LEARNING MINIMAL INTRA-GENRE MULTIMODAL EMBEDDING FROM TRAILER CONTENT AND REACTOR EXPRESSIONS FOR BOX OFFICE PREDICTION

Ming-Ya Ko, Jeng-Lin Li, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
myko@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

ABSTRACT

Movie watching is one of the most popular leisure activities in our daily life. The box office revenue, especially in the first week, is critical for financial planning in the movie industry. Most existing movie box office prediction relies on meta data, viewer's comments, and trailer content. However, when viewers are immersed in a movie experience, they would naturally manifest expressions invoked by the media content. In this work, we propose a novel movie box office prediction framework by joint modeling meta attributes, trailer content, and viewer's natural expressions gathered from YouTube reactor videos. The proposed network learns a discriminability-enhanced content and expression embeddings using a minimal intra-genre distance loss function. The proposed architecture achieves 79.07%, 73.79%, and 76.82% for low/high movie box office tier classification (top 30%, top 10%, and top 5%) on a large scale trailer-reactor database. Furthermore, we provide an analysis on the effectiveness of viewer's reaction and our intra-genre projection.

Index Terms— movie trailer, box office prediction, multimedia, multimodal embedding, YouTube reactor

1. INTRODUCTION

The ability to predict box office revenues and the opening success is essential to assist cost planning and prevent potential financial risk in the movie industry. Research has shown that box office from first week accounts for 25% of the total revenue of a movie, and opening weekend sale impacts theater owner's decision to keep a film running [1]. Prior works have mostly used meta attributes of movies, such as genre, budget, and various ratings for profitability prediction. For example, Chang et al., Sharda et al., and Rhee et al. has shown brand-related objective variables (e.g., actor and director star value, genre, MPAA rating, number of screens) can be used to predict box office success [2] [3] [4]; Wallace et al. demonstrates a differential effect that male actors and female actresses have on the movie revenue [5].

As internet and social network become highly integrated in our daily life, additional information about a movie, e.g., it's trailer content and viewer comments, can now be easily gathered and utilized in predicting box office beyond conventional meta attributes. Movie trailers are one of the most effective means for achieving wide exposure across media platforms. These short-duration clips (2-3 minutes) are professionally edited to showcase the most attractive portion of the movie with a purpose to invoke audience's interest. Hou et al. has used visual and audio information of trailers to evaluate user preferences of a movie on YouTube [6]; Tadimari et al. also recently presents a multimodal framework in combining meta data and features extracted from trailers to improve opening box office prediction [7]. In terms of viewer comments, Twitter chatter and blogger reviews have both been shown to be indicative of a movie's box office [8] [9]. Furthermore, internet popularity, e.g., Google search trends and Wikipedia activities, also provides predictive power in assessing a movie's financial success [10] [11].

These kinds of feedback (e.g., social media reviews, popularity counts, and YouTube likes/dislikes, etc), however, are derived from each viewer's *conscious* acts, often with a purpose, instead of their *natural* reaction to the movie's media content itself. Research has shown that when users watch video clips or listen to music, they would experience feeling and emotion that further triggers their external behaviors [12] [13]; in fact, Soleymani et al. has modeled viewer's evoked emotion while watching movie scenes by capturing their physiological signals including EEG response, pupillary response and gaze distance [14]. While it is well-known that when viewers are immersed in the multimedia experience, they would *naturally* react to the content with expressive facial expressions and even physiological responses, these *natural* reactions have not been considered in the context of box office prediction. Only recently, Deng et al. propose a preliminary framework in leveraging audience facial reaction during in-theatre movie showing [15].

Just-in-time behavior modeling during in-theatre showing, while capturing the most immersive viewer experiences, it limits the range and scale of collectable viewer reactions. Nowadays, people actively share information about their feelings and emotions in public on social media. Many renown

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00310

Box Office (in Millions)			
Min	**Max**	**Mean**	**Median**
0.0045	264	63	40
Box Office Cutoff Point (in Millions)			
Our_Top30%	**All_Top10%**	**All_Top5%**	
53	22	35	

Table 1. Statistics of opening box office in our database and breakpoint box office for three target labels.

