HAL
open science

LIRIS-ACCEDE: A Video Database for Affective Content Analysis

Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, Liming Chen

To cite this version:

Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, Liming Chen. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. IEEE Transactions on Affective Computing, 2015, 6 (1), pp.43-55. 10.1109/TAFFC.2015.2396531 . hal-01375518

HAL Id: hal-01375518

https://hal.science/hal-01375518v1

Submitted on 29 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LIRIS-ACCEDE: A Video Database for Affective Content Analysis

Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen

Abstract—Research in affective computing requires ground truth data for training and benchmarking computational models for machine-based emotion understanding. In this paper, we propose a large video database, namely LIRIS-ACCEDE, for affective content analysis and related applications, including video indexing, summarization or browsing. In contrast to existing datasets with very few video resources and limited accessibility due to copyright constraints, LIRIS-ACCEDE consists of 9,800 good quality video excerpts with a large content diversity. All excerpts are shared under Creative Commons licenses and can thus be freely distributed without copyright issues. Affective annotations were achieved using crowdsourcing through a pair-wise video comparison protocol, thereby ensuring that annotations are fully consistent, as testified by a high inter-annotator agreement, despite the large diversity of raters' cultural backgrounds. In addition, to enable fair comparison and landmark progresses of future affective computational models, we further provide four experimental protocols and a baseline for prediction of emotions using a large set of both visual and audio features. The dataset (the video clips, annotations, features and protocols) is publicly available at: http://liris-accede.ec-lyon.fr/.

Index Terms—Video database, Induced emotion, Computational emotion modeling, Emotion classification, Affective computing

1 Introduction

Affective video content analysis aims at automatic recognition of emotions elicited by videos. It has a large number of applications, including mood-based personalized content delivery [1], video indexing, video indexing, and summarization (*e.g.* [2], [3]). While major progress has been achieved in computer vision for visual object detection, scene understanding and high level concept recognition, a natural further step is modeling and recognition of affective concepts. This has received increasing interest from research communities, *e.g.*, computer vision, machine learning, with an overall goal of endowing computers with human-like perception capabilities. However, while human affective perception is highly subjective, machine-based affective modeling and recognition require large amounts of reliable ground truth data for training and testing. Unfortunately, the subjective nature of "emotions" makes it hard to collect consistent and large volumes of affective annotations suitable for the use as ground truth, while the copyright issues concerning video clips prevent free distribution of existing annotated datasets. Most state of the art work uses a private dataset of a very limited size and content diversity, thus making fair comparisons and results reproducibility impossible, and preventing achievement of major strides in the field.

Horvat *et al.* showed in a survey [4] that, for researchers in the affective science field, current emotionally annotated databases lack at least some stimuli inducing a particular emotion. Participants additionally indicated that they would greatly benefit from large emotionally annotated databases composed of video clips. Soleymani *et al.* also expressed this major need and defined in [5] the specifications to be considered to allow standardized evaluation and to bypass the size and scope of related limitations of existing databases used to train and evaluate computational models in the field of affective content analysis. How can a large and reliable dataset be built that could serve the community as a reliable benchmark? Crowdsourcing is often the recommended solution for creating a large dataset representing a condition. This makes it possible to reach a large number of remunerated annotators, while also guaranteeing reliability of annotators' answers via specific mechanisms.

In this paper, to overcome the limitations of the existing affective video datasets and foster research in affective video content analysis, we release a large dataset of quality video excerpts with high content diversities, along with ground truth affective annotations collected from a wide variety of raters through crowdsourcing. The proposed dataset, namely LIRIS-ACCEDE, contains 9,800 video excerpts shared under Creative Commons licenses, making it possible to release the database without copyright issues. The dataset was first introduced in [6], where we described the experimental protocol for ranking video clips along the induced valence axis. In this paper, we highlight the content diversity of the LIRIS-ACCEDE

- *Y. Baveye and C. Chamaret are with Technicolor, 975, avenue des Champs Blancs, 35576 Cesson Sévigné, France. Y. Baveye is also with the Université de Lyon, Centre National de la Recherche Scientifique, Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France. E-mail: {yoann.baveye, christel.chamaret}@technicolor.com*
- *E. Dellandréa and L. Chen are with the Université de Lyon, Centre National de la Recherche Scientifique, Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France. E-mail: {emmanuel.dellandrea, liming.chen}@ec-lyon.fr*

database and detail its composition: the movies from which the excerpts have been extracted are described and the diversity of the database is demonstrated. We also present a novel experimental protocol for ranking video excerpts of the database along the induced arousal axis. Furthermore, all excerpts in LIRIS-ACCEDE are ranked using crowdsourcing in the widely used 2D valence-arousal space. This dimensional space has been preferred to other categorical approaches that classify emotions into a small number of discrete clusters and may not reflect the complexity, diversity and richness of the emotions which could be induced by such a vast number of videos [7]. To ensure reliability and consistency of raters' affective annotations despite the subjective nature of emotions and the large diversity of their backgrounds, we design a pairwise video excerpt comparison protocol leading to a high inter-rater agreement as revealed by objective measurements. Moreover, to enable fair comparison between future work using the LIRIS-ACCEDE dataset, we also propose four experimental protocols and introduce a baseline using a large set of visual and audio features. The database, containing the 9,800 video clips, the annotations, features and experimental protocols, can be freely downloaded at http://liris-accede.ec-lyon.fr/. We believe that public release of such a database will foster research in the field and benefit various affective computing research communities. The main contributions of this paper are thus three-fold:

- Public release of a large freely accessible database of video excerpts under Creative Commons licenses with a large content diversity;
- Consistent affective annotations collected using a pairwise video excerpt comparison protocol, from a large number of raters with a great variety of backgrounds through crowdsourcing;
- Design of 4 experimental protocols and introduction of a baseline using a large number of visual and audio features.

The paper is organized as follows. Section 2 provides background material on existing affective multimedia databases and computational models of induced emotion. Next, in Section 3, our main LIRIS-ACCEDE contribution is presented, and the process for annotating the database is described in Section 4. In an attempt to enable standardized evaluation of affective computational models, a baseline framework and several protocols are introduced in Section 5. Limitations are discussed in Section 6, while the paper ends in Section 7 with conclusions.

2 Background

2.1 Affective multimedia databases

Creation of an affective database is a necessary step in affective computing studies. While there are many databases composed of facial expression videos for emotion recognition, there are not many databases of video clips annotated according to the emotions they induce in viewers (Table 1). Philippot [8], as well as Gross and Levenson [9], were the first to propose small sets of film excerpts assumed to elicit specific emotions in the laboratory. To achieve this goal, they selected specific excerpts most likely to elicit strong emotions, which thus do not represent the full range of emotions potentially elicited by movies. Even if increased efforts have recently been made to standardize film clip databases, there are no multimedia databases annotated along induced emotional axes dealing with the full spectrum of emotions in movies that are large enough to be used in machine learning and that do not suffer from copyright infringement.

The HUMAINE database [10] created by Douglas-Cowie *et al.* consists of a subset of three naturalistic and six induced reaction databases. The purpose of the database is to illustrate key principles of affective computing instead of applying it to machine learning. It is made up of 50 clips: naturalistic and induced data ranging from 5 seconds to 3 minutes. These have been annotated according to a wide range of labels detailed in Table 1.

Introduced by Shaefer *et al.* in [11], the FilmStim database consists of 70 film excerpts intended to elicit emotional states in experimental psychology experiments. 10 films are selected per emotional category (*i.e.* anger, sadness, fear, disgust, amusement, tenderness and neutral state) and cut into clips ranging from 1 to 7 minutes. 364 participants rated each film clip, and ranking scores were computed for 24 classification criteria displayed in Table 1. Even if it is one of the biggest databases of videos annotated along induced emotional labels, videos are labeled globally. Yet, emotions are a relatively fast phenomenon lasting a few seconds from onset to end [12]. This is why a unique global label is not sufficient to build ground truth data for induced emotion models.

