Deep Sentiment Features of Context and Faces for Affective Video Analysis

Claudio Baecchi, Tiberio Uricchio, Marco Bertini and Alberto Del Bimbo
Media Integration and Communication Center, Università degli Studi di Firenze
{name.surname}@unifi.it

ABSTRACT

Given the huge quantity of hours of video available on video sharing platforms such as YouTube, Vimeo, etc. development of automatic tools that help users find videos that fit their interests has attracted the attention of both scientific and industrial communities. So far the majority of the works have addressed semantic analysis, to identify objects, scenes and events depicted in videos, but more recently affective analysis of videos has started to gain more attention. In this work we investigate the use of sentiment driven features to classify the induced sentiment of a video, i.e. the sentiment reaction of the user. Instead of using standard computer vision features such as CNN features or SIFT features trained to recognize objects and scenes, we exploit sentiment related features such as the ones provided by Deep-SentiBank [4], and features extracted from models that exploit deep networks trained on face expressions. We experiment on two recently introduced datasets: LIRIS-ACCEDE [2] and MEDIAEVAL-2015, that provide sentiment annotations of a large set of short videos. We show that our approach not only outperforms the current state-of-the-art in terms of valence and arousal classification accuracy, but it also uses a smaller number of features, requiring thus less video processing.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Computer vision tasks*;

KEYWORDS

Sentiment analysis; video analysis; affect recognition

ACM Reference format:
Claudio Baecchi, Tiberio Uricchio, Marco Bertini and Alberto Del Bimbo. 2017. Deep Sentiment Features of Context and Faces for Affective Video Analysis. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 6 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079027

1 INTRODUCTION

In the last few years, popular social media platforms have enabled their users to upload an ever increasing amount of video content, that is shared and personally recommended. As a result, the retrieval of relevant content is becoming more and more difficult, due to the large scale quantity of data and the new need of personalizing the subjective experience of the user. Affective video content analysis, i.e. the process of automatically evince the induced sentiment in the viewer of a video, can be helpful in the process of personalizing the user experience. In fact, users can be attracted to videos that reflect their emotional status or look for some specific emotions.

Understanding the emotion induced by a video is useful in many applications, including the delivery of content personalized on the mood of the user [13], video indexing and recommendation [37], summarization [15] or emotional interfaces for impaired users [9].

The main difficulty of this task resides in the *semantic gap* that arises between the low level features and the human interpretation of a video [20]. This poses a set of unique challenges, requiring the abstraction of human concepts like emotion and affect. Typical semantic concept detection regards the recognition of visual concepts (e.g. "duck", "horse") that simply can be present or not in the visual content. In contrast, recognizing abstract concepts like affection, is difficult because, while they are still related to the visual stimuli, the human response is naturally subjective.

Therefore, a large attention from researchers has been dedicated to study features and their relations to the induction of emotions in viewers. They are often based on work on color in art [14], as well as the results of psychological experiments on emotional

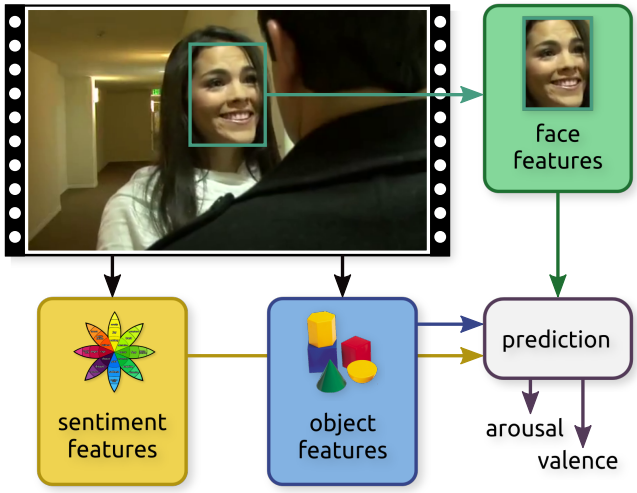

Figure 1: Overview of the proposed method. Prediction of arousal and valence of a video using visual sentiment features related to the context and to the facial expressions.

