行政院國家科學委員會專題研究計畫 成果報告

由電影配樂中探勘音樂情緒之研究(第 2 年)
研究成果報告(完整版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2221-E-004-009-MY2
執 行 期 間 ：96 年 08 月 01 日至 97 年 07 月 31 日
執 行 單 位 ：國立政治大學資訊科學系

計 畫 主 持 人 ：沈錳坤

計畫參與人員：碩士班研究生-兼任助理人員：廖家慧
碩士班研究生-兼任助理人員：黃詰仁
碩士班研究生-兼任助理人員：闕伯丞
博士班研究生-兼任助理人員：邱士銓

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 97 年 12 月 15 日

行政院國家科學委員會補助專題研究計畫成果報告

由電影配樂中探勘音樂情緒之研究

計畫類別：☑ 個別型計畫 □ 整合型計畫
計畫編號：NSC 95－2221－E－004－009-MY2
執行期間：95 年 08 月 01 日至 97 年 07 月 31 日

計畫主持人：沈錳坤
共同主持人：
計畫參與人員：廖家慧、黃詰仁、闕伯丞、邱士銓

成果報告類型(依經費核定清單規定繳交)：□ 精簡報告 ☑完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
☑出席國際學術會議心得報告
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
☑涉及專利或其他智慧財產權，□一年☑二年後可公開查詢

執行單位：政治大學資訊科學系
中　華　民　國　97　年　7　月　31　日

由電影配樂中探勘音樂情緒之研究

ABSTRACT

Movies play an important role in our life nowadays. How to analyze the emotional content of the movies becomes one of the major issues. Based on film grammar, there are many audiovisual cues in movies helpful for detecting emotions of scenes. In this research, we investigate the discovery of the relationship between audiovisual cues and emotions of scenes and automatic annotation of the emotion of the scene is achieved. First, the training scenes are labeled with the emotions manually. Second, six classes of audiovisual features are extracted from all scenes. These classes of features consist of color, light, tempo, close-up, audio, and subtitle. Finally, the graph-based approach, Mixed Media Graph approach is modified to mine the association between audiovisual features and emotions of scenes. The experiments show that the accuracy is up to 70%.

1. Introduction

With the development of content digitalization, the increase of computer storage capacity and the raise of network bandwidth, digital video collections grow rapidly in recent years. Instead of time-consuming and tedious manual search for video clips, various techniques for content-based analysis have been developed to automatically analyze and index multimedia data [12][14][16][20].

The purpose of content-based video analysis is to obtain a structured organization of the original video content or to realize its semantics meaning. Content-based video indexing is the task of tagging semantic video units obtained from content analysis to achieve convenient and efficient content retrieval. The techniques for content-based analysis so far enable us to easily access the events, people, objects, and scenes captured by the camera. For example, we can retrieve the most exciting parts of a sport game [12], or to efficiently generate abstracts or summaries of movies or sports game. In addition, multimedia semantic retrieval is getting popular recently. Today, movies play an important role in our life. How to analyze the emotional content of the movies becomes one of major issues.

A film was composed of many elements. Some important aspects in film production are the color composition in the mise-en-scene, lighting, sound, editing, and narrative et al. The relationships

among these elements are a set of informal rules known as film grammar defined in [3] - “the product of experimentation, an accumulation of solutions found by everyday practice of the craft, and results from the fact that films are composed, shaped and built to convey a certain story.” In other words, film grammar embodies film production knowledge that is found more in history of use. It explains the relationships between many cinematic techniques and their semantic meanings delivering to viewer.

From the book – Understanding Movies [10] – every shot and scene in the movies was arranged painstakingly by the director. The director delivers the emotion of the scene to viewer by the way of using different shots, color distribution, lighting, shot cut rate, and sound effect et al. For example, the excitement of a scene increases as the shot length decreases. Other examples include rules about screen movements, cutting on action, colors and variation of lighting effects etc. By exploiting the constraints afforded by the film grammar, high-level affective meaning can emerge from low-level features such as shot length directly. Thus, it offers computable approach in bridging the difficult transition to high-level semantics such as emotions.

As mentioned above, there are many audiovisual cues in movies helpful for detecting emotions of scenes. In this way, we wish to discover the relationship between audiovisual cues and emotions of scenes and then achieve automatic annotation of the emotion of the scene.

Once all scenes are labeled with emotions, users can retrieve the scenes based on its’ prevailing emotion. For instance, users will be able to search for the funniest or the most exciting parts of a movie. In this way, user can save his (her) time to browse this movie. Besides, one application of this research is automatic music accompaniment for scenes. After the emotion of the scene is determined, the emotion is used as the intermediate to find out the related music delivering the same emotion.

In this report, we develop a system grounded upon film grammar to discover the emotions of scenes in film. The Mixed Media Graph algorithm is employed to classify the scenes.

This report is organized as follows. We introduce the related works of affective classification in Section 2. Section 3 describes the audiovisual features used for affective classification. The scene affinity graph used in this research for emotion discovery is elaborated upon Section 4. Experiments and results are presented in Section 5, followed by the conclusion in Section 6.