The DEAP database is another publicly available database that has been created recently by Koelstra *et al.* [13]. It is composed of 120 one-minute long excerpts of music videos. Each one was rated by at least 14 volunteers from an online self-assessment based on induced arousal, valence and dominance. Physiological signals were recorded from participants, while they rated a subset of 40 of the above music videos in terms of arousal, valence, like/dislike, dominance and familiarity levels. Music videos protected by copyright are not available alongside the annotations. Instead, the YouTube links are given, but some of them are no longer available on YouTube, sometimes due to copyright claims. This shows the need for a database that does not depend on third parties to share its material legally.

The same year, Soleymani *et al.* released MAHNOB-

TABLE 1
Downloadable video databases annotated using labels considering induced emotion

Name	Size	Emotional labels
HUMAINE	50 clips from 5 seconds to 3 minutes long	Wide range of labels at a global level (emotion-related states, context labels, key events, emotion words, *etc.*) and frame-by-frame level (intensity, arousal, valence, dominance, predictability, *etc.*)
FilmStim	70 film excerpts from 1 to 7 minutes long	24 classification criteria: subjective arousal, positive and negative affect, a positive and negative affect scores derived from the Differential Emotions Scale, six emotion discreteness scores and 15 mixed feelings scores
DEAP	120 one-minute music videos	Ratings from an online self-assessment on arousal, valence and dominance and physiological recordings with face video for a subset of 40 music videos
MAHNOB-HCI	20 film excerpts from 35 to 117 seconds long	Emotional keyword, arousal, valence, dominance and predictability combined with facial videos, EEG, audio, gaze and peripheral physiological recordings
EMDB	52 non-auditory film clips of 40 seconds long	Global ratings for the induced arousal, valence, dominance dimensions
VIOLENT SCENES DATASET	25 full-length movies	Annotations include the list of the movie segments containing physical violence according to two different definitions and also include 10 high-level concepts for the visual and audio modalities (presence of blood, fights, gunshots, screams, *etc.*)
LIRIS-ACCEDE	9,800 excerpts from 8 to 12 seconds long	Rankings for arousal and valence dimensions

HCI [14] which is a multimodal database composed of 20 short emotional excerpts extracted from commercially produced movies and video websites. Participants watching these fragments were asked to annotate their own emotive state on a scale in terms of arousal and valence. Facial videos, EEG, audio, gaze and peripheral physiological recordings were also recorded for all 30 participants.

Carvalho *et al.* built in [15] the emotional movie database (EMDB) made up of 52 non-auditory film clips. Film clips are extracted from commercial films and last 40 seconds. They have been selected to cover the entire affective space. 113 participants rated each film clip in terms of induced valence, arousal and dominance on a 9-point scale. Non-auditory clips were used to enhance the scope for future experimental manipulations. However, this clearly modifies how viewers perceive the video clips. Furthermore, multimodal processing is not possible in this case.

Still more recently, the Violent Scene Dataset was made available by Demarty *et al.* [16]. This is a collection of ground truth annotations based on extraction of violent events in movies, together with high level audio and video concepts. This dataset has been used since 2011 in the MediaEval multimedia benchmarking affect task "Violent Scenes Detection". Violent scene detection and prediction of induced emotions are clearly related since they are both part of the affective content analysis field. Violent scenes are most likely to be highly arousing and elicit negative emotions. Due to copyright issues, the 25 annotated movies cannot be delivered alongside the annotations. However, the links to the DVDs used for the annotation on the Amazon web site are provided.

Last but not least, it is worth mentioning the MIT dataset dedicated to animated GIFs [17]. Such kinds of short video footage are becoming increasingly popular by means of social networks. They are so widely adopted that the MIT team is currently and seriously working on predicting perceived emotions from such media support.

All these databases either have different emotional labels or are not representative of the whole range of emotions in movies. Thus, a huge database of videos annotated using induced emotional labels potentially suitable for research, is a requirement of the affective computing community.

2.2 Computational models of emotion

Work on affective video analysis can be categorized into two subgroups: continuous affective video content analysis, which estimates an affective score of each frame of a video, and discrete affective video content analysis, which assigns an affective score to a segment of video.

Hanjalic and Xu pioneered in [18] the analysis of affective video content by directly mapping video features onto the valence-arousal space to create continuous representations. They only offered a qualitative evaluation of their model. Malandrakis *et al.* also proposed in [19] a continuous affective video content analysis relying on audiovisual features extracted on each video frame, combined at an early stage and presented to two Hidden Markov Models (HMMs). These two classifiers are trained independently to model simultaneously the arousal and valence. They

output time series of seven categories interpolated into a continuous-valued curve. Their discrete and continuous curves are compared using the leave-one-movie-out approach to the ground truth collected on 30-min video clips from 12 movies.

Discrete affective video content analysis has been more frequently investigated over the last decade. Kang [20] was the first to propose a model where classifiers are adopted for affective analysis. He suggested detecting affective states in movies including "sadness", "joy" and "fear" from low-level features using HMMs. In the same way, Wang and Cheong introduced features inspired from psychology and film-making rules [21]. One SVM is especially dedicated to audio cues to obtain high-level audio information at scene level. Each video segment is then classified with a second SVM to obtain probabilistic membership vectors for 7 discrete emotional states. Their training data are made up of 36 full-length popular Hollywood movies divided into 2040 scenes labeled with one or two emotional states. In the work of Sun and Yu [22], movie units are first represented in different granularities using an excitement curve based on the arousal curve introduced in [18]. Then, four HMMs are trained independently using features extracted on these granularities to recognize one of the four emotional states among "joy", "anger", "sadness" and "fear". Their ground truth consists of 10 movies labeled at different levels. Xu *et al.* added in [23] a neutral state to these four emotional states. Fuzzy clustering is used to compute the emotion intensity level, then five HMMs are trained using emotion intensity and low-level features to model valence and individually associate the content with five emotion types. They evaluated the efficiency of their method for several movie genres, where the highest accuracy was obtained for action movies. Soleymani *et al.* [24] compared in the valence-arousal space the values obtained automatically from either physiological responses or from audio-visual features. They showed significant correlations between multimedia features, physiological features and spectators' self-assessments. A dataset composed of 64 movie scenes extracted from 8 Hollywood movies was created. The following year, they introduced a Bayesian framework for video affective representation [25] using audiovisual features and textual features extracted from subtitles. The arousal information of each shot is obtained by computing linear weights using a relevance vector machine. Arousal is then used as an arousal indicator feature for scene affective classification. Irie *et al.* [26] proposed an approach based on latent Dirichlet allocation considering the temporal transition characteristics of emotions. The good results obtained may be due to their evaluation protocol. Their data, composed of 206 scenes from 24 movie titles available as DVDs, were randomly selected to form the training and test sets. As a consequence, most films appear both in the training and the test sets, which biases the results. Zhang *et al.* developed in [27] a personalized affective analysis for music videos composed of SVR-based arousal and valence models using both multimedia features and user profiles. Their dataset of 552 music videos is used to train and update the models based on user feedback.

