A SUPERVISED APPROACH TO MOVIE EMOTION TRACKING

N. Malandrakis[1]*, A. Potamianos*[1]*, G. Evangelopoulos*[2]*, A. Zlatintsi*[2]

[1] Dept. of ECE, Technical University of Crete, 73100 Chania, Greece
[2] School of ECE, National Technical University of Athens, 15773 Athens, Greece
[nmalandrakis,potam]@telecom.tuc.gr [gevag,nzlat]@cs.ntua.gr

ABSTRACT

In this paper, we present experiments on continuous time, continuous scale affective movie content recognition (emotion tracking). A major obstacle for emotion research has been the lack of appropriately annotated databases, limiting the potential for supervised algorithms. To that end we develop and present a database of movie affect, annotated in continuous time, on a continuous valence-arousal scale. Supervised learning methods are proposed to model the continuous affective response using hidden Markov Models (independent) in each dimension. These models classify each video frame into one of seven discrete categories (in each dimension); the discrete-valued curves are then converted to continuous values via spline interpolation. A variety of audio-visual features are investigated and an optimal feature set is selected. The potential of the method is experimentally verified on twelve 30-minute movie clips with good precision at a macroscopic level.

Index Terms— Emotion recognition, Multimedia databases, Machine learning, Psychological emotion dimensions

1. INTRODUCTION

Emotion recognition has been a very active field in the past years, since emotional information is highly valuable in applications ranging from human-computer interaction to automated content delivery. Emotion is of particular interest to content delivery systems that provide personalized multimedia content, automatically extract highlights and create automatic summaries or skims. The motivation behind using such technology is simple; humans pick content (movies, music) based on its affective characteristics, therefore a system designed to deliver it should have access to such data. Furthermore, systems aimed at highlight extraction/summarization require detailed representations of emotion in a scalable domain, as well as information of the temporal dynamics of emotion. The process of extracting such information is usually referred to as *emotion tracking* and it is, ideally, a continuous-time continuous-scale representation of the affective content of a movie. A suitable continuous-scale representation is the dimensional representation of valence-arousal. This two-dimensional representation is becoming increasingly popular due to its flexibility and high descriptive power, but also because the representation of emotion in a Euclidean space allows for simpler general-purpose analysis and recognition algorithms. In addition to the two-dimensional valence-arousal model, the three-dimensional valence-arousal-dominance model (or valence-arousal-tension for music) is also popular. In the field of affective multimedia content analysis it has been shown that the two-dimensional model is adequate to represent the range of emotions experienced by viewers/listeners [1]. Adding time as a third dimension, the affective content is represented as two continuous signals, the combination of which can yield an emotional state at any point within a multimedia stream.

There has been very little prior work towards emotion tracking in movies [2], with most researchers instead focusing on the more typical target of classifying large movie segments to a small number of distinct categories [3]. In all cases research has focused in narrow domains, such as specific movie genres [4]. To our knowledge, there has never been an attempt to apply supervised learning techniques to continuous time emotion tracking in movies. A variety of models have been used to classify affective content, including Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) and Neural Networks (NNs). The features used are inspired by the ones used to characterize the modalities that make up a movie; timbre and rhythm to characterize music [5], color and motion to characterize video [6], energy, short-time spectral envelope and prosodic features to characterize speech [7].

One of the most important obstacles facing research in movie emotion and more particularly emotion tracking is the lack of movie databases annotated in an appropriate fashion, which probably explains the limited use of supervised techniques. As such, one of our targets was the creation of such a database, containing emotional responses annotated as continuous curves. Section 2 describes the creation of such a movie database of affect. In Section 3, we implement supervised learning techniques to train a classifier based on HMMs in order to perform emotion tracking, using a variety of audio-visual features. Experimental results are presented in Section 4 and conclusions in Section 5.

2. DATABASE

Before describing the database, it is important to distinguish between three different "types" of movie emotion; *intended*, *expected* and *experienced* emotion. Intended emotion describes the emotional response that the movie attempts to evoke in its viewers, experienced emotion describes the emotion a user actually feels when watching the movie, while expected emotion is the expected value of experienced emotion in a population. Some prior research has assumed that intended and expected emotion match [2], however it is easy to see that a movie can be unsuccessful in conveying the desired effect. In fact the degree of effectiveness with which a movie creates the desired emotion in the viewer is a basic criterion humans use to assess movie quality. Our system attempts to predict intended emotion, however expected emotion is also desirable, since it can potentially be used as a basis for personalized predictions of experienced emotion [8]. This distinction is important for movie selection and annotating procedure definition.