2. Related Work

Nowadays, many approached have been proposed to analyze affective content [11][12][14][15][16][27]. Hanjalic and Xu [11][12][13] exploit motion activity, cut density and sound energy to form the so-called *affect curve*, which maps to the two-dimensional (Arousal-Valence) emotion model. The affect curve on the two-dimensional model will show us the emotion distribution.

A number of works on affective classification of films have been done. Kang [16] proposed a new technique for detecting affective events such as Fear, Sadness, and Joy using Hidden Markov Models (HMM). He performed empirical study on the relationship between emotional events and low-level features of video content. These low level features consisting of color, motion and shot cut rate were computed for each shot. The feature vector of each shot was transformed to observation vector sequences using vector quantization. He introduced two HMM topologies to detect emotional events. The experiments on six thirty-minute video for emotional event classification show about 70.2% and 78.73 separately.

Based on film theories and psychological models, Wei et al. [39] proposed a color content-based system to analyze the color distribution and related feelings brought to viewers. This system uses a set of color features for color-mood analysis and subgenre discrimination. They introduced two color representations for scenes and full films for extracting the essential moods from the films – Movie Palette Histogram and Mood Dynamic Histogram. Movie Palette Histogram is a global measure for the color palette while Mood Dynamic Histogram is a distinguishing measure for the transitions of the moods in the movie. The dominant color ratio and the pace of the movie are also captured for classification. They exploit eight mood types that are defined by psychologist Plutchik - Anger, Fear, Joy, Sorrow, Acceptance, Rejection, Surprise, and Expectancy. In this paper, emotion and mood are regarded as different concepts. Many emotion terms associated with the eight mood types are selected to describe the eight mood types. Each window consist of six video shots is mapped into a mood type according to the emotions of the six shots. c-SVC Support Vector Machine (SVM) [4] is adopted for mood classification. Their experiments on fifteen full-length films for mood type classification of the window (group of six shots) level show about 80% accuracy.

Wang and Cheong [14] proposed an affective scene classification system that is a complementary approach grounded in the fields of cinematography and psychology. This system exploits a number of effective audiovisual cues and an appropriate set of affective categories that are identified for scene classification. For each scene, the audio and the visual signal were processed separately. The visual signal was segmented into shots (represented as key-frames) to computing visual cues. The audio

signal was separated according to audio type (music, speech, environ or silence). The audio segments then were sent into a support vector machine (SVM) based probabilistic inference machine to obtain high-level audio cues at the scene level. The visual and audio cues were finally concatenated to form the scene vectors, which were sent into the same inference machine to acquire probabilistic vectors. The output of every testing scene is expressed probabilistically and each testing scene was classified into one of the output categories. The overall correct classification rate is 74.69%.

3. Feature Extraction

3.1 Introduction

Critics and scholars categorize movies into three main styles: realism, classicism, and formalism. Rather than separate categories, these three styles might be regarded as a continuous spectrum of possibilities. Realism and formalism are general terms used to describe the movies falling into the two styles' extremes, while classicism can be viewed as an intermediate style that avoids the extremes of realism and formalism. In other words, few films are exclusively formalist or realist in style [10].

Realistic movies are unapparent in style. What realism directors concern is how to reproduce the surface of reality with little distortion and make their films seems unmanipulated. Such filmmakers use the camera, as a recording mechanism, to describe the subject matters with as little commentary as possible. They care "what is being shown" rather than "how it is manipulated." Some realists aim for rough look in their movies. Simplicity, spontaneity, and directness are the highest rules.

Formalistic films, on the other hand, are relatively flamboyant in style. Formalism directors are referred to expressionists because their self-expression is at least as important as the subject matter itself. They prefer to express subjective experience of reality. The camera is used to comment on the subject matter, and emphasize its essential rather than its objective nature.

Nowadays, few films are absolutely realistic in style. Most directors use the camera to comment on the subject matter. Even is the famous documentary "Let It Be", the key reason of people are toughed is how the director present the protagonist's character.

As we have stated in Section 1, film grammar explains the relationships between many cinematic techniques and their semantic meanings delivering to viewer. A director's expression is conveyed through film grammar.

In film, the basic film grammar is defined as follows. A film is composed of many scenes, and a scene consists of many shots. A scene is defined as a meaningful story unit while a shot is defined as a stream of many frames continuously recorded by a single camera. Based on how many subject matters or human figures are included within the frame of the screen (not the distance between the camera and the object photographed), most shots can be designed and subsumed under the six basic categories: (1) the extreme long shot, (2) the long shot, (3) the full shot, (4) the medium shot, (5) the close-up, and (6) the extreme close-up. As showed in Table 3.1 [10]. Figure 3.1 shows the examples of shots.

Table 3.1 Six basic categories of shots in cinema [10]

Shot	Description
Extreme Long Shot	The extreme long shot is taken from a great distance, sometimes as far as a quarter of a mile away. It is also called “establishing shots” because of being taken at the exterior space, and serve as spatial frames of reference for the closer shots.
Long Shot	Usually, the distance of long shot is between the audience and the stage in the live theater. The closest distance in long shot is equal to full shot that just includes human body in full, with the head near the top of the frame and the feet near the bottom.
Full Shot	The distance of full shot is between medium shot and long shot. A full shot contains over three figures.
Medium Shot	The medium shot contains a figure from the knees or waist up. It is also called “functional shot”, useful for shooting exposition scenes and for dialogue.
Close-Up	The close-up emphasizes little on the background or external location, but on a relatively small object - the human face, for example. The close-up shot enhances the importance of things, often suggesting a symbolic significance by enlarge the size of an object.
Extreme Close-Up	The extreme close-up is a variation of close-up. Therefore, the extreme close-up might show only a person’s eyes or mouth instead of a face.

