Deep Learning vs. Kernel Methods: Performance for Emotion Prediction in Videos

Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, Liming Chen

Deep Learning *vs.* Kernel Methods: Performance for Emotion Prediction in Videos

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

*Technicolor
975, avenue des Champs Blancs
35576 Cesson Sévigné, France
{yoann.baveye, christel.chamaret}@technicolor.com

†Université de Lyon, CNRS
Ecole Centrale de Lyon
LIRIS, UMR5205, F-69134, France
{emmanuel.dellandrea, liming.chen}@ec-lyon.fr

***Abstract*—Recently, mainly due to the advances of deep learning, the performances in scene and object recognition have been progressing intensively. On the other hand, more subjective recognition tasks, such as emotion prediction, stagnate at moderate levels. In such context, is it possible to make affective computational models benefit from the breakthroughs in deep learning? This paper proposes to introduce the strength of deep learning in the context of emotion prediction in videos. The two main contributions are as follow: (i) a new dataset, composed of 30 movies under Creative Commons licenses, continuously annotated along the induced valence and arousal axes (publicly available) is introduced, for which (ii) the performance of the Convolutional Neural Networks (CNN) through supervised fine-tuning, the Support Vector Machines for Regression (SVR) and the combination of both (Transfer Learning) are computed and discussed. To the best of our knowledge, it is the first approach in the literature using CNNs to predict dimensional affective scores from videos. The experimental results show that the limited size of the dataset prevents the learning or finetuning of CNN-based frameworks but that transfer learning is a promising solution to improve the performance of affective movie content analysis frameworks as long as very large datasets annotated along affective dimensions are not available.**

***Keywords*—continuous emotion prediction; deep learning; benchmarking; affective computing**

I. Introduction

In the last few years, breakthroughs in the development of convolutional neural networks have led to impressive state of the art improvements in image categorization and object detection. These breakthroughs are a consequence of the convergence of more powerful hardware, larger datasets, but also new network designs, and enhanced algorithms [1], [2]. Is it possible to benefit from these progresses for the affective movie content analysis? Large and publicly available datasets composed of movies annotated along affective dimensions start to emerge [3] and, even if they are far from being as large as datasets such as ImageNet [4], tools exist to benefit from the Convolutional Neural Networks (CNN) frameworks, composed of tens of millions parameters, trained on huge datasets [1].

In this work, we aim not to maximize absolute performance, but rather to study and compare the performance of four state of the art architectures for the prediction of affective dimensions. It contributes to the affective movie content analysis field as follows:

- Benchmark of four state of the art architectures for the prediction of dimensional affective scores: fine-tuned CNN, CNN learned from scratch, SVR and transfer learning. To the best of our knowledge, it is the first approach in the literature using CNNs to predict dimensional affective scores from videos;
- Public release of a large dataset composed of 30 movies under Creative Commons licenses that have been continuously annotated along the induced valence and arousal dimensions.

The paper is organized as follows. Section II provides background material on continuous movie content analysis work, as well as CNNs and Kernel Methods. In Section III, the process for annotating the new dataset is described. The computational models investigated in this work are presented in Section IV. Their performance is studied and discussed in Section V, while the paper ends in Section VI with conclusions.

II. Background

A. Dimensional Affective Movie Content Analysis

Past research in affective movie content analysis from audiovisual clues extracted from the movies has focused on the prediction of emotions represented by of a small number of discrete classes which may not reflect the complexity of the emotions induced by movies. However, more and more work describes emotions using a more subtle and dimensional representation: the valence-arousal space.

Interestingly, Hanjalic and Xu who pioneered the affective movie content analysis mapped video features onto the valence-arousal space to create continuous representations [5]. They directly mapped video features onto the valence-arousal space to create continuous representations. More recently, Zhang *et al.* proposed a personalized affective analysis for music videos [6]. Their model is composed of SVR-based arousal and valence models using both multimedia features and user profiles. Malandrakis *et al.* trained two Hidden Markov Models (HMMs) fed with audiovisual features extracted on

the video frames to model simultaneously 7 discrete levels of arousal and valence [7]. Then, these discrete outputs were interpolated into a continuous-valued curve.

In contrast with previous work using handcrafted features, we focus in this work on CNNs to predict affective scores from videos. As a point of comparison for the evaluation of the CNN-based frameworks, we also focus on kernel methods and especially the SVR, commonly used in the state of the art, learned with features from previous work.

