Affective Visualization and Retrieval for Music Video

Shiliang Zhang, Qingming Huang, *Senior Member, IEEE*, Shuqiang Jiang, *Member, IEEE*, Wen Gao, *Fellow, IEEE*, and Qi Tian, *Senior Member, IEEE*

***Abstract*—In modern times, music video (MV) has become an important favorite pastime to people because of its conciseness, convenience, and the ability to bring both audio and visual experiences to audiences. As the amount of MVs is explosively increasing, it has become an important task to develop new techniques for effective MV analysis, retrieval, and management. By stimulating the human affective response mechanism, affective video content analysis extracts the affective information contained in videos, and, with the affective information, natural, user-friendly, and effective MV access strategies could be developed. In this paper, a novel integrated system (i.MV) is proposed for personalized MV affective analysis, visualization, and retrieval. In i.MV, we not only perform the personalized MV affective analysis, which is a challenging and insufficiently covered problem in current affective content analysis field, but also propose novel affective visualization to convert the abstract affective states intuitive and friendly to users. Based on the affective analysis and visualization, affective information based MV retrieval is achieved. Both comprehensive experiments and subjective user studies on a large MV dataset demonstrate that our personalized affective analysis is more effective than the previous algorithms. In addition, affective visualization is proved to be more suitable for affective information-based MV retrieval than the commonly used affective state representation strategies.**

***Index Terms*—Affective content analysis, affective visualization, dimensional affective model, support vector regression.**

I. Introduction

As an increasingly popular technique, affective video content analysis is designed to be an intelligent solution for the problems caused by the explosively increasing video data. Through combining psychological basis and computer science, affective video content analysis identifies the emotional information in videos by extracting affective features and fusing those features in the established affective models [1]–[16]. Compared with the classic content analysis algorithms, affective content analysis combines more psychological basis and are more user-centric. Intuitively, the applications based on it could be more friendly and natural.

Currently, in the academic world, many works have been reported on affective music and movie content analysis [1]–[16]. Music video (MV), which combines the features of music and movies, is also an important entertaining media form. Furthermore, the astonishing increasing number of MVs and the ubiquitousness of mobile sets such as cell phones and music players like iPods and Zune, which can play MVs, have made MV more popular and widely spread. Intuitively, both the fast increasing number and importance of MV have proposed requirements for more efficient and intelligent MV access. Nevertheless, those challenges have not been fully covered in current video content analysis community.

As for the industrial field, most of the commercial music and MV retrieval systems including Shazam, Apple Genius, Pandora, and Google Music, as well as the ones integrated in mobile sets such as iPhone, iPod, and Zune present well-designed interfaces, promising performances, and appealing functions. Notwithstanding their great success, those systems implement retrieval tasks partly according to the metadata such as Artist, Title, or Album. For example, Pandora returns users' favorite songs based on their favorite Artists or Styles. Shazam and iPod allow users to retrieve music with Artist, Album, and Track information. Generally, the traditional metadata-based retrieval presents limitations in three aspects: 1) the metadata describe certain numbers of MVs, rather than each individual MV, resulting in their inaccuracies in describing MVs, e.g., the Artist: "Michel Jackson" corresponds to hundreds of songs with different styles; 2) the metadata such as Artist or Album need to be manually or semi-manually collected and input, which is expensive for large-scale MV databases, in terms of both efficiency and accuracy; and 3) since most users' preferences and requests are originally abstract concepts, the metadata could be too concrete to straightforwardly respond users' queries in some cases. For example, when a user is sad and tired, he/she might want to find some happy and energetic MVs. Traditionally, the user has to do this by MV preview or just selecting some familiar MVs. Obviously, this process is time consuming and ineffective. Recently, Google Music has incorporated a new feature, which allows users to retrieve music according to the rhythm, pitch, and timber information [17]. This new component, although it has brought users novel music retrieval experiences, only covers a small-scale music database. Briefly, developing a novel MV access strategy has become an important topic for both industrial and academic communities.

In this paper, we seek a novel MV access strategy based on the affective content analysis technique. Compared with the traditional methods, affective information based MV access presents advantages in four aspects: 1) affective states such as happy or sad, which are closely related to user experiences, could be informative and accurate descriptions for each MV; 2) affective content analysis can be implemented automatically in large-scale

Manuscript received December 09, 2009; revised March 23, 2010; accepted April 19, 2010. Date of current version September 15, 2010. This work was supported in part by the National Natural Science Foundation of China under Grant 60833006 and Grant 60702035, the National Hi-Tech Development Program under the 863 Program of China under Grant 2006AA01Z117, and by the Beijing Natural Science Foundation under Grant 4092042. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nicu Sebe.

S. Zhang, S. Jiang, and W. Gao are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: slzhang@jdl.ac.cn).

Q. Huang is with the Graduate University, Chinese Academy of Sciences, Beijing 100190, China.

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2059634

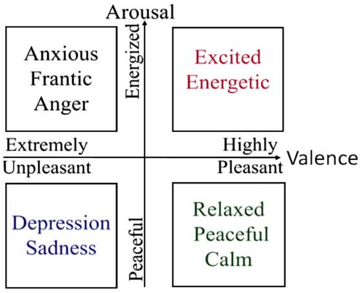


Fig. 1. Illustration of the dimensional affective model.

MV databases; 3) users can retrieve MVs intuitively with abstract concepts without converting them into concrete Titles, Artists, or Styles, e.g, with affective information, users could find energetic MVs even if they do not know which Album contains such MVs; and 4) affective content analysis and traditional metadata describe MVs from different aspects—therefore, they can be complemented to each other. For example, affective information-based MV retrieval can be integrated into current commercial systems to provide users more intelligent and interesting MV services. Because our current work focuses on affective analysis, no metadata information is taken into consideration in this paper. Nevertheless, it should be noted that incorporating more information is certainly promising for more effective and friendly MV retrieval strategies. This will be our future work.

Since computers are employed to identify emotional information contained in videos, psychological theories should be the necessary basis for designing computer affective models. For the moment, the most popular psychological model in affective content analysis is the Arousal-Valence model (Fig. 1) proposed by Thayer [18]–[21], which is also called the dimensional affective model. In this model, human affective responses are represented using two basic components: Arousal and Valence. The Arousal, ranging from “energized” or “excited” to “calm” or “peaceful,” describes the intensity of affective experience. The Valence ranging from “highly pleasant” to “extremely unpleasant” typically denotes the level of “pleasure” that is related to a given affective state [9]. By dividing abstract affective states into two components, the Arousal-Valence model presents significant features: 1) complicated emotions can be expressed by combining A (Arousal) and V (Valence) in different ways and 2) this model is generic. Since A and V can be bridges between emotions and affective features, the dimensional affective model could be applied to different video genres with different Arousal and Valence modeling algorithms. Due to its important features, dimensional affective model is utilized in our work for affective modeling and representation.

To identify the affective states as well as to achieve affective information based MV retrieval with the dimensional affective model, the following problems need to be taken into consideration:

1) Valid affective feature extraction.
2) Affective modeling which bridges the gaps between low level features and affective states (i.e., the affective gaps between affective features and Arousal and Valence).
3) Personalized affective analysis. Users’ understandings of affective states are various, so the computer affective models should be properly established to take users’ personalities into consideration.
4) Affective state representation. Since the affective states are abstract, a user-friendly GUI that can reveal affective states into intuitive forms is desired to put the MV affective analysis into practical use.

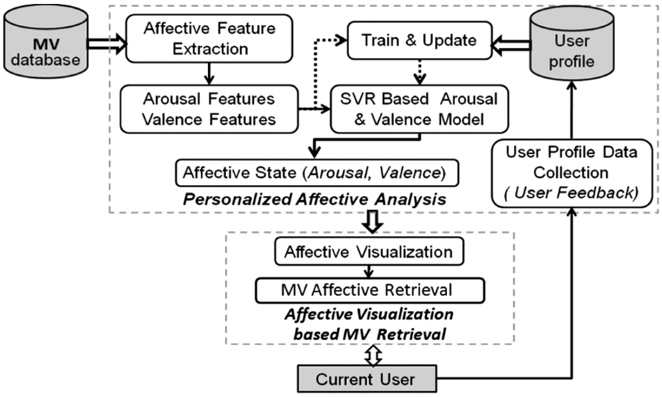

MV database
Affective Feature Extraction
Train & Update
User profile
Arousal Features Valence Features
SVR Based Arousal & Valence Model
Affective State (*Arousal, Valence*)
Personalized Affective Analysis
User Profile Data Collection
(User Feedback)
Affective Visualization
MV Affective Retrieval
Affective Visualization based MV Retrieval
Current User

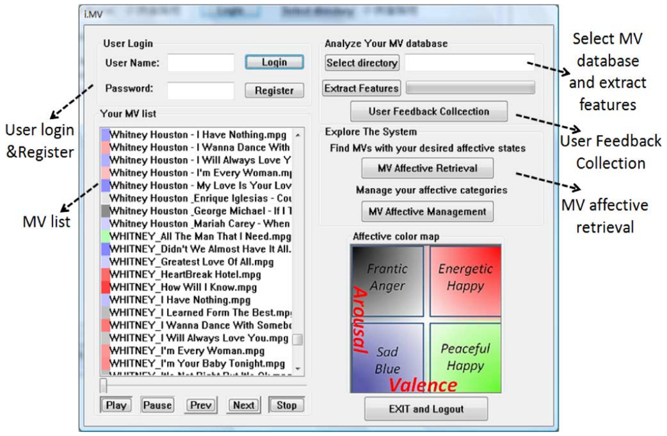
Fig. 2. Framework for i.MV.