ICMR '17, June 6–9, 2017, Bucharest, Romania

DOI: http://dx.doi.org/10.1145/3078971.3079027

response to color [31]. Works like [7, 8, 20] are all dedicated to explore different aspects of such features. These efforts have also been encouraged by the MediaEval community that proposed a competition on measuring the affective impact of movies in recent years [29]. On a different line of research, induction of emotions with face expressions are well known in psychological studies. Human faces are known to induce sentiments in people looking at them [32] and, moreover, basic emotions are constants among cultures [10]. Several works have proposed features for recognizing emotion in faces (e.g. [7, 33]), but optimal features for this task are still unclear.

In this work, we address the problem of predicting the sentiment induced on the audience, focusing on classifying the induced sentiment type (i.e. valence) and its strength (i.e. arousal). Figure 1 shows an overview of the proposed approach. We further advance the study of features for affect video recognition by evaluating recently introduced deep sentiment features and showing better performance than those obtainable with standard deep features derived from object classification. We consider the role of actors as an important element in conveying emotions to the viewer. By combining deep sentiment features with face descriptors, we obtain a performance improvement. Compared to previous work, we demonstrate state-of-the-art performance with an effective pipeline on such features, without the many other features usually exploited.

The paper is organized as follows: in Section 2 we describe relevant previous works in detail. In Section 3 we explain the importance of sentiment-related features and describe how we are going to make use of them, and then, in Section 4, we report the experimental results of the evaluation of these features on two recent datasets, also used in the MediaEval contest. Finally, in Section 5 we give our conclusions of the presented work.

2 RELATED WORK

Visual sentiment classification has received an increasing attention among the scientific community in the last few years. Studies have been extensively conducted on image analysis, as in SentiBank [3] and in its improved version Deep-SentiBank [4] that makes use of deep neural networks. They have been recently extended to the multi-lingual context [16] and to image+text sentiment analysis, as in [1, 34]. However, studies on video are still lacking. Video sentiment analysis is a wide task, encompassing different types of emotion understanding problems. The great majority of works revolve around studying the sentiment expressed by a speaking actor. These works focus on interpreting words, voice and expressions performed in the video. The Acted Facial Expressions in the Wild (AFEW) [6] dataset has been used in several emotion recognition challenges. It consists of 1,426 short sequences (with an average length of 2 seconds) extracted from movies, containing facial expressions, annotated with 6 emotions as identified by Ekman *et al.* [11], plus a neutral class. The baseline uses LBPTOP descriptors and SVR [7] while, typically, solutions that achieve the best results combine audio (e.g. MFCC) and video (e.g. CNN trained on faces) features; for instance, recurrent neural network (RNN) and 3D convolutional networks (C3D), specifically trained on faces, have been combined with audio features by Fan *et al.* [12]. Wöllmer *et al.* [33], try to understand the speaker's sentiment in on-line videos containing movie reviews by leveraging acoustic, visual and linguistic features. Rosas *et al.* [24] use a similar approach to classify the speaker's emotion in Spanish videos. Sentiment analysis is performed on both sentiment polarity and strength by Zadeh *et al.* [36], where they aim at mining opinions in YouTube videos.

Another type of sentiment analysis is the one performed on the audience, that, while watching a video, is induced different kind of emotions. Dumoulin *et al.* [8] have proposed a hierarchical approach for affect classification, using a combination of audio and visual features, including convolutional neural networks for content and sentiment recognition. The proposed method has been tested on the FilmStim dataset [26], that contains 70 movie excerpts annotated with 7 emotions. The task of categorizing the induced sentiment of videos has been thoroughly studied in the MediaEval 2015 challenge [29], extending a new large scale dataset called LIRIS-ACCEDE [2]. Here participants used computer vision features to classify induced emotions. For example, Mironica *et al.* [22] used ColorSIFT and AlexNet's FC6 descriptors. Seddati *et al.* [27] also took Optical Flow into consideration to improve the overall accuracy using motion information. Vlastelica *et al.* [21] used GIST features in combination with HOG and HOF features and AlexNet's FC7 descriptors. Trigeorgis *et al.* [30] used eGeMAPS audio features together with CNN features to improve the classification power of their model. Qi Dai *et al.* [5] also used audio features, specifically the MFCC descriptor, in conjunction with a set of CNN features, LSTM features and texture information. YunYi *et al.* [35] made also use of the Fisher Vector representation of various local spatial and temporal features. Vu Lam *et al.* [19], combining most of the previous features, provided the best classifier in the contest for sentiment classification. Zhu *et al.* [38] proposed to base affective video content analysis on the presence of an actor to identify the most important keyframes and then extract patch features of the scene using CNNs trained on ImageNet, and fusing with audio features such as MFCC. They do not consider a description of the facial expressions of the actors. The authors evaluate the performance of the method on LIRIS-ACCEDE dataset.