B. Convolutional Neural Networks and Kernel Methods

SVM for regression [8], also known as SVR, is one of the most prevalent kernel methods in machine learning. The model learns a non-linear function by mapping the data into a high-dimensional feature space, induced by the selected kernel. Since the formulation of SVM is a convex optimization problem, it guarantees that the optimal solution is found. SVRs have been extensively used in the affective computing field for music emotion recognition [9], as well as spontaneous emotion recognition in videos [10], and affective video content analysis [11]–[13].

Beginning with LeNet-5 [14], CNNs have followed a classic structure. Indeed, they are composed of stacked convolutional layers followed by one or more fully-connected layers. So far, best results on the ImageNet classification challenge have been achieved using CNN-based models [1], [2]. CNNs have been mostly used in the affective computing field for facial expression recognition [15]. Recently, Kahou *et al.* trained a CNN to recognize facial expressions in video frames [16]. Its prediction was then combined with the predictions from three other modality-specific models to finally predict the emotional category induced by short video clips.

The CNN approach disrupts the field of machine learning and has significantly raised the interest of the research community for deep learning frameworks. Generally applied for object recognition, its use will be naturally extended to any recognition task. The contributions using CNNs in the affective computing field will likely show up in the coming months.

III. Continuous Movie Annotation

The LIRIS-ACCEDE dataset proposes 9,800 excerpts extracted from 160 movies [3]. However, these 9,800 excerpts have been annotated independently, limiting their use for learning models for longer movies where previous scenes may reasonably influence the emotion inference of future ones. Thus, as a first contribution, we set up a new experiment where annotations are collected on long movies, making possible the learning of more psychologically relevant computational models.

A. Movie Selection

The aim of this new experiment is to collect continuous annotations on whole movies. To select the movies to be annotated, we simply looked at the movies included in the LIRIS-ACCEDE dataset[1] since they all share the desirable property to be shared under Creative Commons licenses and can thus be freely used and distributed without copyright issues as long as the original creator is credited. The total length of the selected movies was the only constraint. It had to be smaller than eight hours to create an experiment of acceptable duration.

The selection process ended with the choice of 30 movies so that their genre, content, language and duration are diverse enough to be representative of the original LIRIS-ACCEDE dataset. The selected videos are between 117 and 4,566 seconds long ($mean = 884.2sec \pm 766.7sec\ SD$). The total length of the 30 selected movies is 7 hours, 22 minutes and 5 seconds. The list of the 30 movies included in this experiment is detailed in Table I.

B. Experimental Design

The annotation process aims at continuously collecting the self-assessments of arousal and valence that viewers feel while watching the movies.

1) Annotation tool: To collect continuous annotations, we have used a modified version of the GTrace program originally developed by Cowie *et al.* [17]. GTrace has been specifically created to collect annotations of emotional attributes over time. However, we considered that the design of the original GTrace interface during the annotation process is not optimal: the video to be rated is small, the annotation scale is far from it, and other elements may disrupt the annotator's task. That is why we modified the interface of GTrace in order to be less disruptive and distract annotators' attention from the movie as less as possible.

First, we redesigned the user-interface so that the layout is more intuitive for the annotator. During the annotation process, the software is now in full screen and its background is black. The video is bigger, thus more visible, and the rating scale is placed below the video (Figure 1(b)).

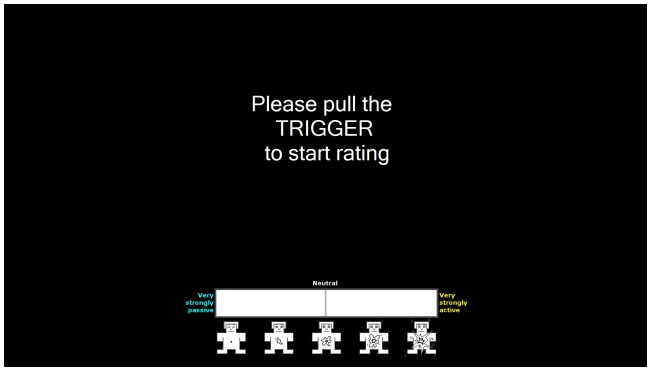
Second, we used the possibility offered by GTrace to create new scales. We designed new rating scales for both arousal (Figure 1(a)) and valence (Figure 1(b)). Under both scale, the corresponding Self-Assessment Manikin scale is displayed [18]. It is an efficient pictorial system which helps understand the affective meaning of the scale.

