Deep Multimodal Features for Movie Genre and Interestingness Prediction

Olfa Ben-Ahmed
EURECOM, Sophia Antipolis, France
Email: olfa.ben-ahmed@eurecom.fr

Benoit Huet
EURECOM, Sophia Antipolis, France
Email: benoit.huet@eurecom.fr

***Abstract*—In this paper, we propose a multimodal framework for video segment interestingness prediction based on the genre and affective impact of movie content. We hypothesize that the emotional characteristic and impact of a video infer its genre, which can in turn be a factor for identifying the perceived interestingness of a particular video segment (shot) within the entire media. Our proposed approach is based on audio-visual deep features for perceptual content analysis. The multimodal content is quantified in a mid-level representation which consists in describing each audio-visual segment as a distribution over various genres (action, drama, horror, romance, sci-fi for now). Some segment might be more characteristic of a media and therefore be more interesting than a segment containing content with a neutral genre. Having determined the genre of individual video segments, we trained a classifier to produce an interestingness factor which is then used to rank segments. We evaluate our approach on the MediaEval2017 Media Interestingness Prediction Task Dataset (PMIT). We demonstrate that our approach outperforms the existing video interestingness approaches on the PMIT dataset in terms of Mean Average Precision.**

***Index Terms*—Interestingness, Genre, Video, Deep features**

I. Introduction

With the exponentially growing amount of multimedia data available in both public and private media archives and sharing platforms, automatic prediction of relevant and interesting content for specific target users is drawing an increasing attention for many applications such as video recommendation and summarisation [1], movie trailer making [2], digital storytelling [3], indexing and retrieval [4], and social media [5].

Traditional works in this direction have focused on learning an interestingness model using low level features extracted from image/video content directly. However, perceptual content analysis for media interestingness prediction is a challenging task due to the gap between low-level audio-visual features and high-level understanding and perception [1]. Within this context, content-based affective understanding and affect-based media recommendation from multimedia documents open new opportunities in predicting media interestingness [6], [7]. Indeed, Humans are attracted to scenes with strong emotional content and may prefer "affective decisions" to find interesting content because emotional factors directly attract the viewer's attention. As a result, automating the process leading to the identification of interesting content within audio-visual documents might benefit from being able to extract affective cues from the media. Hence, an affect-based representation of media content will be useful for identifying the most important/salient parts. In addition, the genre of the audio-visual document also impacts the choice of what is the most interesting segment (i.e. the scariest scene of an horror movie vs the most sentimental scene of a romantic movie).

In this work, we hypothesize that the emotional impact of the movie genre can be a factor for the perceived interestingness of a video for viewers. The multimodal emotional content of the video inferred by the determination of its genre is quantified within a mid-level representation. We propose to represent each movie segment (or shot) as a distribution over five genres (action, drama, horror, romance and sci-fi). For instance, a high confidence for the horror label inside the shot genre distribution could be interpreted as the highly emotional segment (a very scary one in the case of horror). Therefore, this shot might be more characteristic and therefore more interesting than those with a lower confidence. We propose a deep features-based framework in order to learn the genre-based mid-level representations from deep multimodal low-level features. We exploit both the audio and visual modality of videos using respectively Soundnet [8] and ResNet-125 [9] features thanks to their respective pretrained models. In addition, we investigate the importance of the temporal evolution of the visual content toward genre prediction by aggregating the visual features within a Long Short-Term Memory (LSTM) model [10]. We use the learned representations for the interestingness prediction of video. The remainder of the paper is organized as follows: Section 2 goes through the related work. Section 3 describes the deep mid-level representation approach and its use for video interestingness prediction. Section 4 presents experiments and results on the dataset provided by the organisers of the MediaEval 2017 "Predicting Media Interestingness" task. Finally, Section 5 concludes the work and gives some perspectives for interesting research directions.

II. Related Work

Multimedia interestingness prediction aims to automatically analyze media data and identify the most relevant content. Previous works have been particularly focused on predicting media interestingness from the image content [11], [12], [5], [13]. In contrast, video interestingness analysis has received much less attention.

