H-050

Affective Analysis of Films by Low-Level Visual Features

テイシェイラ・ヘネ [1] 山崎俊彦 [1] 相澤清晴 [1,2]
René M A Teixeira Toshihiko Yamasaki Kiyoharu Aizawa

1. Abstract

Many algorithms and works have helped in the understanding and development of affective analysis of films. In spite of the progress done up to now, it is still not very precise how the low-level features of movies shape the resulting affective state of the viewer. In this work we evaluate different visual features and investigate how they impact the evaluated emotion under two paradigms: the dimensional approach (in terms of Pleasure, Arousal and Dominance) and the categorial approach. The evaluation is conducted by using Dynamic Bayesian Networks in the following topologies: a Hidden Markov Model network and Auto Regressive Hidden Markov Model network.

2. Introduction

Every day, thousands hours of visual information are produced [6] providing us with interesting and exciting content. This amount of data requires new approaches and offers new exploration opportunities. Affective analysis is useful considering the new possibilities for search and retrieval and enhanced user experience. The role of the affective content analysis has grown in the past decade [5] and with it a wide range of new possibilities and applications emerged.

Scientists and psychologists developed many models to explain human emotion. Two approaches are suitable for the work here described: The Categorial and the Dimensional models.

From the categorial point of view, an emotion can be decomposed into a set of basic emotions. Emotions are discrete and belong to specific affective families. In order to identify the amount of those basic emotions, Ekman [3] analyzed facial expressions and named six basic emotions, which are: Anger, Disgust, Fear, Happiness, Sadness and Surprise. To this set, Plutchik [12] also added Trust and Anticipation as basic emotions. The Dimensional approach does not have the burden of defining basic emotions. Instead, it splits the affect into a few dimensions that are generic enough to reproduce all affects. One of the most successful models is called Pleasure-Arousal-Dominance (PAD) framework [9]. Pleasure indicates the degree of pleasantness of an emotion. It indicates how "good" or "bad" someone feels. Arousal indicates the level of excitement or activation caused by the emotion. It can indicate a level of boredom or euphoria. The third and more complex is Dominance. It shows the dominion someone has in a situation. It demonstrates how powerful or anguished someone feels. Unfortunately, Dominance has not been widely used due to its variability and lack of understanding. However, we apply dominance in our study in order to identify possible strong features related to it. On the other hand, Pleasure and Arousal tend to be constant in different experiments and cultures. These three dimensions create a nearly orthogonal system.

The relationship between the low level features of a video and the expected emotion elicited on humans is still not well understood [14]. In this work, in order to achieve a better understating of this relationship, we analyzed how much of each basic feature contributes to the final emotion detection. For the analysis two probabilistic models were used. As features, we analyzed color-related features, motion and lighting key as well as the use of saliency maps.

3. Related Works

In the last decade many works contributed to develop the affective analysis [5]. Rasheed et al. [13] presented a framework for video classification into genres, based solely on visual information extracted from movie previews, also known as trailer. Movies are classified using average shot length, color variance, motion content and lighting key. For the lighting key features, the positions of three light sources are considered, the key-light, the back-light and the fill-light. Mean shift clustering is used for classification. The use of shot length is an important feature and is present in almost all of related works. Wang and Cheong [16] used a Support Vector Machine for movie classification. Audio and visual information are used whereas audio is classified according to its type, like music, environment or speech. All visual information is processed in the HSL (hue, saturation, lightness) color space. The use of HSL is justified by the psychological evidence of the human perception of emotion in this color space [15]. Moreover, a color energy feature is defined based on color and contrast information. Arifin and Cheung [1] exploited a hierarchical DBN approach to calculate the PAD indexes. For this method, on which our current work is partially based, each shot is modeled as a node in a Hidden Markov Model (HMM) way. They also introduce the use of saliency maps, used for the detection of the most interesting regions in an image, to help in the affective detection.

