Affective Content Detection using HMMs

Hang-Bong Kang

Dept. of Computer Engineering,
The Catholic University of Korea
#43-1 Yokkok 2-dong Wonmi-Gu
Puchon City, Kyunggi-do, Korea
hbkang@catholic.ac.kr

ABSTRACT

This paper discusses a new technique for detecting affective events using Hidden Markov Models(HMM). To map low level features of video data to high level emotional events, we perform empirical study on the relationship between emotional events and low-level features. After that, we compute simple low-level features that represent emotional characteristics and construct a token or observation vector by combining low level features. The observation vector sequence is tested to detect emotional events through HMMs. We create two HMM topologies and test both topologies. The affective events are detected from our proposed models with good accuracy.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]:

General Terms

Algorithms

Keywords

Emotional event, Hidden Markov Models, Content analysis

1. INTRODUCTION

Recently, various techniques have been developed to automatically analyze and index multimedia data. Much of work, however, has focused on simple low level feature level and semantic analysis level [1]. To deal with a user's preferences effectively in video retrieval and video abstraction, it is desirable to detect affective content from video data. In other words, if affective content is analyzed, a user can retrieve the most interesting video clips or watch most exciting segments of video [1].

Several research works have been done to extract affective content from video. One method is to map low-level video characteristics into emotion space. Hanjalic and Xu[2] used motion and sound information to construct affect curves. From the affect curve, the user's perception may be analyzed. Another method is to use sound energy dynamics to detect and classify affective sound event [3]. In this method, audio cues or sound patterns in horror films are detected automatically only for scene classification.

However, limitations still exist in the detection of emotional event because the measurement of emotion related features or the mapping of low-level features into high level emotions is very difficult. In addition, users with different cultures may not perceive the same emotions from video clips. To deal with these problems, we perform empirical study on the relationship between emotional events and low-level features of video data and construct a unified framework to detect emotional events from video data.

In this paper, we propose a new affective content analysis using Hidden Markov Models (HMM). Section 2 discusses affective feature extraction from color, motion and shot cut rate information. Section 3 presents emotional event detection methods using HMMs. Section 4 shows experimental results.

2. Affective Feature Extraction

To represent affective content, we use a two-dimensional model used by Lang [4,5], where the horizontal axis represents valence and the vertical axis represents arousal. Valence refers to the affective responses ranging from positive state to negative state. Arousal refers to the responses ranging from excited to the calm. Figure 1 shows two-dimensional expression of four basic common emotions such as fear, anger, sadness, and joy [5]. Color and motion information can also be represented in two dimensional emotion space [6,7]. However, it is very difficult to discriminate fear and anger from low level features such as color, motion and shot cut rate. So, in this paper, we detect three affective events such as fear, sadness and joy.

To capture low level features which represent emotional characteristics, we have performed empirical study on the relationship between emotional events and simple low level features in video. The ground truth for three emotional events is manually determined on six 30-minute training video data which

MM'03, November 2-8, 2003, Berkeley, California, U.S.A.

are segmented into scenes. The scenes that belong to three emotional events are labeled by 10 students. If the video scene is labeled with the same emotional event by at least 7 of 10 students, we assign the scene as having one of three emotional events. From emotional events, we extract low level features. Table 1 shows the relationship between emotional events and low level features such as color, motion and shot cut rate. For example, at the fear events the colors are usually “dark and blue” or “dark and red” and “low saturated”. The motion phase may be zoom, tilt or dolly and motion intensity is not related with fear events. The shot cut rate is usually fast. Based on the information like Table 1, we extract color, motion and shot cut rate information in detecting affective events from video data.