2.3 Issues with the existing databases

Due to the constraints on databases presented in Section 2.1, almost all investigations dealing with affective computational models use their own database designed according to their goals and needs (except the DEAP database which has recently been used in [28]). For example, some work represents emotions in the 2D valence-arousal space or in the 3D valence-arousal-dominance space, while other work represents emotions using discrete categories. Furthermore, the models are sometimes dedicated to specific video categories, *i.e.* music videos or a particular movie genre. LIRIS-ACCEDE uses the widely employed 2D valence-arousal space. However, as the database is freely shared, everyone is free to add new modalities, thus enhancing the range of possible applications. Furthermore, this database composed of 9,800 excerpts is very large and diversified, unlike most of the databases presented in the previous sections. As a consequence, we think it could be general enough to be used as a reference in the future. Many studies such as [4] or [29] deplore the lack of a standard affective video database which, combined with the lack of standard evaluation protocols, decreases the efficiency of the affective research community [5]. Indeed, benchmarking and reproducibility both make it easier to know how computational models perform with respect to the state of the art, and to focus on promising research avenues. This is why we introduce LIRIS-ACCEDE and define reproducible protocols in the following sections.

3 Database description

LIRIS-ACCEDE is made up of 9,800 excerpts extracted from 160 feature films and short films. It is the largest video database currently in existence annotated by a broad and representative population using induced emotional labels.

3.1 Movies used in LIRIS-ACCEDE

One of the main requirements of LIRIS-ACCEDE was that it should be freely available to the research community. That is why the 160 movies used for creating the database are shared under Creative Commons licenses. Creative Commons is a non-profit corporation providing standardized free copyright licenses to mark a creative work with the freedom the creator wants it to convey. The CC BY license known as "Attribution" is the most accommodating license since users can reuse

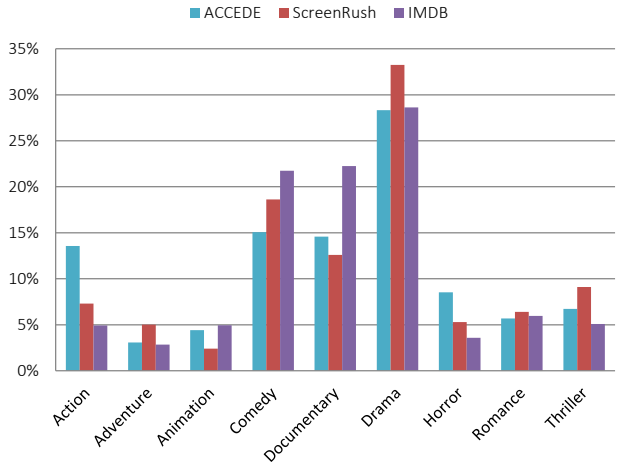

Fig. 1. Normalized distribution of films by genre included in LIRIS-ACCEDE and referenced on ScreenRush and IMDB.

the original creation as long as they credit the creator. Three modules adding more restrictive conditions can be combined. The SA module (ShareAlike) requires that works based on other works shared using this module, have to be licensed under identical terms. The NC module (NonCommercial) prevents original works from being reused for commercial purposes. Last but not least, the ND module (No Derivative Works) prohibits altering, transforming, or building upon original works. To create the database, we have used only movies shared under a Creative Commons license that do not contain the ND module, because our goal was to modify the selected movies by extracting several excerpts from them. Thus, using videos shared under Creative Commons licenses makes it possible to share the database publicly without copyright issues.

Most of the 160 movies used for creating LIRIS-ACCEDE come from the video platform VODO. This references best free-to-share feature films and short films that have been submitted on the website and makes them easily available to millions of people. It is important to notice that free-to-share films do not mean User Generated Contents with low expertise levels. Movies referenced on VODO have been created by filmmakers with excellent technical expertise. Many films in the database have been screened during film festivals including, but not limited to, "RIP! A remix manifesto" directed by Brett Gaylor (Special Jury Prize at the "Festival du Nouveau Cinéma in Montreal"), "Emperor" directed by Juliane Blockand (winner of the Feature Category at the Portable Film Festival) and "Pioneer One" produced by Josh Bernhard and Bracey Smith (winner of the Best Drama Pilot at the New York Television Festival). The "Home" documentary directed by Yann Arthus-Bertrand included in the database is a special case since it is a big budget movie distributed by 20th Century Fox that has no copyright.

In brief, 40 high quality feature films and 120 short films shared under Creative Commons licenses have been collected to create the 9,800 excerpts making up LIRIS-ACCEDE. The total time of all 160 films is 73 hours, 41 minutes and 7 seconds. A list of 9 representative movie genres describes the movies: Comedy, Animation, Action, Adventure, Thriller, Documentary, Romance, Drama and Horror. By displaying the normalized distribution of movies by genre in LIRIS-ACCEDE compared to the normalized distribution of movies by genre referenced on IMDB and on ScreenRush, it can be observed that distributions appear to be similar, as shown in Figure 1. Thus, movies used in LIRIS-ACCEDE are representative of today's movies. Languages are mainly English with a small set of French, German, Icelandic, Hindi, Italian, Norwegian, Spanish, Swedish and Turkish films, subtitled in English. Note that 14 movies are silent movies.

3.2 Characteristics of LIRIS-ACCEDE

The database is made up of 9,800 excerpts extracted from the 160 selected movies presented in Section 3.1.

1,000 excerpts have been manually segmented because they were part of the pilot test to ensure the reliability of the annotations. Subsequently, the other excerpts have been automatically segmented using a robust cut and fade in/out detection, implemented based on the algorithms described in [30]. Because all the segmented excerpts start or end with a cut or a fade, it is very likely that each segment be perceived by users as semantically coherent.

The 9,800 segmented video clips last between 8 and 12 seconds, and the total time of all 9,800 excerpts is 26 hours, 57 minutes and 8 seconds. Even if the temporal resolution, or granularity, of emotions is still under debate, most of psychologists agree that they are part of a complex but very rapid process [31]. They are phenomena with onsets and ends over seconds [12]. Indeed, the length of extracted segments in LIRIS-ACCEDE is large enough to obtain consistent excerpts, making it possible for the viewer to feel emotions. For example Gross and Levenson successfully elicited emotions in the laboratory using short excerpts lasting a few seconds [9]. Moreover, Metallinou and Narayanan have shown in [32] that global ratings of perceived emotion for movies lasting a few minutes are not simple averages over time, but rather are more influenced by highly arousing events with low valence. By using short excerpts, we greatly minimize the probability that annotations are a weighted average of consecutive emotions felt during successive events.

Despite the short duration of excerpts, most are composed of several video-editing features. This is essential since many previous studies, including [18] and [33], have shown that the arousal dimension was

correlated to editing features such as the shot cut rate or the presence of dissolves. Only 1760 excerpts do not include any scene cut or fade in/out. On average excerpts are composed of 2.8 video-editing features (this statistic does not count the editing features on the boundaries).

More generally, we achieved a great variety of excerpts reflecting the variety of selected movies. The excerpts contain scenes of violence, sexuality, murders, but also more common scenes such as landscapes, interviews and many positive scenes of daily life. LIRIS-ACCEDE is currently the only video database annotated along induced emotions that includes such a large range of contexts.