Third, instead of using a mouse, the annotator used a joystick to move the cursor which is much more intuitive. To link the joystick to GTrace, we used a software that simulates the movement of the mouse cursor when the joystick is used.

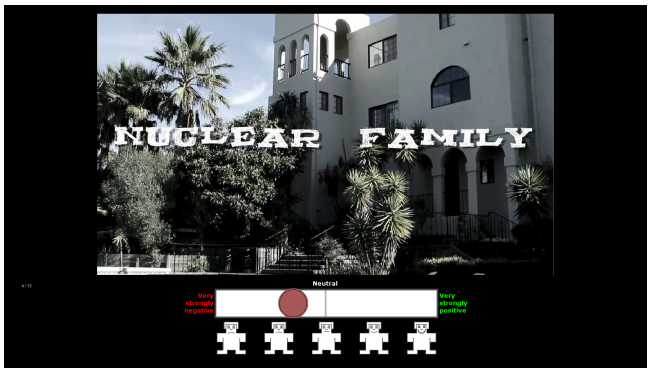
2) Protocol: In the experimental protocol described below, each movie is watched by an annotator only once. Indeed, the novelty criteria that influences the appraisal process for an emotional experience should be taken into consideration [19].

Annotations were collected from ten French paid participants (seven female and three male) ranging in age from 18 to 27 years ($mean = 21.9 \pm 2.5\ \ SD$). Participants

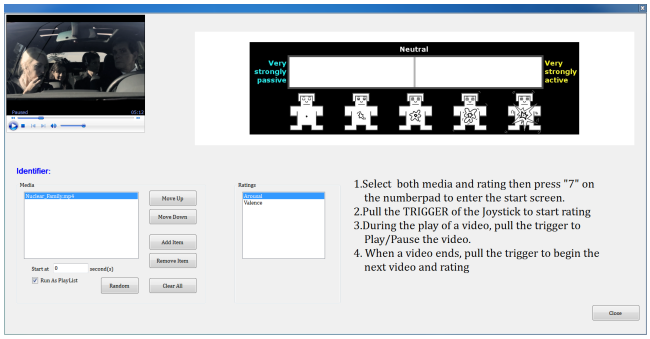
[1]An exhaustive list of the movies included in the LIRIS-ACCEDE dataset as well as their credits and license information is available at: http://liris-accede.ec-lyon.fr/database.php



(a) Screenshot before the annotation along the arousal axis



(b) Screenshot during the annotation along the valence axis



(c) Modified GTrace menu

Fig. 1. Screenshots of the modified GTrace annotation tool. Nuclear Family is shared under a Creative Commons Attribution-NonCommercial 3.0 Unported United States License at http://dominicmercurio.com/nuclearfamily/.

had different educational backgrounds, from undergraduate students to recently graduated master students. The experiment was divided into four sessions, each took place on a different half-day. The movies were organized into four sets (Table I). Before the first session, participants were informed about the purpose of the experiment and had to sign a consent form and fill a questionnaire. Participants were trained to use the interface thanks to three short videos they had to annotate before starting the annotation of the whole first session. The participants were also introduced to the meaning of the valence and arousal scales.

Participants were asked to annotate the movies included in the first two sessions along the induced valence axis and

TABLE I
LIST OF THE 30 MOVIES ON WHICH CONTINUOUS ANNOTATIONS HAVE BEEN COLLECTED

Sets	Duration	Movies
A	01:50:14	Damaged Kung Fu, Tears of Steel, Big Buck Bunny, Riding The Rails, Norm, You Again, On time, Chatter, Cloudland & After The Rain
B	01:50:03	Barely Legal Stories, Spaceman, Sintel, Between Viewings, Nuclear Family, Islands, The Room of Franz Kafka & Parafundit
C	01:50:36	Full Service, Attitude Matters, Elephant's Dream, First Bite, Lesson Learned, The Secret Number & Superhero
D	01:51:12	Payload, Decay, Origami, Wanted & To Claire From Sonny

the movies in the last two sessions along the induced arousal axis. This process ensures that each movie is watched by an annotator only once. The order of the sets with respect to the four sessions was different for all the annotators. For example, the first participant annotated the movies from sets A and B along the induced valence axis and the movies from sets C and D along the induced arousal axis whereas the second participant annotated the movies from sets B and C along the induced valence axis and the movies from sets D and A along the induced arousal axis. Furthermore, the videos inside each session were played randomly. After watching a movie, the participant had to manually pull the trigger of the joystick in order to play the next movie.