4. Feature vector

The first step and also the first feature used in our algorithm is the shot detection (resulting in the shot length feature). There are many algorithms to perform shot detection on videos, ranging in complexity and accuracy [4]. Most of them are robust to a specific kind of transition

In this work a hard cut detection suits well the objectives whilst maintains the simplicity. The detection was performed using color histogram comparison algorithm. For every frame three histograms were built, each one for a color channel (red,

[1] 東京大学電子情報学専攻 – The University of Tokyo, Dept. of Information and Communication Engineering

[2] 東京大学情報学環 – The University of Tokyo, Interfaculty Initiative in Information Studies

green and blue). They are compared bin-by-bin and the resulting color histogram difference index is then compared with a pre-determined threshold for the transition detection. The shot segmentation is an important step of the processing as each shot will be used as the most basic unit of the emotion detection.

With the movie structure defined, the next step is the extraction of the motion feature. "Motion History Images" (MHI) are extracted frame-by-frame [2]. From the difference of two frames, a binary difference image is built. This image shows the areas that have changed and so, can be used to indicate motion. With the succession of frames new images are computed. Finally these images are summed creating the MHI.

Colors play an important role in people's preferences. Many studies have been conducted in order to associate colors with human feelings. Ou et al. [11] investigated different color-emotion scales that were used in previous works in order to obtain a unique model that could explain the relationship between colors and affective characteristics. Investigating different color scales they were able to find those with close relation to the three primary factors of human emotions, described by Osgood [10]. Compiling the scales the following models were obtained:

$$\text{Color Activity} = -2.1 + 0.06\left[(\mathrm{L}^* - 50)^2 + (\mathrm{a}^* - 3)^2 + \left(\frac{\mathrm{b}^* - 17}{1.4}\right)^2\right]^{\frac{1}{2}} \qquad \mathrm{R}^2 = 0.93$$

$$\text{Color Heat} = -0.5 + 0.02(\mathrm{C}^*)^{1.07}\cos(\mathrm{h} - 50^\circ) \qquad \mathrm{R}^2 = 0.74$$

$$\text{Color Weight} = -1.8 + 0.04(100 - \mathrm{L}^*) + 0.45\cos(\mathrm{h} - 100^\circ) \qquad \mathrm{R}^2 = 0.73$$

Where L*, a* and b* are the coordinates of the CIEL*a*b* color space and h and C* are the hue angle and chroma, respectively.

Taking [9] as reference, these color models can be linked to the PAD paradigm in the following way: Color Activity affects Pleasure, Color Heat affects Arousal and Color Weight affects Dominance.

For every frame in the target video clip, a full RGB histogram was extracted. The resulting histograms were averaged in order to produce a dominant color of every shot. Finally, this dominant color was used as input for the color emotion models.

We also used lighting key feature, as described in [16]. In this approach, lighting key is composed of the median of the lightness histogram of the shot and the proportion of dark pixels.

Finally, we used saliency maps [8] as a region of interest mask for the feature extraction. Saliency maps tend to indicate the regions in an image that would draw the attention of the viewer. So, the color and lighting information inside those regions could contribute more to the final affective evaluation. For every frame, a corresponding saliency map was computed. These maps were applied during the color and lighting extraction process but not motion. The reason for that relies on the fact that color and lighting information can be extracted from a single frame whereas motion requires at least two frames. As the salient regions in two frames may be located in different areas, a frame-by-frame comparison for motion extraction was not possible.

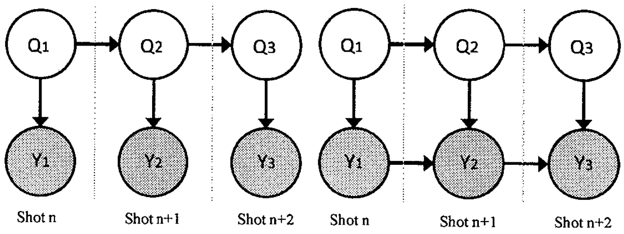

Figure 1 - Probabilistic Graphic Models (Left: HMM, right: ArHMM)

5. Probabilistic Modeling

Two approaches have been used in this work for the affective evaluation, both Dynamic Bayesian Network (DBN) based: Hidden Markov Model (HMM) and Autoregressive HMM (ArHMM).

