A NOVEL VIDEO PARSING ALGORITHM UTILIZING THE PLEASURE-AROUSAL-DOMINANCE EMOTIONAL INFORMATION

Sutjipto Arifin and Peter Y.K. Cheung

Department of Electrical and Electronics Engineering, Imperial College London, SW7 2BT

ABSTRACT

One of the major problems faced when designing a high level video parsing system is that viewers usually have doubts about the exact boundaries of an episode. Moreover, due to the different emotional states that viewers have while viewing a video, it is very difficult to improve the performance of these algorithms using convention methods. To solve this problem, this paper presents a novel spectral clustering based high level video parsing algorithm that utilizes the Pleasure-Arousal-Dominance (P-A-D) [1, 2] emotional content of the video.

Index Terms— Clustering methods, Machine Vision, Modeling, Motion Pictures, Video Signal Processing

1. INTRODUCTION AND RELATED WORK

Extracting video structures is a fundamental task in video content analysis. A video structure consists of video shots, defined as an unbroken sequence of frames recorded in a single camera. While detecting shot boundaries organizes video content at the syntactic level, episode segmentation provides a natural segmentation of video that viewers can associate with.

There are generally four types of segmentation techniques (table 1): overlapping links [3], video coherence [4], time constraint clustering [5] and time adaptive clustering (TAC) [4]. Techniques that employ binary temporal distance functions are more sensitive to the choice of threshold than continuous functions. Sequential comparison techniques perform pair-wise shot visual comparisons, whereas clustering comparison techniques perform group-wise comparisons. TAC is the best method due to its consistency and performance [6].

One of the biggest problems of high level parsing is that viewers usually have doubts about the exact start and end of an episode. In addition, it is almost impossible to significantly improve the performance of these algorithms using convention methods such as the improvement of shot comparison functions. This is due to the difference in the affective states of the viewers that are watching the video. In other words, high level video parsing algorithms can reap benefits from efforts to model the emotional content of the video.

The modeling of the affective content of the video is a difficult problem, as there are very few concrete relations between the low-level features and the high level meaning of the video. Generally, there are two basic approaches to modeling affect. The first approach is the categorial model of affect [7]. In this model, emotions are discrete and belong to one of a few basic groups. However, the number of these basic emotions and the definition of their nature have been contentious questions for some time. Although there are many proposals, the common contenders are found to be "fear", "anger", "sadness", "happiness", "disgust" and "surprise".

Table 1. Video episode segmentation algorithms.

	Binary Functions	**Continuous Functions**
Sequential	Overlapping Links	Video Coherence
Clustering	Time Constraint	Time Adaptive

The second approach is the dimensional model of affect. This model does not reduce emotions into a finite set, but attempts to find a finite set of underlying dimensions into which emotions can be decomposed. One of the most discussed dimensional models is the Pleasure-Arousal-Dominance (P-A-D) model [1]. "Pleasure" stands for the degree of pleasantness of the emotional experience. It is typically characterized as a continuous range of affective responses extending from "unpleasant" to "pleasant". "Arousal" stands for the level of activation of the emotion, and it is characterized as a range of affective responses extending from "calm" to "excited". "Dominance" describes the level of attention or rejection of the emotion. It is useful in distinguishing emotional states that have similar "pleasure" and "arousal". Examples are "violence" and "fear". "Violence" has P-A-D values of $(-0.50,+0.62,+0.38)$, and "fear" has P-A-D values of $(-0.64,+0.60,-0.43)$ [1]. In principle, if the P-A-D dimensions are continuous, this model is able to generate an infinite number of emotional states.

There are very few existing works on video emotional content modeling based on the dimensional emotion model. Hanjalic [8] provided a solid basis for obtaining a reliable dimensional-based approach of affective video content representation. The affective content of the video is represented as set of points in a two-dimension emotion space. Their experiments suggest that the affective content of the video can be more reliably modeled by discovering more concrete relations between the affect dimensions and low-level features.

This paper is presents a new spectral clustering based video episode segmentation algorithm that utilizes the saliency and P-A-D information of a video. It's major contributions include (1) a new visual content feature derived from saliency and P-A-D intensity maps for emotional content representation (section 2.1) and (2) the episode segmentation algorithm that utilizes these information (section 2.2). Results and concluding remarks are presented in section 3 and section 4.