2.1. The data

This emotional database was created as part of a larger project aiming at annotating movie data with affective, sensory and semantic cues. This is a joint project developed by the Technical University of Crete and the National Technical University of Athens, designed to be used by movie summarization systems such as that described in [9]. The database consists of contiguous thirty-minute video clips from twelve movies, featuring their visual, aural and textual data (subtitles). The movies selected are the ten winners of the Academy Award for best picture for the years 1998-2007 and two award winning animation films, namely; "Shakespeare in Love", "American Beauty", "Gladiator", "A Beautiful Mind", "Chicago", "The Lord of the Rings: The Return of the King", "Million Dollar Baby", "Crash", "The Departed", "No Country for Old Men", "Ratatouille" and "Finding Nemo". Using the Academy Award winners list is one way of ensuring the high quality of the movies by a well-acknowledged criterion. One expected effect of this perceived quality is the higher correlation between intended and expected emotion; a high quality movie is expected to be successful in creating the desired emotional experience.

2.2. Annotating Procedure

Annotation was performed on two levels; intended emotion was annotated by experts, while volunteers annotated their individual experienced emotion, from which we derive the expected emotion. The annotations were performed using the FEELTRACE [10] emotion annotation tool. The participants track the annotated emotional response by moving the mouse pointer on a square two-dimensional area representing the valence-arousal emotional space in real-time as they were watching the movie. So far seven volunteers, 20-30 years old, two female and five male have performed the annotation of experienced emotion. All annotators evaluated all clips, with five (out of seven) performing the entire process twice for intra-annotator agreement validation. Furthermore, annotators were presented with their results (curves) and their interpretation in textual terms in order to validate them and filled a questionnaire containing questions regarding their prior knowledge of the movies, their opinion of the movies and clips in regards to informativeness and enjoyability, their own annotating performance and their own perception of some suspected phenomena. Expected emotion is derived from the individual experienced emotion annotations using a correlation-based rejection scheme similar to that in [11] with particularly uncorrelated annotations being rejected as outliers. Validation of the database was done via analyzing the disagreement between users as well as between the users and the intended emotion against the factors suspected of leading to such disagreement from their answers to our questionnaires.

2.3. Annotation Results

The result of each annotation is a pair of curves, one curve for arousal and one for valence. These curves have values in the range $[-1, 1]$ for each dimension and are down-sampled to match the video rate of 25 fps. Overall, including duplicates, 144 annotations of the experienced emotion and 36 annotations of intended emotion were produced, from which twelve annotations of expected emotion and twelve annotations of intended emotion, one of each per movie clip, were created. Fig 1 shows two-dimensional histograms of our annotations for intended and expected emotion. The "V" shape is very similar to that shown in [1] and [2] regarding the response to emotional media, which is reasonable given the similar context. Fig 2 shows some sample frames taken from the extremes of the two emotional dimensions. Table 1 shows agreement statistics in the annotations of experienced emotion. The low agreement is expected, since the participants annotate their own, very subjective, affective response. It is worth comparing these statistics between the two dimensions; distance metrics score higher for valence, while correlation is higher for arousal. That means that agreement in rough terms ("positive", "exciting") is higher for valence than arousal, yet perception of the dynamics ("more", "less") is more uniform for arousal. Factors expressing the viewer's opinion alter agreement as expected; for example, users that like a particular movie agree more with each other and with the intended emotion. Expected and intended emotion end up being highly similar, with correlation coefficients of 0.74 for arousal and 0.70 for valence. Before using for classification, the expected and intended emotion curves are quantized into seven equiprobable bins, using the cumulative distribution function estimated via Parzen windows.