4 Data annotation

4.1 Experimental design

The annotation process aims at sorting the 9,800 excerpts independently along the induced valence and arousal axes. Crowdsourcing is an appropriate choice for achieving this goal requiring a huge amount of annotations, and has proved to be useful in various annotation studies (*e.g.* [34], [35], [36]). To annotate LIRIS-ACCEDE data, video excerpts were presented to annotators, also known as workers, on CrowdFlower.[1]

Rating-by-comparison experiments, *i.e.* ranking approaches, are more suited than rating approaches in experiments conducted on crowdsourcing platforms. Plausibly, asking for pairwise comparisons seems less complex than asking for an absolute value. Indeed, ratings require that annotators understand the range of an emotional scale, which is a sizable cognitive load [37], and it is quite difficult to ensure that the scale is used consistently. Russel and Gray [38] showed that raters using rating scales tend to only use a small subset of the range, while Ovadia [39] pointed out that inter-annotators ratings, *i.e.* ratings from different annotators, and even intra-annotator ratings, *i.e.* ratings from the same annotator, may not be consistent. By choosing pairwise comparisons instead of ratings, the consistency of the annotations is improved, as annotators tend to agree more when describing emotions in relative terms than in absolute terms [32]. Pairwise comparisons are also more appropriate detectors of user states, discarding the subjectivity of rating scales and implicit effects linked to the order of annotations [40]. Yang and Chen also showed in [37] that pairwise comparisons enhance the reliability of the ground truth compared to rating approaches, and simplify emotion annotation. This simplification also makes tasks more attractive and interesting to annotators. From an involved annotator's point of view, because the amount of money they earn is proportional to the quality of their answers and the amount of time they spend on the task, the simpler a task is, the more they are disposed to annotate other comparisons.

Accordingly, the choice of a rating-by-comparison experiment to annotate LIRIS-ACCEDE stands out. For each pair of video excerpts presented to workers on CrowdFlower, annotators had to select the one which conveyed most strongly the given emotion in terms of valence or arousal. The advantage of forced choice pairwise comparisons is that annotators must come to a decision. Forced choice pairwise comparisons enhance the reliability of experiments compared to other protocols such as displaying a single stimulus and a categorical rating scale [41] and encourage more thorough processing of response options [42].

If all possible comparisons had been generated and annotated by three crowdworkers, the experiments would have cost US$2,880,906 each. Thus, it was essential to choose an algorithm to select carefully and efficiently the comparisons judged by the annotators. The quicksort algorithm was used to generate the comparisons and rank the video excerpts according to the annotations gathered from CrowdFlower. This is one of the most efficient sorting algorithms. Indeed, the average computational complexity of this algorithm is $O(n \log n)$, where n is the number of data to sort. In the worst case, complexity is $O(n^2)$, but this performance is extremely rare and in practice the quicksort is often faster than other $O(n \log n)$ algorithms [43]. As the cost of the sorting operation is proportional to the number of comparisons, the quicksort seems the best choice for reducing costs to sort the whole database compared with other sorting algorithms. In practice, the quicksort algorithm allows costs to be reduced to approximately US$10,000 for ranking of the whole dataset along one axis. The principle of the quicksort algorithm is to choose an element, called a pivot, to which all other elements are compared. Thus, two subgroups of unsorted elements are created, one with a higher value than the pivot and the other with a lower value. Each subgroup is then sorted recursively in the same way until every element of each group is sorted.

The subgroups generated by the quicksort algorithm depend on the annotations gathered for a particular axis. Consequently, the pivot and the comparisons vary from one axis to another. That is why the annotation process of LIRIS-ACCEDE was divided into two experiments: one for annotation of valence and another for annotation of arousal. The experimental protocol was virtually the same for each axis and is described in Section 4.2.

4.2 Experimental setup

The annotation of the database along the arousal axis was performed three months after the annotation

1. While we conducted the experiments (summer 2013), CrowdFlower was distributing tasks over 50 labor channel partners, including Amazon Mechanical Turk and TrialPay. Since late 2013, the number of its labor channel partners has been considerably reduced. For example, CrowdFlower does not offer task distribution on Mechanical Turk anymore and it is no longer possible to choose on which labor channel the tasks are distributed.

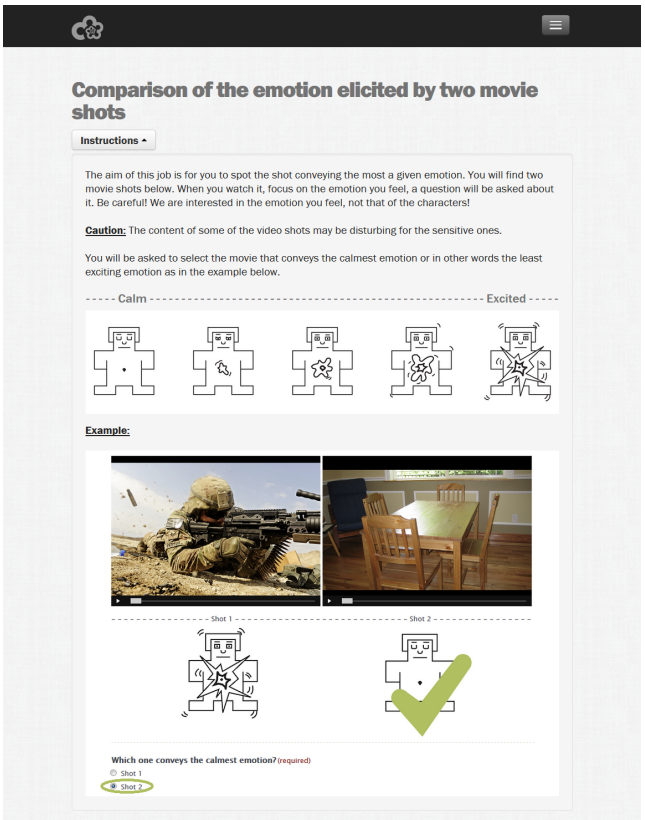

Fig. 2. Interface displayed to workers for annotation of the arousal axis.

along the valence axis. Meanwhile, a new interface for displaying tasks to workers had been released on CrowdFlower. This explains why there are few changes in both protocols to adapt the experimental setup to the new interface.

Given a pair of video excerpts, annotators had to select the one that conveys "the most positive emotion" (for valence) or "the calmest emotion" (for arousal). The words "valence" and "arousal" were not used since they might be misunderstood by the annotators. They were asked to focus on the emotion they felt when watching the video clips, *i.e.* the induced emotion, and not on that of the characters. As the arousal axis was more challenging to annotate, the arousal axis of the Self-Assessment Manikin and an example were displayed at the beginning of the task to make sure that annotators had understood the task properly. The Self-Assessment Manikin [44] is a powerful pictorial system used in experiments to represent emotional valence, arousal and dominance axes. Its non-verbal design makes it easy to use regardless of age, educational or cultural background. The interface displayed to workers for annotation of LIRIS-ACCEDE along the arousal axis is shown in Figure 2.

Video clips were displayed with a size of 280x390 pixels for annotation of the valence dimension and with a size of 189x336 pixels for the arousal dimension, to comply with the width of the new interface and use a more common aspect ratio. These clips were displayed using an embedded video player, meaning that workers were free to play each video clip as many times as they wanted. Workers were paid US$0.05 for answering five comparisons but could exit the task at any time. Despite the low reward for completing tasks, feedback on specialized crowdsourcing forums was very positive. Workers pointed out that the tasks were very easy, fun and enjoyable. Here are a few of their comments: "That's awesome!", "I did that last time, want to do that again, very easy :)".

To ensure the accuracy of annotations, 100 unnoticeable test questions, also called "gold units", were created for each axis and randomly inserted throughout the tasks. This made it possible to test and track annotators' performance by regularly testing them to ensure that they take the video clips comparisons seriously. The gold units correspond to unambiguous pairs of easily comparable video clips. If a wrong answer was given, a small paragraph was displayed explaining the reason why the answer was the other one. Workers were able to question the reason and send a message to explain their point of view. This system made it possible to forgive them when their protest was well-founded and to modify accordingly several gold units that were too subjective. However, if a worker gives too many wrong answers to gold units, none of his answers are considered, he receives no remuneration and his trust level on CrowdFlower drops. Thus, annotators are well aware that they must not answer the questions at random. For annotation of the arousal axis, a new advanced tool called "Quiz Mode" was available on CrowdFlower: annotators first have to answer six test questions and achieve an accuracy threshold of 70% in order to pass the quiz and work on the job. This ensures that only higher performing annotators are allowed to work on the tasks. Test questions were also randomly inserted to test annotators that passed the quiz on an on-going basis as they worked through the job.

