A Large Video Database for Computational Models of Induced Emotion

Yoann Baveye*†, Jean-Noël Bettinelli*, Emmanuel Dellandréa*, Liming Chen*, and Christel Chamaret†
*Université de Lyon, CNRS
Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France
{yoann.baveye, jean-noel.bettinelli, emmanuel.dellandrea, liming.chen}@ec-lyon.fr
†Technicolor, 975, avenue des Champs Blancs, 35576 Cesson Sévigné, France
{yoann.baveye, christel.chamaret}@technicolor.com

***Abstract*—To contribute to the need for emotional databases and affective tagging, the LIRIS-ACCEDE is proposed in this paper. LIRIS-ACCEDE is an Annotated Creative Commons Emotional DatabasE composed of 9800 video clips extracted from 160 movies shared under Creative Commons licenses. It allows to make this database publicly available[1] without copyright issues. The 9800 video clips (each 8-12 seconds long) are sorted along the induced valence axis, from the video perceived the most negatively to the video perceived the most positively. The annotation was carried out by 1517 annotators from 89 different countries using crowdsourcing. A baseline late fusion scheme using ground truth from annotations is computed to predict emotion categories in video clips.**

I. Introduction

Multimedia data, especially videos, are a powerful mean to convey emotions and are used to make the viewer feel what the film-maker wants to express. Emotions are a subjective concept that depend on various factors that affect the perception of videos such as the mental state but also the context and the cultural background. On the other hand, there is a set of 6 basic emotions which are felt and expressed universally [1]. Detecting affective scenes could be useful to enrich recommender systems by adding emotional information, create personalized Video on Demand databases, help film-makers shoot and edit a movie with desired emotion effect and improve movie search engines. For instance, users would be able to view only the funniest scenes of a film, or remove scaring scenes to allow children to view the movie. However due to the subjective nature of emotions, assessing emotions elicited by movies is not a trivial task. The design and public release of a consistent and universal ground truth is a first challenge that we attempt to address.

To our knowledge, there do not exist public emotionally annotated video databases yet. Existing research work mostly relies on private small datasets: each method for affective content recognition of the videos uses its own database for learning and testing. This results in the difficulty to benchmark the efficiency of methods. The goal of the paper is to propose a large database shared under BY-NC-SA Creative Commons license to be used for induced emotion recognition and freely available to the academic community.

[1]The dataset and its annotations are publicly available at: http://liris-accede.ec-lyon.fr/.

The paper is organized as follows: in section II, an overview of existing affective databases, followed by an overview of existing frameworks for video affective tagging are presented. In section III, the proposed database is described including the experiments to gather the results for the valence modality. Finally, section IV presents the results obtained using the database in a common video affective content recognition framework.

II. Background

There exist in the litterature two approaches for the modelling of emotions: discrete or dimensional. The discrete emotions approach is very natural since it goes back to the origin of language and the emergence of words and expressions representing clearly separable states. Many discrete categorizations of emotions have been proposed, such as the six basic universal emotions proposed by Ekman in [1]. This approach faces a granularity issue since the number of emotion classes is too small in comparison with the diversity of emotion perceived by film viewers. In case the number of classes is increased, ambiguities due to language difficulties or personal interpretation appear. Dimensional approaches of emotion have also been proposed, such as Plutchik's emotion wheel [2] and the valence-arousal scale by Russel [3] extensively used in researches dealing with affective understanding. The granularity and ambiguity issues are avoided since emotions are no longer described as discrete categories.

We have thus chosen to use the Russel's valence-arousal scale in the LIRIS-ACCEDE to quantitatively describe emotions. In this scale, each subjective feeling can be described by its position in a two-dimensional space formed by the dimensions of valence and arousal. Arousal can range from inactive (*e.g.*, tired, pensive) to active (*e.g.*, alarmed, excited), whereas valence ranges from negative (*e.g.*, sad, disappointed) to positive (*e.g.*, joyous, elated). While arousal and valence explain most of variations in emotional states, a third dimension of dominance can also be included in the model [3]. Given the difficulty of consistently identifying a third dimension (such as dominance, tension or potency) which differs from arousal, many studies limit themselves to the valence and arousal dimensions.