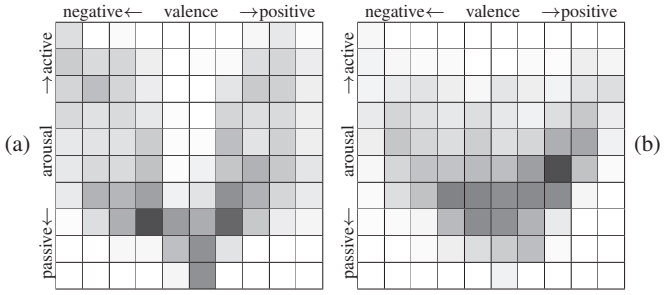

Fig. 1. Joint valence-arousal histograms for (a) intended and (b) expected emotion (darker signifies higher value).

Table 1. Inter-annotator agreement.

metric	valence	arousal
correlation	0.293	0.409
difference of means	0.288	0.411
mean abs. difference	0.445	0.513
Krippendoff's α ordinal (7 levels)	0.308	0.152
Cohen's k (7 levels)	0.035	0.029

Fig. 2. Sample frames for: (a) Low arousal, (b) High arousal, (c) Very negative valence, (d) Very positive valence.

3. SYSTEM DESIGN

Emotion is a dynamic process that evolves rapidly through time. In order to capture the dynamic nature of emotion, we choose to use hidden Markov models that are popular in time series modeling and

have been shown to work to model emotion [3]. The next important modeling issue is how to handle the two affective dimensions. As shown in Fig 2, arousal and valence are correlated. A way to exploit this relation would be either to model arousal and valence jointly, e.g., using 2-D HMMs, or to use a series of classifiers, e.g., the output of the arousal classifier being (one of) the input(s) of the valence classifier. In this paper, we choose to use independent classifiers, one for each dimension, which are also evaluated separately.

HMMs using various numbers of states and Gaussian components were evaluated. We found that increasing the number of states is more beneficial than increasing the number of Gaussian components, particularly when using short-time spectral envelope audio features, e.g. Mel Frequency Cepstral Coefficients (MFCCs), presumably because longer models better capture complex temporal interactions between low level features and emotion. Results are presented next for recognizers that model each affective category with a left-to-right HMM with 32 hidden states and a single Gaussian distribution per state. Inter-category transitions are modeled with a bigram language model that only allows transitions between adjacent categories. Humans don't change affective levels very fast and the language model probabilities are assigned a large exponential weight (40) compared to the acoustic-visual features (1). This weighting results also in smoother curves. The models are trained using the *Baum-Welch* algorithm and classification is achieved via the *Viterbi* algorithm (using the HTK speech recognition package).

A variety of features have been investigated broadly separated into three categories/modalities: audio, music and video features. The low level audio features tested were: fundamental frequency (F0), intensity, log energy, signal zero crossings rate, spectral centroid, spectral flux, spectral roll-off, line spectral pairs, chroma coefficients, MFCCs and Perceptual Linear Prediction (PLP) coefficients. Audio features were extracted via OpenSMILE [12] using a 200ms window and 40ms update. We also created a more extensive feature set by extracting the aforementioned low level features using a 40ms window, 10ms update, then calculating the statistics of these samples (moments, derivatives, extrema) within a 200ms window (and using the statistics as features). High level music features were extracted using the MIR Toolbox [13], namely: tempo, pulse clarity, event density, spectral flatness, rhythm irregularity and inharmonicity. These features must be computed using a larger window in order to be meaningful, so we used a window of 1sec, updated every 40ms. The video features used were the statistics of color, intensity and motion, extracted, per video frame (40ms), through the algorithms described in [14]. All features were evaluated using three models of increasing complexity (states, Gaussian components). The selected feature set was created by hierarchically merging the best performing features. The rejected features did not necessarily perform inadequately, some were simply highly correlated with "more successful" features and therefore provided no additional benefit. Energy and all energy-related features (e.g., 0th order MFCC) performed very well, as expected, for detecting arousal and for separating neutral from non-neutral valence (but were not able to distinguish between positive and negative valence). F0 and rhythm-based features performed poorly; this was perhaps due to the complexity of the audio stimulus containing speech, music, silence and various audio sounds. Visual motion and (musical) tempo performed well individually but failed to provide any additional improvement if the feature set contained energy-based features. MFCCs, PLPs and Chroma coefficients performed similarly in isolation. Color-based video features proved valuable in valence classification. All in all, the selected parsimonious features set that provided the best emotion recognition results can be seen in Table 2.

