, talk shows, news, dramas and sports. They annotated the videos on a five point categorical scale of arousal and valence. Training a single LSTM model with audio and visual features as input, they predicted a single value of arousal and valence for each video clip.

We believe that for real-life applications such as optimal placement of advertisements, continuous prediction of arousal and valence on longer videos is necessary, unlike prediction over short clips mentioned in [11]. A more useful dataset for this purpose is that created by Malandrakis et al. consisting of 30-minute clips from 12 Academy Award winning movies with continuous annotations for arousal and valence [7]. We found that the state-of-the-art results on this dataset reported a Pearson Correlation Coefficient of 0.84 between predicted and annotated arousal, and that of 0.50 between predicted and annotated valence [8], where arousal and valence models were trained independently. This indicated that the independent arousal model could capture the variation in the dataset much better than the independent valence model. Also, the correlation between annotated arousal and absolute annotated valence was relatively high (0.62) for this dataset. We argued that given the high accuracy of arousal prediction models and the high correlation in annotations, we could use the information learned by the arousal models while predicting valence. Furthermore, if we could incorporate the insight from cognitive psychology that typically arousal and valence values lay within the parabolic contour shown in Fig. 1, then we could further improve valence prediction.

1.1 Our Contribution

Zhang and Zhang used a single joint LSTM model to predict arousal and valence simultaneously. We argued that such a model was not adequate to capture the interdependence of arousal and valence for the continuous dataset. In this paper, we use separate LSTM models for continuous prediction of arousal and valence, but incorporate arousal-valence interdependence in two distinct ways: 1. as a joint loss function to optimize the prediction LSTM network, and 2. as a geometric constraint simulating the distribution of arousal-valence observed in psychology literature. Using these models, we improve the baseline for continuous valence prediction by [8] significantly. Since previous work has reported audio being more important to the prediction of valence [5], [8], we use only audio features as input to our models.

2 Dataset and Features

In this paper, we used the dataset described by Malandrakis et al. [7] containing continuous annotations of arousal and valence by experts for 30-minute clips from 12 Academy award winning movies. The annotation scale for both arousal and valence was $[-1, 1]$. The valence annotation of -1 indicated extreme negative emotions, while that of $+1$ indicated extreme positive emotions. Similarly, the arousal annotation of -1 indicated extremely low intensity, while that of $+1$ indicated extremely high intensity. We sampled the annotations of arousal and valence at 5-second intervals as previously suggested Goyal et al. [5]. We found that previous work reported audio being more important to the prediction of valence [5, 8]. So we decided to only audio features as input to our models. We calculated the following audio features for non-overlapping 5-second clips as described by Goyal et al. [5]: Audio compressibility, Harmonicity, Mel frequency spectral coefficients (MFCC) with derivatives and statistics (min, max, mean), and Chroma with derivatives and statistics (min, max, mean). We further used a correlation-based feature selection prescribed by Witten et al. [9] to narrow down the set of input features.

3 Prediction Model

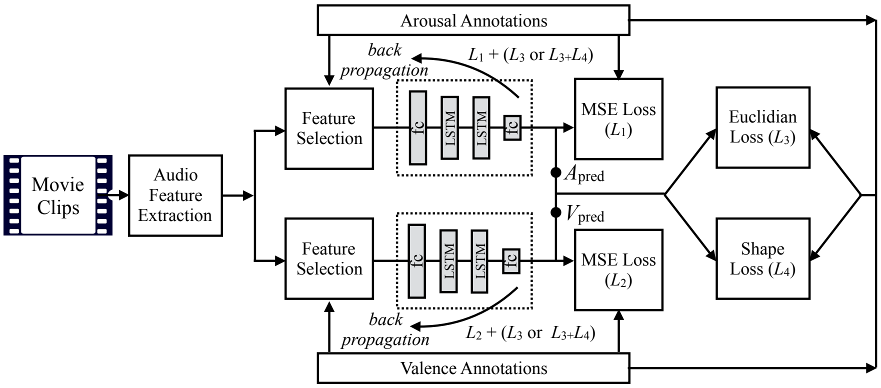


Fig. 2: A schematic diagram for the models employed for continuous prediction of valence.