2. VIDEO EPISODE DETECTION ALGORITHM

Our proposed algorithm consists of two stages. The first preprocessing stage extracts features (section 2.1) to represent each video shot based on the color emotional response, attention and tempo information. In the second stage, spectral clustering (section 2.2) is applied to the features. The clusters are considered as the episode segments of the input video.

2.1. Color Emotional Response, Attention and Tempo

Generally, motion is the primary element of a film and it is often utilized to guide attention. Viewer's attention can be modeled based on motion, which can be computed using any standard block-based motion estimation. However, motion intensity alone is not sufficient to produce a reliable saliency map because of its low sensitivity to low motion energy.

The spatial phase consistencies of the motion vectors can be computed to compensate this weakness. Spatial coherence can be computed by first determining the phase histogram of each macroblock within a $N \times N$ spatial window. The coherence is then measured by entropy,

$$Coh_k(x,y) = -\sum_{i=1}^{i=N_{bins}} \rho_i \times log_2(\rho_i) \qquad (1)$$

where $Coh_k(x,y)$ is the spatial coherence of frame k at macroblock position (x,y) and ρ_i is the phase histogram bin probability within the spatial window at (x,y). The components are combined as $motI_k \times Coh_k$ to produce the motion saliency map, where $motI_k$ is the motion vector field of frame k.

Attention modeled based on motion alone has its limitations, since a film sequence with low motion may still attract attention. We propose a new algorithm that extends [9] to generate our static saliency map. The color, intensity and orientation contrast maps of frame k and $k+1$ are first generated and a winner-takes-all approach is used to compute their respective saliency maps. The salience regions of frame k are treated as objects of interest and are tracked based on their area, centroid coordinates and color histogram by performing a similarity search in frame $k+1$ for each object in frame k. A match is declared if an object pair has maximum similarity.

The reason for this matching scheme is as follows. When an object of interest in frame k has found no match in frame $k-1$, it can mean that a new object of interest has appeared.

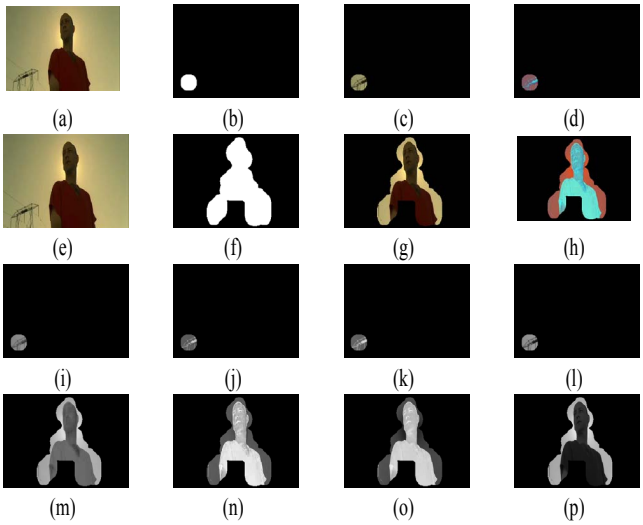

Fig. 1. Image examples: (a) frame 100 of test video 20; (b) motion saliency map; (c) motion saliency image in RGB; (d) I_{mot}^{PAD}; (e) frame 101 of test video 20; (f) static saliency map; (g) static saliency image in RGB; (h) I_{stat}^{PAD} (i) I_{mot}^{P}; (j) I_{mot}^{A}; (k) I_{mot}^{D}; (l) I_{mot}; (m) I_{stat}^{P}; (n) I_{stat}^{A}; (o) I_{stat}^{D}; (p) I_{stat}.

If no match is found for this new object of interest in frame $k+1$, it means that this object of interest may be too fast for the viewer to realize. Therefore, it should not be considered when computing the static saliency map for frame $k+1$.