In concrete terms, the quicksort algorithm was used in both annotation experiments to generate the comparisons. First, an initial video excerpt was randomly chosen to be the first pivot. All the other clips from the database were compared to this excerpt meaning that 9799 pairwise comparisons were generated for the first iteration. Each pair was displayed to workers until three annotations were gathered. We found this was a good compromise between the cost and the accuracy of the experiment. Once three annotations per comparison were made for all the comparisons, all the annotations were collected. In each comparison, the pivot was considered as inducing the most positive emotion for valence or the calmest emotion for arousal if at least two annotators selected the pivot during the annotation process. The final rank of the pivot was thus computed. Assuming that the pivot does not

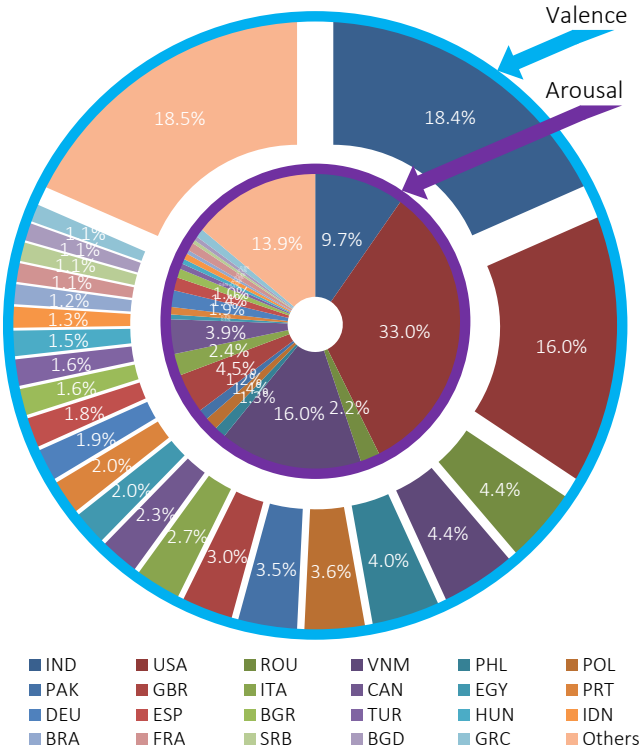

Fig. 3. Countries of the annotators for both the valence (external circle) and arousal (internal circle) annotation experiments. Countries accounting for less than 1% of the total in both experiments are classified as "Others".

induce the lowest or the highest valence or arousal, this process splits the database into two subgroups. For the second iteration, one pivot was selected in each subgroup and the two pivots were compared to the other video clips inside their subgroup, generating 9,797 new comparisons. For the next iteration, four pivots were selected and so on. The process was repeated until a rank was assigned to all the 9,800 video excerpts. Finally, each video excerpt is accompanied by two discrete values ranging from 0 to 9,799 representing its arousal and valence ranks.

4.3 Annotation results

For annotation of the valence axis, more than 582,000 annotations for about 187,000 comparisons were gathered from 1,517 trusted annotators from various countries. Annotators from 89 countries participated in the experiment, reflecting a huge diversity in cultural background. The majority of workers originated from India (18%), USA (16%), Romania (4%) and Vietnam (4%). A more detailed distribution of countries is displayed in Figure 3. Over 90% of data come from 530 of these annotators. The 1,517 trusted annotators showed an accuracy of 94.2% on test questions, whereas this accuracy was about 42.3% for untrusted annotators.

More iterations were needed to fully rank the database along the arousal axis. More than 665,000 annotations for around 221,000 unique comparisons were gathered from 2,442 trusted annotators also from 89 countries. As displayed in Figure 3, the countries of annotators is also diversified but different since most of the workers are American (33%), Vietnamese (16%), Indian (10%) and British (5%). As a point of comparison, this time over 90% of data come from 830 annotators. The accuracies on test questions for trusted and untrusted annotators were approximately the same as for those annotating the database along the valence axis. However, the number of untrusted annotators was slightly lower than for the first experiment thanks to the Quiz Mode.

When creating crowdsourcing tasks, ethical concerns have to be considered and the anonymity of the crowdworkers must be preserved. It is worth mentioning that, in our experiments, crowdworker privacy has been protected since the annotations that led to the ranking of the excerpts are not published. Only the final ranks for valence and arousal are released.

Combination of valence and arousal annotations shows convincing results. Dietz and Lang have shown in [45] that arousal and valence are correlated and that certain areas of this space are more relevant than others. Figure 4 shows the two-dimensional quantized histogram of ranks computed from annotations in the 2D valence-arousal space. Each cell indicates the number of video clips with a ranking for valence and arousal between the values represented on both axes. For example the top-left cell shows the number of excerpts with a ranking between 0 and 700 for valence and between 9,100 and 9,800 for arousal. Similarly to other studies such as [45] and [19], Figure 4 shows that there are relatively few stimuli eliciting responses annotated as low arousal and negative valence and that there are also less excerpts eliciting high arousal and neutral valence. Note that the values displayed in Figure 4 are the relative positions of excerpts in the valence-arousal space and not their absolute position.

4.4 Inter-annotator reliability

Inter-annotator reliability is an indication of how independent annotators participate in an experiment and reach the same conclusion despite the subjectivity of the task. It is essential to evaluate the consistency of the annotations to detect whether the scale is defective or whether the annotators need to be re-trained. Several measures of inter-annotator agreement are used in the literature such as percent agreement, Fleiss' kappa [46] and Krippendorff's alpha [47]. Percent agreement is widely used and intuitive but overestimates inter-annotator reliability since it does not take into account the agreement expected by chance. Most of the annotators who answered randomly have been discarded using gold data, which is why this measure will also be considered in Table 2. Fleiss' kappa and Krippendorff's alpha both take into account observed disagreement and expected disagreement but

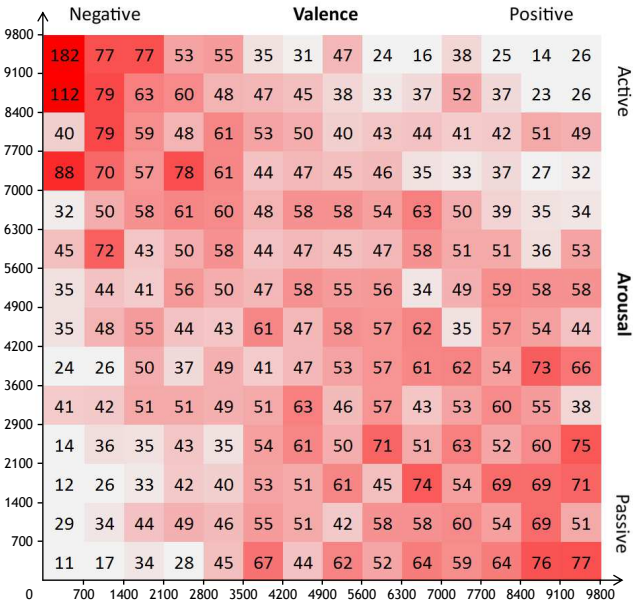

Fig. 4. Joint quantized histogram of ranks for the 9,800 excerpts in the valence-arousal space. For example, the bottom-left cell shows the number of video clips with a valence and an arousal rank between 0 and 700.

