EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Multimedia content analysis for emotional characterization of music video clips

Ashkan Yazdani[1*], Evangelos Skodras[2], Nikolaos Fakotakis[2] and Touradj Ebrahimi[1]

Abstract

Nowadays, tags play an important role in the search and retrieval process in multimedia content sharing social networks. As the amount of multimedia contents explosively increases, it is a challenging problem to find a content that will be appealing to the users. Furthermore, the retrieval of multimedia contents, which can match users' current mood or affective state, can be of great interest. One approach to indexing multimedia contents is to determine the potential affective state, which they can induce in users. In this paper, multimedia content analysis is performed to extract affective audio and visual cues from different music video clips. Furthermore, several fusion techniques are used to combine the information extracted from the audio and video contents of music video clips. We show that using the proposed methodology, a relatively high performance (up to 90%) of affect recognition is obtained.

1 Introduction

With the rapid evolution of digital media, services such as social networks, web-based multimedia archives, and search engines are becoming increasingly popular, and consequently, the amount of multimedia data on the Internet is escalating exponentially. Therefore, developing appropriate schemes for efficiently performing search and retrieval in immense multimedia databases becomes significantly important. In general, there are two approaches to assigning tags to a given content, namely, *explicit* and *implicit* tagging. The former refers to a user's explicit action of manually entering appropriate keywords associated with the content, whereas in the latter approach, users do not necessarily input tags. Instead, automatic analysis of the users' behavior is used to generate tags for the content.

Explicit tagging is used in most of the currently available social network-based systems such as YouTube and Flickr. Nevertheless, it cannot be considered as the ultimate solution for assigning tags to multimedia contents due to the following reasons: Firstly, manual explicit annotation of the enormous amount of multimedia data on the Internet is clearly infeasible, and hence, a large portion of the multimedia data remains untagged, which in turn imposes limits on the search and retrieval process and deteriorates the performance of multimedia search engines. Secondly, users who annotate multimedia contents do not necessarily aim at improving the performance of the current retrieval systems. In fact, personal and social motivations often underlie the annotation [1]: A personal need-driven tag may be meaningless to other users, e.g., 'my lovely granddaughter in my place'. Furthermore, it is also possible that some users generate tags for other purposes, e.g., spam tags for advertisement. Therefore, complementary and/or alternative annotation methods are needed to overcome the aforementioned shortcomings of explicit tagging. Implicit tagging can be considered as one of such solutions. There exist numerous studies on exploring appropriate cues which can be detected by observing and analyzing users' behavior, such as emotion, level of interest, and attention. It has been shown that such cues can thus be employed in implicit tagging applications [2].

Among various kinds of information that can be obtained for the purpose of implicit tagging, emotional information about a given content is of great interest. Emotional information plays an important role for personalized content delivery [3]. For example, users may prefer to watch video clips containing funny contents when they feel sad or depressed in order to improve their

*Correspondence: ashkan.yazdani@epfl.ch
[1]Multimedia Signal Processing Group (MMSPG), Institute of Electrical Engineering (IEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, 1015, Switzerland
Full list of author information is available at the end of the article

mood. As another example, some might not want to watch video clips containing scary or violent scenes.

In such cases, emotional information about the content can be used effectively in search and retrieval. Emotional cues can be obtained by automatic analysis of facial expression, voice and speech [4], physiological signals (e.g., respiration, Galvanic skin resistance, skin temperature, eye blinking rate, electromyogram, electroencephalograph, and blood flow) [5], or their combinations. Such cues can be used to create emotional tags or tags about the different genres (horror, humor, etc.). However, in such methods, the consumers of multimedia contents should be under surveillance or need to go through cumbersome physiological electrode setups, which restricts the usage of these modalities for affective annotation.

Another approach to extracting affective cues is to analyze the multimedia contents themselves. The technology required to achieve this goal is referred to as multimedia content analysis (MCA). MCA aims at bridging the 'semantic gap', that is, to develop models of the relationship between low-level features and the semantics conveyed by multimedia contents [6]. Analyzing a multimedia content at an affective level reveals information that describes its emotional value. This value can be defined as the amount and type of affect (feeling or emotion) of the audience while they consume this multimedia content.

In order to analyze a given multimedia content at an affective level, an appropriate modeling for emotion must be developed. The issue of how to represent and model emotions is, however, a challenging task. To this date, numerous theorists and researchers have conducted research on this subject. Generally, there are two different families of emotion models: the *categorical* and the *dimensional* models. The rationale behind categorical models is to have discrete basic categories of emotions from which every other emotion can be built by combining these basic emotions. The most common basic emotions are 'fear', 'anger', 'sadness', 'joy', 'disgust', and 'surprise', first introduced by Ekman [7]. The dimensional models, on the other hand, describe the components of emotions and are often represented as a two- or three-dimensional space where the emotions are presented as points in the coordinate space of these dimensions [8].