The first two problems have already been discussed in the literature [1]–[16]. However, the third and fourth have not been fully studied. Our framework shown in Fig. 2 is proposed targeting the four problems. In the *Personalized Affective Analysis* module, we first extract affective features by referring to related work, musical and cinematographic studies. Second, machine learning algorithms are applied to bridge the affective gaps by combining the psychological dimensional affective model [18]–[21] and support vector regression (SVR) [22]. Third, to achieve personalized affective analysis, a user interface is constructed to let users play MVs and provide their feedbacks. The two SVR-based affective models (i.e., Arousal model and Valence model) are then trained with the collected user feedbacks to be more personal and effective. Finally, the extracted affective features are fed into the trained affective models to get the personalized affective states. In the *MV Retrieval* module, the novel Affective Visualization [23] is proposed to present affective states in intuitive forms, and thus enable effective and user-friendly MV retrieval. A screen shot of the i.MV user interface for user logging in is shown in Fig. 3.[1]

Our contributions and distinctions from previous works are summarized as follows.

- Personalized affective analysis is proposed and achieved by integrating users’ personal affective understandings into affective models. The proposed personalized affective analysis is proved more effective than the traditional ones [9]–[11].
- Novel intuitive Affective Visualization [23] is proposed to present the abstract affective states intuitively to users. The advantages of Affective Visualization are clearly indicated from the comparisons with the previous affective state representation methods [1]–[16].
- i.MV is successfully constructed as a prototype system for MV applications by integrating the novel techniques proposed in this paper. To the best of the authors’ knowledge,

[1] The demo about i.MV can be downloaded from http://www.jdl.ac.cn/en/project/mrhomepage/iMTV-demo(TMM).avi.



Whitney Houston - I Have Nothing.mpg
Whitney Houston - I Wanna Dance With
Whitney Houston - I Will Always Love Y
Whitney Houston - I'm Every Woman.mp
Whitney Houston - My Love Is Your Lov
Whitney Houston _Enrique Iglesias - Cou
Whitney Houston _George Michael - If I T
Whitney Houston _Mariah Carey - When
WHITNEY_All The Man That I Need.mpg
WHITNEY_Didn't We Almost Have It All.
WHITNEY_Greatest Love Of All.mpg
WHITNEY_HeartBreak Hotel.mpg
WHITNEY_How Will I Know.mpg
WHITNEY_I Have Nothing.mpg
WHITNEY_I Learned Form The Best.mp
WHITNEY_I Wanna Dance With Somebo
WHITNEY_I Will Always Love You.mpg
WHITNEY_I'm Every Woman.mpg
WHITNEY_I'm Your Baby Tonight.mpg

Fig. 3. i.MV user interface.

this is the first of its kinds in MV affective analysis. It has significant potential applications.

The remainder of this paper is organized as follows. Section II introduces the related work. Section III describes the extracted affective features. Section IV presents our proposed personal affective analysis. Affective Visualization based MV retrieval is presented in Section V. Experiments and analysis are discussed in Section VI. Section VII concludes this paper.

II. Related Work

As a newly developing research area, a great deal of attention has been paid to affective video and audio content analysis, and many related works have been reported in recent years. Generally, the existing affective content analysis works can be summarized into two categories: *categorical affective content analysis* and *dimensional affective content analysis*.

In *categorical affective content analysis*, emotions are commonly discrete and belong to one of a few basic categories, such as "fear," "anger," or "happy. " Consequently, many classifiers are adopted for affective analysis. For example, Moncrieff [1] uses four audio clues to detect horror events in videos. Precision rates of 63%, 89%, 93%, and 80% are achieved for "startle," "apprehension," "surprise," and "apprehension" to "climax", respectively. Kang [2] trains two Hidden Markov Models (HMMs) to detect affective states including "fear," "sadness," and "joy" in movies. The classification performances of 81.3%, 76.5%, and 78.4% are reported, respectively, for the three affective states. Similarly, a four-state HMM is adopted in the work of Xu [5] to classify audio emotional events such as laughing or horror sounds in comedy and horror videos. Hierarchical classification framework is utilized in the work of Lu *et al.* [8] for classical music affective content analysis. They extract three types of audio features to represent the mood in music: intensity, timber, and rhythm. The authors classify music segments into four mood categories: contentment, depression, exuberance, and anxious/frantic with a hierarchical classifier. Categorical affective content analysis is more suitable to be called as affective classification. Although the flexibility of these methods is limited, they are simple, and easy to build.

Dimensional affective content analysis commonly employs the dimensional affective model for affective state computation. One representative work is reported by Hanjalic and Xu [9]–[11]. In their work, the Arousal-Valence (A-V) model [18]–[21] is used to express affective states in videos. Modeling Arousal and Valence using linear feature combinations, the authors can obtain the Arousal and Valence values of different video segments and draw affective curves in the A-V space. Consequently, the affective states of video segments can be visualized. A-V model is also utilized in our previous work for MV affective state computation [24]. In [24], linear feature combination is adopted to compute the A and V values of MVs. Then in the A-V space, MVs with similar affective states are clustered together for automatic MV category generation. Another representative work is reported by Arifin, et al. [12]. They adopt the Pleasure Arousal Dominance [25] model (i.e., a dimensional affective model similar to the A-V model) for affective state computation. Different from the Arousal and Valence modeling methods proposed by Hanjalic *et al.*, this work takes the influences of former emotional events and larger emotional events into consideration by employing the Dynamic Bayesian Network for Pleasure, Arousal, and Dominance modeling. The authors finally classify affective states into 6 categories based on the computed Pleasure, Arousal and Dominance values and evaluate the proposed method using 23 test videos. Improvement of 9% over the work of Hanjalic [9] is reported. Similarly, our current work can also be categorized as the dimensional affective content analysis. However, differently, we integrate user's personal affective understandings into the affective models. In addition, novel intuitive Affective Visualization is proposed to serve as the bridge between users and affective analysis, which has proven to be a better affective state representation method than the previous ones.

Most of the reported works on affective analysis focus on the first two problems mentioned in Section I. The third problem (i.e., *personalized affective analysis*) has not been fully investigated and studied. In fact, personalization is gaining more attention in user-oriented applications nowadays. Lots of proposed multimedia systems such as: *Video Scout* [26], *eMediate* [27], *MAGIC* [28], *IMS-TV* [29], *PTVplus* [30], *WebMate* [31], and *SiteIF* [32] have integrated personalization to achieve better performance and usability. An overview of current personalization techniques can be found in [33] and [34]. Among current personalization techniques, feedback and profiling are effective and commonly used for capturing user's personalities (i.e., preferences, intentions, or interests). For example, the personal TV listing system: *PTVplus* [30] exploits user profiles including the user's viewing behavior, playback history, and implicit feedbacks to learn user's preferences and presents personalized daily TV guide. In some webpage recommender systems such as *WebMate* [31] and *SiteIF* [32], the user profile is denoted as a bag of keywords to represent the user's preferences and is updated with both explicit and implicit feedbacks. In [35], Agnihotri *et al.* present a framework for generating personalized video summaries with the user profile which reflects the user's personality traits. More introductions about learning to personalize and user profile acquisition can be found in [36] and [37]. In summary, most of the proposed personalization techniques focus on capturing the different preferences and requirements among users. Differently, our affective analysis learns and integrates the personalized affective understandings of users. Experimental results illustrate that learning personalized understandings is important for improving the affective video content analysis.

As for the fourth problem (*affective state representation*), most reported works on affective analysis express affective states by classifying them into predefined categories described by affective labels such as "happy," "sad," or "frantic" [1]–[8], [12]–[16] (differently, Hanjalic and Xu [9]–[11] visualize the affective states as 2-D curves in A-V space, but individual curves are not sufficiently informative and intuitive for practical applications). Affective labels are simple and effective in expressing common and specific emotions. Nevertheless, they are insufficient for affective information based MV retrieval in that: 1) since the predefined affective categories are fixed, users can only retrieve MVs containing certain affective states—as a result, the flexibility and accuracy of such representation is limited and 2) language is not sufficiently descriptive to abstract and continuously varying emotions. For example, we often cannot find proper words to express our feelings, and different people may have different understandings to the same words describing certain emotions. To solve these problems, we propose the intuitive Affective Visualization to convert abstract affective states into intuitive forms.

In the following sections, we will introduce our personalized affective analysis and Affective Visualization in detail.