Our work addresses the problem of classifying the induced sentiment in the viewer in terms of valence and arousal. Differently from all these works that have used the LIRIS-ACCEDE dataset, we propose to use sentiment-oriented features in addition to the classical computer vision features. Inspired by the studies on sentiment induction from faces, we also specifically consider the emotion showed in faces of the actors, shown for example in close-up and medium shots. In particular we add visual sentiment features from Deep-SentiBank and combine it with descriptions of the facial expressions of the actors, using CNNs trained for face recognition, to better capture the details of the faces.

3 EXPLOITING SENTIMENT-RELATED FEATURES

Given a short video V, the task of affective video recognition is to classify the global valence and arousal of the sequence from its frames. We address the problem using a pipeline based on two steps, feature extraction and classification. We consider three types of features: *sentiment features*, *face features*, for which we consider faces extracted using a face detector, and *object features*. Each feature type is first treated independently. They are extracted from every

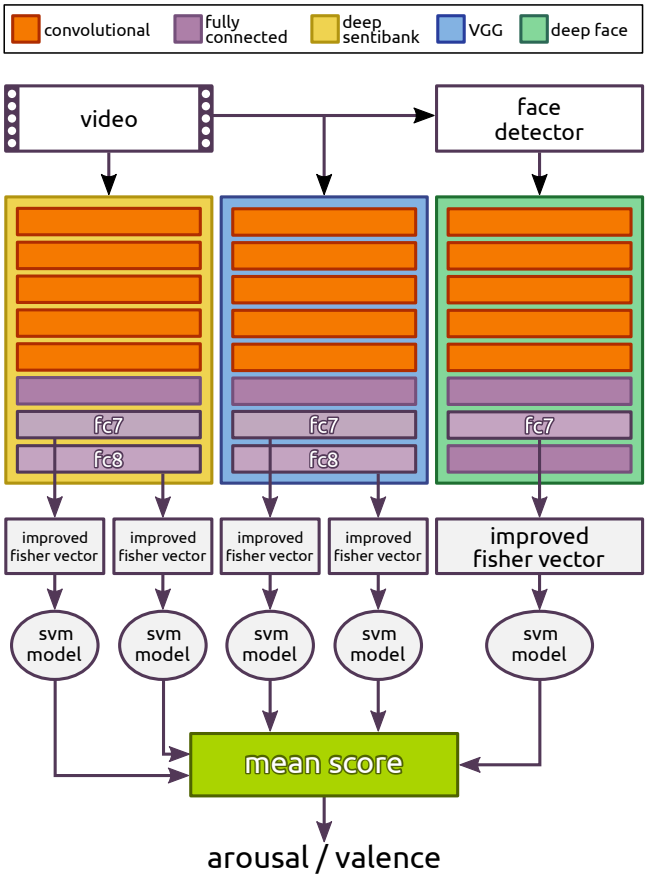

Figure 2: Pipeline of the proposed method. Each feature set is encoded separately into an IFV and a prediction is made by performing late fusion of any subset of the classifiers.

frame and singularly pooled into a global representations of the sequence. Each global representation is used to train a prediction model and lately combined with a late fusion approach. Figure 2 shows the complete pipeline, including each feature extraction and the final classification process.

Sentiment Features. In order to capture sentiment-related information we choose Deep-SentiBank [4] as sentiment related feature extractor. Deep-SentiBank is a deep convolutional neural network trained to discover Adjective-Noun Pairs (ANP) from images and its features can be used as statistical cues to detect emotions. We processed each video $V = \{f_1, f_2, \ldots, f_v\}$ composed of v frames and for each frame f_i we extracted both FC7 and FC8 descriptors from Deep-SentiBank, which we will refer to as Sent-FC7 and Sent-FC8, obtaining two sets of descriptors $D = \{d_1, d_2, \ldots, d_v\}$ and $R = \{r_1, r_2, \ldots, r_v\}$ respectively. To overcome the problem of handling video with different durations and to include the time information into the system, all Sent-FC7 and Sent-FC8 of a video are encoded into an Improved Fisher Vector (IFV) [25] representation $\mathcal{I}$ using a Fisher Encoding $\mathcal{F}$. Each feature type is treated separately to avoid mixing different contributions into a single representation. For each feature type d and r, we first performed an estimation of a Gaussian Mixture Model (GMM) using 32 components obtaining $\mathcal{G}_d$ and $\mathcal{G}_r$. These are then used to encode the features of a video obtaining a single descriptor, that is, $\mathcal{I}_D = \mathcal{F}(D, \mathcal{G}_d)$ and $\mathcal{I}_R = \mathcal{F}(R, \mathcal{G}_r)$.

