Multimedia Tools and Applications (MTAP) manuscript No.
(will be inserted by the editor)

A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material

Esra Acar · Frank Hopfgartner · Sahin Albayrak

Received: 31 October 2015 / Accepted: DD MM YY

Abstract In today's society where audio-visual content such as professionally edited and user-generated videos is ubiquitous, automatic analysis of this content is a decisive functionality. Within this context, there is an extensive ongoing research about understanding the semantics (i.e., facts) such as objects or events in videos. However, little research has been devoted to understanding the emotional content of the videos. In this paper, we address this issue and introduce a system that performs emotional content analysis of professionally edited and user-generated videos. We concentrate both on the representation and modeling aspects. Videos are represented using mid-level audio-visual features. More specifically, audio and static visual representations are automatically learned from raw data using convolutional neural networks (CNNs). In addition, dense trajectory based motion and SentiBank domain-specific features are incorporated. By means of ensemble learning and fusion mechanisms, videos are classified into one of predefined emotion categories. Results obtained on the VideoEmotion dataset and a subset of the DEAP dataset show that (1) higher level representations perform better than low-level features, (2) among audio features, mid-level learned representations perform better than mid-level handcrafted ones, (3) incorporating motion and domain-specific information leads to a notable performance gain, and (4) ensemble learning is superior to multi-class support vector machines (SVMs) for video affective content analysis.

Esra Acar
DAI Laboratory, Technische Universität Berlin, Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
Tel.: +49 30 314 74013, Fax: +49 30 314 74003
E-mail: esra.acar@tu-berlin.de

Frank Hopfgartner
Humanities Advanced Technology and Information Institute, University of Glasgow, UK
E-mail: frank.hopfgartner@glasgow.ac.uk

Sahin Albayrak
DAI Laboratory, Technische Universität Berlin, Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
Tel.: +49 30 314 74001, Fax: +49 30 314 74003
E-mail: sahin.albayrak@dai-labor.de

Keywords Video Affective Content Analysis · Ensemble Learning · Deep Learning · MFCC · Color · Dense Trajectories · SentiBank

1 Introduction

Nowadays, the amount of audio-visual data items available to consumers has attained colossal proportions. Classifying, retrieving, and subsequently delivering personalized video content corresponding to the needs of the consumers is a challenge which still has to be resolved. Video affective analysis – which consists in identifying videos segments that evoke particular emotions in the user [37] – can bring an answer to such a challenge from an original perspective. The recent survey by Wang and Ji [37] provides a thorough overview of the subject and distinguishes two kinds of research. On the one hand, *direct* approaches (mainstream) deduce human emotions directly from audio and/or visual features extracted from the data; on the other hand, *implicit* approaches deduce them from the user's reaction while being exposed to the video [37]. Next to this dichotomy of the research field, emotions themselves can be defined according to discrete categories (i.e., categorical affective analysis) or non-discrete categories (i.e., dimensional affective analysis).

We develop in the present paper a direct and categorical affective analysis framework. In this context, one direction followed by many researchers consists in using machine learning methods (e.g., [18,42,44]). Machine learning approaches make use of a specific data representation (i.e., features extracted from the data) to identify particular events. However, their performance is heavily dependent on the choice of the data representation on which they are applied [6]. As in any pattern recognition task, one key issue is, therefore, to find an effective representation of video content.

Features can be classified according to different schemes. One type is the classification based on the level of semantic information which a given feature carries. In the terminology which we adopt, at one extreme, a feature is said to be "low-level" if it carries (almost) no semantic information (e.g., value of a single pixel or audio sample); at the other extreme, it is said to be "high-level" if it carries maximally semantic information (e.g., a guitarist performing a song in a clip). Between both, "mid-level" feature representations are derived from raw data, but are one step closer to human perception. Another possible type of classification, which is particularly relevant in video analysis, where data items are not a single image but sequences, is the distinction between "static" and "dynamic" (or temporal) features.

In this context, in the field of audio, video and more generally multi-dimensional signal processing, automatically and directly learning suitable features (i.e., mid-level features) from raw data to perform tasks such as event detection, summarization, retrieval has attracted particular attention, especially because such learning kept the amount of required supervision to a minimum and provided scalable solutions. To achieve this, deep learning methods such as CNNs and deep belief networks are shown to provide promising results (e.g., [20, 25, 31]). This recent success of deep learning methods previously incited us to directly learn feature representations from automatically extracted raw audio and color features by deep learning to obtain mid-level audio and visual representations [1]. However, this work was limited to learning

audio and static visual features only. As a consequence, our work could not optimally take into account the motion and temporal coherence exhibited in image sequences compared to a single image. Studies (e.g., [36, 44]) have, indeed, demonstrated that, in addition to audio and static color features, motion plays an important role for affective content analysis in videos. Therefore, we propose to use dense trajectory features to derive a mid-level motion representation obtained via a sparse coding based Bag-of-Words method to boost the classification performance of our system. In our previous work, static color features were based on the representation of images in the RGB space, which is not optimized for affect analysis. Hinted by the prior art evidence that perception of emotions by humans is enhanced in a hue-saturation type color space [34], we close this gap here by working in the HSV color space. In addition to audio, static and dynamic visual features, we use SentiBank domain-specific representations which are shown to be effective for emotional analysis of single images [7, 22]. One additional question arising when using multiple features, is that of fusion of information; we, therefore, also propose an assessment of fusion mechanisms. Finally, we tackle the modeling aspect of video affective content analysis. SVM is the most widely used learning method for the generation of analysis models (employed in e.g., [7, 14, 22, 38, 42]). Distancing ourselves from the prior art, we apply *ensemble learning* (i.e., decision tree based bootstrap aggregating) to generate emotional content analysis models and compare its performance to that of SVM modeling. Consequently, we address the following research questions in this work:

- **(RQ1) What is the discriminative power of learned audio-visual representations?** We review the discriminative power of audio and static visual representations learned from raw data using deep learning.
- **(RQ2) What is the effect of dense trajectories and SentiBank representations?** First, we investigate the discriminative power of dense trajectories and SentiBank. Second, we assess the effect of incorporating dense trajectories and SentiBank representations on the classification performance.
- **(RQ3) How does ensemble learning perform in comparison to SVM modeling?** We investigate the classification performance of decision tree based bootstrap aggregating (i.e., *bagging*) against SVM which is the dominant modeling scheme for video affective content analysis. To the best of our knowledge, the proposed system is the first to adopt ensemble learning for emotional characterization of professionally edited and user-generated videos; and we show that the adopted ensemble learning method outperforms SVM modeling in terms of classification accuracy.
- **(RQ4) How to optimally combine audio-visual features?** We explore optimal late fusion mechanisms and analyze linear and SVM-based fusion for combining the outputs of uni-modal analysis models.

The work presented in the present paper is an extension of our conference publication [2] but includes several contributions. The significant extensions are as follows. First, we apply ensemble learning (bagging) in addition to SVM. Second, we include SentiBank domain specific representations [7] in emotion modeling. Third, we report new results on a more challenging dataset (i.e., VideoEmotion [22]).

The paper is organized as follows. Section 2 explores the recent developments and reviews methods which have been proposed for affective content analysis of video footage. In Section 3, we introduce our method for the affective classification of videos. We provide and discuss evaluation results on a subset of the DEAP dataset [23] and on the VideoEmotion dataset [22] in Section 4. Finally, we present concluding remarks and future directions to expand our method in Section 5.

2 Related Work

The literature on emotion analysis – and in particular on direct methods – can be analyzed from different points of view. As also evoked in the survey by Wang and Jin [37], a direct affective content analysis framework requires two essential elements; these are video feature extraction (i.e., *representation*) and classification or regression (i.e., *modeling*). As, in this work, we focus both on the representation and on the modeling aspects of emotional content analysis, we present a review of existing solutions examined from these two perspectives in Sections 2.1 and 2.2, respectively. In addition, stressing the importance of motion related features enables us to highlight one of our contributions; hence, in Section 2.3, we also position existing studies with respect to the use of motion information.

2.1 From a feature representation point of view

Among video affective content analysis methods, using low-level audio-visual features as video representations is one type of commonplace approach. In [14], Eggink and Bland present a method for mood-based classification of TV Programs on a large-scale dataset. Various low-level features are used as video representations; these are audio (e.g., MFCC, sound energy, spectral rolloff, ZCR) and visual features (luminance, motion, cuts and presence of faces), either taken individually or in combination. SVMs are used as classifiers. In [38], a combined analysis of low-level audio and visual representations based on early feature fusion is presented for facial emotion recognition in videos. The audio features include MFCC and ZCR. The visual features are extracted based on a deformable model fitted on faces and correspond to various face-related information such as facial pose, opening of the mouth or status of eyebrows. The combined feature vectors are classified with SVMs. In order to retrieve videos containing resembling emotions, Niu et al. [27] develop a similarity measure of videos based on affect analysis. They use four low-level features (audio: sound energy, audio pitch average; visual: motion, shot-change rate) to construct Valence-Arousal (VA) graphs. Those VA-graphs are normalized to account for unequal video durations and are used to derive the similarity measures. The set of videos is further partitioned using spectral clustering based on the similarity measure. The result is used in a recommender system to retrieve similar contents. Baveye et al. [3, 5] address the issue of a common emotional database and introduce the LIRIS-ACCEDE dataset which is an annotated creative commons emotional database. In that work, a baseline framework was also presented, which employs low-level audio (Energy,

ZCR, etc.) and still image (Colorfulness, spatial edge distribution, etc.) features, in an SVM framework. Yazdani et al. [44] present a method which uses audio-visual features as representation and k-nearest neighbor classifier as the learning method for the affective analysis of music video clips.