Table 2. List of features used for emotion recognition.

Valence	audio	12 MFCCs and C0, plus derivatives
	video	maximum color value
	video	maximum color intensity
Arousal	audio	12 MFCCs and C0, plus derivatives

4. EXPERIMENTAL RESULTS

(a)

actual \ predicted	passive←						→active
→active	3	4	10	6	9	17	51
	5	9	14	13	13	21	25
	6	13	23	16	9	21	12
	11	13	27	22	10	10	7
	11	18	29	19	11	9	3
	17	16	28	18	8	10	3
passive←	24	18	23	14	6	13	2

(b)

actual \ predicted	negative←						→positive
→positive	2	6	7	10	25	34	16
	5	5	10	13	20	29	18
	3	6	15	18	20	23	15
	6	17	26	24	16	8	3
	8	26	30	20	8	6	2
	13	25	25	15	9	6	7
negative←	18	30	22	11	6	9	4

Fig. 3. Misclassification matrices normalized by row (%) for (a) arousal and (b) valence.

Table 3. Result evaluation metrics.

	metric	arousal	valence	2-D
Discrete (7 levels)	Accuracy	0.24	0.24	0.06
	Accuracy±1	0.57	0.62	0.37
	Mean abs. error	0.52	0.47	0.82
	Mean sq. error	0.48	0.43	0.92
	Correlation	0.43	0.22	-
Continuous	Mean abs. error	0.32	0.37	0.55
	Mean sq. error	0.17	0.24	0.41
	Correlation	0.54	0.23	-

The output of our system is a –usually very noisy– time series of seven categories. The signal is initially filtered with a low pass filter and then passed through a Savitzky-Golay filter [15] that interpolates the affective signal into a continuous-valued curve. To evaluate our system we compare the (seven-level) discrete output of the HMM system with the discretized affective curves. The interpolated continuous output curves are also compared with the reference continuous affective curves. Thus separate results are provided for the discrete-valued and continuous-valued curves.

Experiments are conducted using a "leave one (movie) out" cross-validation scheme. Results are presented as averages across all clips. The following evaluation metrics are shown: classification accuracy, classification accuracy ± 1 (which considers a miss by 1 category as a hit), mean absolute error (MAE), mean square error (MSE) and correlation coefficient. MAE and MSE are calculated after rescaling the curves to a $[-1, 1]$ range. Results are shown in Table 3. Classification accuracy for seven classes is, as expected, rather low at 25%. Accuracy± 1 (equivalent to using fewer categories) is fairly high at 60%; given the variety of movies in our database and the difficulty of the task this is a promising result. Note the very low correlation for valence that is further investigated next. Smoothing the discrete-valued curves further improves our

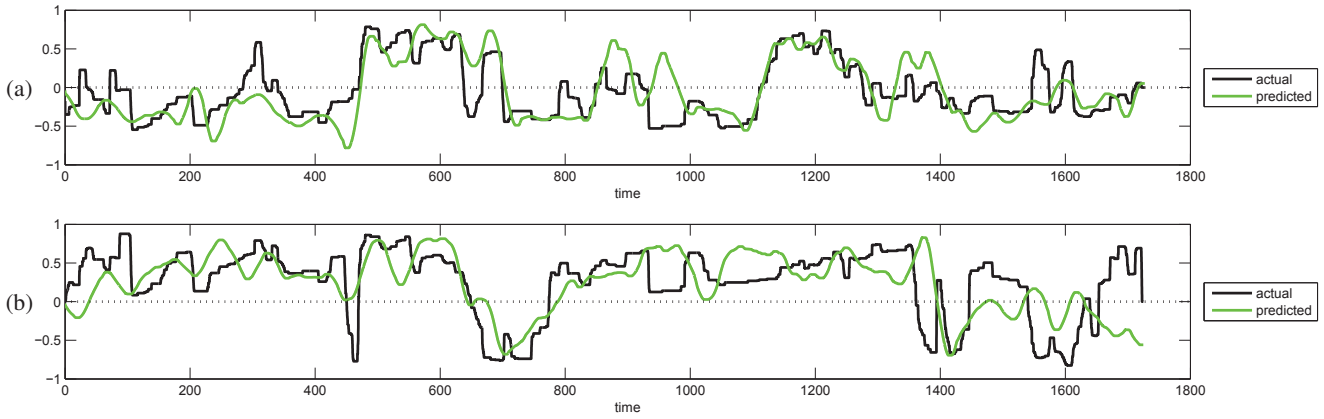

Fig. 4. Continuous intended emotion recognition vs. annotated curves for the "Ratatouille" clip: (a) arousal and (b) valence.

results, as can be seen from the significantly improved MSE and MAE continuous results, especially for arousal.