Face Features. Face related features are extracted using a similar approach to sentiment features. Inspired by the work of Parkhi *et al.* [23] we trained an AlexNet network [18] on the Oxford face dataset provided in the same paper, structuring the problem as an $N-$classification task. The dataset is composed of 2.622 identities which are used to train the deep classifier. We will refer to this network as *deep face*. First we performed face detection using Dlib toolkit [17], extracting for each video a set $A = \{a_1, a_2, \ldots, a_v\}$ of faces. For videos where faces have been detected, similarly to Sent-FC7 and Sent-FC8, we first extracted FC7 descriptors from our *deep face* network for each face of a video, obtaining a set of descriptors $C = \{c_1, c_2, \ldots, c_v\}$, then we estimated a GMM using 32 components obtaining $\mathcal{G}_c$. This model is used to encode the features into an IFV, that is, $\mathcal{I}_C = \mathcal{F}(C, \mathcal{G}_c)$.

Object Features. Since the scene context and objects can be associated to sentiments, we also add object feature information. To this extent we employ features extracted using VGG-16 network [28], taking its FC7 and FC8 descriptors, that we will refer to as VGG-FC7 and VGG-FC8. Following the same procedure used for the visual sentiment features, we extract two sets of features $P = \{p_1, p_2, \ldots, p_v\}$ ans $Q = \{q_1, q_2, \ldots, q_v\}$, we encode them using two GMM $\mathcal{G}_p$ and $\mathcal{G}_q$, and finally we produce the IFV $\mathcal{I}_P$ and $\mathcal{I}_Q$.

Score Prediction. Sentiment prediction is performed separately per feature vector $\mathcal{I}$. We trained five linear SVM, one for each feature, producing five distinct models that can be grouped in three sets: two models for sentiment features; *ii)* one model for facial expression features; *iii)* two models for object features. By computing a separate model for every feature and performing late fusion on the classifier scores we are able to study every feature combination. This way it is possible to study the contribution of each feature, appreciating if one or more share the same information or give different types of contributions to the final prediction. In case a feature is not available for a video, such as in the case of videos where no faces have been detected, no contribution is given for that feature and the score is computed by averaging only the available ones. Prediction is performed distinctly for valence and arousal.

4 EXPERIMENTAL RESULTS

Here we report results of two experimental settings on sentiment annotated videos. First, we will show the effectiveness of single sentiment features for valence and arousal classification tasks. Then, we will report experiments on the combination of two and more features, showing that faces always give a positive contribution, supporting our hypothesis that actors' facial expressions are an important element in conveying emotions to the viewer.

4.1 Datasets

Each year the MediaEval Benchmarking Initiative [1] proposes a number of tasks to be addressed about a variety of multimedia topics. One of these is the affective impact of movies, first introduced in the

[1] https://www.multimediaeval.org/

Figure 3: LIRIS-ACCEDE sample frames: negative *(left)*, neutral *(center)*, positive *(right)*. First row *(top)* shows scene-related samples, second row *(bottom)* shows face-related ones.

MediaEval 2015 challenge [29], that proposes various sentiment-related tasks to be evaluated on a new common dataset, the LIRIS-ACCEDE dataset [2]. Examples of frames from the LIRIS-ACCEDE dataset are shown in Figure 3, we show examples for each class and highlight frames with and without faces with a strong sentiment value. This dataset is the first that directly addresses the problem of affect evaluation of viewers on movies taken from social media. It was in fact created using 160 movies taken from the Vimeo platform [2]. From these movies, 9,800 excerpts of around 10 seconds have been extracted and discretely classified for valence, arousal and violence. The dataset is divided into a training and a test set, both consisting of 4,900 excerpts.