Another type of commonplace approach is to use mid-level or hierarchical representations of videos. These solutions employ mid-level representations created from low-level ones. Irie et al. [17] present an affective video segment retrieval method based on the correlation between emotions and so-called emotional audio events (EAEs) which are *laughter*, *loud voice*, *calm music* and *aggressive music*. The main idea is to use EAEs as an intermediate representation. The detection of EAEs is based on audio information only. Xu et al. [42] present a 3-level affective content analysis framework, in which the purpose is to detect the affective content of videos (i.e., horror scenes for horror movies, laughable sections for sitcoms and emotional tagging of movies). They introduce mid-level representations which indicate dialog, audio emotional events (i.e., horror sound and laughter – similar in concept to the EAEs of Irie et al. [17]) and textual concepts (i.e., informative emotion keywords). In [18], Irie et al. propose to represent movie shots with so-called Bag-of-Affective Audio-visual Words and apply a latent topic driving model in order to map these representations to affective categories, and, hence, to achieve movie classification. The audio-visual words are formed by bringing together audio and visual words, which are constructed based on audio (pitch, short-term energy, MFCC) and visual (color, brightness, motion intensity and shot duration) features, respectively. In [8], Canini et al. introduce a framework where movie scenes are represented in a 3-dimensional connotative space whose dimensions are natural, temporal, and energetic. The aim is to reduce the gap between objective low-level audio-visual features and highly subjective emotions through connotation. As audio-visual representation of movies, they employ low-level audio descriptors (representing intensity, timbre and rhythm), low and mid-level color (e.g., color energy, average saturation of pixels) and motion (average of motion vector modules and standard deviation) descriptors. Ellis et al. [15] introduced a framework for emotional analysis of movie trailers using mid-level representations corresponding to specific concepts (e.g., “gunshot”, “fight”, “sex”). The rationale behind their method is that human emotions are closely related to such concepts rather than low-level features. They define 36 concepts (annotated at video shot level) and build a detector for each concept. Each concept detector is realized with an SVM using low-level audio (e.g., MFCC, pitch) and visual features (e.g., SIFT, number of faces). The concepts are used to infer emotions in the videos. In an attempt to close the so-called emotional gap between low-level features and emotions, Borth et al. [7] construct a visual sentiment ontology containing 3,000 concepts (i.e., mid-level visual representations), each corresponding to an adjective noun pair (ANP). Each pair is composed of a neutral noun associated to a strong sentiment (e.g., beautiful flower). Those concepts are detected using a pool of detectors (SentiBank). Visual features such as color histograms and local binary patterns as well as additional features (e.g., faces or objects) are employed in SVM detectors. The approach was evaluated on images and videos retrieved from online repositories in order to infer emotions. In an improvement to the work by Borth et al. [7], Chen et al. [11] aim at solving two problems related to ANP concepts, namely the localization of objects

and ambiguity of annotations due to adjectives. They solve the localization issue by limiting their study to only six common objects appearing in social media content and by detecting those objects with a parts-based detector. The ambiguity issue between semantically similar concepts is tackled by using a new type of SVMs, which are capable of learning overlapping class labels. The method is reported to outperform conventional SentiBank. In [22], Jiang et al. propose a comprehensive computational framework, where they extract an extensive set of features from the dataset, ranging from well-known low-level audio-visual descriptors (audio: MFCC, energy entropy, etc.; visual: SIFT, HOG, etc.) to high-level semantic attributes such as ObjectBank and SentiBank representations.

All of the above-mentioned works represent videos with low or mid-level hand-crafted features. However, in attempts to extend the applicability of methods, there is a growing interest for directly and automatically learning features from raw audio-visual data rather than representing them based on manually designed features, deep learning being a widespread illustration of such direct and automatic learning. For example, Schmidt et al. [31] address the feature representation issue for automatic detection of emotions in music by employing regression based deep belief networks to learn features from magnitude spectra instead of manually designing feature representations. By taking into account the dynamic nature of music, they also investigate the effect of combining multiple timescales of aggregated magnitude spectra as a basis for feature learning. These learned features are then evaluated in the context of multiple linear regression. Among deep learning solutions, CNNs have become particularly popular in affect analysis in the last years. Li et al. [25] propose to perform feature learning for music genre classification and use CNNs for the extraction of musical pattern features directly from raw MFCC values. Ji et al. [20] address the automated recognition of human actions in surveillance videos and develop a novel 3D-CNN model to capture motion information encoded in multiple adjacent frames. Another CNN-based method is our previous work [1], where we used deep learning to derive mid-level representations directly from the raw data. As an improvement of the SentiBank paper by Borth et al. [7], Chen et al. [10] propose to extend it using deep learning (*DeepSentiBank*). CNNs are used instead of binary SVM classifiers. The increased computational load induced by deep learning is dealt by a GPU-based learning framework. CNNs are shown to provide substantial improvement compared to binary SVMs. Another work making use of DeepSentiBank is the one by Dumoulin et al. [12], where DeepSentiBank is combined with low-level audio-visual and CNN-based features in a hierarchical classification scheme. In [41], Xu et al. predict sentiments in still images using CNNs. CNNs are trained as object classifiers to classify image content according to labels such as “zebra” or “lemon”; subsequently, transfer learning is employed for generating mid-level representations. Logistic regression is then used to predict sentiments using the generated mid-level representations. More recently, Baveye et al. [4] compared CNNs applied on video keyframes against the combination of Support Vector Regression (SVR) and low-level features, and reached the conclusion that CNNs constitute a promising solution. Xu et al. [40] perform emotion recognition in videos using a BoW approach, where the dictionary is constructed by clustering features obtained from so-called auxiliary images by means of CNNs. The same CNNs are then applied on videos frames to encode videos according to

the BoW scheme. Unlike most studies where features are independently extracted from audio and visual modalities, Pang and Ngo [28] propose to extract joint features from the raw data directly using Deep Boltzman Machines (DBM), i.e., they extract mid-level features which simultaneously capture audio, visual and text information. Visual, audio and text features are input to a multi-modal DBM which returns the joint representations. Those joint representations are used to train SVM classifiers with RBF kernels in order to predict emotions.

2.2 From a modeling point of view

A closer look at the methods proposed in the literature (Section 2.1) reveals that most of them (a non-exhaustive list of examples including [7,14,22,38,42]) concentrate on the representation aspect of content analysis by proposing new handcrafted or learned (via deep learning methods) audio-visual descriptors. The modeling part, on the other hand, is rather neglected; we see, indeed, that authors predominantly use SVMs as the learning method in an "Out-of-the-Box" fashion.

Although the SVM-based learning technique is the dominant modeling scheme for video affective content analysis, certain works in the literature employ techniques other than SVM. In [7], Borth et al. use logistic regression in addition to linear SVM and show that logistic regression is superior to SVM. Inspired by the success of logistic regression in [7], Xu et al. also adopt logistic regression as the learning scheme in their work [41]. Dumoulin et al. [12] introduced a hierarchical classification scheme, where input features are classified by traversing a 3-level tree. At the highest level, the features are determined as stemming from an emotional or non-emotional video segment; then those classified as emotional are labeled as corresponding to positive or negative emotions; finally, at the lowest level, the class of emotion is determined. This hierarchical scheme was tested using different types of classifiers (e.g., SVM, random forests).

As already presented in detail in Section 2.1, alternative learning methods utilized in the field of affective content analysis include topic extraction via the latent Dirichlet allocation [17], spectral clustering [27], k-nearest neighbor classifier [44] and SVR [4].

2.3 The issue of motion information

One observation about the works mentioned in Section 2.1 is that the use of the temporal aspect of videos is either limited or totally absent. In other words, videos are generally analyzed as a sequence of independent frames rather than as a whole. A few works (i.e., [8,14,18]) use motion-based features, and these are limited to simple features (e.g., features based on frame differencing). For instance, Canini et al. [8] use average motion vector magnitudes derived from the motionDS descriptor [19]. Eggink and Bland [14] use motion based on the difference between every tenth frame. Irie et al. [18] employ motion intensity as the average magnitude of motion vectors; they also take into account the duration of shots as another feature.

The only notable exception is the work of Ji et al. [20], where multiple adjacent frames are used. However, they take into account only 7 adjacent frames. Increasing this number to higher dimensions would probably render the learning of the 3D-CNNs intractable. Therefore, in our opinion, a more effective mid-level motion representation is needed.

Recently, a new type of video descriptor has emerged, namely dense feature trajectories. These descriptors, which correspond to points which are densely sampled and tracked using dense optical flow fields, were introduced by Wang et al. [35] for the task of action recognition in videos, and have proven robust for action recognition. However, to the best of our knowledge, the applicability of these improved dense trajectories to the task of affective content analysis has not been investigated yet. Distinct from the aforementioned existing works, we suggest combining deep learning based representations with dense motion trajectories.

2.4 Positioning of our work

When considering the papers dealing with emotion recognition from videos, the recent works closely related to ours are those by Baveye et al. [4] and Dumoulin et al. [12] where CNNs are used only in the visual feature extraction. Our approach is novel in that deep learning is used not only with visual data but also with audio data. In addition, our work makes use of advanced motion information, which is another novel aspect. Finally, we not only concentrate on the representation aspects, but we also propose here to comprehensively assess the behavior of two popular classifiers, namely SVM and ensemble learning.

3 The Video Affective Analysis Method

In this section, we present our approach, which is a categorical affective analysis solution. As mentioned in the introduction, affective analysis can either be categorical or dimensional. The choice of categorical or dimensional is not critical, as in practice, categories can always be mapped onto dimensions and vice versa [16]. It is, therefore, possible to map discrete emotions to arousal-valence dimensions.