III. Affective Feature Extraction

Since affective features are the important basis for our affective modeling, we proceed to introduce the affective features applied in our work. Generally, the visual contents of MVs are carefully designed by artists to coordinate with the music. Consequently, both the audio and visual contents are used for affective feature extraction. We extract affective features according to the music theory, cinematography studies and related work on affective analysis. The validity of our extracted features will be tested in Section VI.

A. Arousal Feature Extraction

Arousal represents the intensity of affective states [18]–[21]. The features extracted for Arousal include: *motion intensity*, *short switch rate*, *zero crossing rate*, *tempo*, and *beat strength*.

Motion Intensity: Motion intensity which reflects the smoothness of transition between frames is a commonly used Arousal feature [9]–[11][14]. In our work, motion intensity is acquired based on the motion vectors extracted from MPEG streams. Motion intensity between frames is first computed. Then, the final motion intensity MI is computed and normalized between 0 and 1 with

$$\mathrm{MI}' = \frac{\left(\left[\left(\overline{\mathrm{MI}} - \mu_{\mathrm{MI}}\right) / 3\sigma_{\mathrm{MI}}\right] + 1\right)}{2}$$
$$\mathrm{MI} = \begin{cases} 1, & \text{if } \mathrm{MI}' > 1 \\ 0, & \text{if } \mathrm{MI}' < 0 \\ \mathrm{MI}', & \text{if otherwise} \end{cases} \tag{1}$$

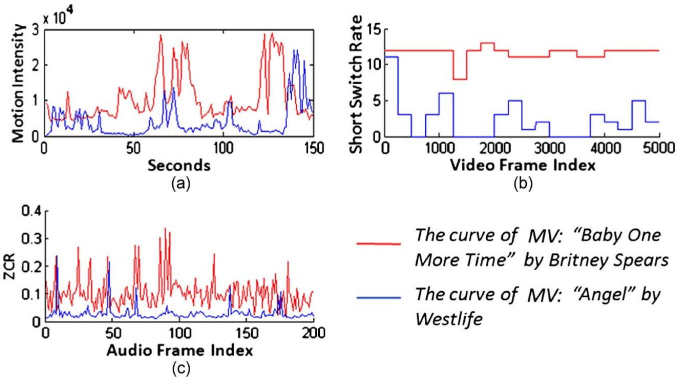


Fig. 4. Comparison of motion intensity, shot switch rate, and ZCR between two MVs.

where $\overline{\mathrm{MI}}$ indicates the mean of inter-frame motion intensity within each MV before normalization, u_{MI} and σ_{MI} denote the mean and standard deviation of motion intensity computed with all MVs in the database, respectively. The same normalization is also applied to other features. Equation (1) is the standard form of Gaussian normalization. It is utilized to format the extracted features into more reasonable distributions. Note that u_{MI} and σ_{MI} can be updated when the MV database is changed.

Shot Switch Rate: Shot is a powerful tool for directors to control the tempo of videos. The duration of shot is used as an important feature describing Arousal in related work [9]–[11]. We take the normalized average shot switch number within each video segment with length of 250 frames as the shot switch rate feature (SSR).

Zero Crossing Rate (ZCR): ZCR is a widely used feature in audio signal analysis. It could be utilized to distinguish harmonic music, speeches and environmental sounds [38]. Our experiments show that intense music generally presents higher ZCR value than the smooth music.

In Fig. 4, the comparisons of Arousal features between two MVs are presented. The red curve stands for "Baby One More Night," which is an energetic MV, while the blue curve denotes "Angel," which is a smooth and sad MV.

Rhythm Based Features, Tempo and Beat Strength: Rhythm is an important characteristic of music, and artists frequently employ rhythm to express their emotions. In general, three aspects of rhythm are closely related with human affective experience: *tempo*, *beat strength*, and *rhythm regularity* [8]. In this paper, we extract the rhythm-related features based on the music onset detection.

Onset detection: onset in music is caused by the instruments like drums which produce high energy and salient sounds [39]. In our onset detection, the audio signal is first divided into five subbands. Each is then smoothed with a Gaussian window. Since onsets are usually shown as energy peaks, we set a threshold to filter the signal with low energy to zero. Then, the local maximums of the signal can be detected as onsets. After computing the first derivative of the audio signal, we confirm the positions of onsets by searching the zero crossing points in the falling edges of the signal.

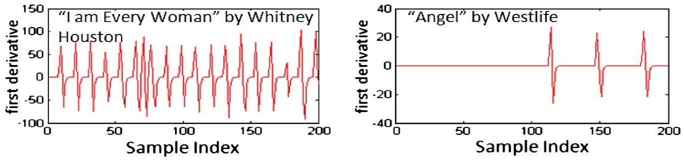


Fig. 5. Comparison between two audio signals' tempo.

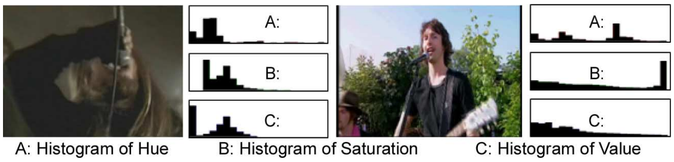


Fig. 6. Comparisons of H, S, and V histograms between two MVs. (a) Histogram of hue. (b) Histogram of saturation. (c) Histogram of value.

Tempo and beat strength calculation: the average onset number within 5 seconds of audio signal is taken as the tempo of the audio (a comparison of tempos between two MVs is shown in Fig. 5). Beat strength is obtained by averaging the audio energy at the onsets' positions. Tempo and beat strength on each subband are fused with linear combinations and then normalized to get the final tempo and beat strength features.

B. Valence Feature Extraction

Valence represents the type of affective state. In this paper, *lighting*, *saturation*, *color energy*, *rhythm regularity,* and *pitch* are extracted to describe Valence.

Lighting: In cinematography, lighting is a powerful tool, used specifically for the purpose of affecting viewers' emotions and establishing the mood of a scene. An abundance of bright illumination is frequently used to generate the lighthearted atmosphere while dim illumination is frequently selected for the opposite purposes [40], [41]. With the histogram containing N bins on V component of hue, saturation, and value (HSV) color space, we calculate the lighting feature L with

$$L^k = \sum_{j=M+1}^{N} v_j \cdot (j - M) - \sum_{j=1}^{M} v_j \cdot (M + 1 - j) \quad (2)$$

(2)where j denotes the index of the histogram bin, v_j is the value of the jth bin of the Value histogram, and k stands for the frame index.

In (2), the lower M bins represent the dark lighting components and the rests denote the bright ones. N is experimentally set to 20. The video frames in MV are commonly dark, which makes the sum of the lower ten bins generally larger than the sum of the upper ten. Consequently, we set M to 7 rather than 10. The final lighting feature L is obtained by computing and normalizing the mean of L^k.

Saturation: It is well known that sad or frightening videos commonly present gray frames which indicate low color saturation. The saturation feature S is calculated in

$$S^k = \sum_{j=M+1}^{N} s_j \cdot (j - M) - \sum_{j=1}^{M} s_j \cdot (M + 1 - j) \quad (3)$$

where s_j denotes the value of the jth bin of the Saturation histogram. N and M are experimentally set to 20 and 10, respectively. The final Saturation feature S is calculated with the same method for computing the Lighting feature.

Color Energy: In most joyous videos, the video frames are colorful and bright, while in sad or frightening videos the video colors are faded and gray. The colorfulness of a video is measured by color energy computed with

$$\mathrm{CEng}^k = \sum_{j=1}^{\mathrm{PixelNum}} s_j v_j \Big/ \mathrm{std}(\mathrm{Hist}_H) \cdot \mathrm{PixelNum} \quad (4)$$

where $\mathrm{PixelNum}$ is the number of pixels in the video frame k, s_j and v_j denote the values of the S and V color component of the pixel j, respectively. $\mathrm{std}(\mathrm{Hist}_H)$ returns the standard deviation of the Hue histogram. The final color energy feature $CEng$ is acquired by computing and normalizing the mean of $CEng^k$.

Fig. 6 illustrates the H, S, and V histograms of video frames from two MVs (a frantic MV and a pleasing MV). From the comparisons, it is clear that the above-mentioned three features would be valid for describing the Valence component.

Rhythm Regularity: Regular rhythm is an obvious characteristic of joyous music [8]. The rhythm regularity feature is calculated based on the regularity of onsets' intervals which can be obtained in the onset detection process.

Pitch: Pitch is a popular audio feature for Valence [9]–[11]. We compute pitch on each audio frame and then utilize formula similar as (1) to obtain the final pitch feature.

IV. Personalized Affective Modeling

Based on the extracted Arousal and Valence features, we present our personalized affective modeling in this section. In our implementation, user profile data is first collected. Then, SVR models are trained with the collected user profile to build personalized affective models.