Figure 3.1 Examples of six kinds of shots in cinema

3.2 Emotion Discovery from Scenes

Based on the film grammar, some audiovisual features extracted from films is helpful for the emotion discovery. The relationship between audiovisual features and emotions is close in films. To automatically label the emotion for the query scene, the association between audiovisual features and emotions in films is discovered from training data (scenes). Training data has been labeled of emotions manually. Audiovisual features are extracted from training data. By using these extracted audiovisual features and labeled emotions from training data, the association between audiovisual features and emotions is discovered (Figure 3.2). The discovered association is therefore utilized to label the emotion for the query scene automatically.

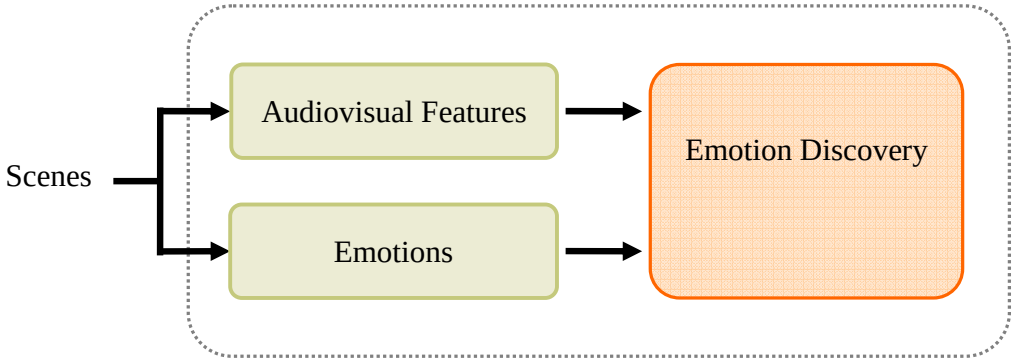

Figure 3.2 Association discovery of scenes

The audiovisual features we adopted consist of several classes of visual features (color, light, tempo, and close-up), one class of audio feature (sound), and one class of textual feature (subtitles). Each class of feature may have more than one feature representation. In total, there are fifteen feature representations which are shown in Table 3.2.

3.3 Visual Features

Visual features include four classes of features – *color*, *light*, *tempo*, and *close-up*.

Color Class

People are often influenced by color in the subconscious. Psychologists have discovered that most people actively try to interpret the lines of a composition, but they accept color passively, permitting it to suggest moods rather than objects.

Table 3.2 Audiovisual features in our work

Types	Classes	Feature Representations
Visual	Color	Movie Palette Histogram (MPH)
		Mood Dynamics Histogram (MDH)
		Family Histogram (FH)
		Dominant Color Ratio (DCR)
	Light	Lightness
	Tempo	Shot Length (SL)
		Number of Shots per Minute (NSPM)
	Close-Up	Close-Up (CU)
		Close-Up Ratio (CUR)
Audio	Sound	Zero-Crossing Rate (ZCR)
		Spectral Roll-off point(SR)
		Spectral Centroid (SC)
		Spectral Flux (SF)
		Mel Frequency Cepstral Coefficients (MFCCs)
Textual	Subtitle	Feeling words

Visual artists have used color for symbolic purposes for a long time. Though Color symbolism is dependent on cultural, in general, cool colors (blue, green, violet) tend to suggest tranquility, aloofness, and serenity. Warm colors (red, yellow, orange) suggest aggressiveness, violence, and stimulation [10]. "Life is beautiful", for example, the movie starts from funny scenes, the color is bright and warm, but when the massacre begin, the color of movie starts receding from images.

In the *color* feature, we adopt four color representations used in [23] - Family Histogram (FH), Movie Palette Histogram (MPH), Mood Dynamics Histogram (MDH), and Dominant Color Ratio (DCR). We modify them as follows,

- Family Histogram: It is a shot level feature, defined as the color histogram of a shot's key-frame. One key-frame is used to represent one shot.
  - Family Histogram H_k of the shot *k*:

$$H_k(i) = \sum_{x,y} \begin{cases} 1, & if \quad Pix(x,y) = i \\ 0, & if \quad Pix(x,y) \neq i \end{cases}$$

  where $Pix(x, y)$ is the pixel value of position (x, y) in the key frame of the shot k;

  i is a bin number in CIELUV color space, $i = 1 \sim 264$;

- Movie Palette Histogram: MPH is a scene level feature in the scene view. It is defined as the color histogram of twelve reference colors in one scene. Main colors of a scene can be captured from MPH.
  1) Before computing MPH, we first compute Dominant Color Palette :

  $DCP(k)$ of the shot *k*,

  $DCP(k) = [\ P_1^k, P_2^k, P_3^k\ ]$, for $k = 1 \sim N$. N is the number of shots in the scene.

  where P_1^k, P_2^k, P_3^k are the three bin numbers of corresponding top three dominant colors in histogram H_k.

$$P_1^k = \arg_i \{ \max_{i=1\sim 264} [H_k(i)] \}$$

$$P_2^k = \begin{cases} P_2^k, & if \quad H_k(P_2^k) > [H_k(P_1^k) \times Th] \\ 0, & otherwise \end{cases}$$

$$P_3^k = \begin{cases} P_3^k, & if \ H_k(P_3^k) > [H_k(P_2^k) \times Th] \\ 0, & otherwise \end{cases}$$

where Th : 0~1, a fixed threshold.