The goal of the underlying dimensional model is not to find a finite set of emotions as in the categorical model but to find a finite set of principal components of emotions. Many theorists have proposed that emotions can be modeled with three underlying dimensions, namely, valence (V), arousal (A), and dominance (D). The dimension 'valence' provides information about the degree of pleasantness of the content and ranges from pleasant (positive) to unpleasant (negative). The dimension 'arousal' represents the inner activation and ranges from energized to calm. The third dimension 'dominance' (control) helps to distinguish between 'grief' and 'rage' and goes from no control to full control [3].

Psycho-physiological experiments have revealed that only certain areas of the three-dimensional V-A-D space are relevant. In these experiments, affective responses of a large group of subjects are included. The stimuli are collected from the International Affective Picture System [9] and the International Affective Digitized Sounds System [10]. Figure 1a shows the relevant areas of the three dimensional V-A-D space.

The V-A model is a simplified two-dimensional model of the V-A-D model. The underlying dimensions, here, are valence and arousal. A large number of emotional states can be represented using this two-dimensional model, and many studies about affective content analysis use this simplified model as, for example, in [3] or in [5]. Similar to the V-A-D model, only some parts of the 2-D space are relevant as shown in Figure 1b. The dimensional models are often used in affect assessment studies since basic emotions can be represented as different areas on V-A or V-A-D coordinates. Figure 2 demonstrates the distribution of different emotional categories on the arousal-valence space.

In this paper, MCA is performed to extract affective cues from different music video clips. The inference of evoked emotions in music videos can be regarded as a better formulated problem than emotion detection in movies, in the following sense: in movies, the perceived emotions are mostly depending on the scenario, the high-level semantics and the speech, which is usually the predominant component. Music usually has a more limited role in movies, used in certain scenes throughout the movie to evoke or enhance particular emotions. On the contrary, in music videos clips, emotions are much more dependent on the characteristics of the audio and video than on the plot or the semantics of the speech. Furthermore, this paper deals with emotions contained in the music video

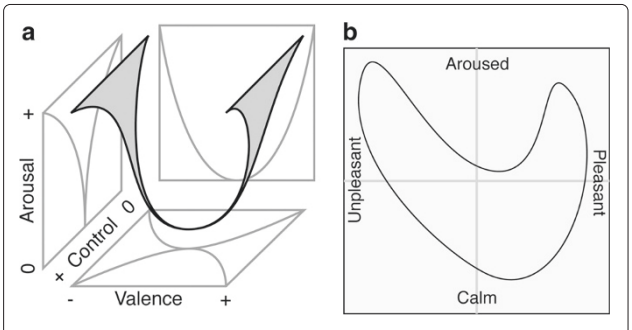

Figure 1 **Relevant areas of the (a) 3-D emotion space and (b) 2-D emotion space [3].**

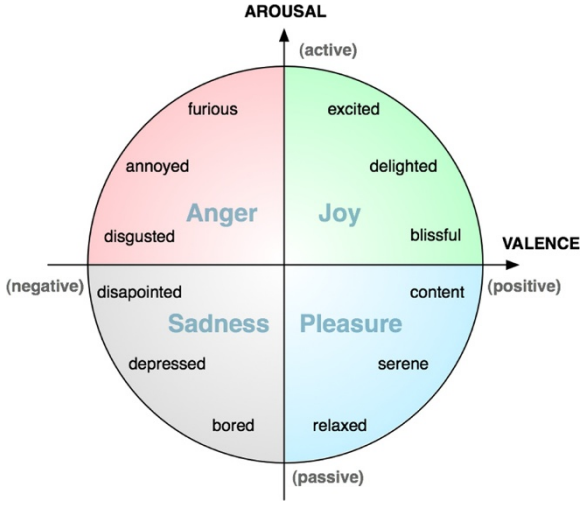

Figure 2 Distribution of different emotions on arousal-valence space.

clips and not the emotions induced in the participants while watching music video clips. In our previous paper [11], we investigated the latter approach, i.e., the possibility of inferring emotional states induced in individual users while watching music video clips. However, in the current paper, we are interested in the emotional characterization of music video clips.

For evaluation of the proposed methodology, the V-A-D emotional model is used. We present the results of affect recognition using audio and video channels of the music video clips, individually. Furthermore, several fusion protocols used to combine the video and audio information are presented. The main contributions of this paper are summarized in the use of music video clips as emotion-inducing stimuli and in the fusion of audio and video information in an efficient manner. The rest of the paper is organized as follows. Section 2 provides the description of the database and the methodology used in this study for feature extraction, classification, and fusion. Section 3 presents the results obtained, and section 4 concludes the paper.

2 Affective multimedia modeling for music videos

Affective multimedia content analysis and modeling is a complex problem. In general, machine learning approaches are used to deal with this problem, and low-level features are extracted from the multimedia contents in order to be mapped to different self-assessed emotions. The aim of this section is to describe how affective multimedia content analysis is performed. In the first part of the section, the database of music video clips used in this study is described. Then, the features which are extracted from the music video clips (audio and video features) are introduced, and the classification method used in this study is explained. Finally, in the last part of this section, different information fusion protocols used in this study are introduced.