We perform categorical affective analysis according to two different classification schemes, namely *VA-based* and *wheel-based* classification. In the VA-based classification, the method classifies a given video into one of the four quadrants of the VA-space, whereas wheel-based classification consists in classifying a video according to an emotion wheel. The available dataset of user-generated videos (Flickr [1] and YouTube [2] videos) only provides annotations according to an emotion wheel. Concerning professionally edited videos (i.e., music video clips), although the dataset we use contains both VA-based and wheel-based annotations, we only report VA-based classification results; the reason is the low number of samples in this dataset (74) considering the high number of classes to discriminate (12), so that the results

[1] `https://www.flickr.com/`

[2] `https://www.youtube.com/`

would not be statistically very relevant. More details concerning these sets can be found in Section 4. The emotion model used for the user-generated videos is based on Plutchik's emotion wheel [29] (Figure 1) which defines a wheel-like diagram of emotions consisting of eight basic emotions and their derivatives.

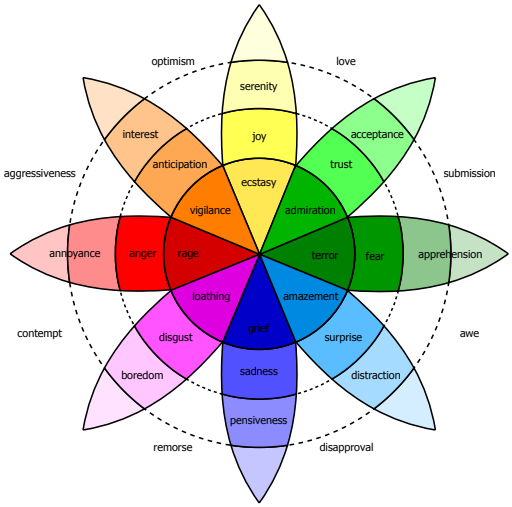


Fig. 1 Plutchik's wheel of emotions [29] used for user-generated videos in this work.

In Figure 2, we provide an overview of the system. As shown in Figure 2, the system consists of the following steps: (1) videos (i.e., one-minute highlight extracts of music video clips or user-generated videos of different length) are first segmented into pieces, each piece lasting 5 seconds (as suggested in [44]); (2) audio and visual feature extraction; (3) learning mid-level audio and static visual representations *(training phase only)*; (4) generating mid-level audio-visual representations; (5) generating an affective analysis model *(training phase only)*; (6) classifying a video segment of 5-second length into one of related emotion categories *(test phase only)*; and (7) classifying a complete video using the results obtained on the 5-second segments constituting the video *(test phase only)*.

The audio and visual feature learning phases are discussed in detail in Section 3.1, whereas the incorporation of temporal information and domain-specific information to the system is explained in Sections 3.2 and 3.3, respectively. The generation of an affective analysis model is discussed in more detail in Section 3.4. This model uses fusion, which is presented in Section 3.5.

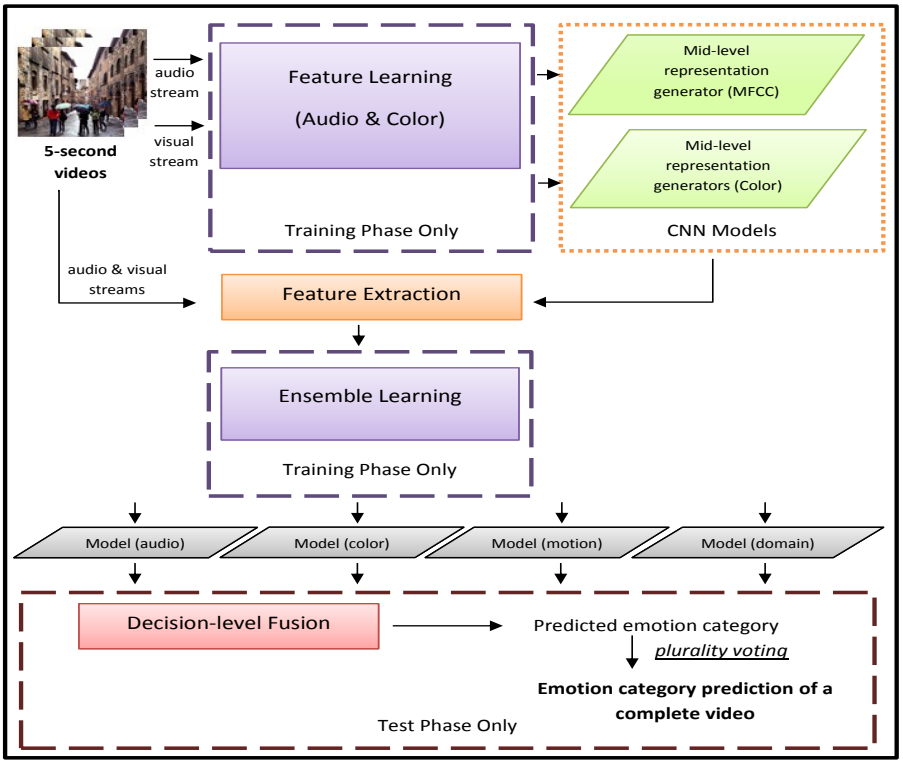


Fig. 2 A high-level overview of the proposed system. Feature learning and extraction, affective analysis model generation and decision fusion parts are explained in detail in the subsections of Section 3.

3.1 Learning Mid-Level Audio and Static Visual Representations

Concerning the learning of audio and visual representations, we improved one of our previous works on affective content analysis [1]. The improvements concern the extraction modalities (e.g., dimensions) of the audio representations, and the use of a different color space which enables deriving more discriminative features. MFCC values are extracted for each video segment. The set of resulting MFCC feature vectors, when concatenated, constitutes an MFCC-time domain representation that can be regarded as an "image", on which we apply 2D CNN modeling. As CNNs were successfully applied to detect patterns from raw image data (e.g., [21, 24]), we considered them for the representation we obtained. The aim is to capture both timbre and temporal information.

The first layer (i.e., the input layer) of the CNN is a 497x13 map which contains the MFCC feature vectors from 125 frames of one segment. In Figure 3, the CNN architecture used to generate audio representations is presented. The CNN has three convolution and two subsampling layers, and one output layer which is fully connected to the last convolution layer (this network size in terms of convolution and subsampling layers has experimentally given satisfactory results). In the VA-based

classification, the output layer consists of four units: one for each quadrant of the VA-space. In the wheel-based classification, the output layer consists of eight units (annotation according to Plutchik's emotion wheel): one unit for each emotion category. Each unit in the output layer is fully connected to each of the 976 units in the last convolution layer. The CNN is trained using the backpropagation algorithm. After training, the output of the last convolution layer is used as the *mid-level audio representation* of the corresponding video segment. Hence, the MFCC feature vectors from 125 frames of one segment are converted into a 976-dimensional feature vector (which constitutes a more abstract audio representation) capturing the acoustic information in the audio signal of the video segment.

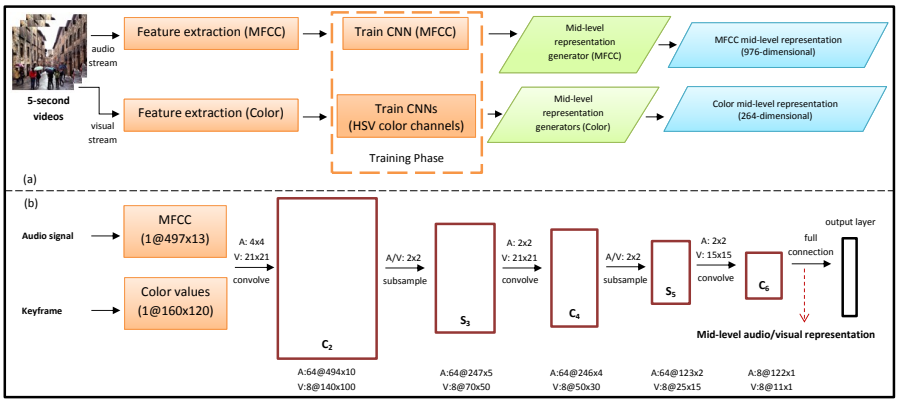


Fig. 3 (a) A high-level overview of our representation learning method, (b) the detailed CNN architectures for audio and visual representation learning. The architecture contains three convolution and two subsampling layers, one output layer fully connected to the last convolution layer (C_6). (CNN: Convolutional Neural Network, MFCC: Mel-Frequency Cepstral Coefficients, A: Audio, V: Visual)

Existing works (e.g., [34]) have shown that colors and their proportions are important parameters to evoke emotions. This observation has motivated our choice of color values for the generation of static visual representations for videos. The frame in the middle of a 5-second video segment is extracted as the keyframe (i.e., representative frame) for the segment. For the generation of *mid-level static visual representations*, we extract color information in the HSV space from the keyframe. The resulting values in each channel are given as input to a separate CNN. Similarly to the audio case, Figure 3 presents the CNN architecture used to generate visual representations, where the first layer (i.e., the input layer) of the CNN is a 160x120 map which contains the values from one channel of the keyframe. The training of the CNN is done similarly to the training of the CNN in the audio case. As a result, the values in each channel are converted into an 88-dimensional feature vector. The feature vectors generated for each of the three channels are concatenated into a 264-dimensional feature vector which forms a more abstract visual representation capturing the color information in the keyframe of the segment.

3.2 Deriving Mid-Level Dynamic Visual Representations

The importance of motion in edited videos such as movies and music video clips, and user-generated videos motivated us to extend our previous approach [1] and to incorporate motion information to our analysis framework. To this end, we adopt the work of Wang et al. on improved dense trajectories [35]. Dense trajectories are dynamic visual features which are derived from tracking densely sampled feature points in multiple spatial scales. Although initially used for unconstrained video action recognition [35], dense trajectories constitute a powerful tool for motion or video description, and, hence, are not limited to action recognition only.