2) After computing Dominant Color Palette, we can get dominant color bin counts :

$$D_1^k = H_k(P_1^k)$$

$$D_2^k = \begin{cases} H_k(P_2^k), & if \ H_k(P_2^k) > [H_k(P_1^k) \times Th] \\ 0, & otherwise \end{cases}$$

$$D_3^k = \begin{cases} H_k(P_3^k), & if \ H_k(P_3^k) > [H_k(P_2^k) \times Th] \\ 0, & otherwise \end{cases}$$

where Th : 0~1, a fixed threshold.

3) Based on dominant color, Representative Dominant Color Sequence for the shot k, $RDCS(k)$, is defined as:

$$RDCS(k) = \sum_{l=1}^{3} P_l^k \times W_l^k \quad , \quad W_l^k = \frac{D_l^k}{D_1^k + D_2^k + D_3^k}$$

for k = 1 ~ N, and l = 1 ~ 3.

4) Next, we get Movie Palette (MP):

$$MP(k) = \arg_{R(m)} \{ \min_{m=1\sim12} [dis(RDCS(k), R(m))] \} \text{ , for } k = 1\sim N$$

$R(m)$: Pre-defined colors, which uniformly divide the CIELUV color space, m = 1 ~ 12.

$dis(*,*)$: the distance between two color, is defined as Euclidean distance.

5) Finally, MPH is derived from MP:

$$MPH(m) = \sum_{k=1}^{N} \begin{cases} 1, & if \ MP(k) = R(m) \\ 0, & if \ MP(k) \neq R(m) \end{cases} \text{ , for } m = 1\sim12$$

where N is the number of shots in the scene.

- Mood Dynamics Histogram: MDH is a scene level feature. Color transitions between shots may lead to mood dynamics [23]. We acquire MDH from the statistics of color transitions in movie palette.
  - $MDH((m1-1)\times 12+m2) =$

$$\sum_{k=2}^{N} \begin{cases} \frac{1}{N}, & if\ [\,MP(k-1) = R(m_1)\,]\ \&\ [\,MP(k) = R(m_2)\,]\ \&\ [\,m_1 = m_2\,] \\ 1, & if\ [\,MP(k-1) = R(m_1)\,]\ \&\ [\,MP(k) = R(m_2)\,]\ \&\ [\,m_1 \neq m_2\,] \\ 0, & otherwise \end{cases}$$

  for $m_1 = 1 \sim 12$, $m_2 = 1 \sim 12$, where N is the number of shots in the scene.
- Dominant Color Ratio: Dominant color ratio is a shot level feature. While MPH signifies main colors, DCR indicates the degree of influence of main colors in a shot. In other words, the higher DCR the shot has, the more representative the main colors are in this shot.
  - Dominant Color Ratio : $DCR = \frac{|P_d|}{|P|}$

  where P_d is the set of dominant color pixels, and P is the set of all pixels in a frame.

Light Class

Light and dark have had symbolic meanings from the dawn of humanity. Light suggests security, justice, and joy. Dark, however, is borrowed to suggest fear, evil, and the unknown. There are various styles of lightings according to different themes and moods of films. In general, comedies tend to be lit in “high key” with bright and little shadows. Mysteries and thrillers are generally in “low key”, with diffused shadows and atmospheric light. Tragedies are usually lit in “high contrast” with harsh shafts of lights and dramatic streaks of blackness.

- Lightness: Lightness is the only representation of the *light* feature. Lightness is defined as the average luminance of the shot’s key-frame.

Tempo Class

From the cinematographic perspective, the pace is manipulated to great effect by editing effects like cuts. As each shot conveys an event, the filmmaker can intensify a scene by increasing the event density via rapid shot changes [40]. To the viewer, rapid shot changes capturing the main action from different angles certainly convey the dynamic and breathtaking excitement far more effectively than a long duration shot [33] [2].

The *tempo* feature consist of two representation, Shot Length (SL) and Number of Shots Per Minute (NSPM), defined as follows,

- Shot Length: Shot length is defined as the number of frames in the shot.
- Number of Shots Per Minute: NSPM is a scene level feature, defined as the number of shots per minute in the scene. The more shots one scene has in a minute, the more excited the audiences feel.

Close-up Class

It is essential for a filmmaker to concern what kind of shot to use to convey the action of a scene. When we see a close-up of the character, it implies that the filmmaker forces us to care about him or her and to identify with his or her feelings. If the character is a villain, the close-up shot can make an emotional revulsion in us. For example, when a threatening character is so close to us, he seems to encroach on our space.

Generally speaking, the greater the distance between the objects and the camera, the more emotionally neutral we remains. One of Chaplin's most famous pronouncements is "Long shot for comedy, close-up for tragedy". This principle appears to make sense for when an action "a person is slipping on a banana peel" is close to us, it's hardly funny because the person's safety will become our first concern. But if we see this event from a greater distance, it seems to be a comical act to us.