We implemented a single model as the one mentioned by Zhang and Zhang [11] to predict valence and arousal simultaneously. We found that such a model was not adequately complex to capture the interdependence of arousal and valence, and performed worse that the baseline results obtained by Sivaprasad et al. [8]. So, we decided to model arousal and valence independently, but utilize a joint loss function to train the models thus allowing the interdependence to be modeled.

In particular, we designed one model for independent prediction of valence, and two models to predict valence using arousal information. Fig. 2 shows a generalized schematic representing these models. For all models, we used the LSTM model architecture proposed by Sivaprasad et al. [8], but designed custom losses to incorporate the arousal-valence relationship in two of them. In the basic model architecture (denoted by the dotted box in Fig. 2), two LSTMs were used, first to build a context around a representation of input (audio features) and second to model the context for the output (arousal or valence). The details of the LSTM models used are available in Sivaprasad et al. [8]. We used one versus all validation strategy with 12 folds (one for every movie in the dataset). Because of the inadequacy of data, we did not use a separate validation set. We instead used the second derivative of training loss as an indicator for early stopping of training. To incorporate the arousal-valence relationship, we used different loss functions giving us three different models as described below:

1. **Independent Model** We created two models to predict arousal and valence independently. We used Mean Squared Error (MSE) as a loss function for

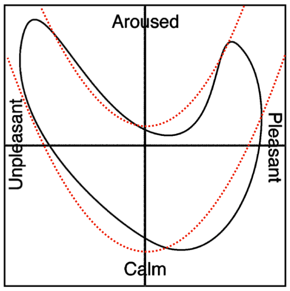


Fig. 3: Indicative parabolas fitted with annotation data on the 2-D Emotion Map for calculating shape loss.

training the arousal and valence models, denoted by L_1 and L_2 in Fig. 2, respectively.

2. **Euclidean Distance-based Model** We first used the independent models of arousal and valence to obtain respective predictions, and then used the independent model weights as initialization for this model. We computed the Euclidean distance between the two points, $P(V_{pred}, A_{pred})$ and $Q(V_{anno}, A_{anno})$, where V_{pred} and A_{pred} are predicted valence and arousal, and V_{anno} and A_{anno} are annotated valence and arousal, respectively. This distance was treated as an additional loss called the Euclidean loss (L_3) while training the LSTM network. We used combined losses to train the models, $(L_1 + L_3)$ for arousal and $(L_2 + L_3)$ for valence. Thus we allowed the Euclidean loss to propagate in both the arousal and valence models ensuring joint prediction.
3. **Shape Loss-based Model** We used the independent models of arousal and valence to obtain respective predictions, and then used the independent model weights as initialization for this model. As can be seen from Fig. 1, the range of valence at any instance is governed by the value of arousal at that instance (and vice versa). It has also been observed that the position of a point in the 2-D emotion map is typically contained within a parabolic contour on this map [6]. We argued that the shape could be described as a set of two parabolas as shown in Fig. 3, one forming an upper limit and another forming a lower limit on the 2-D emotion map. We used annotations of arousal and valence from this dataset as well as from the LIRIS_ACCEDE dataset [2], and fitted these two parabolas as boundaries of convex hulls obtained over the combined datasets. We then incorporated this geometric constraint as an additional loss called the shape loss (L_4) in the prediction model. We measured the distance of point $P(V_{pred}, A_{pred})$ to each of the two parabolas along the direction of line joining $P(V_{pred}, A_{pred})$ and $Q(V_{anno}, A_{anno})$. This distance was computed for both the parabolas and was used as two additional losses to the MSE loss and Euclidean loss described in models 1 and 2. If both predicted and annotated points lay on

the same side of the parabolas, then the shape loss was zero. We used this scheme since considering perpendicular distance of the predicted point to the parabola as an error would not capture the co-dependent nature of arousal and valence. The shape loss was in addition to the Euclidean loss considered above. Thus we used combined losses to train the models, $(L_1 + L_3 + L_4)$ for arousal and $(L_2 + L_3 + L_4)$ for valence.