All videos have been labeled using Amazon Mechanical Turk and fall in one of three categories: negative, neutral or positive for valence classification and calm, neutral or aroused for arousal classification. In the occasion of the MediaEval 2015 contest, the LIRIS-ACCEDE dataset has been enriched with 39 additional videos, annotated in the same manner, from which 1,100 new excerpts have been extracted. These videos, together with the original 9,800 form a new dataset composed by 10,900 videos, referred from now on as the MEDIAEVAL-2015 dataset. The excerpts are divided into two sets, a training set, composed of 6,144 elements from the LIRIS-ACCEDE dataset, and a test set, composed of the remaining 3,656 elements of the LIRIS-ACCEDE dataset and the new 1,100 excerpts introduced by the MEDIAEVAL 2015 challenge, for a total of 4,756 elements. While the valence score are well balanced, we note that arousal values are biased towards the class calm. To address this, we set SVM label weights proportional to the inverse ratio of the training examples for each class.

Following the metrics used in the MediaEval contest [29], we evaluate the performance in terms of accuracy, and compare the proposed method with several state-of-the-art methods.

4.2 Single Feature

To test the discriminative capacity of our features we first use each of them alone. Table 1 reports the results for the task of valence and arousal classification on the MEDIAEVAL-2015 dataset. Considering the Valence column, we observe that sentiment-based CNN features Sent-FC7 and Sent-FC8 have higher accuracy than those obtained from object-based features. This confirm the importance of using networks trained for the specific task at hand. Comparing our single-feature results *(top)* to the ones obtained during the MediaEval 2015 contest *(bottom)*, we can see that both results that make use of Deep-SentiBank features are very close, if not the same, to the state-of-the-art result, which is a remarkable outcome considering that Vu Lam [19] is using a combination of nine features whether our method is using only one feature.

Results for the task of arousal classification are reported in the last column of Table 1. Similarly to the valence classification task, we can observe that sentiment-based CNN features outperform object-based features, and obtain results that are comparable to several of the competing methods that use more features.

Results on the LIRIS-ACCEDE dataset only, to allow the comparison with the method of Zhu *et al.* [38], are reported in Table 2, showing behaviors similar to the MEDIAEVAL-2015 setting. We observe that Sent-FC7 and Sent-FC8 are slightly superior to the Faces feature for both valence and arousal, as seen also in Table 1. They all outperform [38], suggesting that our method, exploiting sentiment features and pooling on all frames, is more powerful than generic CNN features on few frames.

4.3 Feature Fusion

Following the trend of fusing multiple features to incorporate more information, we performed late fusion of all five features to prove that they contain different information. To this end we performed a late fusion of the scores of up to five classifiers, weighting them in equal manner. We performed all possible combinations of fusion between two of the three sentiment features and in the end we combined all the three sentiment features together to further improve the accuracy. Finally we combined sentiment feature with object-based features, using a total of five features. In the Valence column of Table 1 we report results for the related task using the MEDIAEVAL-2015 dataset. Figures reported in the *(middle)* part

[2]https://vimeo.com/

Data	Feature Type	Valence Accuracy	Arousal Accuracy
Single Features	Random	33.29%	33.91%
	VGG-FC7	40.16%	49.14%
	VGG-FC8	40.92%	49.33%
	Sent-FC7	42.14%	52.31%
	Sent-FC8	42.72%	52.29%
	Faces	40.37%	51.96%
Two Features	Sent-FC7 + Faces	43.67%	52.84%
	Sent-FC8 + Faces	44.24%	52.90%
	Sent-FC7 + Sent-FC8	43.50%	53.11%
Three Features	Sent-FC7 + Sent-FC8 + Faces	44.68%	54.52%
Five Features	VGG-FC7 + VGG-FC8 + Sent-FC7 + Sent-FC8 + Faces	**45.31**%	**55.98**%
Other Methods (Contest)	Mironica [22] *(3 features)*	36.10%	45.04%
	Seddati [27] *(1 features)*	37.20%	52.44%
	Vlastelica [21] *(5 features)*	38.50%	51.90%
	Trigeorgis [30] *(2 features)*	41.40%	55.72%
	Qi Dai [5] *(3 features)*	41.80%	48.80%
	YunYi [35] *(7 features)*	41.90%	55.93%
	Vu Lam [19] *(9 features)*	42.90%	55.91%

Table 1: Experimental results on the MEDIAEVAL-2015 dataset. *(top)* our results showing single-feature and feature-fusion accuracies; *(bottom)* other methods that participated in the MediaEval 2015 contest.

