AN IMPROVED VALENCE-AROUSAL EMOTION SPACE FOR VIDEO AFFECTIVE CONTENT REPRESENTATION AND RECOGNITION

Kai Sun, Junqing Yu[+], Yue Huang and Xiaoqiang Hu

School of Computer Science & Technology
Huazhong University of Science & Technology
Wuhan 430074, Hubei, China
[+]*Corresponding author: yjqing@hust.edu.cn*

ABSTRACT

To understand video affective content automatically, the primary task is to transform the abstract concept of emotion into the form which can be handled by the computer easily. An improved V-A emotion space is proposed to address this problem. It unifies the discrete and dimensional emotion model by introducing the typical fuzzy emotion subspace. Fuzzy C-mean clustering (FCM) algorithm is adopted to divide the V-A emotion space into the subspaces and Gaussian mixture model (GMM) is used to determine their membership functions. Based on the proposed emotion space, *the maximum membership principle* and *the threshold principle* are introduced to represent and recognize video affective content. A video affective content database is created to validate the proposed model. The experimental results show that the improved emotion space can be used as a solution to represent and recognize video affective content.

1. INTRODUCTION

Video affective content analysis is one of the latest research areas in CBVR (Content Based Video Retrieval), which can utilize both affective computing [1] and theories of CBVR to understand video affective content [2]. Affective content is an important natural component of human classification and retrieval of information. Recognizing video affective content and using it to automatically label the significant affective features potentially allow a new modality for user interaction with video content.

However, how to represent and recognize video affective content is a big challenge. Based on the well-known 2D valence-arousal (V-A) emotion space, Hanjalic proposed a V-A model [2] which can provide a solid basis to represent the video affective content. The motion, shot length variation, sound energy and pitch information were extracted to generate curves that describe the V-A content of the video. These curves were combined to represent the video affective content as a set of coordinate points in a two-dimension V-A emotion space. Although the V/A curves can smoothly measure the emotion intensity frame by frame and detect the video affective content effectively, it is very hard to relate the computed V-A values to the emotion categories, making it difficult to be utilized for video indexing.

Inspired by the works of Hanjalic [2], an improved 2D V-A emotion space is proposed to represent and recognize the video affective content. The basic idea is to define a set of typical fuzzy emotion subspaces in V-A emotion spaces. By introducing the subspaces, the proposed emotion space can represent discrete affective states in the continuous V-A space. Combined with the V-A model proposed by Hanjalic, the computed V-A values can be related to the typical emotion subspaces, making it convenient to be utilized for video affective content indexing.

2. MODELING METHOD

There are many psychological emotion models. They can be categorized into two classes: a) discrete emotion states [5-6]; b) dimensional continuous emotion space [3-4, 7]. The well-known 2D valence-arousal (V-A) emotion model [3] (Fig. 1) is more precise and general than the discrete emotion states. The V-A emotion model allows a smooth passage from one state to another in an infinite set of values. But the actual affective recognizers usually use a discrete set of typical emotion states depicting the affective experiences. It is desirable to define the typical emotion state areas in the 2D plane of the V-A model for unifying the two main emotion model classes (discrete and dimensional). Another issue of the V-A model is that it does not allow explicitly expressing the intensity of the emotion state. Moreover, it does not cover the personalization issue. The affective experiences between the individuals are inevitably different. It is necessary to find an effectively modeling approach to reflect the audience's personal emotion information.

For convenience of the discussion, the formal descriptions of modeling the emotion space are given as follows: Let $S = \{e \mid e(v,a) \in R \times R, -1 \le v,a \le 1\}$ be the V-A plane, where $e(v,a)$ is one of the affective states (Fig. 1). $\underset{\sim}{E}_i \subseteq S (i = 1,\ldots,k)$ are the typical fuzzy emotion subspaces, which can be expressed as $\underset{\sim}{E}_i = \int_{e \in S} \underset{\sim}{E}_i(e)/e \ \ (i = 1,\ldots,k)$.