Two RGB saliency maps (motion and static) are generated and converted to HSI space. The S and I images are used to generate P-A-D saliency images using the following,

$$I_{x,y}^{P} = \frac{(0.69 \times I_{x,y}) + (0.22 \times S_{x,y})}{0.91} \qquad (2)$$

$$I_{x,y}^{A} = \frac{(-0.31 \times I_{x,y}) + (0.60 \times S_{x,y}) + 0.31}{0.91} \qquad (3)$$

$$I_{x,y}^{D} = \frac{(-0.76 \times I_{x,y}) + (0.32 \times S_{x,y}) + 0.76}{1.08} \qquad (4)$$

The above equations are the expressions of P-A-D responses to brightness and saturation [2], modified with coefficients designed to normalize the values to the range of 0 to 1. Examples of various P-A-D response maps and saliency maps are depicted in figure 1.

Given the saliency images (I_{stat}^{P}, I_{stat}^{A}, I_{stat}^{D}, I_{stat}, I_{mot}^{P}, I_{mot}^{A}, I_{mot}^{D}, I_{mot}), the following equations are computed,

$$\delta_{s\in\{stat,mot\}}^{t\in\{P,A,D\}} = \frac{1}{\sum_{r=1}^{r=R} N_r} \sum_{r=1}^{r=R} \sum_{x=1}^{x=X_r} \sum_{y=1}^{y=Y_r} I_{x,y,r}^{t\in\{P,A,D\}} \qquad (5)$$

$$\epsilon_{s\in\{stat,mot\}} = \sum_{r=1}^{r=R} \sum_{x=1}^{x=X_r} \sum_{y=1}^{y=Y_r} (1-\zeta_r)\left(\frac{I_{x,y,r}}{N_r}\right) \qquad (6)$$

$$\delta_{overall}^{t\in\{P,A,D\}} = (1-\mu)\left(\delta_{stat}^{t\in\{P,A,D\}}\right) + (\mu)\left(\delta_{mot}^{t\in\{P,A,D\}}\right) \qquad (7)$$

$$\epsilon_{overall} = (1-\mu)(\epsilon_{stat}) + (\mu)(\epsilon_{mot}) \quad (8)$$

$\delta_{overall}^{t\in\{P,A,D\}}$ is the estimated P or A or D value computed using the I_{stat}^{P}, I_{stat}^{A}, I_{stat}^{D}, I_{mot}^{P}, I_{mot}^{A} and I_{mot}^{D} images, $\epsilon_{overall}$ is the attention value computed using the I_{mot} and I_{stat} images, ζ_r is the normalized centroid distance of region *r* from the frame center, N_r is the total number of pixels of region *r* and μ is the mean normalized motion intensity of the frame. We assume more attention is placed on the frame center.

Before the definition of our proposed visual feature vector, we first define ΔC_n as the gradient function of $C_n(i)$. The visual content variation function $C_n(i)$ is given by,

$$C_n(i) = \sum_{k=f_{1,n}}^{k=i} D_{k,k+1} \quad (9)$$

where $f_{1,n}$ and $f_{last,n}$ is the first and last frame of shot *n*, *i*= $1 \cdots f_{last,n} - 1$ and *D*(.) can be any standard visual difference metric. We define *D*(.) as the color histogram difference,

$$D_{k,k+1} = \frac{\sum_{b=1}^{b=N_b} \sum_{sp=1}^{sp=N_{sp}} \left| \frac{H_k^{sp}(b)}{\sum_{i=1}^{i=N_b} H_k^{sp}(i)} - \frac{H_{k+1}^{sp}(b)}{\sum_{i=1}^{i=N_b} H_{k+1}^{sp}(i)} \right|}{N_{sp}} \quad (10)$$

where N_b and N_{sp} are the total number of bins and color spaces respectively. ΔC_n is a good indicator of the film's tempo. We also define the shot length variation function,

$$\alpha_n = 1 - \frac{S_n}{\xi} \quad (11)$$

where S_n is the length of the segment in frames and ξ is a normalizing constant. α_n models the fact that different directors use different shot length variations to create tempo and shorter shot lengths usually means higher tempo.