Our dynamic visual representation works as follows. First, improved dense trajectories [35] of length 15 frames are extracted from each video segment. The sampling stride, which corresponds to the distance by which extracted feature points are spaced, is set to 20 pixels. Dense trajectories are subsequently represented by a histogram of oriented gradients (HoG), a histogram of optical flow (HoF) and motion boundary histograms in the x and y directions (MBHx and MBHy, respectively). We learn a separate dictionary for each dense trajectory descriptor (i.e., each one of HoG, HoF, MBHx and MBHy). We employ the sparse dictionary learning technique presented in [26]. In order to learn the dictionary of size k ($k = 512$ in this work) for sparse coding, $400 \times k$ feature vectors are sampled from the training data (this figure has experimentally given satisfactory results). In the coding phase, we construct the sparse representations of dense trajectory features using the LARS algorithm [13]. Given dense trajectory features and a dictionary as input, the LARS algorithm returns sparse representations for the feature vectors (i.e., sparse mid-level motion representations). In order to generate the final sparse representation of video segments which are a set of dense trajectory feature vectors, we apply the *max-pooling* technique.

3.3 Incorporating Domain-Specific Representations

In addition to general-purpose mid-level audio, static and dynamic visual representations, we incorporate mid-level semantic representations of the visual content of videos for high-level video affective content analysis (i.e., domain-specific representations). More specifically, we use SentiBank representations [7]. SentiBank is based on emotion-related concepts. In the first version of SentiBank, there were 1,200 concept detectors; this number increased to 2,089 in the subsequent version [3] (version 1.1). As briefly explained in Section 2, each emotion-related concept is defined as an Adjective-Noun Pair such as “cute baby” or “dark forest”. In these ANPs, adjectives (e.g., “funny”, “peaceful”, “’gorgeous’, “weird”) are strongly connected to emotions, and nouns (e.g., “baby”, “dog”, “car”, “wedding”) are usually objects or scenes that can be automatically detected [7]. Each dimension in the SentiBank representation corresponds to the detection score of the corresponding ANP concept detector. In this work, we use both version 1.0 and version 1.1 of the SentiBank representations (i.e., 1,200 and 2,089 ANP concepts) in order to assess the influence of the number of concepts on the performance of video affective content classification.

[3] `http://www.ee.columbia.edu/ln/dvmm/vso/download/sentibank.html`

3.4 Generating the Affective Analysis Model

In this work, in addition to SVM modeling [2], we apply *ensemble learning* in order to build affective analysis models. For the generation of the models, mid-level audio, static and dynamic visual, and domain-specific representations are fed into separate classifiers. Mid-level audio and static visual representations are created by using the corresponding CNN models for video segments of 5-second length both in the training and test phases. In the training phase, decision trees are combined with bagging. In the test phase, the final prediction which corresponds to the prediction score of the ensemble of the trees for the test data is computed as the average of predictions from individual trees. The prediction score generated by each tree is the probability of a test sample originating from the related class computed as the fraction of samples of this class in a tree leaf.

The prediction scores of the models are merged using one of the fusion strategies presented in Section 3.5. Once all 5-second video segments extracted from a given video are classified, final decisions for the classification of the complete video is realized by a *plurality voting* process. In other words, a video is assigned the label which is most frequently encountered among the set of 5-second segments constituting the video.

3.5 Fusion strategies

When combining results of multiple classifiers, fusion of the results constitutes an important step. In this paper, we investigate two distinct fusion techniques to combine the outputs of the classification models, namely *linear fusion* and *SVM-based fusion*.

3.5.1 Linear fusion

In linear fusion, probability estimates obtained from the classifiers trained separately with one of the mid-level audio, static and dynamic visual, and domain-specific representations are linearly fused at the decision-level. The classifiers that we adopt are all based on the same ensemble learning algorithm. Hence, we are in the presence of homogeneous "learners" (i.e., all of the same type) according to the terminology of [46]. In such a situation, it is advised to directly fuse the probabilities ($h_i(x_j)$) generated by each of the classifiers (i.e., "learners") using the *weighted soft voting* technique [46]:

$$H(x_j) = \sum_{i=1}^{T} w_i h_i(x_j) \tag{1}$$

Classifier-specific weights (w_i in Equation 1) are optimized on the training data. The weights assigned to the classifiers are determined in such a way that they always sum up to 1.

3.5.2 SVM-based fusion

In SVM-based fusion, the probability estimates of the uni-modal classifiers are concatenated into vectors and used to construct higher level representations for each video segment. Subsequently, an SVM classifier which takes as input these higher level representations is constructed. This fusion SVM is then used to predict the label of a video segment. When SVMs are used as uni-modal classifiers, the scores returned by the SVMs are first converted into probability estimates using the method explained in [39].

4 Performance Evaluation

The experiments presented in this section aim primarily at comparing the discriminative power of our method – which is based on mid-level representations learned and derived from audio-visual data as presented in Section 3 – against the works presented in [1, 2, 22, 28, 44]. Accessorily, we provide a comparison of our mid-level audio representations against auditory temporal modulations (ATM) features. These features, which describe temporal modulations in the frequency domain, were recently applied for audio content analysis, in particular in the context of music recommendation [32] and genre classification [33].

An overview of the DEAP and VideoEmotion datasets used for the evaluation is provided in Section 4.1. In Section 4.2, we present the experimental setup. Finally, we provide results and discussions in Section 4.3.

4.1 Dataset and Ground-truth

The experiments are conducted on two types of video content, i.e. professionally edited videos and user-generated videos, which are represented by two different datasets: DEAP and VideoEmotion.

The DEAP dataset is a dataset for the analysis of human affective states using electroencephalogram, physiological and video signals. It consists of the ratings from an online self-assessment where 120 one-minute extracts of music videos were each rated by 14 to 16 volunteers based on arousal, valence and dominance. We have used all the music video clips whose YouTube links are provided in the DEAP dataset and which were available on YouTube at the time when we conducted the experiments (74 music clips). Only one-minute highlight extracts from these 74 videos have been used in the experiments. The extracts of different affective categories downloaded from YouTube equate to 888 music video segments of 5-second length.

In order to evaluate the performance of our method on user-generated videos, we use the recently introduced VideoEmotion dataset [22]. This dataset contains 1,101 videos collected from Flickr and YouTube. The videos in the dataset are annotated according to 8 categories, each category corresponding to a basic emotion represented in a section of Plutchik's wheel of emotions, including "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", and "trust" (Figure 1). In addition to using

the whole VideoEmotion dataset, as suggested in [22] we also provide results on a subset of the dataset which contains only four basic emotions more frequently used in the literature (i.e., “anger”, “fear”, “joy”, “sadness”).

Two different classification schemes are envisaged, one being *VA-based* and the other being *wheel-based*. For the case of the *VA-based classification*, four affective labels exist. These are *high arousal-high valence (HA-HV)*, *low arousal-high valence (LA-HV)*, *low arousal-low valence (LA-LV)* and *high arousal-low valence (HA-LV)*, each representing one quadrant in the VA-space. The evaluation of the *VA-based classification* is performed only on the DEAP dataset. The labels are provided in the DEAP dataset and are determined by the average ratings of the participants in the online self-assessment. Concerning the evaluation of *wheel-based classification*, we use VideoEmotion, for which the labels are also provided as part of the dataset [22]. Tables 1 and 2 summarize the main characteristics of the DEAP and of the VideoEmotion datasets in more detail.

Table 1 The characteristics of the DEAP dataset.

VA-based Category	# Music Videos	# Segments
high arousal-high valence (HA-HV)	19	228
low arousal-high valence (LA-HV)	19	228
low arousal-low valence (LA-LV)	14	168
high arousal-low valence (HA-LV)	22	264
Total	**74**	**888**

Table 2 The characteristics of the VideoEmotion dataset.

Wheel-based Category	# Flickr Videos	# YouTube Videos	Total
Anger	23	78	101
Anticipation	40	61	101
Disgust	100	15	115
Fear	123	44	167
Joy	133	47	180
Sadness	63	38	101
Surprise	95	141	236
Trust	44	56	100
Total	**621**	**480**	**1,101**

4.2 Experimental Setup

The MIR Toolbox v1.6.1[4] is employed to extract the 13-dimensional MFCC features. Frame sizes of 25 ms with 10 ms overlap are used. Mean and standard deviation for each dimension of the MFCC feature vectors are computed, which compose the low-level audio representations (*LLR audio*) of video segments. As *MLR handcrafted*

[4] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

audio, Bag-of-Words (BoW) representations of MFCC features are generated using vector quantization. An audio dictionary of size k (k is equal to 512 in this work) is constructed using $400 \times k$ MFCC descriptors and k-means clustering (figure which was determined experimentally). In order to generate the low-level visual features (*LLR color*) of video segments, we constructed normalized HSV histograms (of size 16, 4, 4 bins, respectively) in the HSV color space. The Deep Learning toolbox[5] is used in order to generate mid-level audio and color representations with a CNN (*MLR audio* and *MLR color*). Wang's dense trajectory implementation[6] is used to extract improved dense trajectories from video segments. Subsequently, BoW representations based on the motion features (i.e., HoG, HoF, MBHx and MBHy) of the dense trajectories (*MLR motion*) are generated as explained in Section 3. SentiBank representations are extracted as explained in Section 3 using both versions of SentiBank, i.e., 1,200 and 2,089 trained visual concept detectors (*MLR attribute1200* and *MLR attribute2089*).

Computationally, the most expensive phase of the representation learning is the training of the CNNs which takes on average 150 and 350 seconds per epoch for MFCC and color features, respectively. The generation of feature representations using CNNs amounts to 0.5 and 1.2 seconds on average per 5-second video segment for MFCC and color features, respectively. The extraction of dense trajectories takes on average 16 seconds per 5-second video segment. All the timing evaluations were performed with a machine with 2.40 GHz CPU and 8 GB RAM.

The multi-class SVMs with an RBF kernel are trained using libsvm [9] as the SVM implementation. Training was performed separately for each audio or visual descriptor extracted at the video segment level. SVM hyper parameters were optimized using a grid search and 5-fold cross-validation. All segmented samples belonging to a specific video were always either in the training set or in the test set during cross-validation. Zero mean unit variance normalization was applied on feature vectors. Fusion of audio and visual features is performed at the decision-level by linear or SVM-based fusion, as explained in Section 3.5. For the DEAP dataset, due to the limited amount of available music video samples, we used the *leave-one-song-out* cross validation scheme; whereas for the VideoEmotion dataset, we used the train-test splits provided as part of the dataset.