3.1 State Reset Noise Removal

Because of the inadequacy of data, for training the models, we used a batching scheme where every training batch contained a number of sequences selected from a random movie with a random starting point in the movie. The length of these sequences was 3 minutes, given the typical scene lengths of 1.5 to 3 minutes [3]. We used stateless LSTMs in our models mentioned above, and reset the state variable of the LSTM model after every sequence, since every sequence was disconnected from the other using the aforementioned training scheme. At prediction time, we observed that sometimes these models introduced a noise in the predicted values at every reset of the LSTM (3 minutes). This was similar to making a fresh prediction without knowing the temporal context of the scene, only from the current set of input features. The noise was more noticeable when the model had not learned adequately from the given data. To remove such noise, we made predictions with a hop length of 1.5 minutes, i.e. half the sequence length. Thus we produced two sets of prediction sequences offset by the hop except first and last 1.5 min of the movie. Since the first half of every reset interval was likely to have the reset noise, we used the second half of every prediction and concatenated these predictions to get the final prediction. This scheme enabled a crude approximation of a stateful LSTM. We used this scheme for all three aforementioned models.

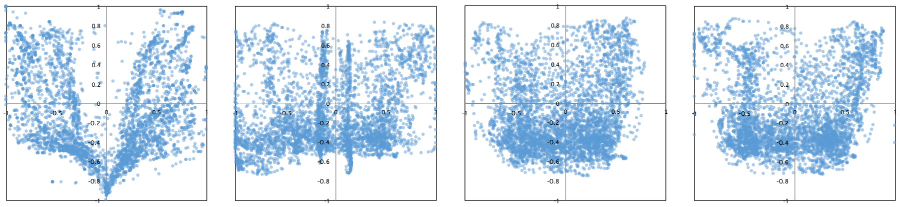
4 Results and Discussion

We treated the valence prediction results obtained by Sivaprasad et al. [8] using only audio input features as the baseline for our experimentation. Table 1 summarizes the comparison of Pearson Correlation Coefficients (ρ) between annotated and predicted valence. We report the following observations:

- We found that Model 3 performed the best of the three models and showed a significant improvement in ρ and MSE over the baseline.
- Fig. 4 shows the 2-D emotion maps for annotations as well as for the three models. We observed that the map for independent models in Fig. 4(b) occupied the entire dimension of valence and did not adhere to the parabolic contour prescribed by Hanjalic and Xu [6]. This was because the independent valence model could not learn enough variations from the audio features. Model 2 with Euclidean loss in Fig. 4(c) could bring the predictions closer to the parabolic contour. Model 3 with shape loss in Fig. 4(d) further improves

Model	ρ_v	M_v	P
Baseline	0.49 ± 0.13	0.24	$--$
Model 1	0.53 ± 0.17	0.27	72.2
Model 2	0.59 ± 0.14	0.12	82.2
Model 3	$\mathbf{0.69 \pm 0.16}$	0.09	87.2

Table 1: A comparison of mean absolute Pearson Correlation Coefficient of valence prediction with annotation (ρ_v), Mean Squared Error (MSE) for valence (M_v) and prediction accuracy for valence polarity (P). The baseline results are from the *model with audio features only* from [8].



(a) Annotation (b) Model 1 (c) Model 2 (d) Model 3

Fig. 4: Comparison of the 2-D emotion maps for different models. On X-axis is valence and on Y-axis is arousal. The ranges for both axes are $[-1, 1]$. Note that Model 1 does not follow the parabolic contours described by [6].

the adherence to the parabolic contour by enforcing the geometric constraints of the parabolic contour.