Data	Feature Type	Valence Accuracy	Arousal Accuracy
Single Features	Random	33.61%	33.92%
	VGG-FC7	41.04%	45.00%
	VGG-FC8	40.94%	44.43%
	Sent-FC7	42.16%	47.55%
	Sent-FC8	41.61%	47.12%
	Faces	40.24%	45.71%
Two Features	Sent-FC7 + Faces	44.53%	48.41%
	Sent-FC8 + Faces	42.40%	49.32%
	Sent-FC7 + Sent-FC8	43.16%	49.52%
Three Features	Sent-FC7 + Sent-FC8 + Faces	45.19%	52.24%
Five Features	VGG-FC7 + VGG-FC8 + Sent-FC7 + Sent-FC8 + Faces	**45.82**%	**53.11**%
Other Methods	Zhu [38] *(3 features)*	30.76%	30.95%

Table 2: Experimental results on the LIRIS-ACCEDE dataset. *(top)* our results showing single-feature and feature-fusion accuracies; *(bottom)* Zhu *et al.* [38].

show that using multiple features yields better results, and the more features we use the higher the improvement, demonstrating that different features contribute with different kind of information. Looking at the numbers, combining two features yield an improvement of 1.5% over our single feature model, obtaining a classification accuracy that surpasses the state-of-the-art result. Moreover, when combining all three sentiment features together the improvement goes up to 2.0%. Adding the two object-based features adds 0.6% performance. These results show that faces carry additional information, in fact adding face features always improves results. This can be also appreciated in Figure 4, where, for the valence task, we show some positive and negative examples of initially misclassified videos that are then correctly classified when face information is included in the method.

Similar considerations apply for the arousal task. Results are reported in the Arousal column of Table 1. In this case the combination of sentiment features improves over object-based features alone, but the gain over single sentiment features is very limited. It is necessary to also add object-based features to the sentiment ones to improve, albeit slightly, over competing methods.

Results on the LIRIS-ACCEDE dataset only are reported in Table 2. Again combining semantic features (*(top)*, lower part) is beneficial, obtaining improvements up to 2%. Similarly to the MEDIAEVAL-2015 dataset using the three sentiment-based features together with the two object-based features provides additional improvements that, in this case, are more pronounced also for the arousal task.

5 CONCLUSIONS

In this paper we proposed a novel approach to sentiment video analysis to classify induced arousal and valence of a video by exploiting sentiment-related features instead of using object related features only. We proposed to exploit sentiment-related features derived from the Deep-SentiBank network [4], together with features related to the actor face expression. To make use of the latter, we trained a neural network for face recognition on the Oxford face dataset. Features are extracted per-frame, and encoded into an Improved Fisher Vector for each feature. A classical linear SVM is used to learn a per-feature classifier, allowing us to perform late fusion on any combination of features. Experiments have been conducted on both the LIRIS-ACCEDE dataset [2], a collection of 9,800 movie excerpts with a variable duration of 8-12 seconds, and the MEDIAEVAL-2015 dataset, an extension of the previous dataset with an additional 1,100 excerpts. We first performed classification of single features, showing that they carry good classification power, then we performed multiple feature combinations by late fusing the classification scores, proving not only that different features give different contributions to the final classification, but also that including information about actor expressions always give a considerable improvement. Results show that with our strategy, using both sentiment related features and object features, we obtain better than state-of-the art results, outperforming the best current methods on both datasets for arousal and valence classification tasks by leveraging a smaller number of features.

ACKNOWLEDGMENTS

This work is partially supported by the MIUR "Social Museum and Smart Tourism" project (Grant No.: CTN01_00034_231545).

REFERENCES

[1] Claudio Baecchi, Tiberio Uricchio, Marco Bertini, and Alberto Del Bimbo. 2016. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications* 75, 5 (2016), 2507–2525.
[2] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.
[3] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. SentiBank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proc. of ACM MM.*

Figure 4: Examples of videos where face information is essential for correct classification. First row *(top)*: positive videos that were classified as negatives, second row *(bottom)*: frames of negative videos that were classified as positives.

[4] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* (2014).

[5] Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang. 2015. Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In *Working Notes Proceedings of the MediaEval Workshop*.

[6] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia* 19, 3 (July 2012), 34–41.

[7] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and Image Based Emotion Recognition Challenges in the Wild: EmotiW 2015. In *Proc. of ACM ICMI*.