$T = \{x_1, x_2, \ldots, x_n\} (n \in N)$ is the training set of video clips, the corresponding emotion coordinates annotated by the audience (i.e. the points in the V-A plane) are $S_T = \{e_1, e_2, \ldots, e_n\}$. The coordinate value (v_i, a_i) of $e_i \in S_T$ can be collected through our software tool developed according the theory of emotional psychology. The modeling objectives are defining the typical fuzzy emotion subspaces $\underset{\sim}{E}_i$ in V-A plane S and determining their affective membership functions $\underset{\sim}{E}_i(e)$, where $i = 1, \ldots, k$.

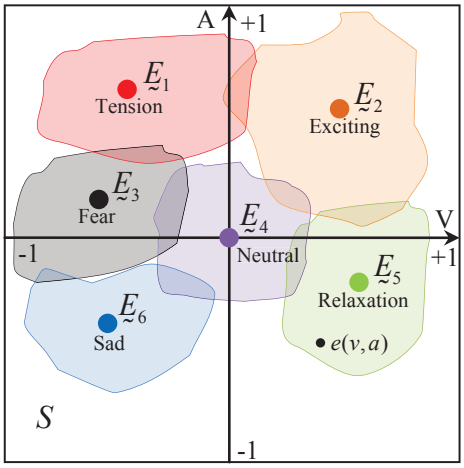

Fig. 1. V-A emotion space and the typical emotion subspaces

2.1. FCM Based Division of V-A Emotion Space

The FCM [8] is adopted to define k typical emotion subspaces $\underset{\sim}{E}_i (i = 1, \ldots, k)$ based on $S_T = \{e_1, e_2, \ldots, e_n\}$. Suppose that the cluster centers of the k subspaces $\underset{\sim}{E}_i$ are $c_1, c_2, \ldots, c_k$. We can use a $k \times n$ membership matrix U to express how well every emotion state belonging to the typical subspaces. To accommodate the introduction of fuzzy partitioning U is allowed to have elements with values between 0 and 1. The summation of membership for an emotion state $e_i \in S_T$ is equal to unity:

$$\sum_{i=1}^{k} u_{ij} = 1, \forall j = 1, 2, \ldots, n. \tag{1}$$

We can use U defining the cost function (or objective function) J for FCM:

$$J(U, c_1, \ldots, c_k) = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^{\ m} d_{ij}^{\ 2}, \tag{2}$$

where u_{ij} is between 0 and 1; c_i is the cluster center of the fuzzy typical affective subspace $\underset{\sim}{E}_i$; $d_{ij} = \|c_i - e_j\|$ is the Euclidean distance between ith cluster center and jth affective state; and $m \in (1, \infty)$ is a weighting exponent.

By differentiating $J(U, c_1, \ldots, c_k)$ with respect to all its input arguments, the necessary conditions for Equation (2) to reach its minimum are

$$c_i = \sum_{j=1}^{n} (u_{ij})^m e_j \Big/ \sum_{j=1}^{n} (u_{ij})^m \tag{3}$$

and

$$u_{ij} = 1 \Big/ \sum_{p=1}^{k} \left(d_{ij} / d_{pj} \right)^{\frac{2}{m-1}}. \tag{4}$$

Based on the above discussions, the steps using FCM to define the typical emotion subspaces $\underset{\sim}{E}_i$ can be concluded as follows:

Step 1: Initialize the membership matrix U with random value between 0 and 1 so that the constraints in Equation (1) are satisfied.

Step 2: Calculate k centers of the typical emotion subspaces c_i, $i = 1, \ldots, k$, using Equation (3).

Step 3: Compute the cost function according to Equation (2). Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Compute a new U using Equation (4). Go to step 2.