A. Emotional databases

Recent advances in affective computing in video data and emotion recognition in human-computer interaction have led

to the creation of new databases containing emotional labels. The HUMAINE database [4] created by Douglas-Cowie *et al.* consists of a subset of three naturalistic and six induced reaction databases. It contains 50 clips described by a structured set of emotional labels attached to the clips both at a global level, and frame-by-frame, showing change over time.
The Database for Emotion Analysis using Physiological signals (DEAP) is a multimodal dataset for the analysis of human affective states proposed by Koelstra *et al.* in [5]. The DEAP dataset consists of two parts: the ratings from an online self-assessment where 120 one-minute excerpts from music videos were each rated by 14-16 volunteers based on arousal, valence and dominance and another experiment where physiological signals from 32 participants were recorded in response to music video clips.
The MAHNOB-HCI database [6] is another recent database consisting of two experiments. In the first experiment physiological signals from 20 participants watching 20 emotional videos between 34.9 and 117s long were recorded. The second experiment was a tag agreement experiment in which images and short videos were displayed with and without a tag. Participants' agreement with the displayed tag was assessed.

None of these databases are large or diverse enough to be used in learning strategies to represent the whole range of existing movie genres. Thus, there is a need for a larger video database which could be used in computational models of emotions.

B. Emotional classification

Hanjalic was one of the first to be convinced that the affective dimension of movies could be explored thanks to the "expected mood" [7] : the set of emotions the filmmaker intends to communicate for a particular audience with a common cultural background. They showed that low level features could be correlated with the affective dimension of media. Xu and Hanjalic proposed an approach based on direct mapping of specific video features onto the Potential-Arousal dimensions [8]. Even though this model was inspired by rules proposed by film theorists to elicit a particular emotion, it has not been validated by psychological experiments.
Wang and Cheong introduced in the same way features inspired from psychology and cinematography rules [9]. Their multimodal features are learned by two SVM based probabilistic inferences. One SVM is especially dedicated to audio cues to obtain high-level audio information at the scene level. Each video segment is then classified with a second SVM to obtain probabilistic membership vectors. Their training data consisted of 36 full-length popular Hollywood movies divided into 2040 scenes.
Arifin and Cheung [10] based their framework on a hierarchical-coupled dynamic Bayesian network to model the dependencies between the Potential-Arousal-Dominance (PAD) dimensions. This model takes into consideration the influence of former emotional events. They used a total of 10970 shots and 762 video segments extracted from 43 videos and labeled with the emotion of the whole video.
Zhang *et al.* used an affinity propagation instead of classification for fast clustering music video clips [11]. Features are first extracted on a segment length of 50 seconds in the center part of the music video. Arousal is computed with a linear combination of "arousal features" (as Motion intensity, shot switch rate) and valence with "valence features" (rhythm regularity, pitch). Their database of 156 English pop music videos is clustered into 8 categories. In another work [12], they introduced a personalized affective model for which they enlarged their database to 552 music videos in different languages and different styles.
Soleymani *et al.* [13] compared the affective grades obtained automatically from either physiological responses or from audiovisual features. They showed significant correlations between multimedia features, physiological features and users self-assessment of valence-arousal. A video dataset of 64 movie scenes extracted from 8 Hollywood movies was created. The following year, they introduced a Bayesian framework for video affective representation [14] using audiovisual features but also 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 arousal indicator feature for the scene affective classification. For this work they built a training set of 21 full length movies annotated by arousal and valence.
Irie *et al.* proposed in [15] a framework to model emotion dynamics with reference to Plutchiks emotion theory [2]. It takes into consideration temporal transition characteristics of human emotion. They first divide a movie into a sequence of movie shots and represent each shot as a histogram of quantized audiovisual features. Their emotion model is then applied to classify the emotions based on the sequence of latent topics corresponding to movie shots. The probability of a given emotion at a given shot, knowing the previous emotion is weighted by the distance between the eight basic emotions on the Plutchiks model. They conducted experiments on 206 scenes extracted from 24 movies titles (average scene length of 1 minute 51 seconds).
Yan *et al.* proposed to use unascertained clustering for video affective recognition [16]. Their experimental dataset was composed of 112 video scenes segmented from four films.
Malandrakis *et al.* presented a model using hidden Markov models to build discrete valued arousal and valence then converted to continuous values via interpolation [17]. Their emotional database consists of contiguous thirty-minute video clips from twelve well known movies of the Academy Award winners list.
In [18], Canini *et al.* modeled the relationship between audiovisual features and connotative rates. A point in the connotative space describes one movie segment in terms of its connotative properties derived from cinematography. The advantage is that there is higher agreement on the judgments of the connotative properties of the movie (by which terms the concept is described) than on the experienced affective reactions. They showed that movie scenes sharing similar connotation are likely to elicit in the same user the same affective reactions. Their ground-truth consisted in 25 movie scenes.