The *close-up* feature comprises Close-Up and Close-Up Ratio (CUR) representations.

- Close-Up: it describes whether the shot is a close-up. *Close-Up* for shot *k*:
  - $Close-Up\,(k) = \begin{cases} 1, & if\ the\ shot\ is\ a\ close-up. \\ 0, & if\ the\ shot\ is\ not\ a\ close-up. \end{cases}$
- Close-Up Ratio: CUR is a scene level feature. Because the audience have an inclination toward identification with the character in a close-up, close-up ratio may be helpful in determining emotion of the scene. CUR is defined as follow:
  - $Close-Up\ Ratio = \frac{The\ number\ of\ close-up\ in\ the\ scene}{Total\ number\ of\ shots\ in\ the\ scene}$

3.4 Audio Features

Sound plays a very essential role in films. In sound, sound effect and music are the most influential on the audience emotion. Famous director Akira Kurosawa had said "Cinematic sound is that which does not simply add to, but multiplies, two or three times, the effect of the image." Although the emotion of audience is usually governed by sounds in films, they are often not aware of it.

Sound Effect

Sound effects can be precise source of meaning in film. The pitch, tempo, and volume of sound effects strongly arouse various emotions of the audience. As high-pitched sounds are usually strident and create a sense of tension, they are often employed in suspense scenes especially just before and during the climax. Low-frequency sounds, on the other hand, are usually used to emphasize the grandeur or solemnity of a scene as their heaviness and fullness. In addition, low-pitched sounds can also suggest anxiety and mystery. For example, usually, a suspense sequence usually borrows the low-pitched sounds first, and gradually turns the sounds to be high-pitched as the scene moves toward its climax.

This is not absolute principle though for silence can be powerful sometimes. In sound movies, complete silence for few minutes or even few seconds can suspend the audience and draw their considerable attention. Moreover, as we tend to fear what we can't see, the filmmakers sometimes prompt a sense of terror in the audience by using off-screen sound effects in horror or suspense films.

Music

In general, music with lyrics is pretty influential since music itself and words convey meanings. However, accompanied with film images, music, no matter with or without lyrics, can be more specific. The theme of a film is usually implied from the cinematic music in the opening. If the filmmaker does not provide the audience with a scene for the dramatic climax, music serves as a foreshadowing. Such hint as Hitchcock, following anxious music, can be one of the warnings to the audience to be prepared. Directors sometimes mislead the audience deliberately by using false musical warnings. Similarly, we can tell actors' internal emotion by music when actors are required to assume neutral expressions.

We exploit five *audio* feature representations that are widely used for audio classification and speech recognition.

- Zero-Crossing Rate: ZCR is a basic acoustic feature that is defined as the number of times the signal value crosses the zero axis in time domain within a frame [36]. ZCR is proved to be useful in characterizing distinct audio signals. It has been popularly used in speech/music classification algorithms [22]. As shown in Figure 3-3, periodical sound tends to have a smaller value of ZCR, while noisy sound tends to have a higher value. Table 3.3 shows the twelve statistics of ZCR in our work.

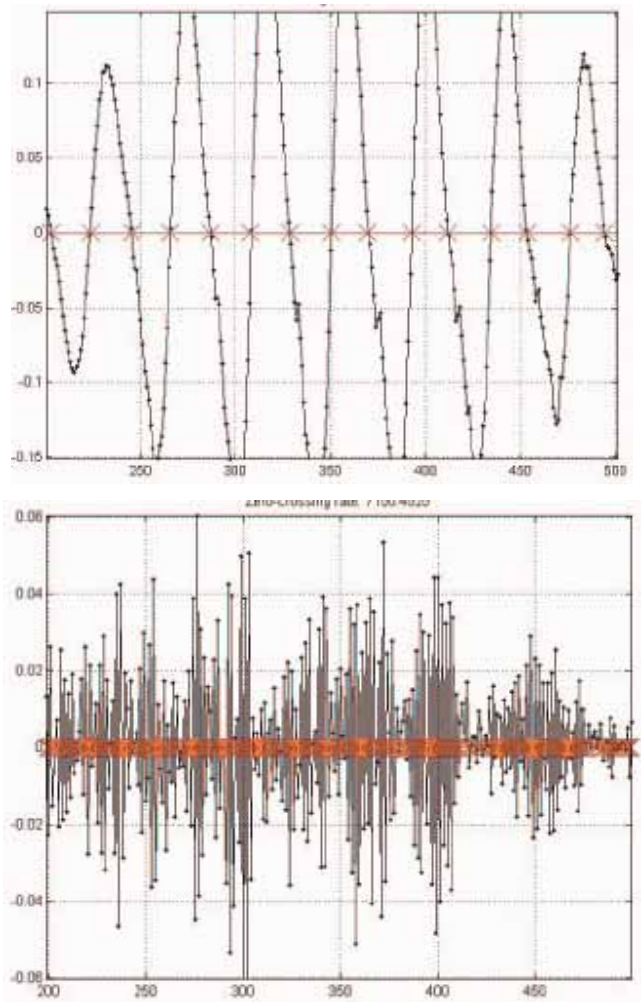

Figure 3.3 Zero-crossing rate during voiced speech region (top, ZCR = 432) and unvoiced speech region (button, ZCR = 7150) [31]