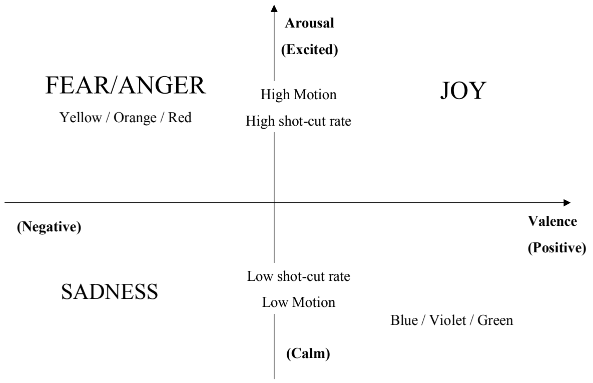

Figure 1. Two dimensional Emotion Space

Since emotional events or scenes consist of consecutive shots, we compute low level features from each shot of video scene. To compute color features, we transform RGB color space into HSV color space and then quantize the pixels into 11 culture colors such as red, yellow, green, blue, brown, purple, pink, orange, gray, black and white [8]. For a keyframe of each shot, we compute the histogram of 11culture colors. We also compute the saturation(S), value(V), dark colors, and bright colors. So, 15 color features are extracted from each shot.

To detect motion information, we compute frame differences between two consecutive frames in the shot and if the frame difference is larger than the threshold value, we divide the frame into nine regions and compute motions using optical flow. The motion phase is quantized into 8 directions and classified into “pan”, “tilt”, “zoom”, and “no camera motion” by template matching [9]. The motion intensity is also computed from dominant motion vectors. In the case of “no camera motion”, object motion intensity or frame difference is computed as motion intensity. Based on the motion intensity, each motion feature is quantized into three levels. The shot length is useful to represent shot cut rate. We compute each shot’s length and compared with median shot length because it shows a better estimate than the average shot length in the presence of outliers.

Table 1. Emotional event and low level features

	Color	Motion (Phase/intensity)	Shot cut rate
Fear	Dark and blue, sometimes dark and red Low saturated	Zoom, tilt, dolly /NA	Fast
Sadness	Dark Low Saturated	No camera motion / Small	Slow
Joy	Bright colors	NA / Large	NA

3. HMM-based emotional event detection

In this section, we will discuss the detection method of emotional events from the video. Our problem is to classify observed low-level feature sequences into emotional events. Unlike most classical pattern classification problems, the data to be classified is time series data. To detect emotional events, we use Hidden Markov Models (HMM) because HMMs are widely used probabilistic models for time series data. In addition, several works using HMMs show good potential for video parsing or segmentation [10-14].

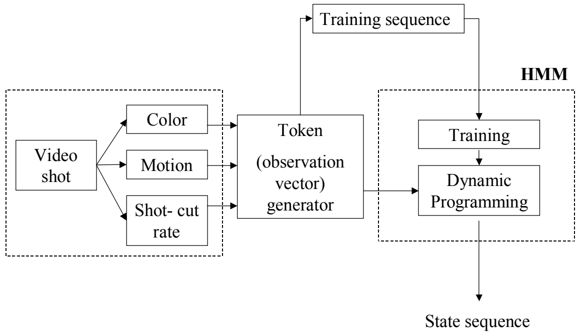

Figure 2. Proposed emotional event detection system

To use HMMs, several things such as topology, observation vectors and statistical parameters of HMM, have to be determined. We construct two different HMM topologies shown in Figure 3 and Figure 4. In Figure 3, a four-state circular HMM model λ = (A, B, π) is created [10]. The states model one of emotional events such as *Fear*, *Sadness*, *Joy* and *Normal* state. For example, in Figure 3, “state 1” represents *Normal* state. The "state 2" represents *Fear* emotional state. "State 3" and "State 4" represent *Sadness* emotional state and *Joy* emotional state, respectively. The arrowed lines indicate possible transitions between states. From the *Normal* state, it is possible to transit to any of the emotional states, but from the emotional state it is only possible to return to the shot state. The parameter A, B, and π are determined during the training process. Here

$A = \{ a_{ij} | a_{ij} = P(s_{t+1} = q_j | s_t = q_i)\}$ (1)

a_{ij} : the probability of transiting from state q_i to the state q_j.

$B = \{b_j(k) | b_j(k) = P(v_k | s_t = q_j)\}$ (2)

$b_j(k)$: the probability of output symbol v_k at state q_j.