Datasets used in these works are task specific, heterogeneous (Table I) and to our knowledge none of these datasets has been made publicly available, making comparison and progress within the field difficult. A large video database publicly available shared with the academic community is thus essential in affective computing.
Our goal is to provide to the community a large dataset of high quality videos and robust relative annotations. Therefore, this dataset could be easily used by researchers to perform their

TABLE I. COMPARISON OF THE DATASETS USED IN PREVIOUS WORKS

Study	Subjects	Dataset Size	Modalities
Wang and Cheong [9]	3	36 full-length popular Hollywood movies (2040 scenes)	One emotion per scene among 7 categories
Arifin and Cheung [10]	14	43 videos (10970 shots and 762 video segments)	One emotion per video among 6 categories
Zhang *et al.* [12]	10	552 music videos in different languages and different styles	Arousal and valence scores
Soleymani *et al.* [13]	8	8 famous Hollywood movies (64 movie scenes)	Arousal and valence self-assessment
Soleymani *et al.* [14]	1	21 full length movies	Arousal and valence changes within scenes
Irie *et al.* [15]	16	24 movies titles (206 scenes)	Scenes scored with Plutchik's emotions
Yan *et al.* [16]	10	4 films (112 scenes)	One emotion among 4 basic categories
Malandrakis *et al.* [17]	7	Twelve well known movies (contiguous 30min video clips)	Intended emotion (by experts) and experienced emotion
Canini *et al.* [18]	240	25 "great" movie scenes	Scores on the 3 connotative dimensions
Proposed database	1517	160 movies (9800 video clips)	Video clips sorted along the valence axis

experiments and to compare fairly their results.

III. PROPOSED DATABASE

A. Composition

The database is composed only of excerpts from movies shared under Creative Commons licenses. The Creative Commons licenses allow creators to use standardized way to give the public permission to share and use their creative work under certain conditions of their choice. Creative Commons licenses consist of four major condition modules: Attribution (BY), requiring attribution to the original author; Share Alike (SA), allowing derivative works under the same or a similar license; Non-Commercial (NC), preventing the work from being used for commercial purposes; and No Derivative Works (ND), allowing only the use of the original work without any modification. Moreover, movies shared under Creative Commons licenses are often little known films, which limits viewers' prior memories. The movies selected to be part of the LIRIS-ACCEDE are shared under Creative Commons licenses with BY, SA or NC modules. Movies shared with ND module are not taken into account in this work since it is not allowed to modify these videos, and therefore it is not allowed to segment them. Thus, using videos shared under Creative Commons licenses allows us to share the database publicly.