2.2. GMM Based Membership Function of Fuzzy Emotion Subspace

Suppose the typical emotion subspace $\underset{\sim}{E}_i = \{e_{ij}\}$ $(i = 1, \ldots, k; j = 1, \ldots, s_i)$, where s_i is the number of elements in $\underset{\sim}{E}_i$. To simplify discussion, we restrict that the covariance matrix of the Gaussian PDF can be expressed as: $\Sigma_i = \sigma_i^2 I, i = 1, \ldots, r$, I is unit matrix. In this case, the single Gaussian PDF can be expressed as:

$$g(e; \mu, \sigma^2) = (2\pi)^{-\frac{d}{2}} \sigma^{-d} \exp\left[-\frac{(e-\mu)^T (e-\mu)}{2\sigma^2} \right] (e \in \underset{\sim}{E}_i). \tag{5}$$

The PDF of the Gaussian Mixture Model (GMM) is:

$$p(e) = \sum_{i=1}^{r} \alpha_i g(e; \mu_i, \sigma_i^2), \tag{6}$$

where r is the number of the Gaussian PDF and $\alpha_1, \alpha_2, \ldots \alpha_r$ should satisfy the constraint condition: $\sum_{i=1}^{r} \alpha_i = 1$.

Based on the above discussions, the steps using GMM to determine the membership functions $\underset{\sim}{E}_i(e)$ of the fuzzy typical emotion subspaces $\underset{\sim}{E}_i$ can be concluded as follows:

Step 1: Initialize the parameter vector:

$$\theta = \{\alpha_1, \ldots, \alpha_r; \mu_1, \ldots, \mu_r; \sigma_1^2, \ldots, \sigma_r^2\}. \tag{7}$$

Step 2: Using θ to Calculate:

$$\beta_j(e) = \alpha_j g(e; \mu_j, \sigma_j^2) \Big/ \sum_{i=1}^{r} \alpha_i g(e; \mu_i, \sigma_i^2), j = 1, \ldots, r. \tag{8}$$

Step 3: Compute the new μ_j according to

$$\tilde{\mu}_j = \sum_{k=1}^{s_i} \beta_j(e_k) e_k \Big/ \sum_{k=1}^{s_i} \beta_j(e_k). \quad (9)$$

Step 4: Compute the new σ_j according to

$$\tilde{\sigma}_j^2 = \frac{1}{d} \sum_{k=1}^{s_i} \beta_j \left(e_k - \tilde{\mu}_j\right)^T \left(e_k - \tilde{\mu}_j\right) \Big/ \sum_{k=1}^{s_i} \beta_j(e_k) \quad (10)$$

Step 5: Compute the new α_j according to

$$\tilde{\alpha}_j = \frac{1}{s_i} \sum_{k=1}^{s_i} \beta_j(e_k) \quad (11)$$

Step 6: Let $\tilde{\theta} = \{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_r; \tilde{\mu}_1, \ldots, \tilde{\mu}_r; \tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2\}$. Stop if $\left\|\theta - \tilde{\theta}\right\|$ is below a certain tolerance value; else let $\theta = \tilde{\theta}$ and go to step 2.

2.3. Video Affective Content Recognition

Based on the proposed emotion space, we introduce *the maximum membership principle* to recognize the typical video affective content and *the threshold principle* to represent and recognize the compound video affective content or non-typical affective content.

***The maximum membership principle for video affective content recognition*:** Suppose there are k fuzzy typical emotion subspaces $\underset{\sim}{E}_1, \underset{\sim}{E}_2, \ldots \underset{\sim}{E}_k$ (i.e. k fuzzy models), which constitute the fuzzy model base. If $\forall e_0 \in S$, $\exists i_0 \in \{1, 2, \ldots, k\}$, satisfy $\underset{\sim}{E}_{i_0}(e_0) = \max_{i=1}^{k} \{\underset{\sim}{E}_i(e_0)\}$, then $e_0 \in \underset{\sim}{E}_{i_0}$.

***The threshold principle for video affective content recognition*:** Suppose there are k fuzzy typical emotion subspaces $\underset{\sim}{E}_1, \underset{\sim}{E}_2, \ldots \underset{\sim}{E}_k$ (i.e. k fuzzy models), which constitute the fuzzy model base. For a fixed threshold of emotion membership $\alpha \in [0,1]$, if $\exists m_1, m_2, \ldots, m_t (t \le k)$, $\forall e_0 \in S$ satisfy $\underset{\sim}{E}_{m_j}(e_0) \ge \alpha$ (where $j = 1, 2, \ldots, t$), then $e_0 \in \bigcup_{j=1}^{t} \underset{\sim}{E}_{m_j}$ (i.e. e_0 is a compound emotion state); else if $\max_{i=1}^{k} \{\underset{\sim}{E}_i(e_0)\} < \alpha$, then e_0 belongs to non-typical emotion state, and e_0 can be expressed by its coordinate (v_0, a_0).