In the assessment of the discriminative power of learned audio representations (i.e., *MLR audio*), we compare them against ATM features. The reason is that both extract conceptually similar information from the audio data, the major distinction being that MLR audio features are learned, while ATM features do not involve a learning phase. As mentioned in Section 3.1, the MLR audio features which we developed capture both timbre and temporal information of an acoustic signal. Similarly, ATM features constitute a representation of slow temporal modulations of acoustic signals. More specifically, motivated by the human auditory and visual systems [33], they describe the power variation over modulation scales in each primary auditory cortex channel. The extraction of the ATM features is realized as follows: (1) the audio signal of a video segment of 5-second length is converted to a mono signal and down-

[5] https://github.com/rasmusbergpalm/DeepLearnToolbox/

[6] https://lear.inrialpes.fr/people/wang/improved_trajectories

sampled to 16 kHz, (2) auditory spectrograms are computed as described in [43], (3) the computed auditory spectrograms are used to find temporal modulations in the frequency domain through (inverse) discrete Fourier transforms. The implementation details of the ATM feature extraction can be found in [33].

4.3 Results and Discussions

In the following two subsections, we first provide and discuss the evaluation results of our method on professionally edited videos (Section 4.3.1) and then on user-generated videos (Section 4.3.2). In Section 4.3.3, we provide a summary of the evaluation results.

4.3.1 Evaluation on Professionally Edited Videos (DEAP)

In this section, we present and discuss the results on the DEAP dataset. We start with the MLR audio-ATM comparison. The classification accuracies for MLR-audio coupled with SVM and ensemble learning are 48.65% and 50.00%, respectively. For ATM coupled with SVM and ensemble learning, they reach 43.24% and 47.30%, respectively. Thus, our experiments have shown that our learned audio representations outperform ATM features, by 2.7% to 5.4%. To verify that the mean of the accuracies obtained using the MLR audio differs significantly from the ones obtained using the ATM features in a statistical sense, a paired Student t-test on classification accuracies from the *leave-one-song-out* cross validation was performed. This t-test showed that the improvement provided by MLR audio over ATM is statistically significant (5% significance level).

We continue by presenting the classification accuracies in the case where only one type of descriptor is employed (Figure 4). The aim of these experiments is to investigate the discriminative power of learned audio-visual, motion and SentiBank representations, and also to compare the bagging approach against SVM modeling (experiments relevant for the research questions *RQ1*, *RQ2* and *RQ3*). These are followed by an evaluation of the performance of different combinations of audio-visual representations with linear and SVM-based fusion (Table 3), where the best performing multi-modal representation and optimal decision-level fusion mechanisms are investigated (relevant for the research questions *RQ1*, *RQ2* and *RQ4*).

Various observations can be inferred based on the overall results presented in **Figure 4**. Concerning the classification methods, ensemble learning, in general, improves the discrimination power of uni-modal representations over SVM-based learning, independently of the features considered (except for *MLR handcrafted audio*). Concerning the features other than domain-specific representations, we first note that the motion representation is the best performing descriptor for both ensemble learning and SVM-based learning. This constitutes evidence that dense motion trajectories are particularly useful (i.e., discriminative) features for visual analysis of videos. The superiority of the dynamic visual feature (i.e., dense motion trajectories) can be explained by the fact that affect present in video clips is often characterized by motion (e.g., camera motion). Another observation concerns the performance gain (around

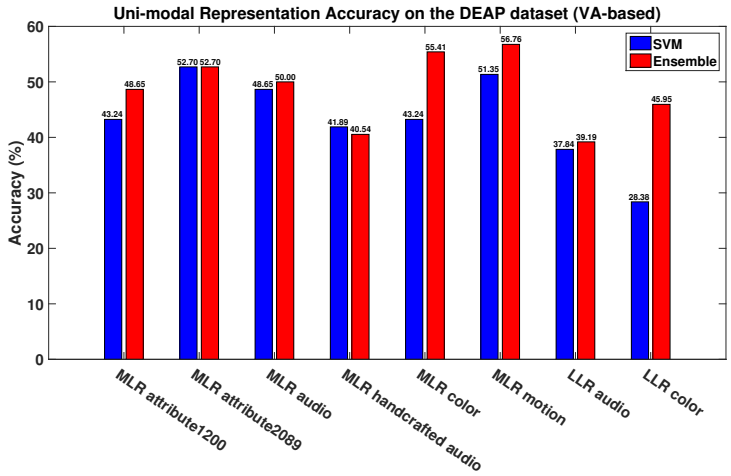


Fig. 4 *VA-based classification* accuracies on the DEAP dataset (with uni-modal representations: audio or visual-only, MLR: mid-level representation, LLR: low-level representation).

10%) of using learned color features compared to low-level ones. When evaluated together with our previous findings about learning color representations in [1], we can conclude that color values in the HSV space lead to more discriminative mid-level representations than color values in the RGB space. Concerning domain-specific representations, using 2,089 trained visual concept detectors instead of 1,200 provides a noticeable increase (around 4%) in terms of classification performance.

Table 3 *VA-based classification* accuracies of multi-modal audio-visual representations on the DEAP dataset (MLR: mid-level representation, LLR: low-level representation).

	Multi-modal Representation	No attribute	MLR attr1200	MLR attr2089
Linear fusion	*MLR motion & LLR audio*	70.27	**74.32**	**74.32**
	MLR motion & MLR audio	75.68	**78.38**	**78.38**
	MLR motion & LLR color	67.57	68.92	**74.32**
	MLR motion & MLR color	70.27	70.27	**75.68**
	MLR motion & LLR audio & LLR color	72.97	74.32	**75.68**
	MLR motion & LLR audio & MLR color	75.68	78.38	**79.73**
	MLR motion & MLR audio & LLR color	72.97	75.68	**77.03**
	MLR motion & MLR audio & MLR color	77.03	78.38	**81.08**
SVM-based fusion	*MLR motion & LLR audio*	67.57	**70.27**	**70.27**
	MLR motion & MLR audio	68.92	71.62	**72.97**
	MLR motion & LLR color	60.81	60.81	**63.51**
	MLR motion & MLR color	64.87	64.87	**66.22**
	MLR motion & LLR audio & LLR color	66.22	66.22	**67.57**
	MLR motion & LLR audio & MLR color	70.27	71.62	**74.32**
	MLR motion & MLR audio & LLR color	66.22	66.22	**70.27**
	MLR motion & MLR audio & MLR color	71.62	71.62	**75.68**

In **Table 3**, we present the classification performances (according to the *VA-based* annotations) of different combinations of audio-visual representations using linear

and SVM-based fusion. As mentioned earlier, dense motion trajectories are discriminative features for representing videos. We, therefore, primarily consider them for assessing feature combinations. Consequently, we combine them with audio, color and domain-specific features of different levels (i.e., low-level or mid-level) in order to evaluate the discriminative power of those combinations. From Table 3, we can derive the following observations which apply both to linear and SVM-based fusion.

First, the performance gain of combining mid-level motion features with mid-level audio and color ones is higher than the gain of combining them with low-level audio and color ones. In addition, the results show that combining low-level color features with mid-level motion and audio ones leads to a decrease in classification accuracy (i.e., leads to more confusion between classes). On the contrary, combining mid-level learned color features with mid-level motion and audio ones leads to less confusion between classes (i.e., increased classification accuracy). We also observe that including domain-specific representations in the final decision process improves classification accuracy; as regards this aspect, using 2,089 trained visual concept detectors instead of 1,200 for domain-specific representations increases classification accuracy in general. As a final remark, concerning fusion schemes, SVM-based fusion leads to poorer results compared to linear fusion. We suspect that this is due to the cascaded classification error introduced by an additional classification (i.e., model generation) layer in the system.

In order to give an overview of the misclassification behavior of the system, we present the confusion matrices on the DEAP dataset in Figure 5. We observe that *HA-LV* (i.e., high arousal and low valence) is the class that can be discriminated at the highest level. The common characteristic of these misclassified classes is that the number of instances in both the training and test sets are limited.

When looking in more detail at the confusion matrices in Figure 5, it appears that the confusion between classes is mostly between neighboring classes, i.e., neighboring emotions more likely to resemble each other. Therefore, plotting the *Cumulative Matching Characteristic (CMC)* curves is a more appropriate choice to present the performance of the system as a function of the *distance between classes*, similar to the approach adopted in [45]. We define the distances between classes as follows: (1) the distance between two classes that are on the same quadrant of the VA-space is 1; (2) the distances between two classes of different quadrants is defined as $d_q + 1$, where d_q corresponds to the number of quadrants encountered when departing from one class in order to reach the other class. For instance, the distance between *HA-LV* and *HA-HV* is 1, whereas it is 2 for *HA-LV* and *LA-HV*. We provide the CMC curve in **Figure 6**. According to this graph, when relaxing the conditions by taking into account the CMC, the accuracies appear less pessimistic.