YouTubers post videos of their own facial expressions as they watch movie trailers. In this work, we integrate this particular source of *user feedback* multimodally with trailer content and movie's meta attributes to improve box office prediction. Specifically, we learn an embedding space each for trailer content and reactor facial expressions using a novel *intra-genre* distance constraint. The intra-genre distance constraint projects features into a space where samples of the same genre would be proximal to each other. This is motivated by the fact that the genre of the movie is a determining factor for the whole trailer atmosphere and the emotion response of the viewers [16], and by introducing this projection, we enhance the discriminability of our multimodal embeddings for the task of low/high movie box office tier classification.

We collect a total of 175 movie trailers with 523 reactor videos of 112 unique movies from YouTube for large scale analyses. Our proposed framework achieves 79.07%, 73.79%, and 76.82% on binary classification task for three different box office tiers (top 30%, top 10%, and top 5%), which is a 6.5%, 10.6%, and 4.2% relative improvement over using conventional meta attributes. We additionally provide an analysis to examine which movie genre would benefit most from our proposed framework. The rest of paper is organized as follows: section 2 describes about the database, features and our proposed network; section 3 details our experiments; finally, section 4 summaries the conclusion.

2. RESEARCH METHODOLOGY

2.1. Movie Trailer-Reactor Video Database

In this work, we collect a database containing movie trailers and reactor videos. Our datasource comes from YouTube Carnatavr T channel[1]. The channel includes many different YouTubers reaction video mashups to each released movie trailer. We collect movie trailers along with 2 to 3 reaction videos for each trailer from the channel. Our database contains a total of 175 movie trailers and 523 reaction videos in total. Our movie trailers come from 112 unique American movies released during the period of 2015 - 2018.

The weekend opening box office figures are obtained from Box Office Mojo website and is adjusted according to inflation. According to the opening box office number distribution of our collected trailer-reactor database, we split our samples into low and high box office classes using 30% percentile as threshold, which makes 78 movies in the low group and 34 movies are in the high group. Besides the threshold determined based on our available database, we additionally set another two thresholds: the top 5 and top 10 percent of all the ranked box office from 13,000 movies on Box Office Mojo to be the cut-off point for our binary classification task. There are 68 movies within Top 10% and 44 movies not in the Top 10%; 64 movies are outside the threshold of top 5% and 48 ones within the threshold. Table 1 lists key statistics and the breakpoints of three cut-off points used in this work. The movie with the maximum box office is *Star Wars: The Force Awakens*, and *The Vault* has the lowest box office.

[1]https://www.youtube.com/channel/UC_00mIU6s9pjR4HGDODaJCw

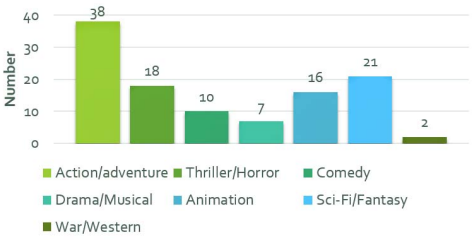

Fig. 1. The distribution of movie genres in our database.

2.1.1. Retrieved Metadata

For each movie, we retrieve the following meta attributes (the same set used in the previous work [7]) collected from Box Office Mojo website.

- Production budget: The total cost for movie production.
- MPAA rating: The Motion Picture Association of America (MPAA) film rating system that is used for evaluating suitability of the movie for certain age audience. The MPAA ratings are given in four different categories: G, PG, PG-13, or R.
- Genre: There are over 10 different categories of genres. A movie may belong to one or two genres. We group similar genres together, and for those genres that have less than 5 movies, they are grouped as "other". The distribution of movie genres used is shown in Fig. 1.
- Release period: The four release periods are: Easter (March and April), Summer (May to August), Christmas (November and December), and Other.

2.2. Trailer Audio-Visual Features

We extract audio and visual features from trailer content that capture movie narrative and emotion-related information similarly to the previous work [7] [17].

2.2.1. Audio Features

We extract eGeMAPS feature set [18] using the OpenSmile toolkit [19]. The eGeMAPS feature set includes 88 dimensional features, which are obtained by computing statistics of