- Spectral Roll-off point: spectral roll-off point is the frequency below which 95th percentile of the power in the spectrum resides [37]. This is a measure of the "skewness" of the spectral shape - the value is higher for right-skewed distributions [37]. SR can tells voiced speech from unvoiced speech as unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum, while most of the energy for unvoiced speech and music is contained in lower bands of spectrum. Table 3.4 shows the twelve statistics of SR in our work. Figure 3.4 shows the energy spectrum (cumulative energy) along frequency with 95% spectral roll-off frequency. The spectral roll-off value for a frame is computed as follows:

- $SR = K$, *where* $\sum_{f=0}^{K} E[f] = 0.95 \sum_{f=o}^{fMAX} E[f]$

E[f] is the energy of the signal at the frequency f.

fMAX is the maximal frequency in the spectrum.

Table 3.3 Zero-Crossing Rate statistics

	Name	Num. of values
1	Zero-Crossings Overall Average	1
2	Zero-Crossings Overall Standard Deviation	1
3	Derivative of Zero-Crossings Overall Average	1
4	Derivative of Zero-Crossings Overall Standard Deviation	1
5	Running Mean of Zero-Crossings Overall Average	1
6	Running Mean of Zero-Crossings Overall Standard Deviation	1
7	Standard Deviation of Zero-Crossings Overall Average	1
8	Standard Deviation of Zero-Crossings Overall Standard Deviation	1
9	Derivative of Running Mean of Zero-Crossings Overall Average	1
10	Derivative of Running Mean of Zero-Crossings Overall Standard Deviation	1
11	Derivative of Standard Deviation of Zero-Crossings Overall Average	1
12	Derivative of Standard Deviation of Zero-Crossings Overall Standard Deviation	1

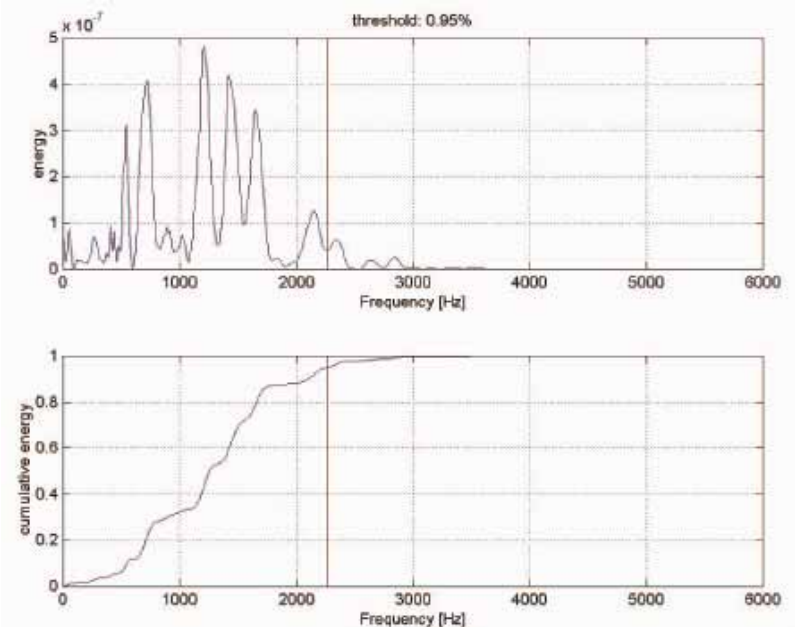

Figure 3.4 [Top] Energy spectrum along frequency with 95% spectral roll-off frequency (vertical red line) [bottom] cumulative energy along frequency with 95% spectral roll-off frequency (vertical red line) [31]

Table 3.4 Spectral Roll-off point statistics

	Name	Num of Values
1	Spectral Roll-off Point Overall Average	1
2	Spectral Roll-off Point Overall Standard Deviation	1
3	Derivative of Spectral Roll-off Point Overall Average	1
4	Derivative of Spectral Roll-off Point Overall Standard Deviation	1
5	Running Mean of Spectral Roll-off Point Overall Average	1
6	Running Mean of Spectral Roll-off Point Overall Standard Deviation	1
7	Standard Deviation of Spectral Roll-off Point Overall Average	1
8	Standard Deviation of Spectral Roll-off Point Overall Standard Deviation	1
9	Derivative of Running Mean of Spectral Roll-off Point Overall Average	1
10	Derivative of Running Mean of Spectral Roll-off Point Overall Standard Deviation	1
11	Derivative of Standard Deviation of Spectral Roll-off Point Overall Average	1
12	Derivative of Standard Deviation of Spectral Roll-off Point Overall Standard Deviation	1

- Spectral Centroid: the "balancing point" of the spectral power distribution. Many types of music involve percussive sounds which push the spectral mean higher by including high-frequency noise [37]. The spectral centroid for a frame is computed as follows:

$$SC = \frac{\sum_k k \cdot X[k]}{\sum_k X[k]}$$

Where: k is an index corresponding to a frequency, or small band of frequencies within the overall measured spectrum, and $X[k]$ is the power of the signal at the corresponding frequency band. Table 3.5 shows the twelve statistics of SC in our work.