Recent research in video interestingness prediction has been focused on the use of low-level audio-visual features. Jiang *et al.* [14] evaluated a large number of hand-crafted visual (i.e. SIFT, GIST, HOG) and audio (i.e. MFCC, audio Spectrum) features and their fusion through building models for predicting interestingness. They introduced a new dataset consisting of short videos collected from YouTube and Flickr. The data were trained using Joachim's SVM Ranking algorithm. Their finding was that fusion of audio and visual features together are effective for video interestingness prediction. Yoon *et al.* [15] extended the study of Jiang et al. by incorporating Hidden Markov Models to capture the temporal dimension of emotions in videos. Besron *et al.* [16] merged contextual information (Word2Vec feature) with audio-visual features (Resnet, LSTM, and MFCC features) in a deep learning framework. They concluded that W2V features slightly improved the results. Similar low-level feature representations approaches have been used in [17], [18], [19], [20], [21], [22], [23] with different learning schemes and decision strategies. Alimeida *et al.* [23] reported the best results on the MediaEval 2016 interestingness Task set [24] using multimodal low-level features and learning-to-rank algorithms. However, low-level features allow to describe only visual and audio characteristics of media content, while interestingness is a high-level perceptual concept [11], [1].

In order to bridge this gap between low-level features and the high-level human perception of interestingness, recent works have introduced an intermediate representation between the video features and the video's affective content. These methods construct mid-level representations based on low-level video features and employ these mid-level representations for affective content analysis of videos. They proved that affect-based representation of media data is closer to human perception than the low level description. Acar *et al.* [25] proposed a mid-level representation of multimodal video content for emotional content analysis of professionally edited and user-generated video. The audio and visual representations are automatically learned from raw data using convolutional neural networks (CNNs). The mid-level motion representation is generated using dense trajectory feature vectors. The affective model for emotion analysis is finally learn using the fusion of all the mid-level representations in an ensemble learning framework. According to their obtained results they find that learned audio-visual representations are more discriminative than hand-crafted ones.

Other works have attempted to infer the affective content of videos directly from the related audio-visual features [26]. Acar *et al.* [27] learn both audio and visual feature representations and fuse these representations at decision level for the affective classification of music video clips. Xu *et al.* [28] combined low-level audio-visual representations with higher-level video representations. Movies of different genres are clustered into different arousal intensities (i.e. high, medium, low) with fuzzy c-means using low-level audio-visual features and then the results from the first step (i.e. higher level representations) are employed along together with low-level audio-visual features in order to perform emotional movie classification. Rayatdoost *et al.* [7] proposed a video interestingness prediction approach based on mid-level semantic visual descriptors and deep learning features with extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) in a regression problem. In the current work, we propose a genre-based mid level presentation for video interestingness prediction.

III. Proposed framework

In this section, we present our method for video interestingness prediction. Figure 1 presents a block diagram of the proposed framework including the CNN-based visual features extraction followed by the LSTM-based temporal dynamics modeling and the deep audio features extraction and learning (**Step 1**). The visual and audio feature learning phases are discussed in details in Section III-A. We call "mid-level representation" the genre prediction result (distribution over the 5 genres) obtained for a given video using the trained genre model. The obtained audio and visual mid-level representations (separately or fused) are fed into another classifier to learn and predict interestingness of video segments (**Step 2**) (Section III-B).

A. Deep multimodal mid-level representation

The multimodal mid-level representation of video is obtained through a visual (section III-A1) and an audio (section III-A2) genre prediction approach.