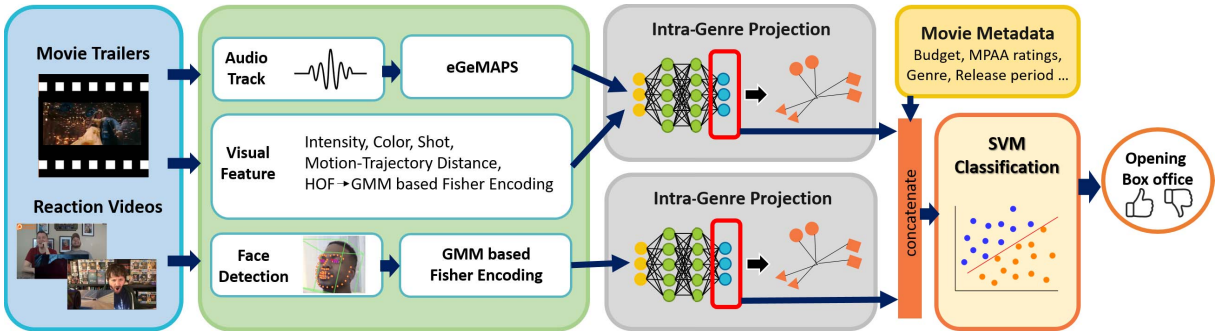

Fig. 2. Illustration of our proposed approach. We extract audio and visual features from movie trailers and reactors, and then we apply an intra-genre projection. Lastly, three different modalities are concatenated and fed in a SVM to perform prediction.

the acoustic low-level descriptors (LLDs). The LLD set consists of frequency, energy related and spectral parameters of acoustic signal computed at a frame rate of 10ms.

2.2.2. *Visual Features*

The psychology of color shows that different colors can affect how people feel [20]. For example, red represents love, energy or blood; white stands for purity and innocence. The color psychology is usually used in film to set a tone of a scene, to trigger the audience's feeling, and to draw focus of attention. The intensity and hue values are two color-based image features that are designed to measure this information. We first compute the average intensity I, which is the frame-wise average value of (R+G+B) for each pixel. Then, we use a normalized 5-bin histogram to represent the average intensity over all the frames in a trailer. Similarly, we convert each image frames from RGB to HSV and represent a trailer using a 12-bin histogram on frame-wise average (H+S+V) values.

In movie, a filmmaker can use frequent cuts and fast motion of actors or camera to create a scene of speed to viewers. The shot changes is another measure that we use in this work. Shot change can be detected using ffmpeg, which computes a change score in a range of 0 to 1 by comparing the difference in pixels between two consecutive frames. It denotes the probability that a current frame is introducing a new scene. For shot changes detection, the threshold of the score is typically set between 0.3 to 0.5. In our database, the average shot length is 7.3 seconds. Since the total length of each trailer varies, we divide the total number of shot changes by the length of trailer in seconds.

To capture the motion activity of trailers, we use a similar approach as the previous work [7], i.e., primarily computed based on dense trajectories to derive motion activity features [21]. Assume in shot n, there are d_n descriptors detected using the method of dense trajectories, and s_n indicates the length of shot n. The i^{th} descriptor has h_i horizontal displacement and v_i vertical displacement. Then, x_i is defined as $\sqrt{{h_i}^2 + {v_i}^2}$, which indicates the moving distance of the i^{th} descriptor. Motion score m_n is then computed as $m_n = \frac{1}{s_n}\sum_{i=1}^{d_n} x_i$ for each shot. We compute the mean, maximum, minimum, and standard deviation of motion scores over a trailer to be used as motion activity features. Additionally, we also use histograms of optical flow (HOF) to capture more granular motion description. For HOF features, we use Gaussian Mixture Model (GMM) based fisher vector encoding to derive the encoded vector of a trailer [22]. When computing optical flow, we delete the trajectories across two different shots to avoid falsely-tracked points.

In summary, we extract three types of visual features from trailers: color-based descriptors, normalized scene changes, and various motion activity encoded descriptors.

2.3. Reactor Expression Features

For the reactor videos, we focus on extracting their facial expressions as they watch each trailer video. We use an open source face processing tool OpenFace [23]. The 2D and 3D facial landmarks, including the contour of the face, eyes, nose and mouth, are extracted as our reactor expression features. The values of facial landmarks are first z-normalized per reactor. Since each reactor video has varying-length frames, we also perform GMM based fisher vector to encode the frame-level features to a fixed-dimension trailer-level representation.

2.4. Intra-Genre Projection

Given that the trailer content and reactor expression can be affected by the genre of the movie genre, we propose to perform *intra-genre* projections on features derived from trailers and reactor videos with a deep embedding network, which groups each of these original feature samples in a new embedding space according to their corresponding genre.