Finally, each movie is annotated by five annotators along the induced valence axis and five other annotators along the induced arousal axis.

C. Post-processing

Defining a reliable ground truth from continuous self-assessments from various annotators is a critical aspect since the ground truth is used to train and evaluate emotion prediction systems. Two aspects are particularly important: there are annotator-specific delays amongst the annotations and the aggregation of the multiple annotators' self-assessments must take into account the variability of the annotations [20].

Several techniques have been investigated in the literature to deal with the synchronisation of various individual ratings. In this work, we combine and adapt the approaches proposed by Mariooryad and Busso [21] and by Nicolaou *et al.* [22] to deal with both the annotation delays and variability.

First, the self-assessments recorded at a rate of 100 values per second are down-sampled by averaging the annotations over windows of 10 seconds with 1 second overlap (*i.e.* 1 value per second). This process removes most of the noise mostly due to unintended moves of the joystick. Furthermore, due to the granularity of emotions, one value per second is enough for representing the emotions induced by movies [20], [23].

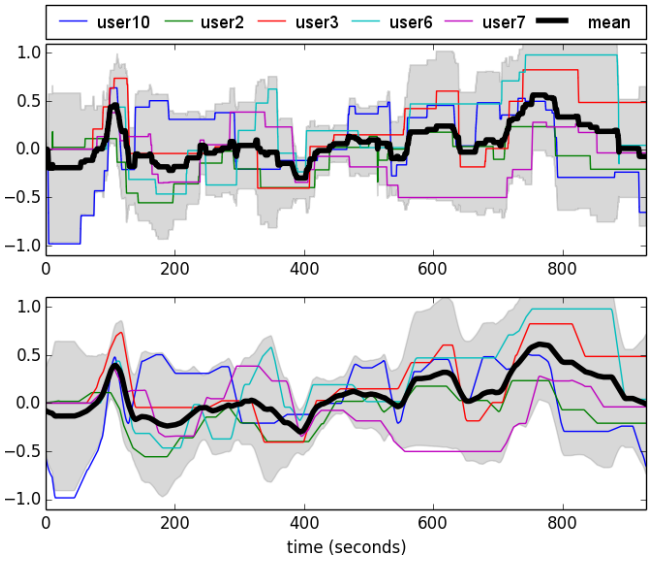
Then, each self-assessment is shifted so that the τ-sec-shifted annotations maximizes the inter-rater agreement between the τ-sec-shifted self-assessment and the non-shifted

self-assessments from the other raters. The inter-rater agreement is measured using the Randolph's multirater kappa free [24]. Similarly to Marioryad and Busso [21], the investigated delay values τ range from 0 to 10 sec. However, in practice, τ ranged from 0 to 6 sec and the largest values (5 or 6 sec) were rarely encountered ($mean = 1.47 \pm 1.53\ SD$). As suggested by Landis and Koch [25], the average Randolph's multirater kappa free shows a moderate agreement for the shifted arousal self-assessments ($\kappa = 0.511 \pm 0.082\ SD$), as well as for the shifted valence self-assessments ($\kappa = 0.515 \pm 0.086\ SD$).