TABLE 2
Inter-annotator reliability

Measure	Arousal	Valence
Percent agreement	0.862	0.835
Fleiss' κ	0.190	0.179
Krippendorff's α	0.191	0.180
Randolph's κ	0.452	0.375

are sensitive to trait prevalence: they consider that annotators have a priori knowledge of the quantity of cases that should be distributed in each category [48] (*e.g.* "Excerpt 1" or "Excerpt 2" conveys most strongly the given emotion). The result is that, especially using binary answers which is the case here, if a value is very rare, reliability is low even if there are few mistakes in the annotations. In our annotation process this is a problem because the rarity of a category (the shot that conveyed most strongly a given emotion) greatly depends on the choice of pivots in the quicksort algorithm. For example, if a pivot with a high valence is selected, most annotators will answer that the pivot (always displayed as "Excerpt 2") has the highest valence. This will result in a low reliability using Fleiss' kappa and Krippendorff's alpha measures. Randolph's multirater kappa free [48] is not subject to prevalence because it does not depend on how many values are in each category. All these reliability coefficients are displayed in Table 2 to ensure a point of comparison.

Both kappa values need a fixed number of annotators per comparison to be computed. However, comparisons can be annotated by different annotators. For this reason, all comparisons that have been annotated by more than three people are discarded to compute both kappa values, corresponding to 7,459 units discarded for valence and 1,539 for arousal. Krippendorff's alpha is more flexible and allows missing data (comparisons can be annotated by any number of workers), thus no comparisons are discarded to compute this measure. The inter-annotator reliabilities for these subsamples are displayed in Table 2. Their values can range from 0 to 1 for percent agreement and from -1 to 1 for the other measures. For Fleiss' kappa, Krippendorff's alpha and Randolph's kappa, a value below 0 indicates that disagreements are systematic and exceed what can be expected by chance, a value equal to 0 indicates the absence of reliability, and a value higher than 1 indicates an agreement between annotators (1 for perfect reliability). In Table 2, all values are positive, which means that agreement is slightly better than what would have been expected by chance and is similar to other emotion annotation studies such as [19] or [36]. The percent agreement indicates that annotators agreed on 83.5% and 86.2% of comparisons. For Randolph's kappa measure, Landis and Koch [49] suggest that a score of 0.375 indicates a fair agreement and that a score of 0.452 corresponds to a moderate agreement. Thus, these results show that annotators have fully understood the tasks and achieved good agreement despite the subjectivity of both annotation experiments.

5 Baseline Framework

The goal of this section is to introduce a baseline similar to what can be found in the state of the art and to define several protocols to assess its performance using LIRIS-ACCEDE in different ways. These reproducible protocols will allow fair comparisons between future models and the baseline described below.

5.1 Regression

SVR has demonstrated good performance in many machine learning problems and, more specifically, in affective content analysis work such as [50], [27] or, more recently, [51]. SVR models construct a hyperplane by mapping vectors from an input space into a high dimensional feature space such that they fall within a specified distance of the hyperplane. Even if we have shown in Figure 4 that arousal and valence are correlated, two independent ϵ-SVRs are used in this work to model arousal and valence separately. The Radial Basis Function (RBF) is selected as the kernel function and a grid search is run to find the C, γ and p parameters. Since the database is ranked along the induced arousal and valence axis, the ground truth is made up of these raw ranks, initially ranging from 0 to 9,799, which are uniformly rescaled to a more common $[-1, 1]$ range. All features are normalized using the standard score before being used in the learning step.

TABLE 3
10 best performing features for estimating arousal and valence dimensions

Arousal	Valence
1. Global activity	1. Colorfulness [54]
2. Number of scene cuts per frame	2. Hue count [55]
3. Standard deviation of the wavelet coefficients of audio signal	3. Audio zero-crossing rate
4. Median lightness	4. Entropy complexity [56]
5. Slope of the power spectrum	5. Disparity of most salient points
6. Lighting	6. Audio asymmetry envelop
7. Colorfulness	7. Number of scene cuts per frame
8. Harmonization energy [53]	8. Depth of field
9. Length of scene cuts	9. Compositional balance [57]
10. Audio flatness envelop	10. Audio flatness

5.2 Feature selection

A large number of features have been investigated before being separated into three modalities: audio, still image and video features.

Audio features are extracted using 40 ms windows with 20 ms overlap. Many audio features were considered: MFCC, energies, flatness, standard deviation and mean of the quadratic spline wavelet coefficients of the audio signal computed using the fast algorithm described in [52], asymmetry, zero-crossing rate, *etc.* all averaged over the signal. Still image features are extracted from the key frame of the excerpts. This is the frame with the closest RGB histogram to the mean RGB histogram of the whole excerpt using the Manhattan distance. We considered many features, which have proven to be efficient in affective image analysis, as well as more uncommon ones including color harmony and aesthetic features related to the composition of the key frame. Video features contain information about the composition (number of scene cuts, fades, etc.) and motion.

We created two feature sets, one for each axis, made up of the most efficient features. Best features are selected by hierarchically merging the best performing ones as long as the mean-square error (MSE) decreases. Using this process, a set of 17 features is obtained for valence and 12 features for arousal. Rejected features were not necessarily inefficient features but features that were strongly correlated with more efficient ones. The 10 best performing features for estimating arousal and valence are summarized in Table 3.

Color features performed well for detecting valence, as five features out of the 17 features were color-related. For valence, colorfulness [54] was the best performing feature followed by "hue count" [55]. The other features in this set, from the third best performing feature to the least efficient feature are: audio zero-crossing rate, entropy complexity [56], disparity of most salient points (standard deviation of normalized coordinates), audio asymmetry envelope, number of scene cuts per frame, depth of field (using the blur map computed in [58]), compositional balance [57], audio flatness, orientation of the most harmonious template [53], normalized number of white frames, the color energy and color contrast [21], scene complexity (area of the bounding box that encloses the top 96.04% of edge energy [55]), number of maximum values in the saliency map and, finally, number of fades per frame.

As expected, motion and energy features were the best performing ones for modeling arousal. The selected features are global activity (average size of motion vectors), standard deviation of the wavelet coefficients of audio signal, the energy corresponding to the most harmonious template [53], the slope of the power spectrum, median lightness, the lighting feature [57], length of scene cuts and the audio flatness envelope. As arousal and valence are correlated, it is not surprising that four features selected among the best performing ones for valence have also been selected for arousal. These features are the number of scene cuts per frame, colorfulness, the normalized number of white frames, and the orientation of the most harmonious template.

5.3 Protocols

The purpose of this section is to introduce standard protocols using the database in different ways to make possible comparisons within the field. The MSE and Pearson's r are computed to quantify the performance of each protocol. The MSE for regression models is widely used to quantify the difference between estimated values and the true values estimated. It measures the amount by which the estimated values differ from the ground truth and assesses the quality of the regression in terms of its variation and degree of bias, while the Pearson product-moment correlation coefficient (or Pearson's r) is a measure of the linear correlation between estimated and true values.

5.3.1 Protocol A: Predefined subgroups

In this protocol, the training and test sets have been manually defined to make sure that they each include 4,900 excerpts from 80 films. In this way, the excerpts extracted from the same film are only in one of the sets, either the training set or the test set. Insofar as possible, we tried to distribute movies equally in the sets according to their genres. We also defined a validation set, should it be needed in future studies, by dividing the training set into two subgroups, each made up of 2,450 excerpts extracted from 40 films. The list of excerpts in each set is available alongside the database and the annotations. The MSE and Pearson's r for this protocol using our baseline model are shown in Table 4.