2.1 Related work

Several studies in literature present strong evidence on the dependence between low-level features and emotions perceived. In [12], a good overview of affective MCA systems is presented, and a roadmap for development of personalized multimedia content recommendation is drawn. In [13], the authors developed a hidden Markov model (HMM)-based affective content analysis system to map low-level features of video data to high-level emotional events. In their study, they discovered a strong relationship between the low-level features of motion, color, and shot cut rate and the emotions of fear/anger, joy, and sadness. In [14], the authors used sound effects from comedy and horror movies to detect audio emotional events. They trained a four-state HMM to detect 'amusement' and 'fear' emotions. Other researchers studied the feasibility of using sound energy dynamics to detect emotional video segments [15]. They showed that using their proposed methodology, four types of emotion can be detected with relatively good accuracy. Sun et al. in [16] used a video affective tree and HMM to recognize the basic emotions of 'joy', 'anger', 'sadness', and 'fear'. A relationship between these four basic emotions and the features of 'color', 'motion', 'shot cut rate', and 'sound information' was found. Rasheed et al. in [17] classified movies into four broad categories: comedies, action, drama, and horror films by using low-level video features such as average shot length, color variance, motion content, and lighting key. They combined the features into a framework to provide a mapping of these four high-level semantic classes to emotional states. Their work can be considered as a step toward high-level semantic film interpretation.

There has been little prior work towards emotion recognition using both audio and visual cues in multimedia contents [16,18,19]. The authors in [16] performed continuous-scale emotion tracking in movies, fusing features from audio, music, and video modalities. An HMM classifier was employed to classify each video frame in the arousal-valence axis. In [18], audio and video features were combined at an early stage in order to form observation vector sequences, which were then presented to four different HMM classifiers, each representing one of the basic emotions of 'joy', 'anger', 'sadness', and 'fear'. In [19], audio and video features were concatenated into row vectors and classified using a specially adapted variant of SVM. Their method was tested with several well-known movies, which were categorized according to the basic emotions that they evoke to the viewers, reporting

promising results and demonstrating the efficacy of using audiovisual cues. The main contribution of the proposed method over existing methods in literature, utilizing features both from audio and video streams, is their combination in a more sophisticated way, permitting the optimal exploitation of these complementary sources of information. Moreover, the problem of recognizing evoked emotion when watching music video clips has never been addressed before.

2.2 Database description

The music video clips used in this study were taken from the Database for Emotion Analysis using Physiological Signals (DEAP; http://www.eecs.qmul.ac.uk/mmv/datasets/deap/) [5]. For selection of emotional music video clips, a web-based subjective test was conducted, where subjects were asked to watch 120 music video clips and rate their perceived emotion. More precisely, the subjects used a discrete nine-point scale to rate the following: valence, with the range going from unhappy or sad to happy or joyful; arousal, with the range going from calm or bored to stimulated or excited; dominance, with the range going from submissive or 'without control' to dominant or 'in control, empowered'; and whether they liked the video or not. Using these subjective data, 40 music video clips were selected so that only the music video clips which induce strong emotions are used. More precisely, ten music video clips from each quadrant or arousal-valence space, which all had the strongest possible volunteer ratings with a small variation, were selected. More information about the selection procedure can be found in [5].

After selecting the test material, 32 participants (50% female), aged between 19 and 37 (mean age 26.9), participated in the experiment to create the DEAP. While they were watching the 40 videos, their physiological signals were recorded. The study conformed with the Ethics of the World Medical Association (Declaration of Helsinki), and an informed consent was obtained from each participant prior to the start of the experiment. After watching each music video clip, they were asked to perform the rating of their perceived emotion. In this paper, we are only interested in the users' self-assessment ratings and the music video clips. In order to create the ground truth for our experiments, the ratings of the participants are thresholded into two classes, thus forming a binary classification problem (high/low arousal and positive/negative valence). For this purpose, on the nine-point rating scales, the threshold was placed in the middle. The mean and standard deviation of the high class per rating scale for arousal 59% (15%) and valence 57% (9%) indicate that the classes are well balanced. The final ground truth for each video was constructed by computing the mean opinion score of ratings performed by all participants. In order to verify the possible inter-rater agreement between the ratings of different participants, the pair-wise Cohen's kappa between self-assessments after quantizing the ratings into nine levels was used. The results of this test revealed a very weak agreement between the self-assessed emotions (mean kappa equals 0.02 ± 0.06 for arousal and 0.08 ± 0.08 for valence). The mean kappa values are very small, which demonstrates the inherent difficulty of the affect recognition problem that is dealt with in this article. More details about the measurements are available in the paper describing the creation of the DEAP [5].

2.3 Feature extraction

This section introduces the features that are extracted from music video clips and provides information on how these low-level features and the emotions relate to each other.

2.3.1 Video feature extraction

Numerous studies in literature report a concrete relationship between the low-level visual features of multimedia contents such as 'lighting key', 'shot length', 'color', 'motion', and emotional states induced to people while watching these contents. Thus, these four visual features were extracted from the music video clips and explained in detail in the following sections.