A. User-Profile Data Collection

In order to obtain user-profile data, we build a user interface and record the users' feedbacks in the user profile database. The profile data of user m is defined as follows:

$$\begin{aligned} \mathrm{Profile}_m &= \{P_1, P_2, \ldots, P_K\} \\ P_i &= [\mathrm{MVid}, \mathrm{Aval}, \mathrm{Vval}] \end{aligned} \quad (5)$$

where $\mathrm{Profile}_m$ denotes the profile database of user m and P_i is the ith item in the database. Each profile item is a vector with three components: MV's unique ID (i.e., MVid), user's descriptions about the MV's Arousal and Valence (i.e., Aval, Vval).

When users retrieve and play MVs, if they are not satisfied with the computed affective states, they can input their feedbacks to the system. Users will then be asked to give two scores to describe their opinions about Arousal (1: peaceful; 2: a little peaceful; 3: a little intense; 4: very intense) and Valence (-2: very negative, unhappy; -1: negative, somewhat unhappy; 1: positive, happy; 2: very positive, very happy) of the played MVs (Fig. 7). For example, if a user thinks a MV with high computed Arousal value is not very energetic, the user can set the MV's Arousal to a lower score (1, 2 or 3) through the interface in Fig. 7. The feedback will be recorded and then utilized to

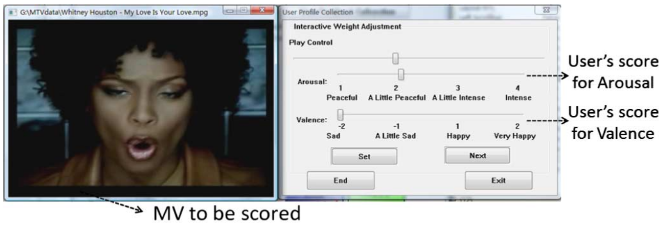


Fig. 7. Interface for user-profile collection.

produce better affective analysis results. We set the feedbacks in four scales because of the following two considerations.

1) Most users want to provide their feedbacks in a simple way.
2) It is not easy for users to confirm affective states' precise positions in affective space but it is easy for them to confirm the rough regions containing them.

With users' feedbacks, new profile items will be generated with the following two steps.

Step 1) Compute the Aval and Vval in the new profile item as

$$\mathrm{Aval} = (\mathrm{Ascore} - 1) \cdot 0.25 + 0.125$$
$$\mathrm{Vval} = \begin{cases} (\mathrm{Vscore} + 2) \cdot 0.25 + 0.125, & \text{if } \mathrm{Vscore} < 0 \\ (\mathrm{Vscore} + 1) \cdot 0.25 + 0.125, & \text{if } \mathrm{Vscore} > 0 \end{cases} \tag{6}$$

where Aval and Vval denote the two elements in profile item (5), Ascore and Vscore denote the user's scores for Arousal and Valence.

Step 2) Assign the MVid in (5) as the ID of the scored MV. If this MV exists in the profile database, overwrite the old profile item; otherwise, add the new item to the profile database.

Consequently, as a user plays MVs and provides feedbacks, the user's profile data will be updated. Then, the personalized affective models can be learned from the collected profile data.

B. Personalized Affective Modeling

The definition of profile based personalized Arousal and Valence modeling is given as follows.

Definition: Profile-based personalized Arousal and Valence modeling is to look for two functions f_V and f_A based on the collected profile data of each user, that map the Valence feature $\mathrm{Feature}_V$ and Arousal feature $\mathrm{Feature}_A$ to real values of Valence and Arousal within 0 and 1, respectively, as

$$f_V : \mathrm{Feature}_V \mapsto [0,1] \quad f_A : \mathrm{Feature}_A \mapsto [0,1].$$

According to this definition, we want to calculate the Arousal and Valence values of each MV based on the underlying relationships between user's affective descriptions and the extracted features indicated by the recorded profile data. Once the functions f_V, f_A can be properly constructed, the affective state of each MV can be computed by feeding their extracted features into the two functions.

Solution to the Problem: To utilize the profile data and learn users' personal affective understandings, we convert the personalized affective modeling into a regression problem, in which a model is trained to give "best fit" of the user's affective descriptions. So, we choose a suitable regression model to achieve this.

SVM enjoys solid theoretical foundations and has demonstrated outstanding performance in many machine learning problems [42]. SVR [22] is based on the SVM theory and presents several advantages, given here.

- SVR provides better prediction on unseen data. It employs the Structural Risk Minimization principle, which is superior to the Empirical Risk Minimization principle in the aspect of generalization [22].
- SVR provides a better solution for the training problem [22], which is important for personalized affective modeling. With effective features and detailed, accurate user affective descriptions, personalized affective models could be constructed by training SVR models.
- The SVR model is psychologically more reasonable. An important basis for previously used linear combination based affective model [9]–[11] is that the relationships between affective features and affective states are linear. This basis is a loose hypothesis because the mechanism of human's affective responses still remains to be further studied by psychologists. Therefore, it is more reasonable to build general nonlinear SVR models for affective state computation.

Due to the above considerations, we build two SVR models for each user: a model for Arousal and the other for Valence.

Train the Default SVR Models: The system does not contain any profile data from the new user. So we train default SVR models for each new user. The MVs suitable for default model training should be the ones to which different users give similar descriptions. Thus, we invited ten users to conduct a user study from which we collected these users' descriptions (scores for Arousal and Valence) for 552 MVs. Then, we ranked the 552 MVs by the times they are similarly described and selected the top 50 ones to generate the default training set. Note that, if more users begin to use i.MV, better default training set can be collected, and thus better default models can be trained. The process for default Arousal SVR model training is given as follows. The default Valence model is trained in similar way.

1) Convert each selected MV's Ascore into Aval with (6) and generate the default training set

$$(\mathrm{Aval}, \mathrm{Feature}_A)_i, \qquad i = 1 \ldots 50.$$

2) The SVR model is established and fivefold cross validation is adopted for parameter selection. Radial basis function (RBF) kernel is experimentally selected as the kernel function for its popularity in pattern recognition problems. Other nonlinear kernels such as Gaussian or Sigmoid can also be adopted.
3) Train the default Arousal model with the selected parameters.

Update the SVR Models: When the system begins to collect more user profile items, the SVR models of each user can be updated. Several rules, given as follows, are proposed to produce new training set by combining the new profile data with the old one.

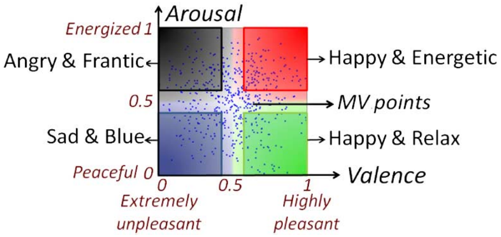


Fig. 8. MV points in dimensional affective space [23].

1) Generate new training items $(\mathrm{Aval}, \mathrm{Feature}_A)$ and $(\mathrm{Vval}, \mathrm{Feature}_V)$. Aval and Vval are acquired with (6).
2) If the newly described MVs exist in the old training set, then overwrite the corresponding old training items with the new ones. Otherwise, add them to the old training set.

Since training new SVR models is time consuming, the models will be updated when the number of newly updated items exceeds a threshold, which is experimentally set as 10 in our implementation. Based on the updated affective models, the affective states of MVs can be computed as 2-D vectors (i.e., Arousal value and Valence value in the range of 0 to 1). Because the affective models are trained to fit the user's affective descriptions, the computed affective states are desired to be more personal and reasonable. The validity of this algorithm will be tested in Section VI.

V. Affective Visualization-Based MV Retrieval

A. *Affective Visualization*

After affective analysis, the affective state of each MV is denoted as a two-dimensional vector which is too abstract to understand intuitively. Consequently, two problems should be considered to achieve user-friendly affective based MV retrieval. i) How to present affective states to users intuitively? ii) How to map users' abstract affective queries back into A-V values accurately? As is discussed in Section II, classifying MVs into predefined categories such as "Happy", "Sad", etc., is not the optimal approach for affective state representation. Consequently, Affective Visualization [23] is proposed as a more effective strategy for affective state representation.

From the psychological points of view, human emotion can be denoted as a 2-D continuous affective space, within which different regions represent different affective states [18]–[21]. Based on this principle, MVs could be visualized as points in the 2-D A-V space (Fig. 8). Meanwhile, the affective states of MVs can be intuitively represented according to their spatial positions in the A-V space. However, similar to the visualization strategy in Hanjalic and Xu's work [9]–[11], such visualization is still not sufficiently informative for retrieval tasks. Therefore, a more informative and efficient solution is desired. Referring to [43], we propose three requirements for the ideal Affective Visualization: *overview*: from the visualized MVs, users should be able to get an overview of the entire MV collection; *affective structure preservation*: the affective relationships between MVs should stay unchanged after visualization; *visibility*: the contents of MV should be visible for convenient MV browsing. According to the three requirements, we propose the Affective Visualization which combines the Video Collage [44] and Dimensional Affective Model.