Specifically, given n samples of trailers, the trailer-trailer pairs can be constructed as $\{(t_i, t_j), y_{i,j}\}_{j=1}^{n}$, where i, j are the indices of samples and $y_{i,j} = 1$ indicates that the two trailers are in the same genre, and $y_{i,j} = 0$ stands for genre-mismatched trailers. The probability of matching trailer t_i to trailer t_j is defined as,

$$p_{i,j} = \frac{\mathbf{exp}{t_i}^T \bar{t}_j}{\sum_{k=1}^{n} \mathbf{exp}{t_i}^T \bar{t}_k}, \bar{t}_j = \frac{t_j}{\|t_j\|} \quad (1)$$

Table 2. Summary of box office prediction results. The left half part shows the results of using only single modality feature, and the right half part shows the fusion of multi-modal features. Metadata, movie trailer, and reactor are denoted as M, T, and R. Trailer audio and visual features are represented as Ta and Tv. Features with *intra-genre* projection are denoted with*.

	M	Ta	Tv	Ta+v	R	M+T	M+T*	M+T+R	M+T*+R*
Low	80.77%	74.36%	74.36%	73.08%	66.67%	78.21%	82.05%	80.77%	84.62%
High	67.65%	55.88%	55.88%	58.82%	50.00%	61.76%	70.59%	70.59%	73.53%
Average	74.21%	65.12%	65.12%	65.95%	58.33%	69.98%	76.32%	75.68%	**79.07%**
Not in Top10%	61.36%	43.18%	40.91%	47.73%	52.27%	50.00%	68.18%	56.82%	68.18%
Within Top10%	72.06%	61.76%	67.65%	63.24%	64.71%	66.18%	73.53%	76.47%	79.41%
Average	66.71%	52.47%	54.28%	55.48%	58.49%	58.09%	70.85%	66.65%	**73.79%**
Not in Top5%	76.56%	67.19%	71.88%	68.75%	62.50%	79.69%	81.25%	76.56%	82.81%
Within Top5%	70.83%	58.33%	58.33%	62.50%	60.42%	62.50%	70.83%	70.83%	70.83%
Average	73.69%	62.76%	65.10%	65.62%	61.46%	71.09%	76.04%	73.69%	**76.82%**

where $\bar{t}_j$ indicates the trailer features. The term $t_i^T \bar{t}_j$ means the projection of t_i onto t_j. The larger value represents that two features are closer to each other. The normalized true genre-matching distribution is defined as,

$$q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{n} y_{i,k}}. \quad (2)$$

To minimize the KL divergence between q_i and p_i, the matching loss of associating t_i with correctly matched trailer samples is defined as,

$$L_i = \sum_{j=1}^{n} p_{i,j} \mathbf{log} \frac{p_{i,j}}{q_{i,j} + \epsilon} \quad (3)$$

where ϵ is a small positive value to avoid numerical problems. The genre-matching loss from trailer to trailer for n samples is then computed by,

$$L = \frac{1}{n} \sum_{i=1}^{n} L_i. \quad (4)$$

We implement the same *intra-genre* projection derive a embedding space for the reactor expression features in the same manner as the trailer's multimodal features.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

In this work, we present prediction results on the three different tasks: 1) binary classification of high and low grosses (top 30% within our database), 2) predicting whether a movie would be in top 10% among all movies or not, and 3) predicting whether a movie would be in top 5% or not. Our models are trained by leave-one-movie-out 5-fold cross validation, and the unweighted average recall (UAR) is used for evaluating the performance.

For the feature extraction part, the mixture number for GMM in fisher vector encode of trailer's HOF features and reactor facial landmarks is selected from the set {4, 8, 16, 32, 64} empirically. The architecture of our intra-genre projection includes four dense feedforward layers. The nodes of hidden layers are optimized from the set of {128, 64, 32, 16} for both trailer and reactor projection. We use ReLU activation function in each layer, and learning rate is set as 10^{-4}. We use linear kernel support vector machine (SVM) with ANOVA feature selection to perform the final box office prediction on the concatenation of meta attributes, trailer's audio-visual features, and reactor's expressions. Since there are multiple trailers for each movie, we take the mean of the predicted probability on each trailer to decide whether the movie falls into *high* or *low* categories.