Lighting key The lighting of the scenes in videos has been greatly exploited as an important agent to evoke emotions. The balance, the direction, and the intensity of light are used to create effects, inducing certain emotions to the viewers and establishing the mood of the scene [17]. High-key lighting denotes an abundance of bright light and usually involves low contrast and a small difference between the brightest and dimmest light. In contrary, in low-key lighting, the scene is predominantly dark with a high contrast ratio. High-key lighting usually communicates activation and positive emotions, while low-key lighting is more dramatic and is often used to evoke negative feelings. Figure 3a,b illustrates examples of high-key and low-key lighting shots, with the respective distributions of their brightness. In order to compute the lighting key features, a 25-bin histogram is computed by analyzing the normalized to $[0, 1]$ value component of the hue, saturation, and value (HSV) color space. The mean and variance scores of the value component are low for low-key lighting shots and high for high-key lighting shots; therefore, the lighting quantity $\zeta_i(\mu, \sigma)$ for a frame i can be defined as

$$\zeta_i = \mu_i \times \sigma_i, \tag{1}$$

where μ_i and σ_i represent the mean and the standard deviation scores computed from the 'value component' of the HSV space of frame i, respectively.

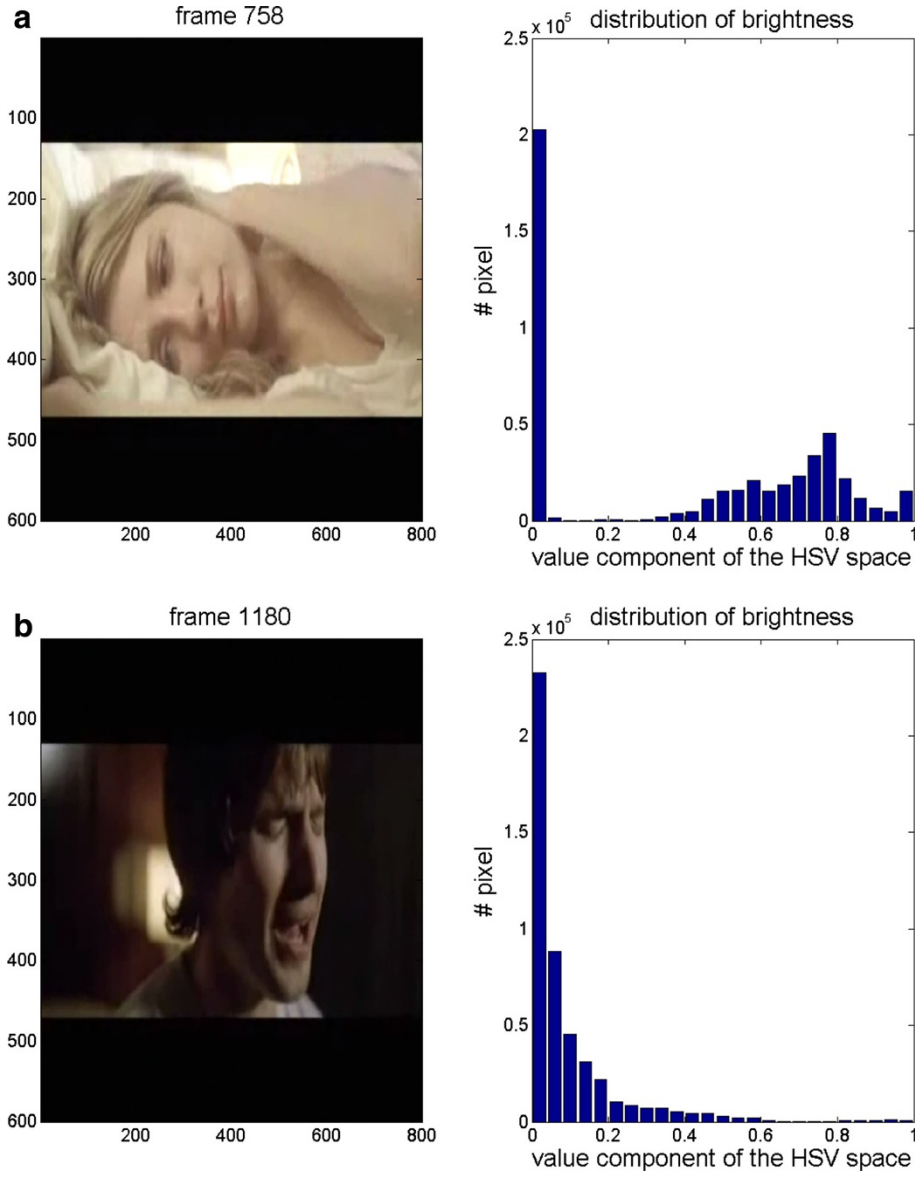

Figure 3 Lighting key. (**a**) High lighting key and (**b**) low lighting key.