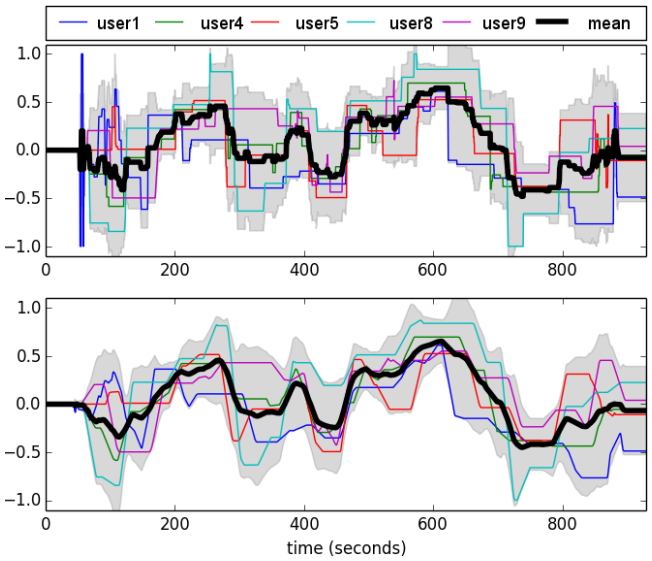
Finally, to aggregate the different ratings we use an approach similar to the one proposed in [22]. The inter-coder correlation is used to obtain a measure of how similar are one rater's self-assessments to the annotations from the other participants. The inter-coder correlation is defined as the mean of the Spearman's Rank Correlation Coefficients (SRCC) between the annotations from the coder and each of the annotations from the rest of the coders. The SRCC has been preferred over other correlation measures since it is defined as the Pearson correlation coefficient between the ranked variables: the SRCC is computed on relative variables and thus ignores the scale interpretation from the annotators. The inter-coder correlation is used as a weight when combining the multiple annotators' annotations. The inter-coder correlation is higher in average for valence ($mean = 0.313 \pm 0.195\ SD$) than for arousal ($mean = 0.275 \pm 0.195\ SD$).

Figure 2 shows the raw ratings and post-processed ones for both induced arousal and valence scales for the movie Spaceman. The bold curves are the weighted average of the continuous annotations computed on the raw ratings or on the smoothed and shifted ones.

To conclude, this post-processing assigns each 1-second segment of a movie two values: one represents the induced arousal and the other the induced valence. Both values are rescaled so that they range from 0 to 1. More precisely, 26,525 1-second segments are extracted from the 30 movies. The full length movies, raw self-assessments as well as post-processed ones are publicly available at: http://liris-accede.ec-lyon.fr/.



(a) Raw and post-processed annotations for arousal



(b) Raw and post-processed annotations for valence

Fig. 2. Annotations collected for the movie "Spaceman". Both subfigures show at the top the raw annotations and at the bottom post-processed annotations for (a) arousal and (b) valence. The shaded area represents the 95% confidence interval of the mean.

IV. Regression Frameworks for Emotion Prediction

In this section, we describe the four frameworks that are compared in Section V. All the models presented in this section output a single value: the predicted valence or arousal score. Thus, they all need to be learned twice: either for predicting induced arousal scores, or for predicting induced valence scores.

A. Deep Learning

Two models using CNNs to directly output affective scores are investigated in this work. Both take as input the key frame of the video segment for which an arousal or valence score is predicted. The key frame is defined as the frame with the closest RGB histogram to the mean RGB histogram of the whole excerpt using the Manhattan distance.

We used data augmentation to enlarge artificially the training set. As in [1], the model was trained using random 224×224 patches (and their horizontal reflections) extracted from the 256×256 input images. These input images were the center crop of the key frames extracted from the video segments in the training set and resized so that the original aspect ratio is preserved but their smallest dimension equals 256 pixels. The training is stopped when the Mean Square Error (MSE), measured every 500 iterations using a validation set, increases for 5 consecutive measurements. At validation and test time, the network makes a prediction by extracting the 224×224 center patch.

1) Fine-tuning: This first framework is based on the fine-tuning strategy. The concept of fine-tuning is to use a model pretrained on a large dataset, replace its last layers by new layers dedicated to the new task, and fine-tune the weights of the pretrained network by continuing the backpropagation. The main motivation is that the most generic features of a CNN are contained in the earlier layers and should be useful for solving many different tasks. However, later layers of a CNN become more and more specific to the task for which the CNN has been originally trained.

In this work, we fine-tune the model proposed in [1] composed of five stacked convolutional layers (some are followed by local response normalization and max-pooling), followed by three fully-connected layers. To adapt this model to our task, the last layer is replaced by a fully-connected layer composed of a unique neuron scaled by a sigmoid to produce the prediction score. The loss function associated to the output of the model is the Euclidean loss. Thus, the model minimizes the sum of squares of differences between the ground truth and the predicted score across training examples. All the layers of the pretrained model are fine-tuned, but the learning rate associated to the original layers are ten times smaller than the one associated with the new last neuron. Indeed, we want the pretrained layers to change very slowly, but let learn faster the new layer which is initialized from a zero-mean Gaussian distribution with standard deviation 0.01. This is because the pretrained weights should be already relatively meaningful, and thus should not be distorted too much.