3.2. Results and Analysis

We compare with the following features in this work:

- M: Movie metadata described in section 2.1.1
- Ta: Audio features extracted from trailers
- Tv: Visual features extracted from trailers
- Ta+v: Fusion of audio-visual features from trailers
- R: Visual features extracted from reactors
- M+T: Fusion of metadata and audio-visual features from trailers
- M+T*: Fusion of metadata and audio-visual features from trailers after *intra-genre* projection
- M+T+R: Fusion of metadata, audio-visual features from trailers, and feature extracted from reactors
- M+T*+R*: Fusion of metadata, *intra-genre* projected audio-visual features from trailers, and *intra-genre* projected feature extracted from reactors

Table 2 lists a summary of our experimental results. The multi-modal scheme of using metadata, audiovisual trailer features and reactor features with our proposed intra-genre projection (M+T*+R*) achieves the best UAR over three tasks, which is 79.07%, 73.79%, and 76.82%, in classifying the top 30% movies in our trailer-reactor video database, top 5% and top 10% among all available movies.

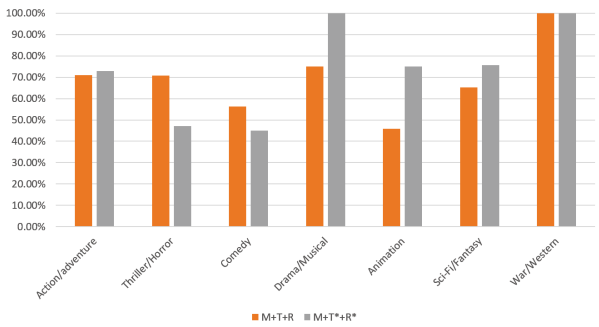

Fig. 3. Effects of projection on box office prediction accuracy

There are several observations to be made. Firstly, the left half part of Table 2 shows the results of using different features as our comparison baseline. We observe a similar trend as previous works that metadata alone provides the most significant impact to movie initial success, which achieves 74.21%, 66.71%, and 73.69% for the 3 tasks. Using trailer content or viewer reactions by themselves is usually insufficient in reliably predicting the box office; specifically, they only achieve approximately 55% to 60% UARs. Secondly, we compare M+T with M+T+R and M+T* with M+T*+R* to evaluate the effectiveness of incorporating reactor's facial expressions. We observe that M+T+R exceeds M+T with 8.15%, 14.74%, and 3.66% relative improvement in the three tasks respectively. Similar boost occurs when comparing M+T*+R* to M+T*, which improves a relative of 3.60%, 4.15%, and 1.03% on the three tasks. These results demonstrate that the facial expression of YouTube reactors, i.e., characterizing their *natural* reactions to the movie trailer, indeed are beneficial to box office prediction.

Lastly, we examine our proposed *intra-genre* projection. The UAR of M+T* improves 9.06%, 21.97%, and 6.96% relative to M+T, and achieves relative improvement of 4.48%, 11.71%, and 4.25% when comparing *intra-genre* projected multimodal representation (M+T*+R*) with the same representation without *intra-genre* projection (M+T+R). The result of M+T*+R* outperforms baseline model using only metadata (a relative gain of 6.55%, 10.61%, and 4.25%).

In summary, multimodally integrating movie meta attributes, trailer content and reactor expressions are important in achieving an improved movie box office prediction. Further, our proposed *intra-genre* projection takes advantage of the fact that genre plays an important factor in shaping the trailer content and viewer response, by appropriately projecting content and expression features onto a minimal *intra-genre* representation space. It effectively mitigates the unwanted variability in the original feature space and, hence, enhances their discriminatory power in box office prediction.

3.2.1. Effects of Intra-Genre Projection

We investigate the effect of our *intra-genre* projection on the box office prediction for each type of genre. Specifically, we show the UARs obtained per genre before and after performing *intra-genre* projection, i.e., comparing M+T+R and M+T*+R* on the best performing task (Top 10%), in Fig.3.

From the Fig. 3, we observe that the improvement of our *intra-genre* projection occurs in genres of drama/musical, animation, and sci-fi/fantasy, which may have been attributed to the distinctive shooting style and similar genre-specific induced emotion of these genres. Hence, when using these projected multimodal trailer content and reactor expression embeddings for box office prediction, the difference between high versus low can be more apparent and less affected by the variability associated with each type genre. However, for genres of thriller and comedy, the predictive accuracy drops slightly. We hypothesize this may be due to the fact that for comedy and thriller trailers and reactor expressions, there exists a *salient* hot spots, not distributed over the viewing experience, during each video that carry predictive information, which are not explicitly modeled in our current framework. However, additional detailed investigation will be required to understand what types of trailer content and reactor expressions are essentially in improving box office prediction beyond movie's own metadata.