The histograms of $\delta_{overall}^{P}$, $\delta_{overall}^{A}$, $\delta_{overall}^{D}$, $\epsilon_{overall}$, ΔC_n and μ of each video shot are computed. The affective content of a video shot is represented using a statistical-based feature vector derived from the histogram of each feature. Defining β_0^x and β_1^x as the normalized counts of the first and second largest peak of the histogram, β_2^x as $\frac{\beta_1^x}{\beta_0^x}$, β_3^x and β_4^x as the bin locations of the two peaks and β_5^x as the entropy of the histogram, where *x* is the feature indicator, a 37-dimension feature vector is defined to represent the emotional information derived from the visual data of the video shot *n*,

$$F_n = [\beta_{0:5}^{\delta_{overall}^{t\in\{P,A,D\}}}, \beta_{0:5}^{\epsilon_{overall}}, \beta_{0:5}^{\Delta C_n}, \beta_{0:5}^{\mu}, \alpha_n] \quad (12)$$

2.2. Video Episode Detection By Spectral Clustering

Feature vectors are computed for each shot using equation 12, and they are treated as data points to be clustered and it is denoted as ν. For each pair of points *i,j* $\in \nu$, a similarity value S_{ij} can be computed using the following expression,

$$S_{ij} = exp(\frac{d_{ij}}{2\sigma^2}) \quad (13)$$

where d_{ij} is the distance between node *i* and *j* and σ is the similarity matrix normalization factor. S_{ij} can be considered as weights on the undirected edges *ij* of a graph *G* over ν. Using *S*, we can compute the stochastic matrix,

$$P = D^{-1}S \quad (14)$$

where *D* is the degree of node *i* defined by the expression,

$$D_i = \sum_{j\in\nu} S_{ij} \quad (15)$$

Once *P* is obtained, the eigenvectors of *P* corresponding to the *K* largest eigenvalues are computed. The eigenvectors are used to form the matrix ψ, where the columns correspond to the eigenvectors in ascending eigenvalue order. *k*-means algorithm is then applied to cluster the rows of the matrix ψ as points in a *K*-dimensional space. Finally, the smallest and largest shot boundary locations of each cluster are treated as the episode boundaries of the movie.

3. EXPERIMENT PROCEDURES AND RESULTS

A pilot panel study based on [10] is first designed to determine the videos that may result in high subject agreement in terms of the elicited emotions. The eight emotion categories are "sadness", "violence", "neutral", "fear", "happiness", "amusement", "exciting" and "surprise". 24 videos were selected. The videos have a total of 4431 shots and 262 episodes. The total time of 24 videos is 220 minutes and 21 seconds. All videos are of XVID format with 5-channel ac3 audio format. 14 postgraduate students were asked to view the videos. They were first asked to sign consent forms. Between each video showing, there is a break time which serves two purposes. Firstly, it allows the emotional state of the participants to "settle" so that it will not affect their emotional experience when the next video begins. Secondly, it allows the participants to complete a post-video questionnaire that aims to find out about the emotions experienced by the participants while watching the video. Finally, the episode boundaries are manually labeled based on the criteria defined as follows: video shots within an episode cannot be too far apart and should have similar tempo, visual and emotional content.

We compared the performance of the proposed algorithm with the time adaptive clustering (TAC) algorithm, which is the best segmentation algorithm so far as evaluated in [6]. The evaluation criterion is defined by the following equation,

$$g = \left(1 - \frac{F_+ + F_-}{\gamma}\right) \times 100\% \quad (16)$$

where F_+ and F_- correspond to the number of false positives (wrongly detected boundaries) and false negatives (missed boundaries) respectively and γ corresponds to the worst case segmentation where every shot is an episode segment. Higher g corresponds to lower F_+ and F_-, thus better performance.

The g values of the TAC algorithm and our proposed algorithm tested on 24 test videos are summarized in table 2. Note that RH2=Rush Hour 2, GD=General's Daughter, JR=Jin Roh the Wolf Brigade, RD=Red Dragon, AHX=American History X, WLB=What Lies Beneath, SC=Sin City, S=Seven, COA=City of Angels, BH=Brave Heart and TR=The Rock.