- Fig. 5 shows the comparison of continuous valence prediction for the movie *Gladiator* for which correlation improved significantly, from 0.33 to 0.84. We observed that models 2 and 3 were much more faithful to the annotation as compared to the independent model. Specifically, we identify two regions in Fig. 5 to discuss the effect of incorporating the arousal-valence interdependence in modeling valence:
  1. Region $R1$ contains a scene that is a largely positive scene featuring discussions about the protagonist's freedom and hope of reuniting with family. The arousal model predicted low arousal for the scene (between 0.0 and -0.4). However, Model 1 predicted it as a scene with extreme negative valence. From Fig 1, we understand that valence can be extremely negative only when arousal is highly positive. The scene had harsh tones and ominous sounds in the background, and independent model predicted it wrongly as negative valence in absence of any arousal information. Model 2 tried to capture the interdependence by predicting a positive valence. Model 3 further corrected the predictions by enforcing the geometric constraint.
  2. Region $R2$ contains a scene boundary between an intense scene where the protagonist walks out victorious from a gladiator fight, and a conversa-

tion between the antagonist and a secondary character. The first part of the scene takes place in a noisy Colosseum with loud background music (high sound energy) and the latter part takes place in a quiet room with no ambient sound (low sound energy). The independent valence model (Model 1) failed to interpret this transition in audio as a change in polarity of valence, as this probably was not a trend seen in other movies in the training set. The independent arousal model interpreted this fall of audio energy as a fall in arousal, which was a general trend in detecting arousal. But this information was available to the valence models 2 and 3, and they could predict the fall in valence accurately. The 2-D emotion map indicates that valence cannot be at an extreme end when arousal is low. Hence both the models with losses incorporating this constraint brought down the valence from extreme positive when arousal fell down.

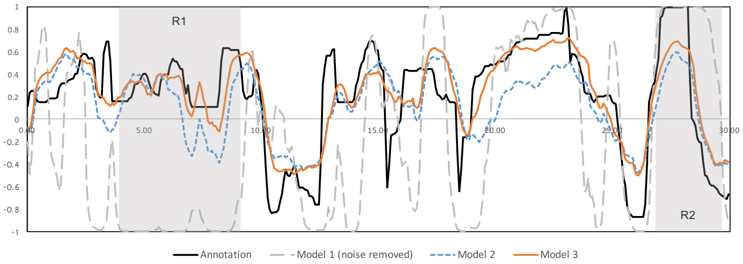


Fig. 5: A comparison of continuous valence prediction for the movie *Gladiator* for different models.

- Predicting polarity of valence is challenging owing to the need for semantic information, which may not always be represented in the audio-visual features [5]. We also calculated the accuracy with which our models predicted the polarity of valence, as summarized in Table 1. We found that Model 3 provided better prediction of polarity (87%) as opposed to Model 1 (72%). Also, the MSE of valence predictions was better for Model 3 (0.09) and Model 2 (0.12) compared to that for Model 1 (0.27). This indicates that incorporating the arousal-valence interdependence better represented polarity as well as value information.
- Fig. 6 shows the improvement in LSTM prediction after the state reset noise correction. We found that this scheme removed the reset noise, seen predominantly as spikes in the prediction without noise removal (the dotted line). This uniformly gave an additional improvement of 0.06 in correlation over the noisy predictions for all valence models. However, for arousal, this improvement was only 0.02, which indicated that the arousal models were already learning well from the audio features.
- We observed that two animated movies in the dataset did not benefit significantly from incorporating the interdependence between arousal and valence.

For *Finding Nemo*, the correlation went down from 0.74 (Model 1) to 0.73 (Model 3), while for *Ratatouille*, it increased slightly from 0.73 (Model 1) to 0.77 (Model 3). We believe this could be because animated movies often use a set grammar of music and audio to directly convey positive or negative emotions. So, the independent model could predict valence using such audio information without the need of additional arousal information.