We trained the new fine-tuned models using the reference implementation provided by Caffe [26] using stochastic gradient descent with a batch size of 256 examples, momentum of 0.9, base learning rate of 0.0001 and weight decay of 0.0005.

2) Learning From Scratch: We also built and learned from scratch a CNN based on the architecture of [1] but much simpler since our training set is composed of 16,065 examples. The model is composed of two convolutional layers and three fully-connected layers. As in [1], the first convolutional layer filters the $224 \times 224 \times 3$ input key-frame with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels. The second convolutional layer, connected to the first one, uses 256 kernels of size $5 \times 5 \times 96$. The outputs of both convolutional layers are response-normalized and pooled. The first two fully-connected layers are each composed of 512 neurons and the last fully-connected layer is the same as the last one added to the fine-tuned model in the previous section. The ReLU non-linearity is applied to the output of all the layers. All the weights are initialized from a zero-mean Gaussian distribution with standard deviation 0.01. The learning parameters are also the same as those used in the previous section.

B. SVR

This model is similar to the baseline framework presented in [3]: two independent ϵ-SVRs are learned to predict arousal and valence scores 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. The SVR is fed with the features detailed in [3], *i.e.*, audio, color, aesthetic, and video features. The features include, but are not limited to, audio zero-crossing rate, audio flatness, colorfulness, hue count, harmonization energy, median lightness, depth of field, compositional balance, number and length of scene cuts per frame, and global motion activity. All features are normalized using the standard score.

C. Transfer Learning: CNN as a feature extractor

The SVR is the same as in the previous section except that the 4,096 activations of the second fully-connected layer called "FC7" of the original model learned in [1] are normalized using the standard score and used as features to feed the SVR in addition to the features used in the previous section. Thus, the CNN is treated as a feature extractor and is used to, hopefully, improve the performance of the SVR.

V. Performance Analysis

In this section, the performance of the four well-known state of the art architectures introduced in Section IV is compared and discussed.

A. The Importance of Correlation

The common measure generally used to evaluate regression models is the Mean Square Error (MSE). However, the performance of the models cannot be analyzed using simply this measure. As a point of comparison, on the test set, the MSE between the ground truth (ranging from 0 to 1) for valence and random values generated between 0 and 1 equals 0.113, whereas the linear correlation (Pearson correlation coefficient) is close to zero. However, the ground truth is biased in the sense that a large portion of the data is neutral (*i.e.* its valence score is close to 0.5) or is distributed around the neutral score. This bias can be seen from Figure 2. Thus, if we create a uniform model that always outputs 0.5, its performance will be much better: its MSE is 0.029. However, the correlation between the predicted values and the ground truth will be also close to zero. The performance for the random and uniform baselines are indicated in Table II. For the random distribution, we generate 100 distributions and report the average MSE and correlation.

To analyze the results and the performance of the computational models, the linear correlation has the advantages not to be affected by the range of the scores to be predicted and to measure the relationship between the predicted values and the ground truth.

B. Experimental Results

To learn and evaluate the various frameworks, the dataset presented in Section III and composed of 26,525 1-second segments extracted from 30 movies is distributed into a training set, a validation set and a test set. Approximately 60% of the data is assigned to the training set and 20% of the data is assigned to both the validation and test sets. More precisely, 16,065 1-second segments extracted from 15 movies are assigned to the training set, 5,310 segments from 8 movies

TABLE II
PREDICTION RESULTS FOR VALENCE AND AROUSAL DIMENSIONS (MSE: MEAN SQUARE ERROR, R: PEARSON CORRELATION COEFFICIENT)

System	Arousal		Valence	
	MSE	r	MSE	r
Random	0.109	0.0004	0.113	-0.002
Uniform	0.026	-0.016	0.029	-0.005
CNN – Fine-tuned	**0.021**	0.152	**0.027**	0.197
CNN – From scratch	0.023	0.157	0.031	0.162
SVR – Standard	0.023	0.287	0.035	0.125
SVR – Transfer learning	0.022	**0.337**	0.034	**0.296**

to the validation set and finally, 5,150 segments from 7 movies to the test set. This distribution makes also sure that the genre of the movies in each set is as diverse as possible.