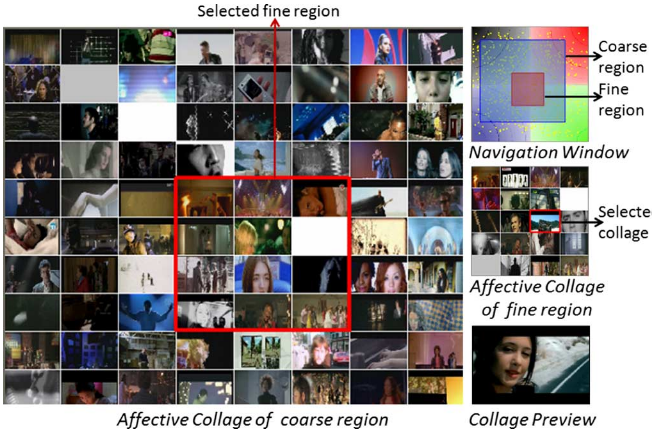


Fig. 9. Overview of affective visualization.

As shown in Fig. 9, our Affective Visualization contains four components: *Navigation Window*, *Affective Collage of coarse region*, *Affective Collage of fine region*, and *Collage Preview*. Navigation Window is constructed based on the affective space for three purposes: 1) mapping MVs into points in the A-V space to give users an intuitive overview of their MV collections; 2) allowing users to select coarse regions in Navigation Window as their coarse affective queries; and 3) indicating users the positions of their selected coarse and fine regions in affective space. As a result, with Navigation Window, users not only can conveniently input their coarse affective queries, but also can be indicated the positions of their selected affective regions. In the selected coarse region, Affective Collage is then adopted to provide users more information. As Fig. 9 shows, Affective Collage consists of video key frames extracted from the most centered MVs in the corresponding regions. From the Affective Collage of coarse region, user can select fine regions, within which he/she can browse, choose, and preview MVs in finer scales. Therefore, combining the affective space and collages, users can intuitively browse their MV collections and conveniently input their affective queries.

Compared with the affective label-based affective state representation [1]–[16], Affective Visualization presents the following advantages.

- Affective states are represented in the continuous space rather than the discrete space. Therefore, more affective states can be presented and there would be no burden to define boundaries for affective categories. Additionally, since human mood and emotion are continuous, Affective Visualization is more psychologically reasonable.
- Affective Visualization is intuitive, informative, and less language-dependent. Because language is not good at describing abstract emotions, the less-language-dependent property makes Affective Visualization more suitable for affective state representation. Meanwhile, the Affective Collage is also informative and helpful for MV browsing.

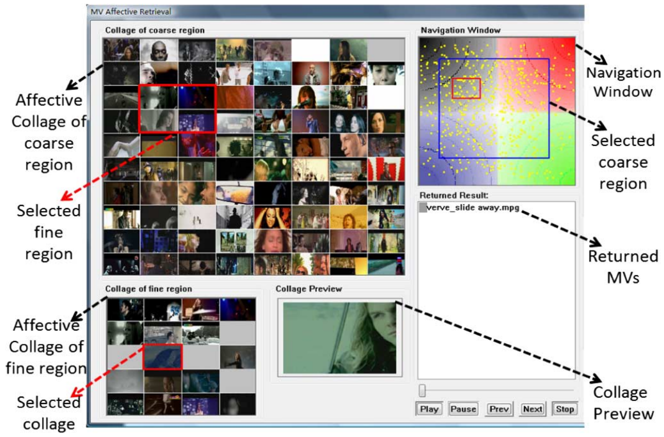


Fig. 10. Interface of i.MV for MV retrieval.

- From psychological perspective, the user's emotions are composed of two components, thus it is theoretically reasonable for users to map their abstract affective queries into the 2-D affective space for MV retrieval.
- Since Affective Visualization needs no textual input, it can be straightforwardly transplanted into mobile sets like cell phones and music players to improve their entertainment properties. Moreover, it will present more significant advantages in the sets equipped with touch-screens.

B. *Visualization-Based MV Retrieval*

Most existing information retrieval techniques require users to convert their demands into concrete words or sentences as queries. It is easy for users to describe explicit objects such as sky or desk with words or sentences. However, abstract concepts like affective states, emotions, or feelings are hard to be accurately described with language. In order to achieve natural affective state representation, accurate affective query input, and intuitive MV collection browsing, Affective Visualization is adopted in i.MV to serve as the bridge between users and the abstract affective states. As shown in Fig. 10, the typical retrieval procedure in i.MV can be summarized as follows.

1) Users first map their abstract queries into certain regions in the 2-D affective space. Then, they drag the mouse and confirm the coarse regions from the Navigation Window.
2) With the help of the contents and positions of Affective Collages in coarse regions, users can drag the mouse and select the fine regions in which they are more interested.
3) Through selecting, browsing, and previewing collages in the fine regions, users can confirm their desired collages. MVs covered by these collages will be returned to users.

Consequently, with the help of Affective Visualization, users would be able to retrieve MVs in an easy and intuitive way. Because our current work focuses on affective information-based MV retrieval, no other cues such as Album or Artist is considered. However, because affective analysis and metadata describe the characteristics of MVs from different aspects, they can be combined for novel and more effective retrieval applications. Accordingly, introducing more information into i.MV is certainly helpful for a better MV retrieval solution. This will be our future work.

TABLE I
OPTIONS, DESCRIPTIONS AND QUANTIFIED GROUND TRUTH

Arousal			Valence		
Score	Description	**Ground Truth**	Score	Description	**Ground Truth**
1	very peaceful	**[0,0.25)**	-2	sad	**[0,0.25)**
2	a little peaceful	**[0.25,0.5)**	-1	a little sad	**[0.25,0.5)**
3	a little intense	**[0.5,0.75)**	1	a little pleasant	**[0.5,0.75)**
4	very intense	**[0.75,1]**	2	very pleasant	**[0.75,1]**

VI. EXPERIMENTS AND ANALYSIS

A. *MV Dataset*

Our dataset consists of 552 MVs of MPEG format (about 25 GB in total and each MV lasts for about 4 min). These MVs are collected through two ways: downloading from the Internet and converting from DVDs. Thus, our dataset presents different resolutions and visual qualities. Moreover, MVs in the dataset are recorded in different languages (English, Chinese, French, Korean, and Japanese), different periods (containing classical songs such as "Tears in Heaven" by Eric Clapton, "Right Here Waiting" by Richard Marx, and the latest ones such as "Girl Friend" by Avril Lavigne, and "What I Have Done" by Linkin Park), and different styles (e.g., country, jazz, rock, and R&B). Therefore, this MV dataset is fairly representative to conduct experiments.

B. *User Study for Profile and Ground Truth Preparation*

In order to capture enough profile data and ground truths to test the proposed techniques and system, we carry out two user studies. The first one, which was introduced in Section IV, is designed to collect default affective model training set. Ten users consisting of one female and nine males aging from 21 to 28 are invited. Each participant is required to score 150 evenly selected MVs. Based on the collected scores, we generate the default affective model training set, which contains 50 MVs and is then used to train the default SVR models.

The second user study is separately designed to collect user feedbacks and personalized ground truths. As shown in Table I, the ground truth is denoted as one of the four equal intervals. For example, if the Arousal ground truth for a MV is [0, 0.25) (i.e., the score for Arousal is 1), then the correct Arousal value of this MV should be within [0, 0.25). In this user study, 27 users consisting of 8 females and 19 males aging from 20 to 31 are invited. They are required to use i.MV for MV retrieval and provide their feedbacks to the system. Before the study, the users are required to use i.MV for at least 30 min to get familiar with it. In the study, if users satisfy with the computed affective states of certain MVs, the ground truth intervals which the computed A and V values within will be considered as the personalized ground truths for corresponding users. For example, suppose the computed Arousal value of a MV is 0.6. If this result satisfies a user, then [0.5, 0.75) will be taken as the personalized ground truth for the user. While, if users are not satisfied with the computed Arousal or Valence, they are required to give scores corresponding to their affective understandings from Table I as feedbacks. Their feedbacks will be recorded and then the personalized ground truths are generated. For instance, if the user scores 1 for Arousal, the interval [0, 0.25) will be taken as the personalized ground truth of Arousal for this user. Meanwhile, the inputted score 1 will also be recorded as the feedback. Each user is required to play 250 MVs; thus, for each user, we can collect his/her personal ground truths for 250 MVs and feedbacks to some of these MVs. The distribution of the collected user profile and groundtruth across different Arousal and Valence values are illustrated in Fig. 11.

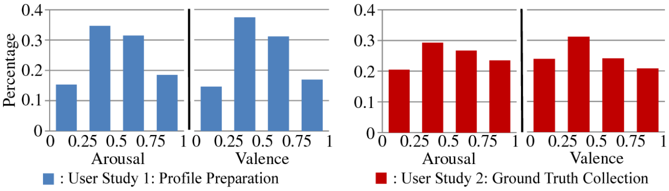


Fig. 11. Distribution of collected user profile and ground truth.