160 films and short films with different genres are used and segmented into 9800 video clips. The total time of all 160 films is 73 hours 41 minutes and 7 seconds, and a video clip is extracted on average every 27s. The 9800 segmented video clips last between 8 and 12 seconds and are representative enough to conduct experiments. Indeed, the length of extracted segments is large enough to get consistent excerpts allowing the viewer to feel emotions [19] and is also small enough to make the viewer feel only one emotion per excerpt. The content of the movie is also considered to create homogeneous, consistent and meaningful excerpts not to disturb the viewers. A robust shot and fade in/out detection has been implemented using [20] to make sure that each extracted video clip start and end with a shot or a fade. Furthermore, the order of excerpts within a film is kept, allowing to study the temporal transitions of emotions. Several movie genres are represented

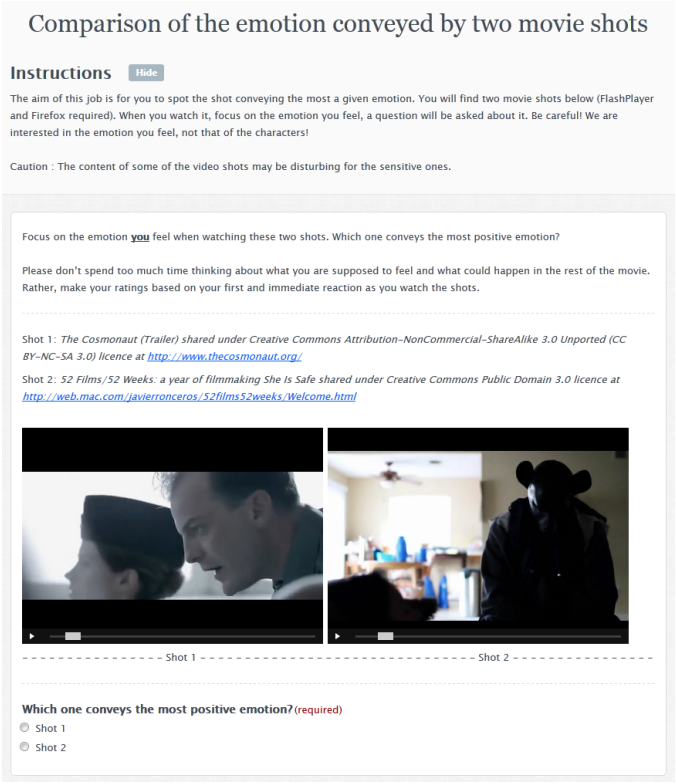

Fig. 1. Interface displayed to workers.

in this collection of movies such as horror, comedy, drama, action and so on. Languages are mainly English with a small set of Italian, Spanish, French and others subtitled in English.

B. Protocol

In order to sort the database along the induced valence axis, pairs of video clips were presented to annotators on Crowdflower (Figure 1). CrowdFlower is a crowdsourcing service which has over 50 labor channel partners, among them Amazon Mechanical Turk and TrialPay. CrowdFlower differs from these individual networks because they offer enterprise solutions and a higher degree of quality control, called "Gold Standard Data", to ensure the accuracy on the tasks.

The choice of a rating-by-comparison experiment has been motivated by several arguments. Metallinou and Narayanan showed in [21] that people agree more when describing emotions in relative terms rather than in absolute terms. Rating emotion is more difficult because of the internal scale of annotators. Indeed, requesting a score to annotators requires to give scoring references to get them understand the rating scale. By doing so, annotators tend to rate video clips according to scoring examples or previous ratings instead of their true feeling. The work of Yannakakis and Hallam in [22] runs along the same lines as the previous remark saying that pairwise preferences are more appropriate detectors of user states, eliminating the subjective notion of scaling and effects related to order of play. Yang and Chen also used a ranking-based approach to annotate a song in terms of valence and arousal [23]. They showed that the ranking-based approach simplifies emotion annotation and enhances the reliability of the ground

truth compared to rating approaches. That is why the rating-by-comparison process used in this work is intuitively easier, more robust and more suited for an experiment using crowdsourcing. This protocol leads in the relative position of video clips along the induced valence axis, from the video perceived the most negatively to the video perceived the most positively.

For each pair of video clips, annotators had to select the video clip which conveyed the most positive emotion. The word "valence" was not used since it 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 that of the characters.