$\pi = \{\pi_i | \pi_i = P(s_i = q_i)\}$ (3)

π_i = initial state probability.

To apply HMM to time-sequential video shots consisting of a scene, the features of each video shot must be transformed to observation vector sequences in learning and recognition phase. We use vector quantization technique in generating observation vector sequences. Each feature vector is transformed into the symbol that is assigned to the codeword nearest the vector in the feature space. Once the model topology and observation vectors are determined, the next step is to train the model by training data and determine the initial, state-transition and emission probabilities. Parameter estimation for the HMM is done using Baum-Welch algorithm [10].

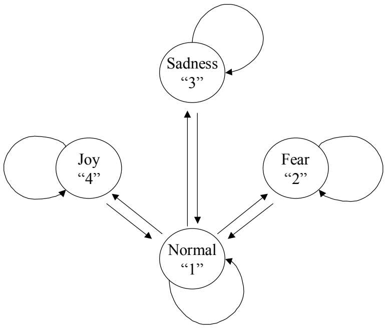

Figure 3. HMM for modeling emotional events.

To detect emotional events through HMM like Figure 3, we compute observation vector sequences first. The observation vector or token v_k is computed from three low level features. Then, we decode observation vector sequences into the most likely sequence of hidden states by dynamic programming [10]. Emotional events like *Fear* are detected by identifying sequences of hidden states "1-2s".

Another HMM topology for emotional event detection is shown in Figure 4. Each HMM consists of two states and represents one of emotional events. Let us denote 4 HMMs by HMM-FEAR, HMM-SAD, HMM-JOY, HMM-NORMAL. Four HMMs are learned from a training data set, respectively. Given a set of observation vectors obtained from test data, we compute the likelihood of each HMM and assign the test data to one of emotional events which has the largest likelihood.

4. Experimental Results

Preliminary experiments have been done with the HMM-based emotional event detection model proposed in Section 3. We made two data sets from six video data such as "I Know What You Did Last Summer", "Dying Young", "Autumn in New York", "Mask", "Scream" and "When Harry met Sally". The ground truth for three emotional events is manually determined. One data set is used for training set and the other set is used for test set. The video data was segmented into shots using color histograms and key frames were selected for each shot using color and motion information [15]. For each key frame, we transformed RGB color space into HSV space and finally computed 15 color features(11 culture colors, saturation, value, dark colors and bright colors). The motion phase and intensity for each shot is computed. The relative shot length is also computed. From color, motion and shot length features, observation vectors are generated by vector quantization.

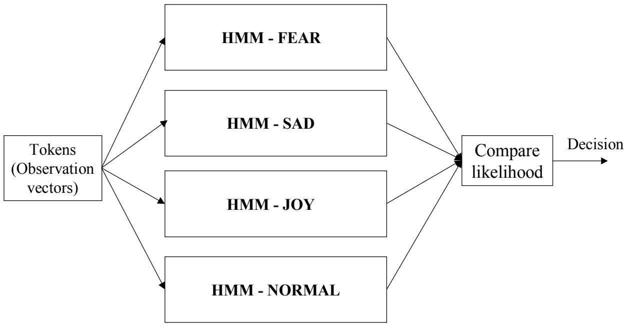

Figure 4. The 4-HMM Structure

Two HMM topologies like Figure 3 and Figure 4 are created. For both HMM topologies, three different sets of observation vectors as color-only, motion+ shot length, and color + motion + shot length, are trained using Baum-Welch algorithm. To test our method, we compute a sequence of observation vectors from test video. To detect emotional events through HMM like Figure 3, we decode state trajectory by dynamic programming. We decide whether emotional events exist or not by counting the number of emotional states. The emotional event detection through HMMs like Figure 4 is performed by comparing the likelihood of each HMM. The observation vector sequence is assigned to an appropriate event depending on which likelihood is larger. The experimental results of color-only are shown at Table 2. Table 3 and Table 4 show the experimental results of motion + shot length and color + motion + shot length, respectively. While both topologies are quite successful for detecting emotional events, the experimental results reveal that the HMM topology like Figure 4 performs better compared to the HMM topology like Figure 3. In addition, the performance improvement due to motion and shot length is not significant. Color-only observation vectors provide reasonable results.