- Spectral Flux statistic: the average variation value of spectrum between the adjacent two frames in one second window [21]. Table 3.6 shows the twelve statistics of SF in our work.

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k-1}^{K-1} [\log(X(n,k)+\delta) - \log(X(n-1,k)+\delta)]^2$$

where X means the magnitude of FFT coefficients.

N is the number of frames in one window.

δ is a little value to avoid log zero.

n is the frame index, and k is frequency index.

Table 3.5 Spectral Centroid statistics

	Name	Num of Values
1	Spectral Roll-off Point Overall Average	1
2	Spectral Roll-off Point Overall Standard Deviation	1
3	Derivative of Spectral Roll-off Point Overall Average	1
4	Derivative of Spectral Roll-off Point Overall Standard Deviation	1
5	Running Mean of Spectral Roll-off Point Overall Average	1
6	Running Mean of Spectral Roll-off Point Overall Standard Deviation	1
7	Standard Deviation of Spectral Roll-off Point Overall Average	1
8	Standard Deviation of Spectral Roll-off Point Overall Standard Deviation	1
9	Derivative of Running Mean of Spectral Roll-off Point Overall Average	1
10	Derivative of Running Mean of Spectral Roll-off Point Overall Standard Deviation	1
11	Derivative of Standard Deviation of Spectral Roll-off Point Overall Average	1
12	Derivative of Standard Deviation of Spectral Roll-off Point Overall Standard Deviation	1

Table 3.6 Spectral Flux statistics

	Name	Num of Values
1	Spectral Flux Overall Average	1
2	Spectral Flux Overall Standard Deviation	1
3	Derivative of Spectral Flux Overall Average	1
4	Derivative of Spectral Flux Overall Standard Deviation	1
5	Running Mean of Spectral Flux Overall Average	1
6	Running Mean of Spectral Flux Overall Standard Deviation	1
7	Standard Deviation of Spectral Flux Overall Average	1
8	Standard Deviation of Spectral Flux Overall Standard Deviation	1
9	Derivative of Running Mean of Spectral Flux Overall Average	1
10	Derivative of Running Mean of Spectral Flux Overall Standard Deviation	1
11	Derivative of Standard Deviation of Spectral Flux Overall Average	1
12	Derivative of Standard Deviation of Spectral Flux Overall Standard Deviation	1

- Mel Frequency Cepstral Coefficients (MFCCs): MFCCs is the feature popularly used for speech recognition and audio classification due to it's effectiveness in representing the spectral variations of audio. The MFCCs stands for the shape of the spectrum with few coefficients. The cepstrum is the Fourier Transform (or Discrete Cosine Transform DCT) of the logarithm of the spectrum. Instead of the Fourier spectrum, the Mel-cepstrum is the cepstrum computed on the Mel-bands. By using of the mel scale, the mid-frequencies part of the signal is taken better into account. The MFCCs are the

coefficients of the Mel cepstrum. MFCCs are commonly derived as follows:

1. Take the Fourier Transform of (a windowed excerpt of) a signal.
2. Map the log amplitudes of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the Discrete Cosine Transform of the list of mel log-amplitudes, as if it were a signal.
4. The MFCCs are the amplitudes of the resulting spectrum.

Table 3.7 shows the twelve statistics of MFCCs in our work.

Table 3.7 MFCCs statistics

	Name	Num of Values
1	MFCCs Overall Average	13
2	MFCCs Overall Standard Deviation	13
3	Derivative of MFCCs Overall Average	13
4	Derivative of MFCCs Overall Standard Deviation	13
5	Running Mean of MFCCs Overall Average	13
6	Running Mean of MFCCs Overall Standard Deviation	13
7	Standard Deviation of MFCCs Overall Average	13
8	Standard Deviation of MFCCs Overall Standard Deviation	13
9	Derivative of Running Mean of MFCCs Overall Average	13
10	Derivative of Running Mean of MFCCs Overall Standard Deviation	13
11	Derivative of Standard Deviation of MFCCs Overall Average	13
12	Derivative of Standard Deviation of MFCCs Overall Standard Deviation	13

3.5 Textual Feature

Monologue and dialogue are two types of spoken languages in films. Monologue happens when the character talks to self or the off-screen person narrates the background story to facilitate our understanding about the scene or the film. Dialogue can be the conversation among more than two actors.

The textual feature for each scene is defined as a set of the feeling words appearing in the caption stream of the scene. The feeling word list we used for textual feature extraction is collected by Hein [42]. There are over 3000 words in total. We divided these feeling words into two classes – positive and negative feelings for similarity measure. Each feeling word belongs to one class.

4. Emotion Discovery

In this section, we will introduce the output emotion categories we used in this report, the graph-based algorithm – Mixed Media Graph (MMG) and two topologies revised from MMG will also be introduced.

4.1 Emotion Taxonomy

Affective perception includes emotion, feeling, and mood. Emotion and feeling are different responses. Emotion is directed outwards whereas feeling is inward [6]. Also, emotion and mood are different concepts. Emotion is aroused by some events or objects and usually lasts for a few minutes, while mood is an emotional state and lasts for a longer period of time compared with emotion [17].