A Bayesian Network is a graph, with no cycles, that can be used to represent probabilistic models of random variables and their interconnections. The DBN has in its structure two distinct sub-networks, or slices. The first slice represents the initial conditions and the second slice, the state in which the model finds itself at the moment. The structure repeats itself throughout the time. This representation is called 2TBN (two-slice temporal Bayesian Network) [7]. HMM is considered the most simple case of DBN and carries the assumption that the current state of the system being modeled depends only on the previous state.

We model the video as a sequence of shots, represented by nodes in each time-slice. Slices are composed of three hidden nodes (representing the PAD dimensions, that we want to obtain) and as many observable nodes as the number of features extracted from the video. The model showed in Figure 1 (left) assumes that the current affective state of a viewer depends only on the previous affective state. Qi are the hidden nodes, the information we intend to get. Yi are the features we extract from the video. If we break the assumption that all observations are not conditionally independent (e.g.: the current measure depends on the previous measure), the so called Autoregressive Hidden Markov Model (ArHMM) can be built. This assumption is reasonable if we consider that the shots contained in a scene are not very different one from another. Five discrete observable nodes represent the extracted features, namely: color activity, color heat, color weight, shot frequency and amount of motion.

6. PAD Mapping

In order to create a reliable ground-truth from which accuracy and reliability could be established, a user study was conducted aiming at the determination of the relationship between the displayed videos and the affective reaction of viewers. A total of 24 commercial, cinemas screened, Hollywoodian style movies were chosen. They were split into small clips with 1 minute and 52 seconds length in average, summing up to 346 clips and 10 hours 46 minutes and 15 seconds of play time. The clips were cut in a way that they would represent, most of the time, one unique emotion, empirically determined. Test subjects were 16 undergraduate students, being 8 men and 8 women. They are all

Japanese speakers and, with the exception of the movies narrated in English, all clips were displayed in its original language with Japanese subtitles. The presence of subtitles instead of a dubbed version was chosen to keep the affective influence of the movie as close as possible to what was expected by the movie director.

After each clip, viewers were requested to fill an electronic answer sheet with their appreciation about the affective content of the video. Users were asked to evaluate the eight Plutchik emotion families: joy, trust, fear, surprise, sadness, disgust, anger and anticipation. In the next section of the answer sheet the viewer had to evaluate their impression according to the PAD dimensions paradigm.

We assumed that the relationship between PAD values and the actual user's emotions can be represented as a linear combination of the PAD values, for simplicity reasons. More complex models could enhance the detection performance.

Fifteen clips that were not included in the learning process and were randomly chosen were used for tests. The resulting PAD evaluation was compared to the affective evaluation done by the test subjects. Table 3 and Table 4 display the results of the correlations between human evaluated clips versus computer evaluated clips, the values range from zero to one, being one the best possible case.

For all the test shots, two affective labels (primary emotion and secondary emotion) were evaluated by the system and compared against the affective labels evaluated by the test subjects. Table 1 and Table 2 show the accuracy of the identification performed by each model.

The process though which these results were obtained will now be described.

7. Analysis Process

To measure the effect of the features in the final result, the work of each model, HMM and ArHMM, was tested individually. The same measures were conducted for both models.

Figure 2 shows the sequence used. Initially, a model with all features was prepared. The results for this model were used as reference. In the figure, Subtracted Model stands for the models from which some features were removed. As previously explained on section 4, saliency maps were used to select regions of interest on the analyzed frames. Measures were performed using this filter and also without it. For each case, with and without saliency, models missing color features, lighting features and motion features were tested. Motion is one exception in this sequence as it was not analyzed using saliency maps.