Table 2. Experimental Results (Color Only)

	Detection Rate (Topology like Fig. 3)	Detection Rate (Topology like Fig. 4)
Fear	58.3 %	87.6 %
Sadness	72.1 %	71.4 %
Joy	65.5 %	69.3 %

Table 3. Experimental Results (Motion + Shot Cut Rate)

	Detection Rate (Topology like Fig. 3)	Detection Rate (Topology like Fig. 4)
Fear	58.3%	65.1%
Sadness	76.8%	74.3 %
Joy	77.4%	85.7%

Table 4. Experimental Results (Color + Motion + Shot Cut rate)

	Detection Rate (Topology like Fig. 3)	Detection Rate (Topology like Fig. 4)
Fear	61.9 %	81.3%
Sadness	75.1%	76.5%
Joy	73.6%	78.4%

5. Conclusions

In this paper, we propose a new method to detect affective events from video data using HMMs. We compute color, motion and shot cut rate from video data and construct feature vectors as observation vectors. We create two HMM topologies to detect emotional events. For both HMM topologies, the affective events are detected with good accuracy. Currently, we are experimenting with complex features to improve the detection results. We also develop a learning method to reflect user's preferences in detecting emotional events. Furthermore, if we incorporate audio features, the detection results for emotional events will be improved.

ACKNOWLEDGEMENTS

This work was supported by the University Research Program of Ministry of Information & Communication in Republic of Korea.

6. REFERENCES

[1] Hanjalic, A. Video and Image Retrieval beyond the Cognitive Level: The Needs and Possibilities. Proc. SPIE Storage and Retrieval for Media Databases 2001, San Jose, CA, pp.130-140, 2001.

[2] Hanjalic A. and Xu, L. User-oriented Affective Video Content Analysis, Proc. IEEE Workshop on CBAIBL'01, Kauai, HI, pp.50-57, Dec. , 2001.

[3] Moncrieff, S.,Dorai, C. and Venkatesh, S.: Affect Computing in Film through Sound Energy Dynamics, Proc. ACM MM'01, pp. 525-527, 2001.

[4] Lang, P.: The emotion probe: Studies of motivation and attention, American Psychologist, 50(5), pp. 372-385, 1995.

[5] Picard, R. Affective Computing. MIT Press, 1997

[6] Valdez, P. and Mehrabian, A.: Effects of color on emotions, Journal of Experimental Psychology: General, pp. 394-409, 1994.

[7] Scheirer, J. and Picard, R.: Affective Objects, MIT Media lab Technical Rep. No 524.

[8] Goldstein, E. : Sensation and Perception, Brooks/Cole, 1999.

[9] Lee, S., and Hayes, M.: Real-time camera motion classification for content-based indexing and retrieval using templates, Proc. ICASSP, pp.3664-3667, 2002.

[10] Rabiner, L. and Juang, B.: Fundamentals of Speech Recognition, Prentice Hall PTR , 1993.

[11] Boreczky, J. and Wilcox, E.: A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features, Proc. ICASSP' 98 , 1998.

[12] Eickeler, S. and Muller, S.: Content-based Video Indexing of TV Broadcast News Using Hidden Markov Models, Proc. ICASSP'99 , 1999.

[13] Li, B. and Sezan, M.: Event Detection and Summarization in Sports Video, Proc. IEEE CBAIBL'01, Kauai, HI , 2001.

[14] Naphade, M., Garg A. and Huang, T.: Audio-Visual Event Detection using Duration dependent input output Markov models, Proc. IEEE CBAIBL'01, Kauai, HI, 2001.

[15] Zhang, H., Wu, J., Zhong, D., and Smoliar, S.: An integrated system for content-based video retrieval and browsing, " *Pattern Recognition*, Vol. 30, pp.643-58, 1997.