4. CONCLUSIONS

In this work, we propose a novel multimodal framework using metadata, trailer content and reactor expressions to improve opening weekend box office prediction. Our framework considers both filmmakers (metadata and trailer content) and audience perspectives (reactor expressions). Specifically, we introduce intra-genre projection to derive both trailer content and reactor expressions embeddings to enhance the prediction accuracy. Our results show that our minimal intra-genre multimodal embeddings achieve promising accuracy, and the *intra-genre* projection improves on genres of drama/musical, animation, and sci-fi/fantasy. To the best of our knowledge, this is one of the first works in leveraging large scale *in-the-moment* natural viewer responses for box office prediction. We plan to extend the scale of the database to include reactors across a wider range of ages, cultures, different preference of movies to further improve the predictive power. At the same time, we will develop modeling approach to characterize the *moment-to-moment* changes of reactors expressions as a function of trailer content to better understand the relationship between content as input and expression as output as viewer immerses in multimedia experiences.

5. REFERENCES

[1] Jeffrey S. Simonoff and Ilana R. Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," *CHANCE*, vol. 13, no. 3, pp. 15–24, 2000.

[2] Byeng-Hee Chang and Eyun-Jung Ki, "Devising a practical model for predicting theatrical movie success: Fo-

cusing on the experience good property," *Journal of Media Economics*, vol. 18, no. 4, pp. 247–269, 2005.

[3] Ramesh Sharda and Dursun Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, vol. 30, no. 2, pp. 243–254, 2006.

[4] T. G. Rhee and F. Zulkernine, "Predicting movie box office profitability: A neural network approach," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2016, pp. 665–670.

[5] W. Timothy Wallace, Alan Seigerman, and Morris B. Holbrook, "The role of actors and actresses in the success of films: How much is a movie star worth?," *Journal of Cultural Economics*, vol. 17, no. 1, pp. 1–27, 1993.

[6] Y. Hou, T. Xiao, S. Zhang, X. Jiang, X. Li, X. Hu, J. Han, L. Guo, L. S. Miller, R. Neupert, and T. Liu, "Predicting movie trailer viewer's like/dislike via learned shot editing patterns," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 29–44, Jan 2016.

[7] A. Tadimari, N. Kumar, T. Guha, and S. S. Narayanan, "Opening big in box office? trailer content can help," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016, pp. 2777–2781.

[8] Gilad Mishne and Natalie Glance, "Predicting movie sales from blogger sentiment," *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 155–158, Jan 2006.

[9] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke, "Predicting imdb movie ratings using social media," in *Advances in Information Retrieval*, Berlin, Heidelberg, 2012, pp. 503–507, Springer Berlin Heidelberg.

[10] S. Asur and B. A. Huberman, "Predicting the future with social media," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Aug 2010, vol. 1, pp. 492–499.

[11] Mrton Mestyn, Taha Yasseri, and Jnos Kertsz, "Early prediction of movie box office success based on wikipedia activity big data," *PLOS ONE*, vol. 8, no. 8, pp. 1–8, Aug 2013.

[12] Hideo Jose, Joemon M. Jose, Roberto Valenti, and Nicu Sebe, "Exploiting facial expressions for affective video summarisation," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, Jan 2009.

[13] J. Kim and E. Andr, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, Dec 2008.

[14] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, Apr 2012.

[15] Z. Deng, R. Navarathna, P. Carr, S. Mandt, Y. Yue, I. Matthews, and G. Mori, "Factorized variational autoencoders for modeling audience reactions to movies," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017, pp. 6014–6023.

[16] Hee Lin Wang and Loong-Fah Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, June 2006.

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

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr 2016.

[19] Florian Eyben, Martin Wllmer, and Bjrn Schuller, "opensmile the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*, Oct 2010, pp. 1459–1462.

[20] Patricia Valdez and Albert Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology*, vol. 123, no. 4, pp. 394–409, 1994.

[21] H. Wang, A. Klser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, Jun 2011, pp. 3169–3176.

[22] Jorge Sanchez, Florent Perronnin, Thomas Mensinka, and Jakob Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, Dec 2013.

[23] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 59–66.