TABLE 4
Performance for Protocols A (Predefined subgroups) and B (Leave-One-Movie-Out). Ground truth and estimated scores range from -1 to 1

	Protocol A		Protocol B	
Metric	Arousal	Valence	Arousal	Valence
MSE	0.303	0.302	0.326	0.343
Pearson's r	0.308	0.310	0.242	0.221

TABLE 5
Performance for Protocol C (Same genre)

Genre	Arousal MSE	Valence MSE
Action	0.278	0.326
Adventure	0.389	0.363
Animation	0.336	0.335
Comedy	0.297	0.295
Documentary	0.326	0.308
Drama	0.313	0.327
Horror	0.331	0.364
Romance	0.324	0.361
Thriller	0.355	0.337

5.3.2 Protocol B: Leave-One-Movie-Out

This protocol is a standard protocol used in numerous studies in affective analysis. It consists in selecting the excerpts of one movie for testing while using the rest for training. This process is repeated for the 160 movies in the database. The final averaged results for this protocol are presented in Table 4.

5.3.3 Protocol C: Same genre

It could also be interesting to focus on specific genres to study the efficiency of models and the effect of features depending on the movie genre. The protocol is the leave-one-movie-out protocol for movies that share the same genre. The final averaged MSEs for each genre, still using the same sets of features defined in Section 5.2, are shown in Table 5.

5.3.4 Protocol D: Same movie

The purpose of this last protocol is to gain insight into the regularity of the movie in terms of affective impact. Indeed, by learning on samples from the first half of a movie and testing on the remaining excerpts, the results can provide information on how well the first part of a movie is able to model and to be generalized to the induced valence and arousal of the whole movie. The results of this protocol, applied to some movies of the database, are displayed in Table 6.

5.4 Regression results

The results are promising given the huge variety of movies in the database. They indicate that regression models perform well in modeling of both induced valence and arousal, but with varying degrees of success depending on which protocol is used. Globally, MSE values are significantly smaller than MSE values computed using random sets (around 0.667 and estimated by generating large random samples made up of values between -1 and 1). As pointed out in previous sections, it is not possible to directly compare the performance of our model to previous state of the art models. They use different test sets and, in most cases, different performance metrics and output scales. On the other hand, researchers using one of the protocols defined in Section 5.3 will be able to know how their model performs not only with respect to this baseline but also to all future work using one of these protocols.

There are several unexplored leads in this work that we would like to share with the reader. One way to improve performance is to check inter-annotator reliabilities to detect outliers and thus remove them from the process. Furthermore, protocol C shows different levels of performance depending on the movie genre, indicating that our model misses some information that could potentially be added by higher level features. Last but not least, as arousal and valence are correlated, it seems legitimate to model them jointly.

TABLE 6
Performance for Protocol D (Same movie)

Movie	Arousal MSE	Valence MSE
20 Mississippi	0.305	0.317
Dead Man Drinking	0.309	0.274
Decay	0.330	0.321
Home	0.176	0.401
Lionshare Legacy	0.443	0.273
Monolog	0.290	0.395
Sweet Hills	0.206	0.197
The Master Plan	0.303	0.344
You Again	0.089	0.098

6 Discussion

One of the main limitations of the proposed database lies in the fact that the video clips have been ranked relatively to each other. Thus the rankings provide no information on distribution of the database in the 2D valence-arousal space. In other words, it is uncertain whether the extreme cases with the lowest or highest ranks elicit extreme emotions. Furthermore, these ranks are relative to this particular database, which prevents comparison with other video clips annotated with absolute valence and arousal scores.

To address this limitation, we carried out an experiment [59] in which annotators were asked in a controlled environment to rate on a 5-point discrete Self-Assessment Manikin scale some excerpts from the database for arousal and valence. The results have shown that the Spearman's rank correlation coefficient (SRCC) between affective ratings and crowdsourced rankings is significantly high for both arousal ($SRCC = 0.751, t(44) = 7.635, p < 1 \times 10^{-8}$) and valence ($SRCC = 0.795, t(44) = 8.801, p < 1 \times 10^{-10}$), thus cross-validating the overall dataset. The controlled affective ratings also make it possible to estimate the distribution of the database and understand the range of emotions elicited by the database.

Several other unknown factors can potentially affect the ratings and would require further research.

First, crowdworkers were asked to focus on what they felt in response to the video excerpts. Contact with the crowdworkers was quite limited. As such, it was not possible to ascertain that annotators were annotating the induced emotion and not the perceived emotion or even the emotion they thought they should feel, since it is possible to make judgments on the basis of conventional characteristics without experiencing any emotion [60]. If some crowdworkers did not distinguish between felt and perceived emotions, noisiness could potentially be introduced in our data as Zenter *et al.* showed that ratings of perceived emotion differ significantly from ratings of felt emotion [61]. The distinction between ratings of perceived or felt emotion is outside the scope of this paper. Thus, in this work, we do not try to distinguish ratings of felt emotion from ratings of perceived emotion.

Second, there was no way to make sure that crowdworkers really turned on the volume to judge the videos. While creating the gold data, sound was taken into consideration. Thus, we assume that most workers passing the gold data turned the volume on. Furthermore, the correlation between affective ratings collected in a controlled environment where the sound was turned on and crowdsourced rankings is significantly high [59]. As a consequence, we hypothesize that most crowdworkers turned the volume on to rate the pairwise comparisons.

Third, the crowdworkers made the annotations in various uncontrolled environments under different conditions. However, elicitation of an emotion is a subtle process depending on a large number of factors (e.g. listener, performance or contextual features) [62]. Despite this, inter-annotator reliability indicates that an overall agreement was achieved among crowdworkers and that annotations tend to be stable. Moreover, these results have been compared to ratings gathered in controlled conditions in order to validate the annotations made in uncontrolled conditions and to detect potential outliers [63]. The correlation between affective ratings and crowdsourced rankings is significantly high, thus cross-validating the overall database for future uses in research work. These affective ratings also make it possible to enhance the range of applications for automatic approaches capable of predicting the affective impact. Indeed, it will be easier to create new evaluation protocols, such as separating data to create two or more meaningful categories to evaluate the efficiency of classifiers for which precise affective ratings are not necessary.

7 Conclusion

This paper has addressed the lack of large video databases for affective video analysis, as current existing databases are limited in size and not representative of today's movies. Following the work began in [6], we proposed LIRIS-ACCEDE, a large video database freely shared to be used by the research community. The database is made up of 9,800 excerpts lasting from 8 to 12 seconds, extracted from 160 diversified movies. All the 160 movies are shared under Creative Commons licenses, thus allowing the database to be shared publicly without copyright issues. It is available at: http://liris-accede.ec-lyon.fr/.

All the excerpts have been ranked along the induced valence and arousal axes by means of two experiments conducted on a crowdsourcing platform. Both experiments were highly attractive. A large number of annotators performed each experiment, making it possible to collect large volumes of affective responses from a wide diversity of annotators and from a large spectrum of contexts. With this experimental design, high inter-annotator reliabilities were achieved considering the subjectivity of the experiments. We also introduced standard protocols using the database in an attempt to perform standardized and reproducible evaluations to fairly compare future work within the field of affective computing. Four protocols were proposed corresponding to different goals and needs. Moreover, we implemented a baseline and used these protocols to assess its performance, showing promising results. Note that all the audio and visual features used for the baseline are also released alongside the database.

By creating this database, we aim at helping compensate for the lack of large database availability for affective video analysis and to create a database that could be easily shared and used by other researchers in future work dealing with affective computing. We encourage other researchers to annotate the database according to other modalities they need, and to extend its full range of applications in affective video analysis and, even more globally, in affective computing.

Acknowledgments

This work was supported in part by the French research agency ANR through the VideoSense Project under the Grant 2009 CORD 026 02 and the Visen project within the ERA-NET CHIST-ERA framework under the grant ANR-12-CHRI-0002-04.

REFERENCES

[1] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Processing Magazine*, 2006.

[2] S. Arifin and P. Y. K. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1325–1341, Nov. 2008.