Shot boundary detection A shot is defined as a series of interrelated consecutive frames taken by a single camera, without any major change in the color content, representing a continuous action in time and space. Shot boundary detection implies the automated detection of transitions between shots. The average shot length as a feature conveys the tempo of a scene and is inextricably linked with the rhythm and dynamism of the video. The direct relationship between the average shot length and the induced affective states is demonstrated in several experimental studies (e.g., [13,16]). In [20], a comparison of the most well-known techniques for automatic shot boundary detection was carried out. The authors in [21] proposed a method based on the differences between color histograms of frames belonging to a video sequence in order to detect hard cuts, fades, or dissolves. In the current work, the proposed method of [22] is used. In this approach multivariate, low-cost color features are extracted to construct a feature matrix which is then processed using sliding window singular value decomposition. The features extracted are the average, the longest, and the shortest shot lengths.

Color The selection of colors, their proportions, and contrasts play more than an aesthetic role in videos, and many researches in psychology demonstrate that color

constitutes a non-verbal communication code, which is very effective and precise. Each color conveys concrete messages, acting upon the spectators' emotions and provoking subconscious reactions. For example in [13], the warm colors 'orange' and 'red' were shown to correspond to aggressive and dynamic emotions such as 'fear' and 'anger', while the colors 'blue', 'violet', and 'green' communicate 'calmness' and 'relaxation', emotions of high valence combined with low arousal. In this work, the color features were computed from the histogram of the hue component in the HSV color space. The features extracted are the median of maximum, minimum, and mean values of the hue component for every frame.

Motion The motion in a video can be caused by the objects moving in a scene or the motion of the camera. Research studies show that motion has a significant impact on the viewers' affective responses. This intimate correlation seems to derive from the natural association of fast motion with activation and excitement, and limited motion with serenity and calmness. As a feature, motion proves to be especially helpful in differentiating between emotions in different halves of the arousal axis. The contribution of the motion component for several emotions such as 'joy', 'anger', 'sadness', and 'fear' is studied by Sun et al. in [16]. In the current work, the motion features extracted are the mean of median and mean absolute value of the motion vector, computed for every fourth frame of the video sequence.

2.3.2 Audio feature extraction

As reported in several research studies, sound can have a close relationship with the affective content of a music video clip [23]. In general, arousal is believed to be correlated with tempo (fast/slow) and pitch, while valence is associated with energy, harmony, and scale (major/minor). In order to capture some of these characteristics, the following features are computed for each music video clip using the Matlab music information retrieval MIR toolbox [24].

Zero-crossing rate The zero-crossing rate is defined as the number of times the signal crosses the zero line (x - axis) per unit time. In other words, it is the number of times the signal changes its sign per time unit.

Mel-frequency cepstral coefficients The mel-frequency cepstral coefficients (MFCC) can be seen as the description of the spectral shape of the sound. Most of the signal information can be found in low-frequency coefficients. Therefore, only the 13 first coefficients are used as features.

Delta MFCC coefficients (ΔMFCC) The ΔMFCC provide quantitative measures of the movement of the MFCC and can be derived as follows:

$$\Delta MFCC_i(\nu) = MFCC_{i+1}(\nu) - MFCC_i(\nu), \tag{2}$$

where $MFCC_i(\nu)$ represents the νth MFCC of frame i.

Tables 1 and 2 present an overview of the extracted video and audio features, respectively. Each video is first segmented into 20 sequences of 3 s in duration. The features, listed in Tables 1 and 2 are then extracted from each frame of a sequence. To construct the final feature vector of a sequence, the mean value of the features extracted from all frames of the sequence is computed.

2.4 Classification

After the feature extraction stage, the extracted features are fed into Gaussian mixture model (GMM) classifiers. For the training of the GMM classifiers, the expectation maximization algorithm is utilized, while eight Gaussian mixtures are created to describe each class. Experiments are conducted using the leave-one-video-out cross-validation scheme, which involves training the classifier with samples from all but one video and testing its performance on the left-out video (test set). In other words, for performing this cross validation, all sequences of the test video clip are kept out, and a classifier is trained based on the sequences of training video clips. This procedure is repeated until each video clip is considered as a test sample once. The test set is then compared to the GMMs trained for each class, resulting in a matching score. The benefits of the GMM classifier include superior classification performance and low computational complexity, as well as producing soft output results which can be later used for the information fusion stage.

Table 1 Summary of the extracted video features

Feature	Description
Lighting key (3 features)	Average value of lighting key of frames of a video segment;
	number of high- and low-key frames
Shot boundary (3 features)	Average shot length of a video;
	longest and shortest length of shots in a video
Color (3 features)	Median score of maximum, minimum, and median values of the hue
	component of frames of a video segment
Motion (2 features)	Mean of median and mean absolute values
	of the motion vector

Table 2 Summary of the extracted audio features

Description	Number of features
Zero crossing	1 feature
MFCC	13 features
ΔMFCC	13 features

2.5 Fusion of audio and visual features

Information fusion can be generally discussed in two main categories: feature fusion and decision fusion. In feature fusion (or early integration), feature vectors extracted from the respective signals are combined together before the classification stage, and the classifier learns the statistics of the joint observation. In decision fusion (or late integration), on the other hand, the feature vectors are classified separately, and the outputs of the classifiers are then combined into a single belief.