As a final evaluation on the DEAP dataset, **Table 4** provides the classification accuracies of our method (i.e., ensemble learning using MLR audio, motion, color and domain-specific representations linearly fused at the decision-level) compared to the works [1], [2] and [44] to position our approach in relation to these prior approaches (evaluation pertaining to the research question *RQ3*). Our method outperforms the works [1], [2] and [44] by achieving 81.08% accuracy. The paired Student t-test on classification accuracies from the *leave-one-song-out* cross validation showed that

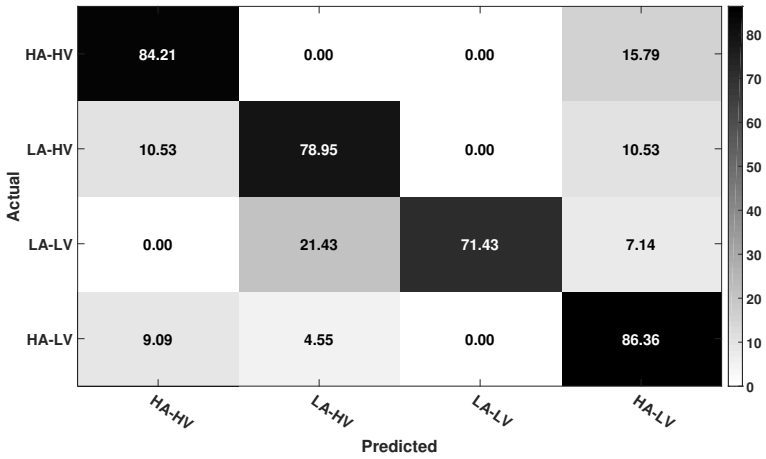


Fig. 5 Confusion matrices for the *VA-based classification* on the DEAP dataset with the best performing multi-modal audio-visual representation (i.e., MLR audio, motion, color and domain-specific representations). Fusion method is linear fusion. Mean accuracy: 81.08%. Darker areas along the main diagonal correspond to better discrimination.

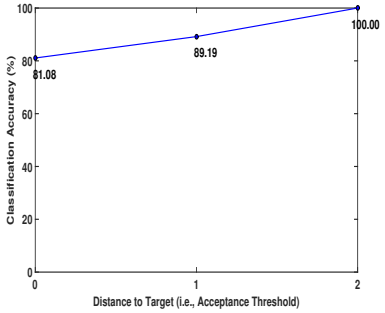


Fig. 6 Cumulative Matching Characteristic (CMC) curve for the *VA-based classification* on the DEAP dataset with the best performing multi-modal audio-visual representation (i.e., MLR audio, motion, color and domain-specific representations). Fusion method is linear fusion.

the improvement over the works [1] and [2] is statistically significant at the 5% significance level.

Table 4 *VA-based classification* accuracies on the DEAP dataset (with audio-visual representations, *MLR audio-visual features*: MLR audio, motion, color and domain-specific).

Method	**Accuracy (%)**
Our method *(MLR audio-visual features & ensemble learning & linear fusion)*	81.08
Acar et al. (MLR audio, color and motion & SVM learning and fusion) [2]	66.22
Acar et al. (MLR audio and color & SVM learning and fusion) [1]	50.00
Yazdani et al. [44]	36.00

The results in Table 4 demonstrate the potential of our approach for video affective content analysis. Only a subset of 40 video clips from the DEAP dataset form the basis of the experiments in [44]. Therefore, a comparison is biased towards our approach due to the increased dataset (74 clips). On the other hand, the 40 music video clips used in [44] were selected so that only the music video clips which induce strong emotions are used. Therefore, the dataset we used in this paper can be regarded as more challenging. Another difference with the setup of [44] is that, therein, the authors used the user ratings from laboratory experiments instead of the online self-assessment ratings mentioned in Section 4.1 as the ground-truth.

4.3.2 Evaluation on User-generated Videos (VideoEmotion)

In this section, we address the case of the VideoEmotion dataset. VideoEmotion is a dataset made of user-generated Flickr and YouTube videos, and, hence, is more challenging compared to the DEAP dataset which is made of professionally edited videos. Further, the videos in VideoEmotion are annotated only according to Plutchik's wheel, which implies that the results presented in this section only cover *wheel-based classification*.

As previously, we start with the MLR audio-ATM comparison. The findings for VideoEmotion are in accordance with the comparison realized for DEAP. MLR audio features provide an improvement ranging from 1.66% to 3.90% in classification accuracy over ATM. On the entire VideoEmotion set, we obtained for MLR-audio associated with SVM and ensemble learning 34.19% and 40.33%, respectively. For ATM associated with SVM and ensemble learning, the figures are 30.29% and 37.20%, respectively. On the VideoEmotion subset, the results are 42.17%, 50.14%, 40.51%, 46.96%. The paired Student t-test revealed again statistically significant results (5% significance level).

We continue with the classification accuracies in the case where only one type of descriptor is employed (Figures 7 and 8). This enables to assess the discriminative power of learned audio-visual, motion and SentiBank representations, and also to compare the bagging approach against SVM modeling (assessment relating to the research questions *RQ1*, *RQ2* and *RQ3*).

Figure 7 presents the performance of each descriptor on the entire VideoEmotion dataset which is annotated according to eight emotion categories as explained in Section 4.1. When compared to the classification performance reported for the DEAP dataset in Figure 4, classification accuracies are lower for the VideoEmotion dataset. This can be explained by the fact that VideoEmotion contains videos which are not necessarily recorded or edited by professionals, and therefore, is more challenging compared to DEAP.

The conclusions which can be drawn from Figure 7 for VideoEmotion are mostly in concordance with the ones drawn in Section 4.3.1 for DEAP. Concerning classifiers, ensemble learning improves the discrimination power of uni-modal representations over SVM-based learning in general. Concerning features, SentiBank domain-specific representations are the best performing descriptors (especially when 2,089

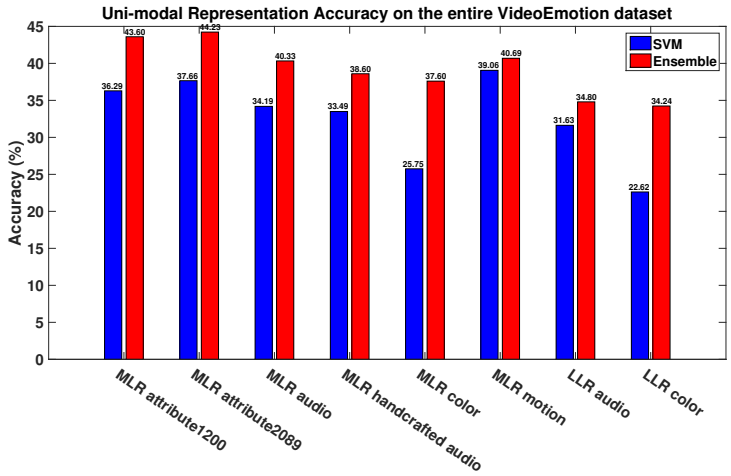


Fig. 7 *Wheel-based classification* accuracies on the entire VideoEmotion dataset (with uni-modal representations: audio or visual-only, MLR: mid-level representation, LLR: low-level representation).

trained visual concept detectors are used instead of 1,200), followed by mid-level motion and audio descriptors. In addition, all mid-level representations (learned or handcrafted) outperform the low-level audio and visual representations. Further, learned audio representations outperform handcrafted mid-level audio ones.

However, there is one significant difference in comparison to the results for DEAP. Although mid-level motion descriptors are still one of the most discriminative descriptors, they are no longer the best performing ones. This can be explained by the fact that motion present in a professionally edited video is "deliberately" present to elicit specific emotions in the audience, whereas this might not be the case for user-generated videos.

Figure 8 presents the performance of each descriptor on a subset of the VideoEmotion dataset, where we have four basic emotion categories as explained in Section 4.1. As a preliminary remark, we note that the results on Figure 8 are globally better than those on Figure 7. The explanation for this discrepancy lies in the lower number of classes to be discriminated (4 against 8), which means that the risk of confusion is reduced.

The conclusions concerning the subset of VideoEmotion (Figure 8) are similar to the ones derived for the whole set (Figure 7). Although the improvement it provides is lower on the subset (in comparison to the whole set), ensemble learning still outperforms SVM-based learning. SentiBank and mid-level motion representations are the best performing uni-modal descriptors.

In the following paragraphs, we present the *wheel-based classification* performances of different combinations of audio-visual representations using linear and SVM-based fusion (**Table 5** for the entire VideoEmotion dataset; **Table 6** for the VideoEmotion subset), where the best performing multi-modal representation and

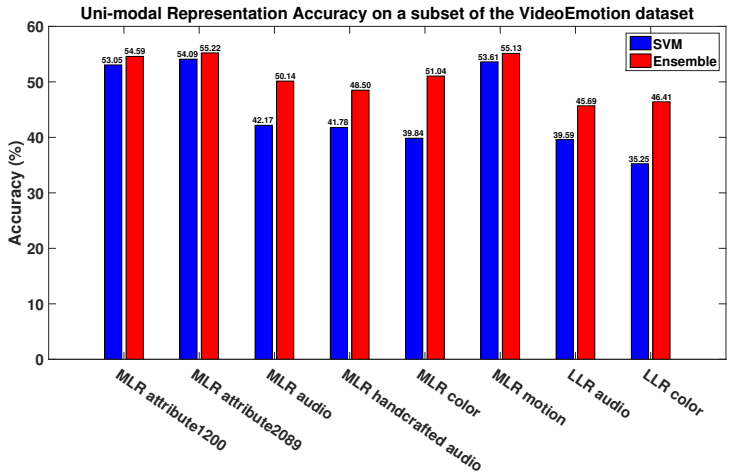


Fig. 8 *Wheel-based classification* accuracies on the VideoEmotion subset (with uni-modal representations: audio or visual-only, MLR: mid-level representation, LLR: low-level representation).

optimal decision-level fusion mechanisms are investigated (experiment relevant to the research questions *RQ1*, *RQ2* and *RQ4*).

The feature combinations we consider involve dense motion trajectories and SentiBank domain-specific representations. This choice is justified by the findings concerning individual features. The experiments have indeed demonstrated that dense motion trajectories and SentiBank domain-specific representations are more discriminative features. Further, mid-level learned or handcrafted audio-visual representations are more discriminative than low-level ones. Therefore, we combine the dense motion trajectory and SentiBank representations with mid-level audio and color features in order to evaluate different multi-modal audio-visual representations.

Table 5 *Wheel-based classification* accuracies of multi-modal audio-visual representations on the entire VideoEmotion dataset (MLR: mid-level representation, hc: handcrafted).