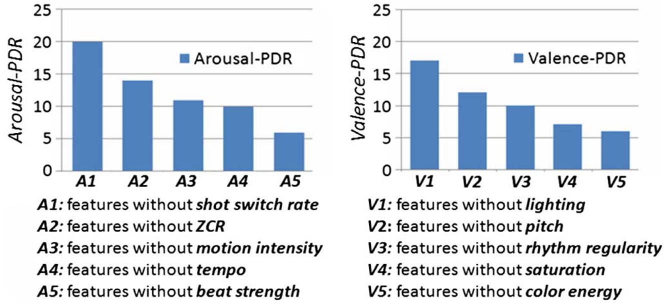


Fig. 12. PDR after removing each affective feature.

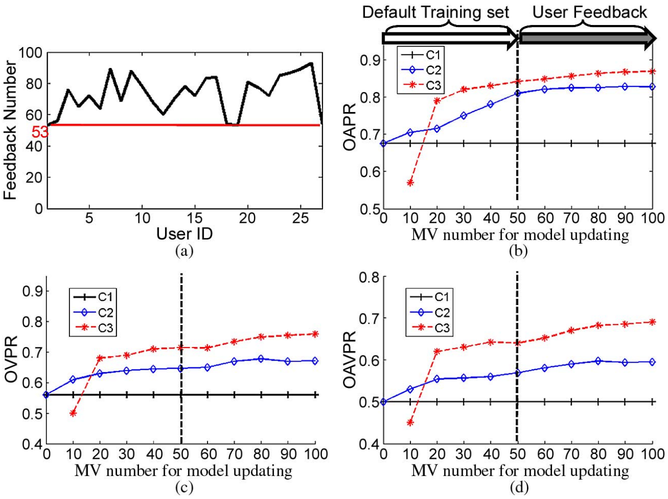


Fig. 13. Performance comparisons between *C1*, *C2*, and *C3*. (a) Each user's feedback number. (b) Comparisons between OAPR. (c) Comparisons between OVPR. (d) Comparisons between OAVPR.

C. Proving the Validity of the Extracted Affective Features

In this experiment, we test a feature's performance by measuring the performance descent rate if it is removed from the complete feature set. For each user, the feedbacks obtained in the first 180 played MVs are used for training the personalized affective models. The ground truths of the other 70 MVs are used to compute the Arousal and Valence precision rate with

$$\mathrm{PR} = \mathrm{CorrectNum}/\mathrm{GTruthNum} \cdot 100\% \tag{7}$$

where PR denotes the precision rate, $\mathrm{GTruthNum}$ is the number of MVs in the ground truth set (i.e., 70), $\mathrm{CorrectNum}$ stands for the number of MVs whose computed affective states (i.e., Arousal or Valence) fall within the ground truth intervals. Different affective models are trained based on the same profile data but with different feature combinations, within which one of the features is removed. Then, the precision descent rate (PDR) caused by removing each single feature can be computed with

$$\mathrm{PDR} = (\mathrm{PR} - \mathrm{PR}')/\mathrm{PR} \tag{8}$$

where PR and PR' denote the precision rate computed with all features and the precision rate computed by removing one feature, respectively. Finally, the average *Arousal-PDR* and *Valence-PDR* of all users caused by removing each feature are computed and shown in Fig. 12.

It is clear from Fig. 12 that the precision rates of both Arousal and Valence are dropped if any one of the extracted affective features is removed. Thus, our affective features can be considered valid. The experimental results show that shot switch rate and lighting are the most important features for Arousal and Valence, respectively.

D. Proving the Validity of Personalized Affective Analysis

In this experiment, we compare our affective model with other reported ones. The compared models are listed as follows.

1) *Comparison1 (C1): classic linear feature combination*. Linear feature combination is a commonly used affective modeling method proposed by Hanjalic and Xu [9]–[11]. Linear feature combination with equal feature weights is used as a baseline for comparison.
2) *Comparison2 (C2): linear feature combination with weight adjustment*. To make the comparisons fair, we introduce a weight adjusting method to update each feature's weight according their importance. The basic idea is to enlarge the weights of more effective features and decrease the weights of less important ones. The initial weight of each feature is set equal. Weights are then updated according to the default training set and each user's feedbacks, respectively.
3) *Comparison3 (C3): SVR-based affective mode*: This is the model introduced in this paper.

For each user, the feedbacks collected from the former 180 played MVs are used for updating models; the left 70 MVs are used for evaluation. Note that the models are first trained with the default training set which contains 50 feedbacks collected in the first user study (i.e., the "Default Training Set" shown in Fig. 13) and are then updated each time when ten new feedbacks are collected (i.e., the "User Feedback" shown in Fig. 13). Among the 27 users, the minimum feedback number obtained from the 180 MVs is 53 [Fig. 13(a)]. Thus, overall arousal precision rate (OAPR), overall valence precision rate (OVPR) and overall arousal-valence precision rate (OAVPR) of *C1*, *C2*, and *C3* across all users are computed when up to 50 new feedbacks are used for model updating. Fig. 13 presents the precision rate curves obtained when the models are updated. Note that the *C3* model needs at least ten profile items to train the model, thus it cannot output values before ten items are used for model training. The reason why *C3* initially performs worse than *C1* and *C3* is because 10 profile items are too few to train valid SVR models

Fig. 13(b) illustrates the OAPR curves. After updating models with the default training set, the performances of *C2* and *C3* remain relatively stable and do not show obvious enhancements. This might be because most users' opinions about Arousal are similar and the default training set is representative enough of most users' understandings about Arousal. It is obvious that the performances of *C2* and *C3* are similar. This could be explained by the fact that the relationships between Arousal and its features are more likely to be linear and straightforward, thus linear feature combination is valid in modeling such simple relationships. Differently, Fig. 13(c) shows *C3* outperforms *C1* and *C2* by large margins. This is mainly because linear model is not sufficient to represent the complicated relationships between Valence and the extracted features. Thus, our non-linear model is more suitable in learning such relationships and capturing user's personalities. Since both Arousal and Valence are considered for OAVPR computation, which poses a stricter requirement, the OAVPR curves in Fig. 13(d) are lower than the OAPR and OVPR curves. However, *C3* still show significant improvements over *C2* and *C1*. Consequently, we can draw the conclusion that the SVR-based affective model is more reliable in learning user's personalities than the classic linear feature combination [9]–[11].

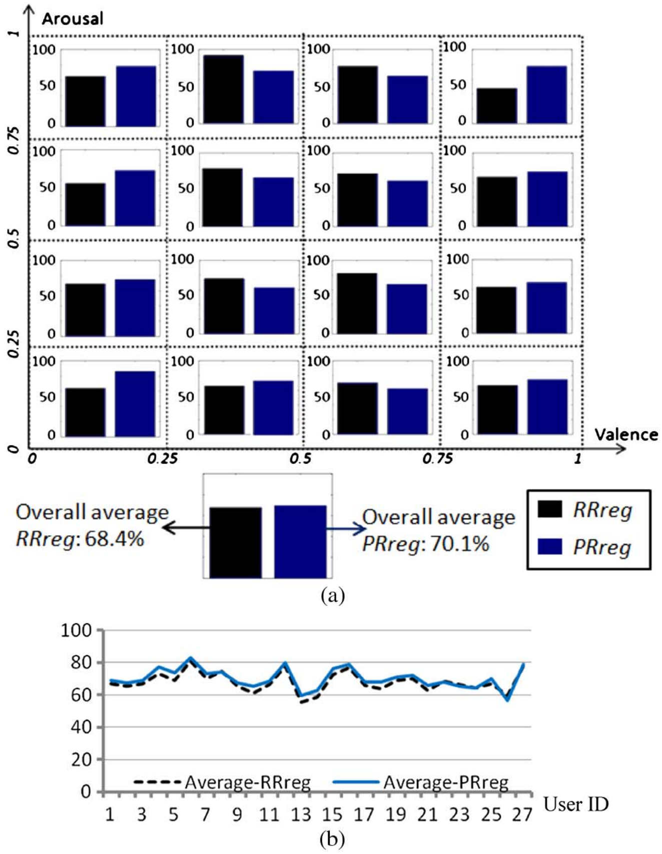
E. Measuring User's Satisfaction for MV Retrieval in i.MV

To conduct MV retrieval in i.MV, users need to input affective regions as their queries. Obviously, users will be satisfied with MV retrieval if the MVs falling in their selected regions present high precision rates and recall rates. Therefore, the regional precision rate (PRreg) and regional recall rate (RRreg) computed within certain affective regions are used to measure user's degree of satisfaction.

The ground truths in the user study divide the affective space into 4 × 4 (16) regions. Accordingly, we will measure user's degree of satisfaction in these regions. Similarly, each user's models are updated with the feedbacks in the former 180 MVs, and the remaining 70 MVs are used for evaluation. For each user, we calculate the $rmRreg$ and RRreg with

$$\begin{aligned}\mathrm{RRreg}_R &= \mathrm{CorrectNum}_R/\mathrm{GTruthNum}_R \cdot 100\% \\ \mathrm{PRreg}_R &= \mathrm{CorrectNum}_R/\mathrm{ComputedNum}_R \cdot 100\% \end{aligned} \tag{9}$$

where $\mathrm{GTruthNum}_R$ is the number of MVs with ground truths falling in region R, $\mathrm{CorrectNum}_R$ and $\mathrm{ComputedNum}_R$ denote the number of correctly computed and totally computed MVs falling in region R, respectively, in the 4 × 4 (16) regions divided by the two four-scale scores. Then, the average PRreg and RRreg of all users are presented in Fig. 14(a).