1) Visual mid-level representation: Deep visual features are computed from video frames by extracting features from the global average pooling layer of the Resnet-152 model [9], trained on Imagenet [29]. The obtained features are represented by 2048-D vectors which capture important statics information for the understanding of the scene such as its background and basic objects. However, certain emotions and hence the genre are difficult to grasp, even for humans, from a single image. Indeed, the rhythm of movies is different from one genre to another. For example, an action movie with car chases, fighting and explosion scenes always has a faster tempo. Accordingly, the temporal evolution of video frames may encode additional information which could be useful in making more accurate predictions. In this work, we use the Long Short-Term Memory (LSTMs) architecture to model the temporal dynamics in movies with ResNet features of video frames as inputs. The proposed Resnet-LSTM architecture contains one LSTMs layer with 1024 hidden units for each LSTM block and a fully-connected layer with softmax activation to produce genre prediction. The network is trained with mini-batch Stochastic Gradient Descent (SGD). We use categorical cross-entropy as a loss function. In addition, the learning rate and momentum are set to 0.01 and 0.9 respectively. We use a dropout of 0.5 in the LSTM layer to avoid overfitting. The final output of the Resnet-LSTM block is a 5 dimensional vector representing the probability distribution of the video segment over the genres. This vector is considered as our visual mid-level representation of a movie segment.

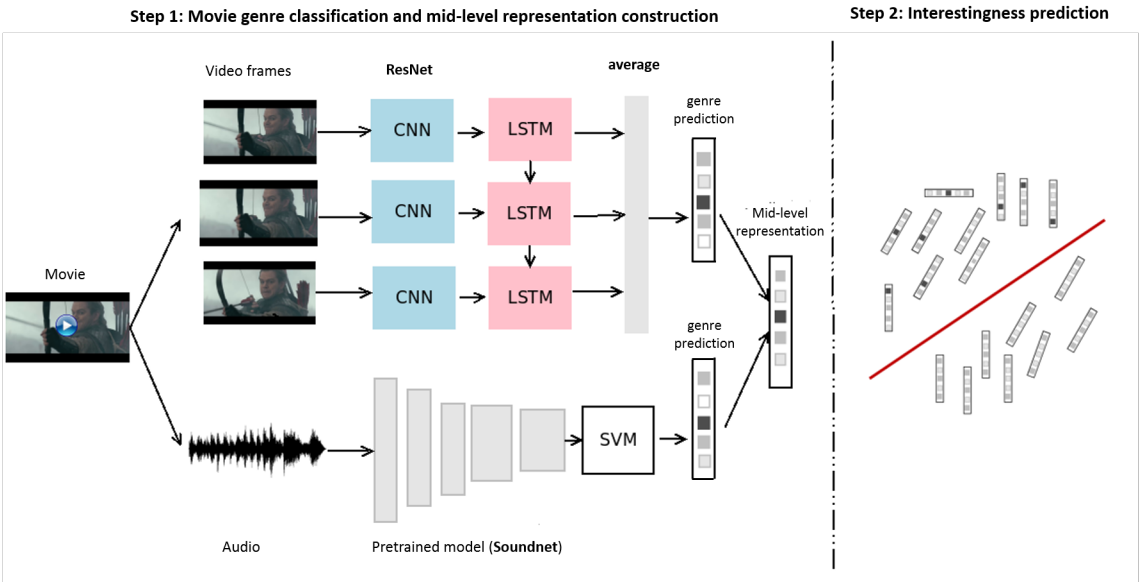

Fig. 1. Interestingness based-genre prediction pipeline

2) Audio mid-level representation: Adding the audio information surely plays an important role for perceptual content analysis in videos. Most of the multimodal approaches in related work (Section II) only focus on hand-crafted audio features, with either traditional or deep classifiers. However, those audio features are rather low-level representations and are not designed for semantic video analysis. Instead of using such classic audio features, we extract deep audio features from a pretrained model; Soundnet [8]. The latter has been learned by transferring knowledge from vision to sound to ultimately recognize objects, scenes and events from sound data. According to *Ayter et al.* [8], an audio feature representation using Soundnet attains state-of-the-art accuracy on three standard acoustic scene recognition datasets. In this work, the audio segment corresponding to each video sequence (or shot) is used to extract deep acoustic features. Finally, an SVM classifier is trained to classify the extracted deep audio features from all the segments (corresponding directly to video shots provided by the task organizers) and produces the probability distribution of the audio segment over the 5 movie genres.