Each video clip was displayed with a size of 280x390 pixels and there was no limit of time: annotators (also called workers) were free to play each video clip as many times as desired. A worker had to rank 5 pairs of video clip before being paid 0.05$ and was able to exit the task at any time. A total of 1517 annotators performed the task from 89 different countries. Most of the workers are Indian (18%), American (16%), Romanian (4%) and Vietnamese (4%). Over 90% of our data comes from 530 of these annotators. To test and track the performance of annotators, undetectable test answers also called "Gold Standard Data" were created and randomly inserted throughout the tasks. They correspond to pairs of video clips easily comparable pre-labeled with known answers. If a worker gives too many wrong answers to these test questions, none of his answers are considered, he does not receive any remuneration and its statistical confidences on CrowdFlower fall. Thus, Internet users have little interest in trying to answer the questions at random. Each pair is displayed to annotators until the same answer has been given three times. With the described protocol, the 1517 annotators provided more than 582 000 comparisons. The 1517 trusted annotators had a gold accuracy of 94.2% whereas this accuracy was about 42.3% for untrusted annotators. Annotators took approximately 23s to perform a task (*i.e.* to compare two videos and give the result) and the inter-annotator agreement was about 81.3% which means that the task was well understood by annotators.

C. Rating management

The quick sort algorithm was used to create and sort the comparisons, which is one of the most efficient sorting algorithm. Indeed, the average computational complexity of this algorithm is $N \times \log(N)$, where N is the number of data to sort. Hence, the cost of the sorting operation being proportional to the number of comparisons, the quick sort algorithm seems to be the best choice to reduce the costs as well as the time needed for the annotators to sort the database in comparison with other sorting algorithms. The principle of the quick sort algorithm is to chose randomly a pivot data to which all other data are compared, creating two subgroups of data: one having a higher value than the pivot, the other having a lower value. Each of the subgroups are then sorted in the same fashion recursively until every element of each group is sorted.
The choice of the considered comparisons and the pivot clip, as well as the processing of the data are done thanks to a program we made, called CPS (Crowd Powered Sort) which will be released as an open source program for those who would like to improve or expand the database with new videos or new axis. Some examples of key frames from the video clip with the lowest valence to the video clip with the highest valence (left-right, top-bottom) are showed in Figure 2.

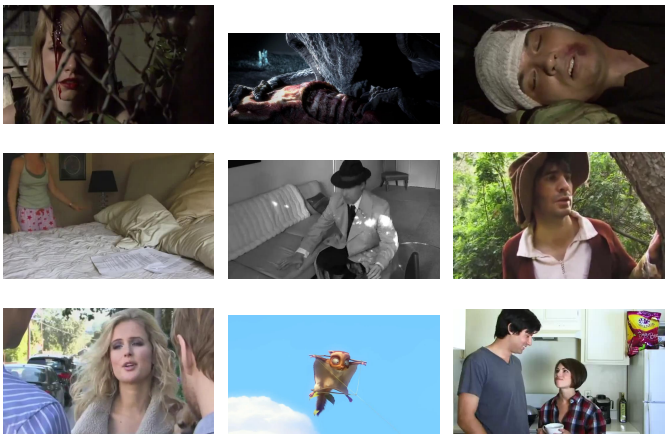

Fig. 2. Key frames from video clips sorted from the video clip with the lowest valence (top-left) to the video clip with the highest valence (bottom-right).

In a near future, it is envisaged to score some video clips of the database, labeled with their induced valence values, to get an estimation of the distribution of the LIRIS-ACCEDE.

IV. Baseline framework

In this section, the experimental protocol for using LIRIS-ACCEDE dataset is defined and a baseline framework using various audio and visual features is presented. This will be made available along with the dataset for later comparisons with other approaches. Multimodal features are extracted from each video clip and grouped into several types. A late fusion scheme is then computed to distinguish low valence and high valence excerpts.