(a)

(b)

Fig. 14. Measurement of users' satisfaction for MTV retrieval. (a) PRreg, RRreg of each region and the overall average PRreg, RRreg. (b) Each user's average regional precision rate and average regional recall rate.

As shown in Fig. 14(a), with the personalized affective models, the average PRreg and RRreg of each user are stable in different regions. The overall average PRreg and overall average RRreg reach 70.1% and 68.4%, respectively. Considering fine quantization (16 regions denoting 16 different affective states are considered), the result is still promising for affective information-based MV retrieval. Fig. 14(b) presents the average precision rate and average recall rate of each user. From the figure, it can be seen that the precision of MV retrieval is stable for different users although they may have various affective understandings. Therefore, we can conclude that the proposed affective model is effective and valid. In our future work, by acquiring more precise affective descriptions with better feedback strategy and extracting more powerful affective features, we expect to improve the usability and accuracy of i.MV to a higher ground.

F. User Study for Affective Visualization

As mentioned above, Affective Visualization is proposed mainly for two reasons: 1) affective state expressing (ASE) and 2) mapping user's affective queries into a form that can be used for retrieval (MAQ). Twenty participants are invited to compare the ASE and MAQ performance between Affective Visualization and the classic affective label-based affective

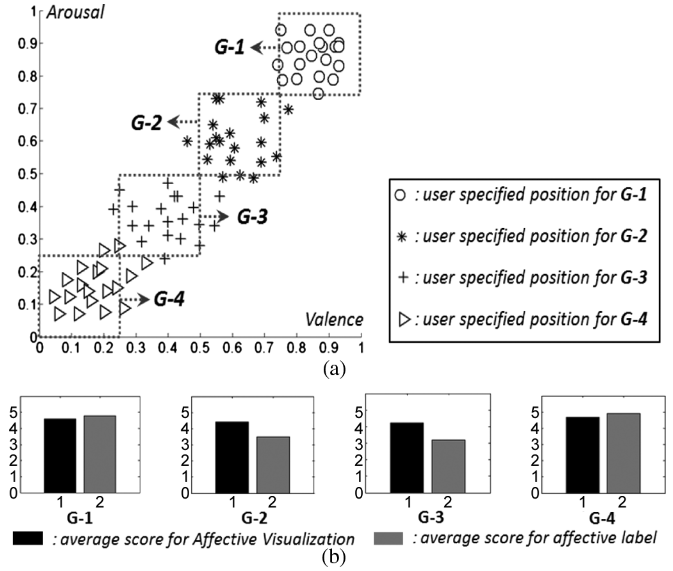


Fig. 15. Positions of MV groups and the results of user study. (a) Affective region of each MV group and the specified affective positions. (b) Users' scores for Affective Visualization and affective label.

TABLE II
AFFECTIVE LABELS SPECIFIED BY USERS

Group ID	specified affective labels
G-1	**joy**, climax, **high**, exciting, **happy**, energetic, exhilarating
G-2	contentment, amusement, **happy**, less happy, **joy**, less exciting, **high**, surprise, glee
G-3	**sorrow, sadness**, **depression**, disgust, **gloomy**, **tearful**
G-4	**sorrow, sadness**, very sad, **depression**, **tearful**, blue, weepy, **gloomy**

state representation [1]–[8], [12]–[16]. Before the study, we collected the affective labels appearing in related work [1]–[8], [12]–[16] (18 in total, including, joy, startle, frantic, happy, sadness, surprise, contentment, depression, and amusement). We also explain Affective Visualization to the participants and ask them to use i.MV for at least 10 min.

Four groups of MVs from different affective regions [G-1, G-2, G-3, and G-4 in Fig. 15(a)] are selected and each group contains 5 MVs with similar affective states (all of the ten users' scores for them are the same in the user study in Section VI-B). Participants have no idea how the four groups are generated, and the selected MVs are removed from the MV database so they also cannot browse them from Affective Visualization. In the user study, the participants are asked to watch each group of MVs and describe their affective states through first confirming affective positions with Affective Visualization and then assigning affective labels (they can select from the labels we collect or define labels themselves), respectively. They are also asked to give five-scale scores (1: extremely unsure and difficult; 2: unsure and difficult; 3: neutral feeling; 4: relatively confident and easy; 5: very confident and easy) to these two methods about their confidences and overall experience. The affective positions and labels specified by the 20 users are presented in Fig. 15(a) and Table II, respectively. Users' average scores for Affective Visualization and affective label are illustrated in Fig. 15(b).

It is clear from Fig. 15(a) that 95%, 75%, 80%, and 75% of the affective positions specified by users are within the G-1, G-2, G-3, and G-4 regions, respectively. This illustrates that users can reflect their affective states correctly with Affective Visualization. From Table II, it can be seen that some users describe different MV groups with the same affective labels (these labels are presented in bold), which could indicate the limitations of language's describing ability for emotions. For example, "happy" appears as description for both G-1 and G-2, though the affective states of these two groups are different. In addition, Fig. 15(b) shows that users' scores for Affective Visualization are generally similar to their scores for affective labels in G-1 and G-4. Accordingly, it can be concluded that Affective Visualization is valid for users to describe their affective queries. Furthermore, compared with G-1 and G-4 which contain specific emotions (very happy and energetic, and very sad and peaceful), G-2 and G-3 present relatively obscure and neutral emotions. From users' scores for G-2 and G-3, it is obvious that Affective Visualization is more effective in representing obscure affective states than affective labels. Consequently, the ASE of Affective Visualization can also be proved better. From the above analysis, it can be concluded that the proposed Affective Visualization is more promising than the commonly used affective labels in affective state representation [1]–[8], [12]–[16].

G. Discussions about Limitations and Solutions

The human affective response mechanisms still need to be further studied by psychologists. In addition, the current affective features are still not strong enough to accurately capture the affective cues in videos. Thus, affective content analysis is still very challenging. Therefore, we must address the limitations and challenging issues with our schemes, as well as provide feasible directions for solutions in our future work.

The first limitation is the computational complexity, which is mainly caused by the property of SVR. In addition, as the profile database increases, the training set will be augmented, which will also slow down the training speed. To solve this problem, we will investigate incremental machine leaning algorithms and will introduce better training set update schemes to generate more compact and effective training set.

The second limitation is about the profiling method in this work. First, the presence of noise is inevitable when feedbacks are collected. Second, the feedbacks are provided in four fixed scales, which is a tradeoff between precision and user experience. Better training set update scheme and more robust machine learning method are helpful in overcoming the first problem. We will solve the second one by studying more effective and user-friendly feedback strategies for continuous affective ground truth acquisition. With continuous affective ground truth, affective models can be trained to output the continuous affective values in the affective space.

Third, the Affective Visualization still needs to be improved to be more intuitive and informative. For example, the music signal in MVs can be visualized to let users browse both the audio and visual content intuitively. In this paper, MVs are represented as points in the affective space, thus, the temporal information and the affective state variations are ignored. However, it is helpful to visualize such information in proper forms.

In addition, it is also desirable to incorporate the metadata such as Album, Artist, and Style into Affective Visualization. All of these investigations will make the Affective Visualization more informative and usable and thus will improve the efficiency of retrieval tasks in large-scale MV databases.

Finally, it is necessary to discuss the reasonability of stimulating the human affective response with regression models. In our work, only audio-visual features are taken into consideration for affective state computation. However the human affective responses can be influenced by other factors such as short-term memory and environment. Ideally, more factors should be modeled to make the affective analysis more reasonable. Though incorporating such factors makes the affective analysis more complex and expensive, it is worthwhile to pursue a tradeoff between efficiency and performance in our future work. Moreover, it is also interesting to investigate the importance of various features for different affective regions for MV affective analysis. This study would be utilized to find an effective and compact feature set; meanwhile, it will be helpful for improving the final performance of affective computation.

VII. Conclusion

In this paper, an integrated framework for MV affective analysis, visualization, and retrieval is proposed. Valid affective features are extracted. Through analyzing the user profile data which reflects users' affective understandings, personalized affective modeling is achieved. Furthermore, to make the affective state intuitive and achieve user-friendly affective information-based MV retrieval, Affective Visualization is proposed to serve as a bridge between users and affective analysis. Both comprehensive experiments and subjective user studies on a representative MV dataset demonstrate the effectiveness of the proposed techniques.

Built upon the proposed techniques, an MV affective retrieval system named i.MV is constructed. As the first of its kind in MV affective analysis, i.MV provides promising applications to MV lovers. For example, i.MV can be transplanted into music players and cell phones to improve their entertaining properties. i.MV can also be integrated into the internet music search engines to enable efficient and user-friendly MV search.

Besides the advantages of the proposed techniques, the shortcomings and possible solutions are also discussed. Future work will be carried out to overcome the current limitations and make the proposed methods more effective and usable.