In general, fusion at the feature level is easy and straightforward to implement, with the simplest case being the concatenation of features across different modalities (i.e., audio, video) into a single vector. The greatest advantage of integration at feature level is the feasibility of capturing the low-level inter-modal dependencies and the temporal relationships of the signal, which can be partially obscured in late integration. Nevertheless, feature fusion is very difficult if the feature sets are incompatible or if a degree of asynchrony exists between them, as it imposes a rigid coupling between them [25]. Moreover, due to the fact that classification occurs at such a low level, if one or more modalities are corrupted, then the entire system is affected.

On the other hand, fusion at the decision level detaches the different sources and permits formulations that address issues such as feature asynchrony and channel weighting. Decision fusion also represents the optimal choice when having small datasets since it does not explicitly increase the dimensionality. Last but not the least, one of the most prominent characteristics is the ability to mitigate the effect of errors occurring in different modalities. This advantage holds as long as the outputs of the classifiers are uncorrelated and independent from all the other classifier outputs. A drawback of decision fusion methods is that the integration at this level may obscure intra-model dependencies and correlation between features.

There is still no consensus in the literature as to the level on which fusion should take place. On theoretical grounds and on the necessity of maintaining temporal relationships and low-level dependencies, many argue for feature fusion [26]. On the other hand, comparative empirical studies have demonstrated that decision fusion techniques are performing better than feature fusion methods, having more practical benefits [27]. As a common practice, the level at which the fusion occurs is chosen based on the characteristics of each specific application.

In this paper, a multilevel fusion scheme is proposed. In other words, given the inputs of the multiple classifiers, a fusion scheme that can best exploit the information of the individual modalities is searched for. Existing approaches address the fusion of the different modalities in either the feature level or decision level, with very few of them considering a hybrid approach that takes advantage of both methodologies [25]. The multilevel fusion approach can be regarded as an extension to decision fusion (Figure 4). In addition to the audio and video modalities, a joint audio and video modality derived from feature fusion is considered to form an additional modality *i*. Each of the dedicated classifiers delivers a quantity y_i that represents the likelihood of the observed data in each modality. The final decision is granted using the sum rule over the quantities y_i.

The sum rule constitutes a judicious choice as a data-independent, static combination function that does not require any training. Moreover, it is shown theoretically and empirically to be a more beneficial rule compared to other combination rules (such as the product rule) when there is minimal knowledge about testing conditions and when confidence errors are present [28]. The sum rule is based on the following assumptions: (a) statistical independence over the multiple modalities and (b) posterior probabilities do not deviate much from the prior probabilities. An input pattern is assigned to class *c* such that

$$c = \arg\max_j \sum_{i=1}^{R} P(w_j \mid \overrightarrow{x_i}), \tag{3}$$

where $\overrightarrow{x_i}$ denotes the feature vector derived from an input pattern presented to the i_{th} classifier, *R* denotes the number of the different modalities, $P(w_j \mid \overrightarrow{x_i})$ represents the pseudo-probability of the classifier considering a pattern belonging to class w_j given the feature vector $\overrightarrow{x_i}$, and $c \in \{1, 2, \ldots, m\}$ is the class to which the input pattern is finally assigned to, with *m* being the total number of classes.

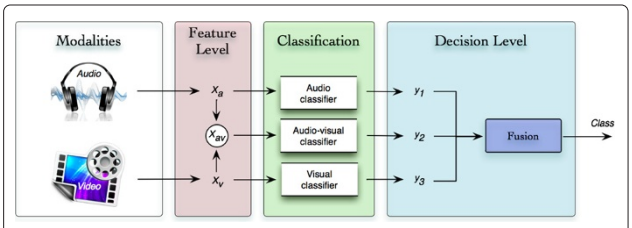

Figure 4 Multilevel fusion scheme. This scheme allows the integration of the audio and visual modalities in both the feature level and the decision level, taking advantage of both.

3 Results

This section presents the evaluation results of the different fusion schemes conducted at different levels of integration as well as the multilevel fusion scheme. The final ground truth for each video was constructed by computing the mean opinion score of ratings performed by all participants. For feature fusion, the direct concatenation method is utilized, and for the decision fusion, the sum rule is used. The results are also compared against the unimodal audio-only and video-only systems. The unimodal systems are denoted as 'A' and 'V' respectively. The fusion at feature and decision levels are represented by 'AV' and 'A+V' , respectively, and the multilevel fusion scheme is indicated as 'A+V+AV'. The results for arousal, valence, and dominance are depicted in Figure 5.

For implementation of the sum rule, different modalities are equally weighted under all conditions. Empirically, this 'democratic' approach usually leads in sub-optimal results. This can be improved by adjusting the contribution of each individual channel to the overall decision, using appropriate weights. In the scope of the particular application, weights based on measures like environmental conditions or front-end confidence cannot be well defined, and weights based on dispersion measures [29] did not yield any improved results in this case. Therefore, optimal stationary weights are assigned to each channel, determined by grid search. In order to compute these optimal stationary weights, the leave-one-video-out cross-validation scheme is used. In other words for computing the optimal wights for any given music video clip, all feature vectors of the corresponding music video clip are left out, and the weights are trained with the features of the remaining 39 video clips.