[3] S. Zhao, H. Yao, X. Sun, P. Xu, X. Liu, and R. Ji, "Video indexing and recommendation based on affective analysis of viewers," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11, 2011, pp. 1473–1476.

[4] M. Horvat, S. Popovic, and K. Cosic, "Multimedia stimuli databases usage patterns: a survey report," in *Proceedings of the 36nd International ICT Convention MIPRO*, 2013, pp. 993–997.

[5] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1075–1089, Jun. 2014.

[6] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, "A large video data base for computational models of induced emotion," in *Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 13–18.

[7] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[8] P. Philippot, "Inducing and assessing differentiated emotion-feeling states in the laboratory," *Cognition & Emotion*, vol. 7, no. 2, pp. 171–193, 1993.

[9] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[10] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction*, 2007, vol. 4738, pp. 488–500.

[11] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, Nov. 2010.

[12] J. Rottenberg, R. D. Ray, and J. J. Gross, "Emotion elicitation using films," *Handbook of emotion elicitation and assessment*, p. 9, 2007.

[13] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: a database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[14] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[15] S. Carvalho, J. Leite, S. Galdo-Álvarez, and O. Gonçalves, "The emotional movie database (EMDB): a self-report and psychophysiological study," *Applied Psychophysiology and Biofeedback*, vol. 37, no. 4, pp. 279–294, 2012.

[16] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "A benchmarking campaign for the multimodal detection of violent scenes in movies," in *Proceedings of the 12th International Conference on Computer Vision*, ser. ECCV'12, 2012, pp. 416–425.

[17] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated GIFs," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14, 2014, pp. 213–216.

[18] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[19] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 2376–2379.

[20] H.-B. Kang, "Affective content detection using HMMs," in *Proceedings of the eleventh ACM international conference on Multimedia*, ser. MULTIMEDIA '03, 2003, pp. 259–262.

[21] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[22] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *Affective Computing and Intelligent Interaction*, 2007, vol. 4738, pp. 594–605.

[23] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceedings of the 16th ACM international conference on Multimedia*, ser. MM '08, 2008, pp. 677–680.

[24] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *IEEE International Symposium on Multimedia*, Dec. 2008, pp. 228–235.

[25] M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Affective Computing and Intelligent Interaction*, Sep. 2009, pp. 1–7.

[26] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[27] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.

[28] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 26, 2013.

[29] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 566–569.

[30] R. W. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Electronic Imaging*, 1998, pp. 290–301.

[31] H. Leventhal and K. Scherer, "The relationship of emotion to cognition: A functional approach to a semantic controversy," *Cognition & Emotion*, vol. 1, no. 1, pp. 3–28, 1987.

[32] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–8.

[33] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in visual information retrieval," *IEEE Multimedia*, vol. 6, no. 3, pp. 38–53, Sep. 1999.

[34] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[35] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010*, Jul. 2010.

[36] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, Aug. 2013.

[37] Y.-H. Yang and H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.

[38] P. A. Russell and C. D. Gray, "Ranking or rating? some data and their implications for the measurement of evaluative response," *British Journal of Psychology*, vol. 85, no. 1, pp. 79–92, Feb. 1994.

[39] S. Ovadia, "Ratings and rankings: reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.

[40] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: A comparative study of self-reporting," in *Affective Computing and Intelligent Interaction*, 2011, vol. 6974, pp. 437–446.

[41] R. Mantiuk, A. M. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment." *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[42] J. D. Smyth, D. A. Dillman, L. M. Christian, and M. J. Stern, "Comparing check-all and forced-choice question formats in web surveys," *Public Opinion Quarterly*, vol. 70, no. 1, pp. 66–77, 2006.

[43] S. S. Skiena, *The algorithm design manual*, 2nd ed. London: Springer, 2008.

[44] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[45] R. B. Dietz and A. Lang, "Aefective agents: Effects of agent affect on arousal, attention, liking and learning," in *Cognitive Technology Conference*, 1999.

[46] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[47] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, Apr. 1970.

[48] J. J. Randolph, "Free-marginal multirater kappa (multirater κ_{free}): An alternative to fleiss fixed-marginal multirater kappa," Paper presented at the *Joensuu University Learning and Instruction Symposium*, Oct. 2005.

[49] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[50] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 1853–1856.

[51] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 636–647, 2013.

[52] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710–732, Jul. 1992.

[53] Y. Baveye, F. Urban, C. Chamaret, V. Demoulin, and P. Hellier, "Saliency-guided consistent color harmonization," in *Computational Color Imaging Workshop*, 2013, vol. 7786, pp. 105–118.

[54] D. Hasler and S. Suesstrunk, "Measuring colourfulness in natural images," in *Proc. SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, 2003, pp. 87–95.

[55] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 419–426.

[56] O. Le Meur, T. Baccino, and A. Roumy, "Prediction of the inter-observer visual congruency (IOVC) and application to image ranking," in *Proceedings of the 19th ACM International Conference on Multimedia*, 2011, pp. 373–382.

[57] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proceedings of the 10th International Conference on Computer Vision*, 2008, vol. 5304, pp. 386–399.

[58] Y. Baveye, F. Urban, and C. Chamaret, "Image and video saliency models improvement by blur identification," in *Computer Vision and Graphics*, vol. 7594, 2012, pp. 280–287.

[59] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "A protocol for cross-validating large crowdsourced data: The case of the LIRIS-ACCEDE affective video dataset," in *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, ser. CrowdMM '14, 2014, pp. 3–8.

[60] J. A. Sloboda, "Empirical studies of emotional response to music." in *Cognitive bases of musical communication.* American Psychological Association, 1992, pp. 33–46.

[61] M. Zentner, D. Grandjean, and K. R. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement." *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[62] K. R. Scherer and M. R. Zentner, "Emotional effects of music: Production rules." *Music and emotion: Theory and research*, pp. 361–392, 2001.

[63] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "From crowdsourced rankings to affective ratings," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.

Yoann Baveye is working toward his Ph.D. degree at Technicolor Research & Innovation in Rennes, France, in partnership with the Ecole Centrale de Lyon, France. He was awarded his Master's degree in computer science from the University of Rennes, France, in 2012. His research interests include modeling and extraction of emotional impact in movies.

Emmanuel Dellandréa was awarded his Master and Engineering degrees in Computer Science from the Université de Tours, France, in 2000 followed by his Ph.D. in Computer Science in 2003. He then joined the Ecole Centrale de Lyon, France, in 2004 as an Associate Professor. His research interests include multimedia analysis, image and audio understanding and affective computing, including recognition of affect from image, audio and video signals.

Christel Chamaret is currently working as a Senior Scientist at Technicolor Research & Innovation in Rennes, France. Her research focuses on color image processing and, more particularly, on color harmonization and color grading. She has previously worked on visual attention models, quality metrics as well as on aesthetic models. She has also designed and conducted a number of user studies (pairwise protocol, eyetracking, etc.) to validate computational models and image processing algorithms. She holds two master's degrees defended in 2003 from the University of Nantes and the Ecole Centrale de Nantes, France.

Liming Chen was awarded his B.Sc. degrees in joint mathematics-computer science from the University of Nantes, France, in 1984, and his M.S. and Ph.D. degrees in computer science from the University of Paris 6, France, in 1986 and 1989, respectively. He first served as an Associate Professor with the Université de Technologie de Compiègne, France, before joining the Ecole Centrale de Lyon (ECL), France, as a Professor in 1998, where he leads an Advanced Research Team in multimedia computing and pattern recognition. Since 2007, he has been the Head of the Department of Mathematics and Computer Science at ECL. His current research interests include computer vision and multimedia, and in particular 2-D/3-D face analysis, image and video categorization, and affective computing. He is a Senior Member of the IEEE.