Another important point to observe is that we considered color feature as one unique feature. The consequence is that during the feature removal test, when color feature was removed, the three PAD dimensions were directly affected.

In order to avoid fluctuations in the results, each test was conducted three times, having the results averaged.

8. Results

Results were analyzed under two perspectives: the affective labels precision rate and the PAD coefficients accuracy, for HMM and ArHMM models as well as the presence or absence of saliency filter. The six Ekman basic emotion were used as affective lables. From Table 1 we observe a slight improvement of the affective labels detection for the ArHMM model when using saliency filter. On the other hand the HMM model shows a small precision decrease in the presence of saliency. The connection between observations in the ArHMM turns it into a more complex chain and thus, more susceptible to noise and error propagation. The use of saliency filter may help reduce these factors.

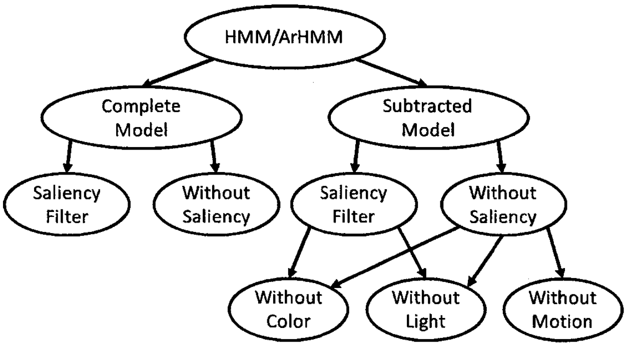

Figure 2 - Schematic of the experiments

Subtracting features from the models, both HMM and ArHMM behave in similar ways in the presence or absence of saliency. From the Table 2 it is possible to observe that the two models tend to have its accuracy reduced when a certain feature is removed. The exception to this statement is the subtraction of the color features that tends to improve the affective label detection, or in the case of Salient ArHMM, keep the same precision. It is important to notice that three color features were used and they were removed altogether so it was not possible to determine what is the contribution of each of these features. Still about affective labels, all the results were relatively close to each other with advantage of the complete, non-salient HMM model. There is no real gain in using saliency except if ArHMM is desired.

The results for the PAD coefficients, introduced on Table 3 and Table 4, showed to be very different one from another and require a separate discussion for each:

The Pleasure domain behaved in a non-stable fashion. The most probable explanation resides in the fact that the used features may not represent well the pleasure paradigm. Previous works, for example, those cited on section 3, showed good results applying audio features for pleasure calculation and this should be considered on future implementations. We observed an antagonist behavior between color features and lighting features for pleasure and thus, should they not be used at the same time. A more elaborated lighting feature could help solve this issue. Moreover, the use of saliency filters showed better results when the HMM model is used.

Arousal showed itself vulnerable to the subtraction of features in the two evaluated models. Motion features were not evaluated using saliency filter. The HMM model was more stable to the subtraction of motion features whilst the ArHMM heavily relies on it. From the three coefficients, arousal was the most robust to subtractions and the one with higher detection rates.

Dominance was very dependent on the lighting feature. The subtraction of this feature causes a drastic decrease in dominance detection. Saliency should not be used for dominance calculation as a clear behavior pattern could not be observed from the data.

9. Conclusion

The results lead to a more precise use of low-level features, although increasing the complexity of the system. It is important to evaluate how much precision is needed and how much complexity is acceptable to be added to the systems.

The present analysis should be further extended to include audio and other kind of features, like textual information. The addition of these features could turn Pleasure and Dominance estimation more robust.

Also, as the low level features are a direct representation of the affective content of the video, it is possible that the same analysis could be conducted using other kinds of models, respecting the interrelation between the variables.