B. Interestingness prediction

In section III-A, we trained the separate genre classifiers (i.e. one based on visual and one based on audio features). Therefore, we end up with two probability vector outputs for visual and audio data respectively. In order to obtain a global genre distribution for the video shots, we average the obtained audio and visual probability vectors (as illustrated in figure 1). For Interestingness prediction, the obtained probabilistic genre distribution is used as mid-level representation to train an Interestingness classifier. A Support Vectors Machine (SVM) [30] binary classifier is trained on these representations to predict, with a confidence score, whether a shot is considered interesting or not.

Consider $\{(x_i, y_i), i = \{1, 2, ..., n\}\}$ where $y_i \in \{1, 0\}$, for "interesting" and "not interesting" classes respectively. The mid-level representation is noted by the vector $x_i \in R^5$. The MediaEval 2017 "Predicting Media Interestingness Task" training data provide a confidence information for each training instance i. Hence, in our learning process, we use those values as weights to increase the probability of correct classification samples with high confidence values.

IV. Experiments and Results

A. Movie genre prediction

1) Movie trailer data for genre prediction: The dataset used to train the movie genre model and thus to build the mid-level representation of movie contains originally 4 different movie genres: *action*, *drama*, *horror* and *romance*. We extended the original dataset [31] with an additional genre (*sci-fi*) to obtain a more sophisticated genre representation for each movie trailer shot. Our final dataset comprises 415 movies trailers of 5 genres (69 trailers for *action*, 95 for *drama*, 99 for *horror*, 80 for *romance* and 72 for *sci-fi*). The movie trailers are segmented into visual shots using the PySceneDetect tool [1]. The final dataset contains 22340 shots in total (14300 for training and 8040 for testing). Before data processing, we removed the first frames of movie trailer that contains potential production logos and credits. The trailer black bars are cropped out. Video shots have an average duration of 3 seconds on this dataset. This is rather short but very much expected from this type of media (i.e. movie trailers).

2) Results and discussion: The deep audio features extracted from the audio segment of the video shot using Soundnet model are then trained using a probabilistic SVM classifier with a linear kernel and a regularization value of

[1] http://pyscenedetect.readthedocs.io/en/latest/

$C = 1.0$. The output of the classifier contains the probability distributions of movies scenes over the genres (Figure 1). Final decisions for the movie trailers classification are realized by a majority voting scheme. A trailer is classified as a particular genre when most of the scenes in it are classified as this genre. In order to evaluate the performance of deep audio features against hand-crafted audio ones, we compare the deep Soundnet features classification performance with MFCC features. Hence, MFCC features are extracted from the audio shots of the trailer using OpenSmile [32] audio feature extraction tool. Since the shots have different duration and therefore a different number of MFCC samples, the number of extracted features is not the same for all the scenes. Therefore, we aggregate the MFCC samples for each shot using a *Bag of MFCC* representation. The obtained audio signatures are fed to the SVM classifier to perform audio-based genre classification. Obtained results are shown in Table I.

TABLE I
GENRE PREDICTION RESULTS WITH DIFFERENT FEATURES

Features/Metric	**Average Precision**	**Average Recall**
Deep Soundnet	62%	57%
Bag-of-MFCC	57%	53%
One Frame-VGG + Soundnet	86%	85%
ResNet-LSTM	87%	85%
ResNet-LSTM+Soundnet	**90%**	**87%**

We obtained for the deep audio features respectively 62% and 57% as average precision and average recall for the movie trailer testing data. We can see that the deep audio features perform better than the MFCC ones (57% and 53%). This may be justified by the fact that deep audio features extracted from soundnet are more adapted for acoustic scene and object recognition and thus for movie genre prediction.

In our work, the fusion of mid-level audio and visual representations (Resnet-LSTM+Soundnet) further improves the performance (Average precision) of genre prediction by 3%. We also note an improvement compared to our previously proposed framework for genre prediction (using VGG features extracted from only one Key-frame and audio extracted with Soundnet) [33]. It can be explained by the fact that our new model uses the temporal dynamic of scenes with more frames, adding interesting additional information for predicting movie genres.