References

[1] S. Moncrieff, C. Dorai, and S. Venkatesh, "Affect computing in film through sound energy dynamics," *ACM Multimedia*, pp. 525–527, 2001.

[2] H. B. Kang, "Analysis of scene context related with emotional events," *ACM Multimedia*, pp. 311–314, 2002.

[3] H. B. Kang, "Emotional event detection using relevance feedback," in *Proc. IEEE ICIP*, 2003, pp. 721–724.

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

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

[6] M. Xu, S. Luo, and J. Jin, "Video adaptation based on affective content with MPEG-21 DIA framework," in *Proc. IEEE SCIISP*, 2007, pp. 386–390.

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

[8] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[9] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, 2006.

[10] A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[11] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1114–1122, Dec. 2005.

[12] S. Arifin and P. Y. K. Cheung, "A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information," *ACM Multimedia*, pp. 68–77, 2007.

[13] S. Arifin and P. Y. K. Cheung, "User attention based arousal content modeling," in *Proc. IEEE ICIP*, 2006, pp. 433–436.

[14] H. L. Wang and L. F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[15] C. H. Chan and G. J. F. Jones, "Affect-based indexing and retrieval of films," *ACM Multimedia*, pp. 427–430, 2005.

[16] Y. H. Yang, Y. F. Su, Y. C. Lin, and H. H. Chen, "Music emotion recognition: the role of individuality," in *Proc. ACM Int. Workshop Human-Centered Multimedia*, 2007, pp. 13–22.

[17] [Online]. Available: http://www.google.cn/music/songscreener

[18] P. J. Lang, "The network model of emotion: motivational connections," in *Advances in Social Cognition*. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1993, pp. 109–133.

[19] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, pp. 273–294, 1977.

[20] H. Schlosberg, "Three dimensions of emotion," *Psycholog. Rev.*, vol. 61, no. 2, pp. 81–88, Mar. 1954.

[21] K. Isbister, K. Hook, M. Sharp, and J. Laaksolahti, "The sensual evaluation instrument: Developing an affective evaluation tool," in *Proc. SIGCHI Conf. Human Factors in Computing Syst.*, 2006, pp. 1163–1172.

[22] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Stat. Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

[23] S. L. Zhang, Q. Huang, Q. Tian, S. Jiang, and W. Gao, "i.MTV—An integrated system for MTV affective analysis," *ACM Multimedia*, pp. 985–986, 2008.

[24] S. L. Zhang, Q. Tian, S. Q. Jiang, Q. M. Huang, and W. Gao, "Affective MTV analysis based on arousal and valence features," in *Proc. IEEE ICME*, 2008, pp. 1369–1372.

[25] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual difference for describing and measuring individual differences in temperament," *Current Psychol.*, vol. 14, no. 4, pp. 261–292, Dec. 1996.

[26] N. Dimitrova, L. Agnihotri, R. Jasinschi, J. Zimmerman, G. Marmaropoulos, T. McGee, and S. Dagtas, "Scouting demonstration: smart content selection and recording," *ACM Multimedia*, pp. 499–500, 2000, BVideo.

[27] B. Adams and S. Venkatesh, "An embedded suggestive interface for making home videos," in *Proc. IEEE ICME*, 2006, pp. 2025–2028.

[28] C. Dorai, R. Farrell, A. Katriel, G. Kofman, Y. Li, and Y. Park, "BMAGICAL demonstration: System for automated metadata generation for instructional content," *ACM Multimedia*, pp. 491–492, 2006.

[29] R. Jana, J. Murray, and C. W. Rice, "IMS-TV: An IMS-based architecture for interactive, personalized IPTV," *IEEE Commun. Mag.*, vol. 46, no. 11, pp. 156–163, Nov. 2008.

[30] D. O. Sullivan, B. Smyth, D. C. Wilson, K. McDonald, and A. Smeaton, "Improving the quality of the personalized electronic program guide," *User Modeling and User-Adapted Interaction*, vol. 14, no. 1, pp. 5–36, Feb. 2004.

[31] L. Chen and K. Sycara, "WebMate: Personal agent for browsing and searching," in *Proc. Int. Conf. Autonomous Agents*, 1998, pp. 132–139.

[32] B. Magnini and C. Strapparava, "User modeling for news web sites with word sense based techniques," *User Modeling and User-Adapted Interaction*, vol. 14, no. 2–3, pp. 239–257, Jun. 2004.

[33] S. Venkatesh, B. Adams, D. Phung, C. Dorai, R. G. Farrell, L. Agnihotri, and N. Dimitrova, ""You Tube and I Find"—Personalizing multimedia content access," *Proc. IEEE*, vol. 96, no. 4, pp. 697–711, Apr. 2008.

[34] N. Sebe and Q. Tian, "Personalized multimedia retrieval: the new trend?," in *Proc. ACM Int. Workshop MIR*, 2007, pp. 299–306.

[35] L. Agnihotri, J. Kender, N. Dimitrova, and J. Zimmerman, "Framework for personalized multimedia summarization," in *Proc. ACM Int. Workshop MIR*, 2005, pp. 31–38.

[36] G. I. Webb, M. J. Pazzani, and D. Billsus, "Machine learning for user modeling," *User Modeling and User-Adapted Interaction*, pp. 19–29, Mar. 2001.

[37] H. Hirsh, C. Basu, and B. Davison, "Learning to personalize," *Commun. ACM*, vol. 43, no. 8, pp. 102–106, Aug. 2000.

[38] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 9, no. 4, pp. 441–457, Dec. 2001.

[39] N. C. Maddage, C. S. Xu, M. S. Kankanhalli, and X. Shao, "Content-based music structure analysis with applications to music semantics understanding," *ACM Multimedia*, pp. 112–119, 2004.

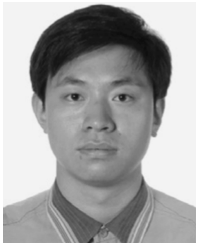
[40] D. Bordwell and K. Thompson*, Film Art: An Introduction*, 7th ed. New York: McGraw-Hill, 2004.

[41] H. Zettl*, Sight Sound Motion: Applied Media Aesthetics*, 3rd ed. Belmont, CA: Wadsworth, 1998.

[42] C. J. C. Burges, "A tutorial on support vector machine for pattern recognition," *Data Mining Knowledge Discovery*, pp. 121–167, Jun. 1998.

[43] G. P. Nguyen and M. Worring, "Optimization of interactive visual similarity based search," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 4, no. 1, Jan. 2008.

[44] X. L. Liu, T. Mei, X. S. Hua, B. Yang, and H. Q. Zhou, "Video collage," *ACM Multimedia*, pp. 461–462, 2007.



Shiliang Zhang is currently working toward the Ph.D. degree at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His research interests include image and video processing, large-scale image retrieval, multimedia content analysis, computer vision, and pattern recognition.

Qingming Huang (M'04–SM'08) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1994.

He was a Postdoctoral Fellow with the National University of Singapore from 1995 to 1996 and was with the Institute for Infocomm Research, Singapore, as a Member of Research Staff from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, under the Science100 Talent Plan in 2003, and currently is a Professor with the Graduate University, Chinese Academy of Sciences. His current research areas are image and video analysis, video coding, pattern recognition, and computer vision.

Shuqiang Jiang (M'06) received the M.S. degree from the College of Information Science and Engineering, Shandong University of Science and Technology, Shandong, China, in 2000, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Faculty Member with Digital Media Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision.

Wen Gao (M'92–SM'05–F'08) received the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

He was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo, in 1992, and a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 1993. From 1994 to 1995, he was a Visiting Professor with the AI Laboratory, Massachusetts Institute of Technology, Cambridge. Currently, he is a Professor with the School of Electronic Engineering and Computer Science, Peking University, Peking, China, and a Professor of computer science with the Harbin Institute of Technology. He is also the Honor Professor of computer science with the City University of Hong Kong and the External Fellow of International Computer Science Institute, University of California, Berkeley. His research interests are in the areas of signal processing, image and video communication, computer vision, and artificial intelligence.

Qi Tian (SM'04) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Urbana, IL, in 2002.

He is currently an Associate Professor with the Department of Computer Science, University of Texas at San Antonio (UTSA). He has authored or coauthored over 100 refereed journal and conference papers. His research projects were funded by ARO, DHS, HP Lab, SALSI, CIAS, and CAS. He has served as guest co-editor of *Journal of Computer Vision and Image Understanding*, *ACM Transactions on Intelligent Systems and Technology,* and *EURASIP Journal on Advances in Signal Processing* and is a member of the Editorial Board of the *Journal of Multimedia.* His research interests include multimedia information retrieval and computer vision.

Dr. Tian is a member of the Association of Computing Machinery (ACM).He has been serving as Program Chairs, Session Chairs, Organization Committee Members and TPC for over 120 IEEE and ACM Conferences, including ACM Multimedia, SIGIR, ICCV, and ICASSP. He has served as a guest coeditor of the IEEE TRANSACTIONS ON MULTIMEDIA and is an associate editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.