Fig 3 shows the misclassification matrices for arousal (a) and valence (b) normalized by the sum of each row, i.e, each cell (i, j) indicates the percentage of samples that belong to category i (actual) and are classified in category j (predicted). Best emotion recognition results are obtained for high arousal values, over 50% of the high activity frames are classified correctly (level 7). Note that frames are rarely misclassified to very distant categories, while neighboring categories are highly confusable.

Overall, the classifiers on both dimensions perform very well in classifying the mood of large segments, with the arousal classifier also performing well in describing detailed dynamics. The valence classifier fails at describing the continuous curve in detail, as revealed by the low correlation coefficient. Interestingly, this observation, as well as the overall relative performance of the classifiers in the two dimensions (prior to interpolation) also holds true for the performance of human annotators when evaluating their own experience (see Section 2). Note that a typical error in valence recognition is the misclassification of a contiguous area to entirely wrong valence categories, very positive scenes being identified as very negative and vice versa. This seems to happen in scenes where there is a conflict of modalities (e.g., "joyous" music, but "angry" video) or a conflict of sensory and semantic information. Our system lacks such semantic information, so it can not understand that a dark and gloomy battle will be perceived as positive if the viewers know that the hero is going to win. An example actual vs. predicted annotation for a 30 minute movie clip is shown in Fig 4.

5. CONCLUSIONS

We have briefly presented an annotated database of affect and our experiments in tracking the affective contents of the movies using HMMs. Evaluation of a large number of audio-visual features yielded somewhat surprising results, with many popular features being rejected before selecting the "optimal" feature set. Two independent HMM recognizers were used for arousal and valence, each utilizing a small number of low level features and a large number of states. The recognizers work well at a macroscopic level, capturing the general mood of the vast majority of scenes across movies. On the arousal dimension, the model also does well in capturing fine detail, subtle transitions, as revealed by the average correlation coefficient of 0.54. On the valence dimension, the model is successful at capturing the mood but sometimes fails at capturing the valence sign and transitions. Overall this is a first step towards continuous emotion recognition in movies. Further research in feature extraction, high-level semantic analysis, modeling and modality fusion is required to improve these results.

6. REFERENCES

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

[2] A. Hanjalic, "Extracting moods from pictures and sounds," *IEEE Signal Processing Magazine*, pp. 90–100, Mar. 2006.

[3] H.B. Kang, "Affective content detection using HMMs," in *Proc. ACM Multimedia*, 2003, pp. 259–262.

[4] M. Xu, L.T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Proc. ICME*, 2005, pp. 622–625.

[5] A. Austin, E. Moore, U. Gupta, and P. Chordia, "Characterization of movie genre based on music score," in *Proc. ICASSP*, 2010, pp. 421–424.

[6] M. Xu, J. Jin, S Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proc. ACM Multimedia*, 2008, pp. 677–680.

[7] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[8] 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, June 2006.

[9] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis, "Video event detection and summarization using audio, visual and text saliency," in *Proc. ICASSP*, 2009, pp. 3553–3556.

[10] R. Cowie, E Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Workshop on Speech & Emotion*, 2000, pp. 19–24.

[11] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 381–385.

[12] F. Eyben, M. Wollmer, and B. Schuller, "Openear – introducing the munich open-source emotion and affect recognition toolkit," in *Proc. ACII 2009*, 2009, pp. 1–6.

[13] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," in *Data Analysis, Machine Learning and Applications*, pp. 261–268. Springer Berlin Heidelberg, 2008.

[14] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, and S. Kollias, "Bottom-up spatiotemporal visual attention model for video analysis," *Image Processing, IET*, vol. 1, no. 2, pp. 237 –248, June 2007.

[15] A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, 1964.