Figure 5 provides supporting evidence that the integration of information from the audio and visual modalities leads to an improved performance compared to unimodal approaches for all the fusion schemes examined. The best fusion scheme reported in our results outperforms the best single modality by 7.5% up to 12.5%. Furthermore, the classic feature fusion generally provides better results (or the same in the case of optimal weighting) than the decision fusion. This can be attributed to the fact that early integration allows the modeling of the complete audio-visual observation. Therefore, detailed correlation between audio and video features can be accurately modeled, without having to deal with corruptions of either the audio or visual modalities, which would deteriorate its performance. Moreover, issues such as the audio and video asynchrony or 'the curse of dimensionality' are not present in this case, justifying the superior performance of the early integration. Optimal weighting, obtained through grid search, improves the performance of the system in almost all cases (except from the A+V+AV case in dominance), as it reflects the contribution of each of the different modalities in terms of measurement uncertainty and informativeness. Our results indicate that audio features carry more information on arousal, whereas for valence and dominance, visual information is of higher significance (cf. Figure 5). Most importantly, we observe that the multilevel approach achieves significant improvement in performance over the rest of the approaches examined. The results strongly imply that the merged audio-visual modalities at feature level provide important cues which are not captured in the conventional decision fusion approach. Additionally, fusion at the decision level allows to perform weighting of the channels which reflect the contribution of each, in terms of informativeness and output certainty. This constitutes the reasoning why the multilevel fusion scheme consistently outperforms all the examined approaches.

4 Conclusions and future work

In this paper, multimedia content analysis of different music video clips was performed to extract affective information about them. Several features were extracted from audio and video channels of the music video clips, and the classification performance using each of these modalities was assessed separately. Furthermore, several fusion

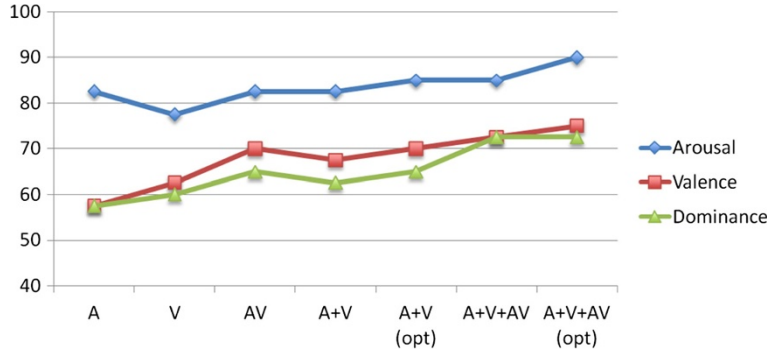

Figure 5 Comparison of performance (vertical axis) of different fusion schemes (horizontal axis) for arousal, valence, and dominance.

techniques were employed to combine this information. The results obtained throughout this paper have revealed that recognition and extraction of emotional information are very challenging tasks in machine vision, as emotions are highly person- and context-dependent. Although our system lacks semantic and contextual information, as it utilizes only low-level audio and video features, it is still able to capture with relatively high precision the emotional states evoked on the spectators. Moreover, because of the low-level feature extraction, the system is expected to work equally well when other than music videos are presented as input, and it can also be easily extended to cover other application domains. It is commonly accepted that the integration of information emanating from various sources leads to improved performance over single modal approaches. However, the question of how to combine the information from the different sources and on which level still remains an open issue. Although some theoretical framework has been developed, most of the work done in this area is quite heuristic and application-dependent. In this work, we demonstrate that a hybrid approach which combines the advantages of both levels of integration shows superior performance. The proposed methodology can be used for determining the arousal, valence, and dominance values of any given music video clip. These values can be used as emotional tags, which can be used for search and retrieval applications. Future efforts are directed towards exploring features, which can potentially capture semantics, and also using dedicated classifiers for each modality which will better reflect their specificities. Accommodating new modalities may also constitute a straightforward way for further improving the performance of the system.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The research leading to these results has been performed in the frameworks of Swiss National Foundation for Scientific Research (FN 200020-132673), the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET, and the NCCR Interactive Multimodal Information Management (IM2).

Author details
[1]Multimedia Signal Processing Group (MMSPG), Institute of Electrical Engineering (IEL), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, 1015, Switzerland. [2]Artificial Intelligence Group, Wire Communications Laboratory, Department of Electrical and Computer Engineering, University of Patras, Patras, 265 04, Greece.