B. Interestingness prediction

1) MediaEval2017 Media Interestingness Prediction Task Dataset (PMIT): The Media Interestingness Dataset contains Interestingness annotations for 103 Hollywood like movies trailers and 4 continuous extracts of ca. 15 min from full-length movies. This data is annotated by human assessors as interesting (class 1) or not interesting (class 0) with a degree of confidence for each annotated sample. It is worth pointing out that the PMIT dataset and the dataset used for genre prediction ([31] extended with sci-fi trailers) are not related. The PMIT dataset is split into development data, intended for designing and training the algorithms which is based on 52 trailers; and testing data based on 26 trailers. The final data for the video is obtained by segmenting the trailers into video shots [34] conducting to 5054 and 2342 shots for respectively training and testing.

TABLE II
VIDEO INTERESTINGNESS PREDICTION RESULTS ON MEDIAEVAL 2017 PMIT TEST DATA

Model	**MAP**	**P@5**	**P@10**	**P@20**	**P@100**
ResNet	0.1913	0.1467	0.14	0.1317	0.089
ResNet-LSTM	0.1991	0.1133	0.1633	0.1756	0.101
Audio	0.1806	0.10	0.15	0.1444	0.0893
ResNet-LSTM+Audio	**0.2122**	0.148	0.1612	0.152	0.103

2) Results and discussion: To evaluate the performance of the proposed interestingness model, we tested several SVM kernels (linear, RBF and sigmoid) with different parameters on the development dataset. Best results, reported in Table II, are obtained with a sigmoid kernel.

Different models have been tested to investigate the contribution of each modality (Visual, Audio and evolution of high level features). Table II presents the obtained results in terms of Mean of Average Precision (MAP) , Precisionn at 5 (P@5), at 10 (P@10), at 20 (P@20) and at 100 (P@100). According to results presented in Table II, adding the temporal evolution of visual features increases the interestingness prediction MAP from 0.1913 to 0.1991. Moreover, adding the audio information improves the results even further to achieve 0.2122 of MAP and 0.103 of P@100.

Figure 2 presents some interesting video shots, from the MediaEval 2017 PMIT dataset and their corresponding genre scores. These examples present, according to the predicted genre, the scariest scenes of an horror movie, the most exciting shot in an action movie or the most sentimental shot in a romantic movie. Hence, the genre-based representation of videos content helps in identifying the most important part of a given video.

Table III presents a comparison of our framework with state-of-the-art methods in video interestingness prediction on the MediaEval2017 PMIT dataset in terms of MAP and MAP@10. MAP@10 is the Mean Average Precision computed over the top 10 best ranked video segments. We use MAP@10 because it reflects well the interestingness prediction problem (i.e. finding a set of pertinent shots representing the entire video) and because it is the official evaluation metric used for the Media Interestingness prediction task [20] so it allows us to perform direct performance comparisons with existing approaches.

The best reported results in the state-of-the-art are based on the use of low-level features to learn an interestingness model. Our approach produces to-date the best reported results on the MediaEval 2017 PMIT with 0.2122 and 0.0841 as MAP and MAP@10 performance. Our work shows that the learned mid-level audio-visual representations are more discriminative and provide more precise results than low-level audio-visual ones proposed in [16] [18] [19] and [17]. Affective media description, which defines the human perception of media