Table 1 - Detection rate of affective labels

Detection of Affective Labels (Precision %) Complete Models		
	Without Saliency	With Saliency
HMM	58.28	53.85
ArHMM	53.84	55.36

Table 2 - Detection rate of affective labels

Detection of Affective Labels (Precision %) Incomplete Models				
	Without Saliency		With Saliency	
Subtracted feature	HMM	ArHMM	HMM	ArHMM
Color	65.7	57.1	59.2	55.5
Lighting key	55.2	50.7	53.1	44.4
Motion	58.8	50.5	NA	NA

Table 3 - PAD coefficients detection rate

Detection of PAD coefficients (Correlation) Without saliency						
	HMM			ArHMM		
	P	A	D	P	A	D
Complete Model	0.06	0.68	0.43	0.06	0.58	0.30
No Color	0.29	0.63	0.47	0.23	0.50	0.35
No Lighting key	0.08	0.61	0.28	0.16	0.58	0.16
No Motion	0.09	0.79	0.43	0.08	0.46	0.18

Table 4 - PAD coefficients detection rate

Detection of PAD coefficients (Correlation) With saliency						
	HMM			ArHMM		
	P	A	D	P	A	D
Complete Model	0.13	0.69	0.15	0.15	0.51	0.34
No Color	0.11	0.62	0.39	0.01	0.59	0.28
No Lighting key	0.31	0.78	0.27	0.13	0.62	0.30

10. References

[1] Arifin, S.; Cheung, P. A Computation Method for Video Segmentation Utilizing the Pleasure-Arousal-Dominance Emotional Information. MM'07 Germany 2007.

[2] Bradski, G. R; Davis J.W. "Motion segmentation and pose recognition with motion history gradients " Machine Vision and Applications. Publisher Springer. Berlin/Heidelberg. ISSN0932-8092.Volume 13, Number 3 / July, 2002.

[3] Ekman P. Basic Emotions. In T. Dalgleish and M. Power (Eds.). Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons, Ltd. 1999.

[4] Gargi, U.; Kasturi, R.; Strayer, S.H., "Performance characterization of video-shot-change detection methods," Circuits and Systems for Video Technology, IEEE Transactions on ,vol.10, no.1, pp.1-13, Feb 2000.

[5] Hanjalic, A., "Extracting moods from pictures and sounds: towards truly personalized TV," Signal Processing Magazine, IEEE , vol.23, no.2, pp.90-100, March 2006.

[6] Howhard D. W. The challenges of continuous capture contemporaneous analysis and customized summarization of video content, 2001. Defining a Motion Imagery Research and Development Program Workshop.

[7] Intel PNL User Guide. (2009, April 15) Retrieved from http://sourceforge.net/projects/openpnl/

[8] Itti, L.; Koch, C.; Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.20, no.11, pp.1254-1259, Nov 1998.

[9] Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology: Developmental, Learning, Personality, Social, 14:261–292.

[10] Osgood CE, Suci GJ, Tannenbaum PH. The Measurement of Meaning. University of Illinois Press. 1957; 31-75.

[11] Ou L.C, Luo M. R., Woodcock A., Wright A. "A study of colour emotion and colour preference. Part I: Colour emotions for single colours". Color Research & Application. VL: 29. NO: 3. PG: 232-240. YR: 2004

[12] Plutchik, R. and Hope R.C. Circumplex Models of Personality and Emotions. Washington, DC: American Psychological Association, 1997.

[13] Rasheed, Z.; Sheikh, Y.; Shah, M., "On the use of computable features for film classification," Circuits and Systems for Video Technology, IEEE Transactions on , vol.15, no.1, pp. 52-64, Jan. 2005.

[14] Tang, J., Song, Y., Hua, X., Mei, T., and Wu, X. 2006. To construct optimal training set for video annotation. ACM international Conference on Multimedia MULTIMEDIA '06. ACM, New York, NY, 89-92.

[15] Valdez, P. & Mehrabian, A. (1994). Effects of color on emotions. Journal of Experimental Psychology: General. 123, 394-409.

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