A. Features extraction

Features are extracted using the video and audio streams of input data. Still image features, *i.e.* aesthetics and complexity features, are computed on the key frame of the video clip. The key frame is the frame with the closest RGB histogram to the mean RGB histogram of the whole clip using the Manhattan distance. A comprehensive list of extracted features grouped into three types is given in Table II.

TABLE II. List of extracted features

Type of features	Extracted features
Audio features	Energy, Flatness, Zero-crossing rate, Frequency Centroid, Spectral roll-off, Asymmetry
Complexity	Variance of salient pixels, Number of maximum salient pixels, Standard deviation of local maxima, Spatial edge distribution area, Entropy complexity
Color harmony	Colorfulness, Orientation of the harmonious template

A total of 6 commonly used low-level audio features were determined for each of the audio signals.
The audio energy contains information about the volume of the signal. The spectrum flatness measures the noisiness character of the spectrum. It is defined as being the ratio of the geometric mean and the arithmetic mean of the spectrum. The zero-crossing rate is the rate of sign-changes along the audio signal. The frequency centroid is a measure of the balance of the spectrum and an estimation of the tone. It corresponds to the

first order spectral moment. The spectral roll-off is defined as the frequency below which 85% of the magnitude distribution is concentrated. The last audio feature called asymmetry is a measurement of the symmetry of the spectrum around its mean value (centroid).

Color and more generally aesthetics are important parameters to elicit emotions [24]. A total of 2 color harmony features and 5 complexity features analyzing the color combinations and visual cues were extracted on the key frame of the video stream.
The colorfulness is computed using [25]. The harmony feature is the orientation of the harmonious template which has the lowest harmony energy, defined in [26]. Pavlova *et al.* [27] showed that a strong positive correlation was found between negative emotions and perceived instability. The saliency map provides a representation of the most visually attractive pixels. Therefore, the perceived instability is linked to the variance of most salient pixels in a saliency map from the visual attention model of [28] and to the standard deviation of the euclidean distance between local maxima coordinates and the centroid of local maxima. The number of pixels having high values in the saliency map is also computed.
Finally, the complexity of the key frame is computed using two features: the compactness of the spatial distribution of edges (area of the bounding box that encloses 69.04% of the edge energy [29]) and also the entropy based scene complexity (sum of the entropies of wavelet subbands [30]).

B. Late fusion scheme

The fusion scheme implemented in this work is a late fusion scheme. Late fusion schemes have been shown to be more efficient compared to early fusion methods in several comparative studies such as [31].
The late fusion scheme is implemented has follows. For each type of features (*i.e.* audio, complexity and color harmony features), a Support Vector Machine binary classifier (SVM) is trained. The SVM can map vectors from an input space into a high dimensional feature space and find the best separating hyperplane. This classifier has only two adjustable parameters that do not depend on the dimensionality of feature space and it tends to be less susceptible to the curse of dimensionality [32]. The chosen kernel is the radial basis function (RBF) and a grid search is computed to find the C and γ parameters with the best cross validation accuracy. Each classifier returns the distance from the separating hyperplane. These scores are then normalized and combined using a sum rule to predict if the input video has a low or high valence.

The ground truth of the training set and test set is achieved by binarizing the LIRIS-ACCEDE into two subsets: lowest valence (below the median), and highest valence (above the median). The training set and test set are each composed of 4900 video clips from 80 movies of the database such that the genre of movies in the training set and in the test set have the same distribution. By doing so, clips from the same movie are only in one of the sets, either training set or test set. The experimental protocol to build the training set and the test set will also be shared, allowing future studies to be compared to this work.

C. Results

The accuracy and *f1* measure (1), widely used in information retrieval, are computed to evaluate the performance of emotion classification:

$$f1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (1)$$

Sensitivity, also called recall, relates to the classifier's ability to identify positive results. Precision, or Positive predictive value, is the fraction of video clips classified with positive valence that are relevant.
Accuracies and *f1* measures are shown in Table III with respect to the type of features.