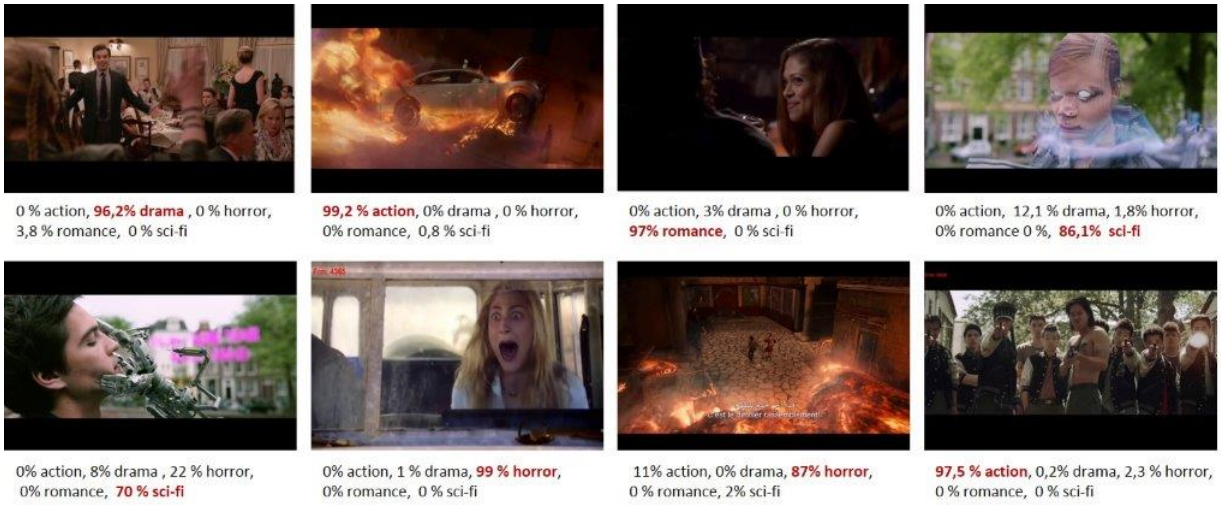

Fig. 2. Examples of interesting video shots from the MediaEval 2017 PMIT dataset with their corresponding genre detection scores

TABLE III
COMPARISON WITH STATE-OF-THE-ART RESULTS REPORTED ON THE MEDIAEVAL 2017 PMIT DATASET

Methods	MAP	MAP@10
Berson et al.[16]	0.1918	0.0609
Berson et al. [16]	0.1878	0.0641
Constantin et al[18]	0.2028	0.0732
Gupta et al. [19]	0.1885	0.064
Shuai et al. [17]	0.1897	0.0637
Our method	**0.2122**	**0.0841**
Baseline [24]	0.1716	0.0564

genre, is more accurate in predicting interestingness compared to the use of low-level subjective features.

V. CONCLUSION

In this paper, we proposed a deep multimodal framework for video shot interestingness prediction based on the genre and affective impact of movie content. We evaluated our approach on the MediaEval 2017 Predicting Media Interestingness Task dataset. We obtained a Mean Average Precision (MAP) of 0.2122 which is to our knowledge the best performance to date on this dataset. This result outperforms state-of-the art approaches which are based on low-level features. Future works include proposing a more integrated method for audio and visual genre recognition and the addition of emotion prediction to further improve video segment interestingness prediction.

ACKNOWLEDGMENT

The Titan Xp used for this research was donated by the NVIDIA Corporation. This work was partially funded by the NexGenTV project under the FUI program (supported by the Banque Publique d'Investissement (BPI)) and the European Unions Horizon 2020 research and innovation programme via the project MeMAD (GA780069).

REFERENCES

[1] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3090–3098.

[2] J. R. Smith, D. Joshi, B. Huet, H. Winston, and J. Cota, "Harnessing a.i. for augmenting creativity: Application to movie trailer creation," in *Proceedings of ACM Multimedia*, October 23-27, Mountain View, CA, USA, 2017.

[3] S. Rudinac, T.-S. Chua, N. Diaz-Ferreyra, G. Friedland, T. Gornostaja, B. Huet, R. Kaptein, K. Lindén, M.-F. Moens, J. Peltonen *et al.*, "Rethinking summarization and storytelling for modern social multimedia," *24th International Conference MultiMedia Modeling, Bangkok, Thailand 5-7 February*, 2018.

[4] P. Le Callet and J. Benois-Pineau, *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Cham: Springer International Publishing, 2017, pp. 1–10.