3. EXPERIMENTAL RESULTS

3.1. Video Affective Content Database

There is still lacking a standard video affective content database (VACDB for short) to validate our proposed video affective semantic space. We chose movies to create our VACDB because they have rich affective content. Based on the statistical figures of the IMDB [9], we select 46 typical movies as the source of affective video clips in VACDB. The total length of these movies is 84 hours 43 minutes 4 seconds. These movies can be classified into 6 genres: 9 animations, 10 actions, 11 dramatics, 7 science fictions, 3 horrors and 6 comedies. The ground truth for the 6 typical affective contents, i.e. joy, tension, fear, relaxation, sad and neutral, is manually determined within the 46 movies. If one of the video clips is labeled with the same affective content by at least 6 of 9 researchers, we assign the clips as having one of 6 affective contents. Finally, we select total 1037 video clips from 46 movies to create the VACDB. The total length of the 1037 clips is 10 hours 42 minutes 41 seconds. We choose 4 audiences (A_1, A_2, A_3 and A_4) from different fields, gender, age and backgrounds to record the coordinate values of these 1037 video clips. The coordinate values are recorded with the affective content annotation tool designed by us. We don't tell the audience the emotion labels of these movie clips. The audiences watch every movie clip and record its emotion coordinate based on their affective experience, i.e. the intensities of their valence and arousal (V-A). Finally, we get 4 sets of the 1037 emotion coordinates of these movie clips, which are used to validate our proposed emotion space.

3.2. Experimental Results

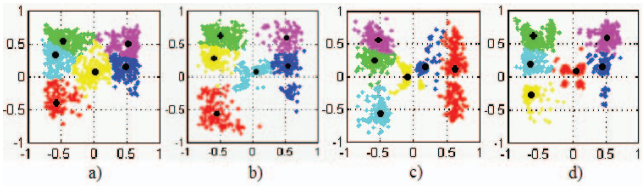

Fig. 2. The 4 partitioning results of V-A plane based on FCM. (a), (b), (c) and (d) are plotted according to the coordinate values recorded by A_1, A_2, A_3 and A_4 respectively.

At the first step of our experiment, the inputs of the FCM are the 1037 coordinates. The number of the typical affective sunspaces (i.e. clusters) is 6. The weighting exponent m of the cost function J is 2. The tolerance value and the threshold of the iterations in FCM are 100 and 10^{-5}. Fig. 2 (a), (b), (c) and (d) demonstrate the division results of the V-A plane. The centers, borders, shapes and densities of the 6 typical affective subspaces in Fig. 2 (a), (b), (c) and (d) are different, which proves that our modeling method can characterize the personalized emotion experience of the audiences.

The second step is using the GMM to formulate the membership functions of the subspaces. Fig. 3 demonstrated the 6 membership functions determined by the GMM in our experiments (the coordinate values for modeling are recorded by A_1 and the membership functions based on the coordinate values recorded by A_2, A_3 and A_4 are similar). Bringing the 1037 coordinate values into these 6

membership functions, the average recognition rate is up to 97.8% based on *the maximum membership principle* for video affective content recognition.