TABLE III. AFFECTIVE VIDEO CLIPS CLASSIFICATION ACCURACIES AND *f1* MEASURES WITH RESPECT TO THE TYPE OF FEATURES

Type of features	Accuracy	*f1* measure
Audio features	0.537	0.519
Complexity	0.536	0.441
Color harmony	0.593	0.583
Late fusion	0.600	0.565

The results are promising given the variety of movies in the database and the difficulty of the task and are better than chance which legitimates the use of low level features to evaluate induced emotion. Indeed it has been shown in previous works that the valence dimension is more difficult to classify than the arousal dimension because some features are more directly linked to arousal (audio energy, motion, shot cut rate). It may be worth mentioning that if the video clips having an absolute combined score below 0.5 (corresponding to 580 video clips) are not taken into account, an accuracy of 0.681 is achieved.
These results also show the consistency of the database to predict valence. Although the order of magnitude of these results is almost the same, it is difficult to legitimately compare our results to other papers since a different database is used for testing in each paper from the state of the art. That is why we hope that the LIRIS-ACCEDE we propose will be a useful database to be used in future studies to compare the efficiency of computational models of emotions.

V. CONCLUSION

In this paper, a large video database for video affective representation frameworks shared under Creative Commons license is proposed. The 9800 video clips that makes up the database are sorted along the valence axis thanks to a user study on CrowdFlower.

Furthermore, content based multimedia features were extracted from the video clips to be used in a computational model of emotions to accurately estimate users' affect in response to video clips. Results showed the efficiency of the combined database and proposed scheme to predict low and high valence. These features as well as the experimental protocol will be provided along with the database to allow researchers to compare fairly their models[1].

In a near future, we plan to sort the database along the arousal axis and we intend to create another user study to score some video clips of the database, labeled with their arousal and valence values, to get an estimation of the distribution of the LIRIS-ACCEDE.

ACKNOWLEDGMENT

This work was supported in part by the French research agency ANR through the VideoSense Project under the Grant 2009 CORD 026 02.

REFERENCES

[1] P. Ekman, "Facial expression and emotion." *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[2] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, p. 344, 2001.

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

[4] 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*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4738, pp. 488–500.

[5] 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.

[6] 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.

[7] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia, March*, 2006.

[8] 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.

[9] 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.

[10] 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.

[11] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective MTV analysis based on arousal and valence features," in *Multimedia and Expo, 2008 IEEE International Conference on*, Apr. 2008, pp. 1369–1372.

[12] S. Zhang, Q. Huang, Q. Tian, S. Jiang, and W. Gao, "Personalized MTV affective analysis using user profile," in *Advances in Multimedia Information Processing - PCM 2008*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, vol. 5353, pp. 327–337.

[13] 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 *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, Dec. 2008, pp. 228–235.

[14] M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, Sep. 2009, pp. 1–7.

[15] 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.

[16] L. Yan, X. Wen, and W. Zheng, "Study on unascertained clustering for video affective recognition," *JICS 2011*, vol. 8, no. 13, pp. 2865–2873, 2011.

[17] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2376–379.

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

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

[20] R. W. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Electronic Imaging'99*, 1998, pp. 290–301.

[21] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, Apr. 2013.

[22] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: A comparative study of self-reporting," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6974, pp. 437–446.

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

[24] P. Valdez and A. Mehrabian, "Effects of color on emotions." *Journal of experimental psychology General*, vol. 123, no. 4, pp. 394–409, 1994.

[25] D. Hasler and S. Susstrunk, "Measuring colourfulness in natural images," in *Proc. IS&T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, 2003, pp. 87–95.

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

[27] M. Pavlova, A. A. Sokolov, and A. Sokolov, "Perceived dynamics of static images enables emotional attribution," *Perception*, vol. 34, no. 9, pp. 1107–1116, 2005.

[28] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[29] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," vol. 1. IEEE, 2006, pp. 419–426.

[30] O. Le Meur, T. Baccino, and A. Roumy, "Prediction of the inter-observer visual congruency (IOVC) and application to image ranking," in *ACM Multimedia*, Phoneix, tats-Unis, 2011.

[31] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05, 2005, pp. 399–402.

[32] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.