Our proposed algorithm managed to achieve an average improvement of 8.1% over the TAC algorithm. Table 2 shows that there are 5.7%, 4.2% and 6.6% reductions in g for test videos 5, 9 and 22 respectively. After some investigation, we discovered that this is because a temporal distance mechanism is not introduced in our algorithm. In other words, shots that are too far apart in terms of temporal distance may still be clustered into the same episode cluster. For test videos 5, 9 and 22, a few shots that should belong to the last episode cluster are clustered to the first cluster due to similar content, thus increasing the number of false positives and negatives. To justify this claim, we modified equation 13 by introducing a continuous temporal distance function,

$$S_{ij} = max(0, 1 - \frac{\phi_{ij}}{\varphi}) \times exp(\frac{d_{ij}}{2\sigma^2}) \qquad (17)$$

where ϕ_{ij} is the temporal distance between *i* and *j* in terms of frames and φ is the average length of an episode in frames. The temporal function basically reduces the similarity S_{ij} to 0 if ϕ_{ij} is too large. Using equation 17, we reapplied our algorithm on test videos 5, 9 and 22, and we managed to achieved g values of 96.4%, 87.5% and 96.6% respectively.

4. CONCLUSIONS

In this paper, we presented a novel episode segmentation algorithm that utilizes the emotional content of the video. On average, our proposed algorithm managed to achieve a performance improvement of 8.1% over the TAC algorithm, the best episode segmentation algorithm so far according to [6]. Future works include the improvement of our algorithm's performance by providing a more reliable P-A-D value estimation, which may be achieved by Dynamic Bayesian Networks.

Table 2. The g values of the TAC and proposed algorithm.

No.(Origin)	Emotional Content	min:sec	TAC	Ours
1.(RH2)	Amusement, Neutral	9:27	28.6	42.9
2.(RH2)	Amusement, Neutral	8:23	54.6	63.6
3.(GD)	Neutral, Sad	10:11	73.5	82.3
4.(GD)	Sad, Violent	8:47	65.9	77.2
5.(JR)	Fear, Neutral	10:43	96.4	90.7
6.(JR)	Fear, Violent	10:01	50.0	77.8
7.(JR)	Fear, Neutral	9:05	66.7	73.3
8.(JR)	Fear, Violent	9:29	71.4	76.2
9.(RD)	Fear, Neutral	9:21	87.5	83.3
10.(AHX)	Neutral, Violent	10:18	58.3	66.7
11.(AHX)	Neutral, Happy	7:31	72.1	73.3
12.(AHX)	Fear, Violent	7:47	74.3	85.7
13.(AHX)	Neutral, Happy	8:33	80.0	87.5
14.(AHX)	Neutral, Violent	7:56	41.7	66.7
15.(WLB)	Amusement, Neutral	9:25	50.0	62.5
16.(WLB)	Fear, Neutral	9:41	86.1	88.9
17.(SC)	Fear, Violent	8:36	78.1	88.6
18.(SC)	Neutral, Fear	8:20	61.1	66.7
19.(SC)	Fear, Violent	7:10	61.1	77.8
20.(S)	Neutral, Fear	11:07	40.9	68.1
21.(S)	Fear, Violent	10:38	83.8	83.8
22.(COA)	Neutral, Sad	10:45	93.3	86.7
23.(BH)	Sad, Violent	10:53	76.5	76.5
24.(TR)	Exciting, Surprise	6:14	81.8	86.4

5. REFERENCES

[1] Albert Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psycology: Development, Learning, Personality, Social*, vol. 14, no. 4, pp. 261–292, Dec. 1996.

[2] Patricia Valdez and Albert Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology*, vol. 124, no. 4, pp. 394–409, Dec. 1994.

[3] Alan Hanjalic and R. Lagendijk, "Automated high level segmentation for advanced video retrieval systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 580–588, June 1999.

[4] J. Kender and B. L. Yeo, "Video scene segmentation via continuous video coherence," in *IEEE Conference on Computer Vision and Pattern Recognition*.

[5] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features," in *International Conference of Multimedia Systems*.

[6] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *IEEE Transactions of Multimedia*, vol. 4, no. 4, pp. 492–499, Dec. 2002.

[7] Rosalind W. Picard, "Building affective computing," in *Affective Computing, The MIT Press: Cambridge, MA, 1997*.

[8] Alan Hanjalic and Li Qun Xu, "Affective content representation and modeling," *International Conference of Multimedia Modeling*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[9] Laurent Itti, Christof Koch, and Ernst Niebur, "A model for saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[10] James J. Gross and Robert W. Levenson, "Emotion elicitation using films," *Cognition and Emotion*, vol. 9, no. 1, pp. 87–108, Jan. 1995.