Tab. 1. Labels describing the content of the test sequence excerpted from the movie "*Saving Private Ryan*" [2]

Segment	Content description
1	US Army HQ, typists make letters for soldiers' families, male voices reading the letters
2	Colonel's office, finding out about private Ryan
3	Bad news brought to Ryan's home
4	General's offices, decision is being made to search for Ryan
5	Omaha Beach, US Army HQ, an officer gets the order to search for Ryan
6	Omaha Beach, US Army HQ, preparation for the search action
7	Beginning of the action, walking through the fields
8	It starts to rain and gets dark, the suspense grows, the actual beginning of the action

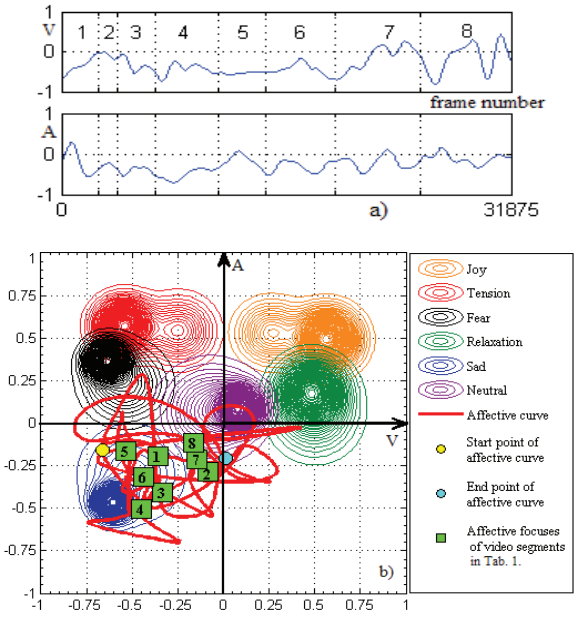

Fig. 3. a) V/A curves (1-8 are the labels of video segments in Tab. 1); b) GMM based membership functions of typical emotion subspaces and the *affective curve* generated from "*Saving Private Ryan*" within the V-A emotion space.

We use the V/A affective curves (Fig. 3 (a)) proposed by Hanjalic [2] to estimate the emotion coordinates (v_i, a_i) of the corresponding video frames (Tab. 1). Relating the affective curve in Fig. 3 (b) to the typical emotion subspaces, we can recognize video affective content intuitively.

3.3. Discussions

The improved emotion space has four prominent characteristics: 1) The model originates from the well-known 2D V-A emotional model, which is a continuous model and can express infinite emotion states in V-A plane. By introducing the typical fuzzy emotion subspaces, the model can represent both discrete and continuous emotion states in the V-A plane; 2) The model can explicitly express the intensities of the typical emotion states according *the maximum membership principle*; 3) The model can describing the compound emotion states according *the threshold principle*; 4) At every step of modeling, the inputs rely completely on the affective experiences recorded by the audiences. The centers, borders, shapes and densities of these subspaces can truthfully reflect the emotional tendencies of audiences (Fig. 2), which means that the model can also be used to cover the personalization issues.

4. CONCLUSION

An improved V-A emotion space for video affective content representation and recognition is presented. By introducing the typical fuzzy emotion subspace, the model can represent discrete affective states in the continuous V-A plane. Combined with the model proposed by Hanjalic [2], we can relate the computed V-A values to the typical emotion states, making it convenient to be utilized for video affective content indexing. To understand video affective content more accurately, it is desirable to design a set of video affective features to relate the video clips with the emotion coordinate values based on the proposed emotion space.

5. ACKNOWLEDGMENT

We gratefully acknowledge the supports received from the National Science Foundation of China (60703049), and Wuhan "Chen Guang" Foundation for Young Scientists of China (200850731353).

6. REFERENCES

[1] R. Picard. *Affective Computing*. Cambridge. Cambridge: MIT Press, 1997.
[2] A. Hanjalic, L.Q., Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1): 143-154, 2005.
[3] R. Plutchik. *The Psychology and Biology of Emotion*. Harper Collinns Ed., New York, 1994.
[4] J.A., Russell. The circumplex model of affect. *Journal of Personality and Social Psychology*, 39: 1161-1178, 1980.
[5] P., Ekman. *An Argument for Basic Emotions, Cognition and Emotion*. 169-200, 1992.
[6] A., Ortony, G.L., Clore, and A., Collins. *The cognitive structure of emotion*. Cambridge University Press, Cambridge, UK, 1988.
[7] A., Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology: Developmental, Learning, Personality, Social*, 14, 1996.
[8] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.
[9] http://www.imdb.com/chart/top.