	Multi-modal Representation	No attrib	MLR attr1200	MLR attr2089
Linear fusion	*MLR motion & MLR hc audio*	44.60	45.85	46.87
	MLR motion & MLR audio	45.16	46.78	48.05
	MLR motion & MLR color	43.87	45.32	46.32
	MLR motion & MLR hc audio & MLR color	45.50	46.85	47.45
	MLR motion & MLR audio & MLR color	46.66	47.33	**49.19**
	MLR motion & MLR attribute1200	43.08	N/A	N/A
	MLR motion & MLR attribute2089	43.33	N/A	N/A
SVM-based fusion	*MLR motion & MLR hc audio*	43.78	43.87	44.87
	MLR motion & MLR audio	44.32	44.78	46.05
	MLR motion & MLR color	43.24	43.71	45.08
	MLR motion & MLR hc audio & MLR color	44.34	45.17	46.27
	MLR motion & MLR audio & MLR color	45.15	46.39	**47.18**
	MLR motion & MLR attribute1200	42.60	N/A	N/A
	MLR motion & MLR attribute2089	43.14	N/A	N/A

The first observation (according to **Table 5**) is that learned audio-visual representations when combined with motion representations perform better than the combination of motion and SentiBank domain-specific features. In addition, combining motion and SentiBank representations with learned audio features achieves better results than combining them with mid-level handcrafted audio ones. Concerning domain-specific representations, the use of 2,089 trained visual concept detectors instead of 1,200 provides again an increase in classification accuracy. Our final observation is that the best classification performance (highlighted in the related tables) is achieved by combining learned audio-visual representations with motion and SentiBank (i.e., *MLR attribute2089*) at decision-level and that simple linear fusion is superior to SVM-based fusion. As evoked earlier, this is likely caused by the cascaded classification error due to an added classification (i.e., model generation) layer in the system.

Table 6 presents the performance of multi-modal audio-visual representations on the VideoEmotion subset using linear and SVM-based fusion. We can draw conclusions which match those deduced for the entire VideoEmotion dataset. The only important difference is that SentiBank performs much better than learned audio-visual representations when combined with motion representations. This result is actually on par with the results presented in [22], where attribute features that include SentiBank (*MLR attribute1200*) are shown to outperform audio-visual representations. The difference between classification accuracies (when compared to Table 5) can be explained by the increased risk of confusion due to the number of classes.

Table 6 *Wheel-based classification* accuracies of multi-modal audio-visual representations on the VideoEmotion subset (MLR: mid-level representation, hc: handcrafted).

	Multi-modal Representation	No attrib	MLR attr1200	MLR attr2089
Linear fusion	*MLR motion & MLR hc audio*	54.59	57.47	61.81
	MLR motion & MLR audio	55.39	58.53	62.36
	MLR motion & MLR color	55.63	57.01	61.14
	MLR motion & MLR hc audio & MLR color	56.00	59.30	62.16
	MLR motion & MLR audio & MLR color	57.01	60.34	**63.65**
	MLR motion & MLR attribute1200	57.77	N/A	N/A
	MLR motion & MLR attribute2089	60.53	N/A	N/A
SVM-based fusion	*MLR motion & MLR hc audio*	52.25	56.73	60.40
	MLR motion & MLR audio	54.82	57.21	61.77
	MLR motion & MLR color	55.20	56.37	59.91
	MLR motion & MLR hc audio & MLR color	56.50	58.88	61.34
	MLR motion & MLR audio & MLR color	57.58	59.56	**62.16**
	MLR motion & MLR attribute1200	55.59	N/A	N/A
	MLR motion & MLR attribute2089	59.67	N/A	N/A

We present the confusion matrices on the entire VideoEmotion dataset in Figure 9(a) and on the VideoEmotion subset in Figure 9(b). In **Figure 9(a)**, we observe that *Surprise* is the class that can be discriminated the most. The *Anticipation* and *Trust* classes are difficult to differentiate; it seems that these classes do not contain clear audio-visual cues. In **Figure 9(b)** (where only four basic emotions are considered), we observe that the *Joy* and *Fear* classes can be very well discriminated from

other classes, whereas *Sadness* is the class with the lowest recognition rate. As in the results on professionally edited videos (Section 4.3.1), the confusion between classes mostly occurs between neighboring classes. Therefore, we again plot the CMC curves in order to show the performance of the system as a function of the *distance between classes*. As Plutchik's wheel is used for the VideoEmotion dataset, the distances between classes are defined slightly differently than in Section 4.3.1. Here, the distance between any two classes is defined as the minimum number of emotion leaves encountered when going from one class to the other in the emotion wheel (Figure 1). For instance, the distance between *Fear* and *Anger*, which are opposite emotions, is 4. Similarly, the distance between *Sadness* and *Anticipation* is 3.

We provide the CMC curves on the entire VideoEmotion dataset in **Figure 10(a)** and on the VideoEmotion subset in **Figure 10(b)**. As stated earlier, with the use of these distances, classification accuracies achieved on both datasets are revealed to be higher, when the acceptable threshold for predictions is set to 2.

As a final evaluation on the VideoEmotion dataset, **Table 7** provides the classification accuracies of our method (i.e., ensemble learning using MLR audio, motion, color and domain-specific representations linearly fused at the decision-level) compared to the works [22] and [28] to position our approach in relation to these prior approaches (addressed research question *RQ3* in Section 1). Our method outperforms the works [22] and [28] by achieving 49.19% and 63.75% accuracies for the entire VideoEmotion dataset and the VideoEmotion subset, respectively.

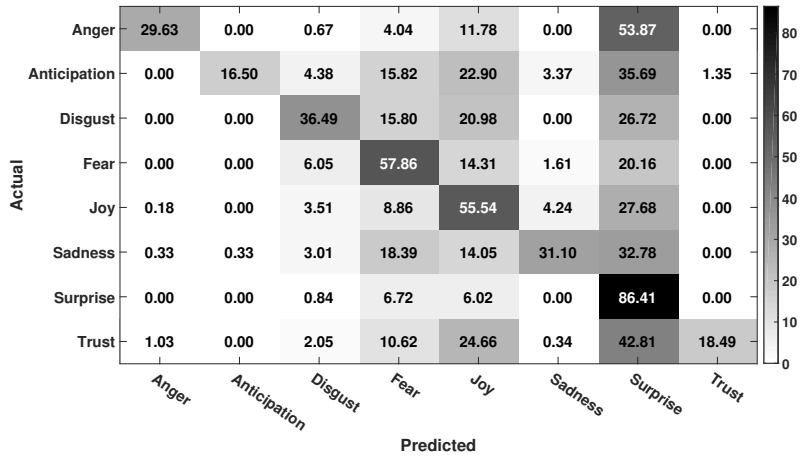
Table 7 *Wheel-based classification* accuracies on the VideoEmotion dataset (with audio-visual representations).

Method	**Accuracy – Entire (%)**	**Accuracy – Subset (%)**
Our method *– MLR audio-visual features & ensemble learning & linear fusion*	49.19	63.75
Jiang et al. [22]	46.10	60.50
Pang et al. [28]	37.10	-

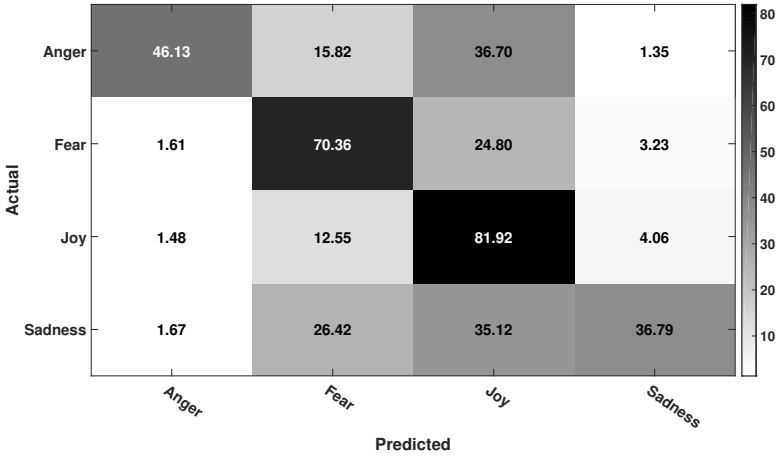
4.3.3 Summary of Evaluation Results: The Bottom Line

In this section, we identify the common denominator of the evaluation results on both professionally edited and user-generated videos, i.e., the findings consistent across both datasets. This also enables us to answer the research questions posed in Section 1.

- Regarding *RQ1* and *RQ2*, we investigated the discriminative power of uni and multi-modal representations. Our findings from the perspective of feature representations are as follows:
  1. When considering uni-modal representations, we observe that SentiBank and dense trajectory-based motion representations are the most discriminative features for emotional content analysis of both professionally edited and user-generated videos.



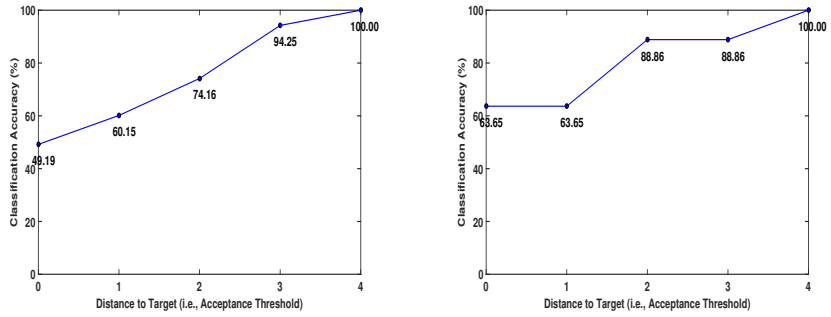
(a)



(b)

Fig. 9 Confusion matrices for the *wheel-based classification* (a) on the entire VideoEmotion dataset and (b) on the VideoEmotion subset with the best performing multi-modal audio-visual representation (i.e., MLR audio, motion and domain-specific representations) using linear fusion. Darker areas along the main diagonal correspond to better discrimination. Mean accuracy: (a) 49.19%, (b) 63.65%.