[5] X. Amengual, A. Bosch, and J. L. de la Rosa, "Review of methods to predict social image interestingness and memorability," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2015, pp. 64–76.

[6] M. Soleymani, "The quest for visual interest," in *Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, October 26-30, 2015*, New York, NY, USA, 2015, pp. 919–922.

[7] S. Rayatdoost and M. Soleymani, "Ranking images and videos on visual interestingness by visual sentiment features," in *Proceedings of the MediaEval 2016 Workshop, Hilversum, Netherlands, October 20-21, 2016*, 2016.

[8] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proceedings of Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[11] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, December 1-8, 2013*, 2013, pp. 1633–1640.

[12] M. Gygli and M. Soleymani, "Analyzing and predicting gif interestingness," in *Proceedings of ACM Multimedia, Amsterdam, The Netherlands, October 15-19, 2016*, New York, NY, USA, 2016, pp. 122–126.

[13] C. Chamaret, C.-H. Demarty, V. Demoulin, and G. Marquant, "Experiencing the interestingness concept within and between pictures," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–12, 2016.

[14] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence, July 14-18*.

[15] S. Yoon and V. Pavlovic, "Sentiment flow for video interestingness prediction," in *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia*, ser. HuEvent '14, New York, NY, USA, 2014, pp. 29–4.

[16] E. Berson, C.-H. Demarty, and N. Duong, "Multimodality and deep learning when predicting media interestingness," in *Proc. MediaEval 2017 Workshop*, 2017.

[17] S. Wang, S. Chen, J. Zhao, W. Wang, and Q. Jin, "Ruc at mediaeval 2017: Predicting media interestingness task."

[18] M. G. Constantin, B. Boteanu, and B. Ionescu, "Lapi at mediaeval 2017-predicting media interestingness," 2016.

[19] R. Gupta and M. Narwaria, "Da-iict at mediaeval 2017: Objective prediction of media interestingness," in *MediaEval*, 2017.

[20] C.-H. Demarty, M. Sjöberg, M. G. Constantin, N. Q. Duong, B. Ionescu, T.-T. Do, and H. Wang, "Predicting interestingness of visual content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 2017, pp. 233–265.

[21] Y. Liu, Z. Gu, Y.-m. Cheung, and K. A. Hua, "Multi-view manifold learning for media interestingness prediction," in *Proceedings of ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, June 6-9, 2017*, New York, NY, USA, 2017, pp. 308–314.

[22] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao, "Interestingness prediction by robust learning to rank," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 488–503.

[23] J. Almeida, L. P. Valem, and D. C. G. Pedronette, "A rank aggregation framework for video interestingness prediction," in *Image Analysis and Processing - ICIAP 2017*, 2017, pp. 3–14.

[24] C.-H. Demarty, M. V. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. Duong, F. Lefebvre *et al.*, "Mediaeval 2016 predicting media interestingness task," in *MediaEval 2016 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2016 Workshop*, 2016.

[25] E. Acar, F. Hopfgartner, and S. Albayrak, "A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 11 809–11 837, May 2017.

[26] S. Wang and Q. Ji, "Video affective content analysis: a survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[27] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *MultiMedia Modeling*, C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. O'Connor, Eds., 2014, pp. 303–314.

[28] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 677–680.

[29] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.

[30] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[31] K. S. Sivaraman and G. Somappa, "Moviescope: Movie trailer classification using deep neural networks," *University of Virginia*, 2016.

[32] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838.

[33] O. Ben-Ahmed, J. Wacker, A. Gaballo, and B. Huet, "Eurecom@ mediaeval 2017: Media genre inference for predicting media interestingnes," in *the Proceedings of the MediaEval 2017 Workshop, Dublin, Ireland*, 2017.

[34] C.-H. Demarty, M. V. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. Duong, F. Lefebvre *et al.*, "Media interestingness at mediaeval 2017," in *Proceedings of MediaEval 2017 Workshop, Dublin, Ireland, September 13-15, 2017*.