Table II presents the results of using CNNs (fine-tuned and learned from scratch), SVR and transfer learning for the prediction of valence and arousal dimensions based on the MSE and the Pearson's r correlation coefficient. For the four frameworks, the predicted scores as well as the ground truth for valence and arousal range from 0 to 1. Table II shows that for valence and arousal, the highest correlation is obtained by the transfer learning approach. Once again, this result reveals that CNNs provide generic mid-level image representations that can be transferred to new tasks, including the transfer from the classification of 1,000 ImageNet classes to the prediction of the valence and arousal affective scores. Transfer learning improves by 50% the performance in terms of correlation of the second best performing framework for predicting valence, and by 17% for arousal. However, no clear gain is obtained for MSE. For valence, the MSE is even higher than the MSE of the uniform strategy.

The fine-tuned CNN outperforms the other models in terms of MSE for both valence and arousal. The gain in terms of MSE is more important for valence. For arousal, the MSE value is close to the performance obtained by the transfer learning strategy. However, for both arousal and valence, the correlation is much lower than the performance obtained with transfer learning. Nevertheless, it is a promising result given that the performance of this model on the training set indicates that, despite the use of a validation set to stop the learning phase if the performance on the validation set increased for 5 consecutive measurements, the size of the dataset is not big enough to prevent overfitting. Indeed, previous work has shown that overfitting and training set size are closely related [27]. For example, the performance of the fine-tuned model on the training set for the prediction of valence is much better ($MSE = 0.012$, $r = 0.79$). It may also explain why the performance of the CNN learned from scratch is lower than the performance of the fine-tuned CNN.

Regarding the arousal dimension, it is interesting to note that the correlation of the SVR is almost twice the correlation of the pure deep learning frameworks. This could be explained by the fact that both deep-learning models lack audio and motion information, unlike the SVR framework which uses features extracted from the audio signal and from statistics for consecutive frames of a video segment. However, Nicolaou *et al.*, among others, showed that the prediction of arousal is greatly enhanced by the use of audio and motion cues [22]. Thus, we plan to investigate the use of audio cues to produce more accurate affective predictions for videos and to take into account more than one frame to predict the induced affective score of a 1-second length video segment.

VI. CONCLUSION

This work presents the performance of Convolutional Neural Networks for affective movie content analysis and introduces a new dataset composed of 30 movies continuously annotated along the induced valence and arousal axes split into 25,525 1-second length video segments. This new dataset is publicly available at: http://liris-accede.ec-lyon.fr/, and is complementary to the original LIRIS-ACCEDE dataset. We have found that the fine-tuned CNN framework is a promising solution for emotion prediction. However, the limited size of the training set (16,065 samples) prevents the pure CNN-based frameworks to obtain good performances in terms of correlation. Nevertheless, intermediate layers, originally trained to perform image recognition tasks, are generic enough to provide mid-level image representations that can greatly improve the prediction of affective scores in videos. As long as very large datasets annotated along affective dimensions are not available, transfer learning is a convenient trade-off to improve the performance of affective movie content analysis frameworks.

In future work, we plan to treat the prediction of valence and arousal as a 2D regression problem to take into account the correlation of valence and arousal. We also plan to investigate the use of audio cues that are known to be important to model the arousal in particular. Finally, all the frameworks investigated in this work are static frameworks that do not model the dynamic of the videos. We hope that using this temporal information may help to produce more accurate affective predictions for videos.

ACKNOWLEDGMENT

This work was supported in part by the French research agency ANR through the Visen project within the ERA-NET CHIST-ERA framework under the grant ANR-12-CHRI-0002-04. We further would like to thank Xingxian Li for his help on the modification of the GTrace program.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[3] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, Jan 2015.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 248–255.

[5] 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.

[6] 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.

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

[8] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik *et al.*, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[9] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, pp. 1664–1078, 2013.

[10] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," in *16th European Signal Processing Conference, Lausanne, Switzerland*, 2008.

[11] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *16th IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 1853–1856.

[12] 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.

[13] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[15] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5, pp. 555–559, 2003.

[16] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13, 2013, pp. 543–550.

[17] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept 2013, pp. 709–710.

[18] 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.

[19] K. R. Scherer, "Appraisal considered as a process of multi-level sequential checking," in *Appraisal processes in emotion: Theory, Methods, Research*, K. R. Scherer, A. Schorr, and T. Johnstone, Eds. Oxford University Press, 2001, pp. 92–120.

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

[21] S. Marioryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, Sept 2013, pp. 85–90.

[22] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[23] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 2013, pp. 1–6.

[24] 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.

[25] 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.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.