Received: 28 August 2012 Accepted: 19 March 2013
Published: 30 April 2013

References

1. O Nov, M Naaman, C Ye, in *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*. What drives content tagging: the case of photos on Flickr (ACM New York, 2008), pp. 1097–1100
2. M Pantic, A Vinciarelli, Implicit human-centered tagging [Social Sciences]. Signal Process. Mag. IEEE. **26**(6), 173–180 (2009)
3. A Hanjalic, L Xu, Affective video content representation and modeling. Multimedia, IEEE Trans. **7**, 143–154 (2005)
4. Z Zeng, M Pantic, G Roisman, T Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions. Pattern Anal. Mach. Intell., IEEE Trans. **31**, 39–58 (2009)
5. S Koelstra, C Muehl, M Soleymani, J Lee, A Yazdani, T Ebrahimi, T Pun, A, Nijholt, I Patras, DEAP: a Database for Emotion Analysis using Physiological Signals. IEEE Trans. Affect. Comput. **3**, 18–31 (2011)
6. Y Wang, Z Liu, J Huang, Multimedia content analysis-using both audio and visual clues. Signal Process. Mag., IEEE. **17**(6), 12–36 (2000)
7. P Ekman, in *Handbook of Cognition and Emotion*. Basic emotions (Wiley Hoboken, 1999), pp. 45–60
8. J Kim, E André, Emotion recognition based on physiological changes in listening music. IEEE Transactions in Pattern Analysis and Machine Intelligence. **30**, 2067–2083 (2008)
9. P Lang, M Bradley, B Cuthbert, International Affective Picture System (IAPS): instruction manual and affective ratings. The Center for Research in Psychophysiology, University of Florida (1999)
10. M Bradley, P Lang, in *The International Affective Digitized Sounds (IADS): Stimuli, Instruction Manual and Affective Ratings*. Center for the Study of Emotion and Attention and National Institute of Mental Health (NIMH Center for the Study of Emotion and Attention Florida, 1999)
11. A Yazdani, K Kappeler, T Ebrahimi, in *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*. Affective content analysis of music video clips (ACM New York, 2011), pp. 7–12
12. A Hanjalic, Extracting moods from pictures and sounds: towards truly personalized TV. Signal Process. Mag., IEEE. **23**(2), 90–100 (2006)
13. H Kang, in *Proceedings of the Eleventh ACM International Conference on Multimedia*. Affective content detection using HMMs (ACM, 2003), pp. 259–262. New York
14. M Xu, L Chia, J Jin, in *2005 IEEE International Conference on Multimedia and Expo*. Affective content analysis in comedy and horror videos by audio emotional event detection (IEEE Washinton, DC, 2005), pp. 4–10
15. S Moncrieff, C Dorai, S Venkatesh, in *Proceedings of the Ninth ACM International Conference on Multimedia*. Affect computing in film through sound energy dynamics (ACM New York, 2001), pp. 525–527
16. K Sun, J Yu, Video affective content representation and recognition using video affective tree and hidden Markov models. Affect. Comput. Intell. Interact. **4738**, 594–605 (2007)
17. Z Rasheed, Y Sheikh, M Shah, On the use of computable features for film classification. Circuits Syst. Video Technol., IEEE Trans. **15**, 52–64 (2005)
18. N Malandrakis, A Potamianos, G Evangelopoulos, A Zlatintsi, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A supervised approach to movie emotion tracking (IEEE Washington, DC, 2011), pp. 2376–2379
19. HL Wang, LF Cheong, Affective understanding in film. Circuits Syst. Video Technol., IEEE Trans. **16**(6), 689–704 (2006)
20. R Lienhart, in *Proceedings of SPIE, Vol. 3656*. Comparison of automatic shot boundary detection algorithms (SPIE Bellingham, 1999), pp. 290–301
21. J Mas, G Fernandez, in *Notebook Pap TRECVID2003*. Video shot boundary detection based on color histogram (NIST Gaithersburg, 2003)
22. W Abd-Almageed, in *15th IEEE International Conference on Image Processing, 2008 ICIP 2008*. Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing (IEEE Washington, 2008), pp. 3200–3203
23. Y Yang, H Chen, Ranking-based emotion recognition for music organization and retrieval. Audio, Speech, Lang. Process., IEEE Trans. **19**(4), 762–774 (2011)
24. O Lartillot, P Toiviainen, T Eerola, A matlab toolbox for music information retrieval. Data Anal., Mach. Learn. Appl, 261–268 (2008)
25. S Chu, V Libal, E Marcheret, C Neti, G Potamianos, in *IEEE International Conference on Multimedia and Expo, 2004 ICME'04 Volume 3*. Multistage information fusion for audio-visual speech recognition 2004 (IEEE Washington, 2004), pp. 1651–1654
26. M Hennecke, D Stork, K Venkatesh Prasad, Visionary speech: Looking ahead to practical speechreading systems. NATO ASI Ser. F: Comput. Syst. Sci. **150**, 331–350 (1996)
27. S Lucey, T Chen, S Sridharan, V Chandran, Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. Multimedia, IEEE Trans. **7**(3), 495–506 (2005)

28. J Kittler, M Hatef, R Duin, J Matas, On combining classifiers. Pattern Anal. Mach. Intell., IEEE Trans. **20**(3), 226–239 (1998)
29. T Lewis, D Powers, in *Proceedings of the 27th Australasian Conference on Computer Science-Volume 26*. Sensor fusion weighting measures in audio-visual speech recognition (Australian Computer Society Inc. Sydney, 2004), pp. 305–314

doi:10.1186/1687-5281-2013-26
Cite this article as: Yazdani *et al.*: Multimedia content analysis for emotional characterization of music video clips. *EURASIP Journal on Image and Video Processing* 2013 **2013**:26.