- There was a slight decrease in performance of the independent model for arousal (correlation of 0.81 for baseline compared to 0.78 for model 1, 0.75 for Model 2 and 0.78 for Model 3). While arousal can be modeled well independently using audio information [8], in our models, it also had to account for the error in valence thus reducing its accuracy. For all practical purposes, we recommend using the independent arousal model (Model 1), as while it gave equal performance to Model 2 or Model 3, it is more robust owing to less complexity.

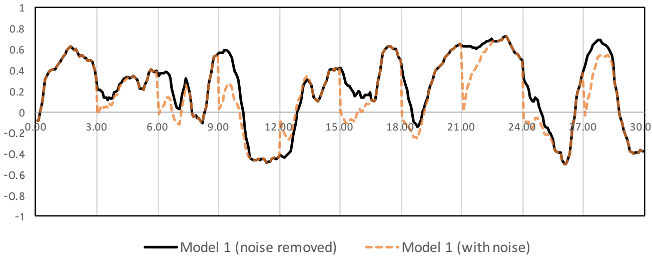


Fig. 6: A comparison of continuous valence prediction with model 3 with and without state reset noise removal for the movie *Gladiator*.

5 Conclusion

In this paper, we proposed a way to model the interdependence of arousal and valence using custom joint loss terms for training different LSTM models for arousal and valence. We used only audio features to model arousal and valence. We found the method to be useful in improving the prediction of valence. We believe that a correlation of 0.69 with annotated values is a practically important result for applications involving continuous prediction of valence.

In future, we would like to improve the accuracy of valence prediction models by utilizing semantic information such as events and characters. We would also like to incorporate scene boundaries to allow LSTMs to learn more complex semantic information such as effect of scene transitions on emotion. This necessitates creation of a larger dataset of continuous annotations for movies. We believe it to be a research direction worth pursuing making use of crowdsourcing, wearables and machine/deep learning.

References

1. Baveye, Y., Chamaret, C., Dellandréa, E., Chen, L.: Affective video content analysis: A multidisciplinary insight. IEEE Transactions on Affective Computing (2017)
2. Baveye, Y., Dellandrea, E., Chamaret, C., Chen, L.: Liris-accede: A video database for affective content analysis. IEEE Transactions on Affective Computing **6**(1), 43–55 (2015)
3. Bordwell, D.: The way Hollywood tells it: Story and style in modern movies. Univ of California Press (2006)
4. Canini, L., Benini, S., Leonardi, R.: Affective recommendation of movies based on selected connotative features. IEEE Transactions on Circuits and Systems for Video Technology **23**(4), 636–647 (2013)
5. Goyal, A., Kumar, N., Guha, T., Narayanan, S.S.: A multimodal mixture-of-experts model for dynamic emotion prediction in movies. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. pp. 2822–2826. IEEE (2016)
6. Hanjalic, A., Xu, L.Q.: Affective video content representation and modeling. IEEE Transactions on multimedia **7**(1), 143–154 (2005)
7. Malandrakis, N., Potamianos, A., Evangelopoulos, G., Zlatintsi, A.: A supervised approach to movie emotion tracking. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 2376–2379. IEEE (2011)
8. Sivaprasad, S., Joshi, T., Agrawal, R., Pedanekar, N.: Multimodal continuous prediction of emotions in movies using long short-term memory networks. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. pp. 413–419. ACM (2018)
9. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)
10. Yadati, K., Katti, H., Kankanhalli, M.: Cavva: Computational affective video-in-video advertising. IEEE Transactions on Multimedia **16**(1), 15–23 (2014)
11. Zhang, L., Zhang, J.: Synchronous prediction of arousal and valence using lstm network for affective video content analysis. In: 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). pp. 727–732. IEEE (2017)
12. Zhang, S., Huang, Q., Jiang, S., Gao, W., Tian, Q.: Affective visualization and retrieval for music video. IEEE Transactions on Multimedia **12**(6), 510–522 (2010)