2. For SentiBank, using 2,089 trained visual concept detectors instead of 1,200 provides a noticeable increase in terms of classification performance. We closely looked at the importance of individual ANPs to check if some of them were consistently playing an important role across all videos. There appears to be no ANP which clearly stands out and which is represented in all videos of both datasets. Therefore, we can only conclude that increasing the number of ANPs helps increasing classification accuracies.



(a) (b)

Fig. 10 Cumulative Matching Characteristic (CMC) curves for the *wheel-based classification* (a) on the entire VideoEmotion dataset and (b) on the VideoEmotion subset with the best performing multi-modal audio-visual representation (i.e., MLR audio, motion and domain-specific representations) using linear fusion.

3. Learned audio and color representations are more discriminative than handcrafted low and mid-level representations.

– Regarding *RQ3* and *RQ4*, we explored the modeling perspective of video affective content analysis. Our findings from the perspective of modeling are as follows:
  1. As a result of the extensive experiments conducted on both datasets, ensemble learning (i.e., decision tree based bagging) is superior to SVM-based learning in emotion modeling of videos.
  2. Fusing the outputs of ensemble learning models using a simpler fusion method (i.e., linear fusion) has proven to be more effective than an advanced fusion mechanism (i.e., SVM-based fusion).

5 Conclusions

In this paper, we presented a promising approach for the affective labeling of professionally edited and user-generated videos using mid-level multi-modal representations and ensemble learning.

We concentrated both on the representation and modeling aspects. Concerning the aspects relating to representation, higher level representations were learned from raw data using CNNs and fused with dense trajectory based motion and SentiBank domain-specific features at the decision-level. As a basis for feature learning, MFCC was employed as audio feature, while color values in the HSV space formed the static visual features. Concerning the aspects relating to modeling, we applied ensemble learning, *viz.* decision tree based bagging, to classify each video into one of the predefined emotion categories.

Experimental results on the VideoEmotion dataset and on a subset of the DEAP dataset support our assumptions (1) that learned audio-visual representations are more discriminative than handcrafted low-level and mid-level ones, (2) that including dense trajectories and SentiBank representations contribute to increase the classifica-

tion performance, and (3) that ensemble learning is superior to multi-class SVM for video affective content analysis. In addition, we have demonstrated that fusing the outputs of ensemble learning models using a simpler fusion method (i.e., linear fusion) is more effective than an advanced fusion mechanism (i.e., SVM-based fusion).

As future work, we plan to focus on the optimization of the ensemble learning method in order to further improve performance. In particular, we will experiment more advanced classifiers within the ensemble learning framework, such as SVMs [30].

Acknowledgements The research leading to these results has received funding from the European Community FP7 under grant agreement number 261743 (NoE VideoSense).

References

1. Acar, E., Hopfgartner, F., Albayrak, S.: Understanding affective content of music videos through learned representations. In: International Conference on MultiMedia Modelling (MMM), pp. 303–314 (2014)
2. Acar, E., Hopfgartner, F., Albayrak, S.: Fusion of learned multi-modal representations and dense trajectories for emotional analysis in videos. In: IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–6 (2015)
3. Baveye, Y., Bettinelli, J., Dellandréa, E., Chen, L., Chamaret, C.: A large video database for computational models of induced emotion. In: Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 13–18 (2013)
4. Baveye, Y., Dellandréa, E., Chamaret, C., Chen, L.: Deep learning vs. kernel methods: Performance for emotion prediction in videos. In: International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 77–83 (2015)
5. Baveye, Y., Dellandréa, E., Chamaret, C., Chen, L.: LIRIS-ACCEDE: A video database for affective content analysis. IEEE Transactions on Affective Computing **6**(1), 43–55 (2015)
6. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **35**(8), 1798–1828 (2013)
7. Borth, D., Chen, T., Ji, R., Chang, S.: Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In: ACM International Conference on Multimedia (ACMMM), pp. 459–460 (2013)
8. Canini, L., Benini, S., Leonardi, R.: Affective recommendation of movies based on selected connotative features. IEEE Transactions on Circuits and Systems for Video Technology **23**(4), 636–647 (2013)
9. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2**(3), 1–27 (2011)
10. Chen, T., Borth, D., Darrell, T., Chang, S.: Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. Computing Research Repository (CoRR) **abs/1410.8586** (2014)
11. Chen, T., Yu, F.X., Chen, J., Cui, Y., Chen, Y., Chang, S.: Object-based visual sentiment concept analysis and application. In: ACM International Conference on Multimedia (ACMMM), pp. 367–376 (2014)
12. Dumoulin, J., Affi, D., Mugellini, E., Khaled, O.A., Bertini, M., Bimbo, A.D.: Affect recognition in a realistic movie dataset using a hierarchical approach. In: First International Workshop on Affect & Sentiment in Multimedia (ASM), pp. 15–20 (2015)
13. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of statistics **32**(2), 407–499 (2004)
14. Eggink, J., Bland, D.: A large scale experiment for mood-based classification of tv programmes. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 140–145 (2012)
15. Ellis, J.G., Lin, W.S., Lin, C., Chang, S.: Predicting evoked emotions in video. In: IEEE International Symposium on Multimedia (ISM), pp. 287–294 (2014)

16. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image and Vision Computing **31**(2), 120–136 (2013)
17. Irie, G., Hidaka, K., Satou, T., Yamasaki, T., Aizawa, K.: Affective video segment retrieval for consumer generated videos based on correlation between emotions and emotional audio events. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 522–525 (2009)
18. Irie, G., Satou, T., Kojima, A., Yamasaki, T., Aizawa, K.: Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. IEEE Transactions on Multimedia **12**(6), 523–535 (2010)
19. Jeannin, S., Divakaran, A.: Mpeg-7 visual motion descriptors. IEEE Transactions on Circuits and Systems for Video Technology **11**(6), 720–724 (2001)
20. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **35**(1), 221–231 (2013)
21. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia (ACMMM), pp. 675–678 (2014)
22. Jiang, Y., Xu, B., Xue, X.: Predicting emotions in user-generated videos. In: The AAAI Conference on Artificial Intelligence (AAAI) (2014)
23. Koelstra, S., Mühl, C., Soleymani, M., Lee, J., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (2012)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
25. Li, T.L., Chan, A.B., Chun, A.H.: Automatic musical pattern feature extraction using convolutional neural network. In: International MultiConference of Engineers and Computer Scientists (IMECS) (2010)
26. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research **11**, 19–60 (2010)
27. Niu, J., Zhao, X., Abdul Aziz, M.A.: A novel affect-based model of similarity measure of videos. Neurocomputing (in press) (2015)
28. Pang, L., Ngo, C.W.: Multimodal learning with deep boltzmann machine for emotion prediction in user generated videos. In: ACM International Conference on Multimedia Retrieval (ICMR), pp. 619–622 (2015)
29. Plutchik, R., Kellerman, H.: Emotion: theory, research and experience, vol. 3. Academic press New York, NY (1986)
30. Safadi, B., Quénot, G.: A factorized model for multiple SVM and multi-label classification for large scale multimedia indexing. In: 13th International Workshop on Content-Based Multimedia Indexing, CBMI 2015, Prague, Czech Republic, June 10-12, 2015, pp. 1–6 (2015)
31. Schmidt, E., Scott, J., Kim, Y.: Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In: International Society for Music Information Retrieval Conference (ISMIR), pp. 325–330 (2012)
32. Soleymani, M., Aljanaki, A., Wiering, F., Veltkamp, R.C.: Content-based music recommendation using underlying music preference structure. In: Multimedia and Expo (ICME), 2015 IEEE International Conference on, pp. 1–6 (2015)
33. Sturm, B.L., Noorzad, P.: On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In: International Symposium on Computer Music Modeling and Retrieval, pp. 379–394 (2012)
34. Valdez, P., Mehrabian, A.: Effects of color on emotions. Journal of Experimental Psychology: General **123**(4), 394–409 (1994)
35. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 3551–3558 (2013)
36. Wang, H.L., Cheong, L.: Affective understanding in film. IEEE Transactions on Circuits and Systems for Video Technology **16**(6), 689–704 (2006)
37. Wang, S., Ji, Q.: Video affective content analysis: A survey of state-of-the-art methods. IEEE Transactions on Affective Computing **6**(4), 410–430 (2015)
38. Wimmer, M., Schuller, B., Arsic, D., Rigoll, G., Radig, B.: Low-level fusion of audio and video feature for multi-modal emotion recognition. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 145–151 (2008)
39. fan Wu, T., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5**, 975–1005 (2003)

40. Xu, B., Fu, Y., Jiang, Y., Li, B., Sigal, L.: Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. Computing Research Repository (CoRR) **abs/1511.04798** (2015)
41. Xu, C., Cetintas, S., Lee, K., Li, L.: Visual sentiment prediction with deep convolutional neural networks. Computing Research Repository (CoRR) **abs/1411.5731** (2014)
42. Xu, M., Wang, J., He, X., Jin, J.S., Luo, S., Lu, H.: A three-level framework for affective content analysis and its case studies. Multimedia Tools and Applications (MTAP) **70**(2), 757–779 (2014)
43. Yang, X., Wang, K., Shamma, S.A.: Auditory representations of acoustic signals. IEEE Transactions on Information Theory **38**(2), 824–839 (1992)
44. Yazdani, A., Kappeler, K., Ebrahimi, T.: Affective content analysis of music video clips. In: ACM International Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM), pp. 7–12 (2011)
45. Yucel, Z., Salah, A.A.: Resolution of focus of attention using gaze direction estimation and saliency computation. In: International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–6 (2009)
46. Zhou, Z.: Ensemble methods: foundations and algorithms. CRC Press (2012)