[8] Joël Dumoulin, Diana Affi, Elena Mugellini, Omar Abou Khaled, Marco Bertini, and Alberto Del Bimbo. 2015. Affect Recognition in a Realistic Movie Dataset Using a Hierarchical Approach. In *Proc. of ASM*. 6.

[9] Joël Dumoulin, Diana Affi, Elena Mugellini, Omar Abou Khaled, Marco Bertini, and Alberto Del Bimbo. 2015. Movie's Affect Communication Using Multisensory Modalities. In *Proc. of ACM MM*. 739–740. DOI: http://dx.doi.org/10.1145/2733373.2807965

[10] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124.

[11] Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53, 4 (1987), 712–717.

[12] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. In *Proc. of ACM ICMI*.

[13] Alan Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* 23, 2 (2006), 90–100.

[14] Johannes Itten. 1973. The Art of Color: The Subjective Experience and Objective Rationale of Color (trans. Ernest van Haagen). (1973).

[15] Hideo Joho, Joemon M Jose, Roberto Valenti, and Nicu Sebe. 2009. Exploiting facial expressions for affective video summarisation. In *Proc. of CIVR*.

[16] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology. In *Proc. of ACM MM*.

[17] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *J. Mach. Learning Research* 10 (2009), 1755–1758.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. of NIPS*.

[19] Vu Lam, Sang Phan, Duy-Dinh Le, Shinichi Satoh, and Duc Anh Duong. 2015. NII-UIT at MediaEval 2015 Affective Impact of Movies Task. In *Working Notes Proceedings of the MediaEval Workshop*.

[20] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proc. of ACM MM*.

[21] P Marin Vlastelica, Sergey Hayrapetyan, Makarand Tapaswi, and Rainer Stiefelhagen. 2015. KIT at MediaEval 2015 – Evaluating Visual Cues for Affective Impact of Movies Task. In *Working Notes Proceedings of the MediaEval Workshop*.

[22] Ionut Mironica, Bogdan Ionescu, Mats Sjöberg, Markus Schedl, and Marcin Skowron. 2015. RFA at MediaEval 2015 Affective Impact of Movies Task: A Multimodal Approach. In *Working Notes Proceedings of the MediaEval Workshop*.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *Proc. of BMVC*.

[24] Veronica Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems* 28, 3 (2013), 38–45.

[25] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision* 105, 3 (2013), 222–245.

[26] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion* 24, 7 (2010), 1153–1172.

[27] Omar Seddati, Emre Kulah, Gueorgui Pironkov, Stéphane Dupont, Saïd Mahmoudi, and Thierry Dutoit. 2015. UMons at MediaEval 2015 Affective Impact of Movies Task including Violent Scenes Detection. In *Working Notes Proceedings of the MediaEval Workshop*.

[28] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[29] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quang, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-Hélène Demarty, and Liming Chen. 2015. The MediaEval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*.

[30] George Trigeorgis, Eduardo Coutinho, Fabien Ringeval, Erik Marchi, Stefanos Zafeiriou, and Björn Schuller. 2015. The ICL-TUM-PASSAU Approach for the MediaEval 2015 "Affective Impact of Movies" Task. In *Working Notes Proceedings of the MediaEval Workshop*.

[31] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of Experimental Psychology: General* 123, 4 (1994), 394.

[32] Barbara Wild, Michael Erb, Michael Eyb, Mathias Bartels, and Wolfgang Grodd. 2003. Why are smiles contagious? An fMRI study of the interaction between perception of facial affect and facial movements. *Psychiatry Research: Neuroimaging* 123, 1 (2003), 17–36.

[33] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28, 3 (2013), 46–53.

[34] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks. In *Proc. of ACM MM*.

[35] Gang Yu and Junsong Yuan. 2015. Fast action proposals for human action detection and search. In *Proc. of CVPR*.

[36] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[37] Sicheng Zhao, Hongxun Yao, Xiaoshuai Sun, Pengfei Xu, Xianming Liu, and Rongrong Ji. 2011. Video indexing and recommendation based on affective analysis of viewers. In *Proc. of ACM MM*.

[38] Yingying Zhu, Zhengbo Jiang, Jianfeng Peng, and Sheng-hua Zhong. 2016. Video Affective Content Analysis Based on Protagonist via Convolutional Neural Network. In *Proc. of PCM*.