Many emotion models were proposed in psychology. For dimensional approach to describe emotions, VAD is the most popular model proposed by Osgood et al. [26] and also Russell and Mehrabian [34]. The VAD comprises three basic dimensions: *Valence*, *Arousal*, and *Dominance*. Valence is characterized as a continuous range of affective responses or states ranging from pleasant to unpleasant, while Arousal is characterized by affective states ranging on a continuous scale from excited to calm. We can also say that Arousal stands for the "intensity" of emotion, while Valence stands for the "type" of emotion. The third dimension – Dominance – is useful to distinguish between emotional states with similar Arousal and Valence (e.g., "grief" and "rage") and ranges from "no control" to "full control". Therefore, the entire scope of human emotions can be represented as a set of points in the three-dimensional VAD coordinate space, as shown in Figure 4.1. The influence of Dominance, however, is truly small so that the control dimension can be ignored. Consequently, the emotion space is reduced to the projection of three-dimensional surface onto the arousal-valence plane as shown in Figure 4.2.

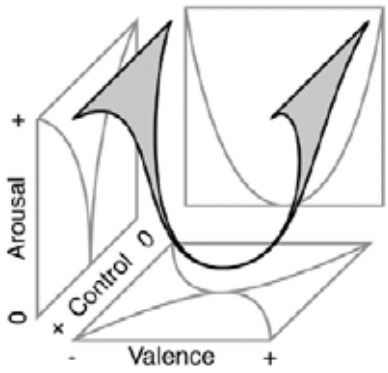

Figure 4.1 Illustration of the 3-D emotion space (from Dietz and Lang [7])

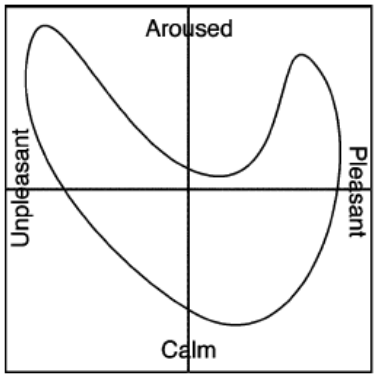

Figure 4.2 Illustration of the 2-D emotion space (from Dietz and Lang [7])

Thayer [38] adapted Russell's model to divide music mood into four classes based on a two-dimensional Energy-Stress mood model. The Energy dimension corresponds to the arousal, while Stress corresponds to valence in Russell's model. As shown in Figure 4.3, *Contentment* refers to happy and calm music, such as Bach's "Jesus, Joy of Man's Desiring"; *Depression* refers to tired and tense music, such as the opening of Stravinsky's "Firebird"; *Exuberance* refers to happy and energetic music such as Rossini's "William Tell Overture"; and *Anxious*/*Frantic* refers to tense and energetic music, such as Berg's "Lulu".

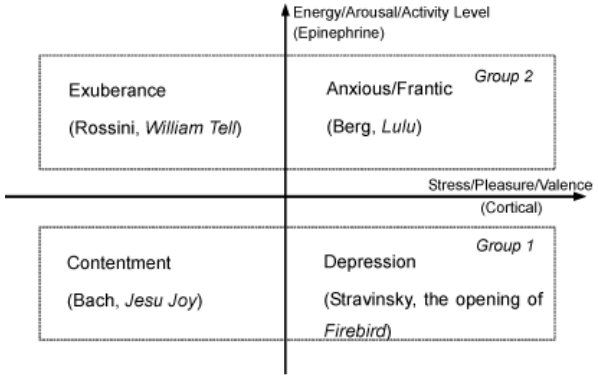

Figure 4.3 Illustration of Thayer's two-dimensional mood model [38]

Based on human facial expressions of emotion, Ekman [9] identified the six basic emotions: *Happy*, *Surprise*, *Anger*, *Sad*, *Fear* and *Disgust*. The set of emotions is found to be universal among humans.

Ortony et al. [25] proposed an emotion model based on the assumption that an emotion is a reaction to events (pleased versus displeased), agents (approving versus disapproving) or objects (liking versus disliking). There are twenty-two emotions in this model: *Happy-for*, *Resentment*, *Gloating*, *Pity*, *Hope*, *Fear*, *Satisfaction*, *Fears-Confirmed*, *Relief*, *Disappointment*, *Joy*, *Distress*, *Pride*, *Shame*, *Admiration*, *Reproach*, *Gratification*, *Remorse*, *Gratitude*, *Anger*, *Love*, and *Hate*.

In our work, we mainly use the emotion categories proposed by Wang and Cheong [14]. In order to fit emotion categories for film domain application, Wang et al modified Ekman's six basic emotions as follows:

1. *Disgust* is dropped due to the lack of scenes that seek to evoke "pure" *Disgust*. In addition, scenes that evoke *Disgust* often contain a strong element of *Fear*.
2. Add a *Neutral* emotion category to emotion list because many scenes in films are emotionally neutral.
3. Partition *Happy* emotion category into *Joyous* and *Tender Affections*. *Happy* includes many sub-families of positive feelings because these sub-families don't have their own unique facial expression besides smiling expression. Therefore, it is useful and cinematically relevant to partition *Happy*.

Therefore, the emotion categories are *Joyous*, *Tender Affections* (abbreviated as *TA*), *Anger*, *Sad*, *Fear*, *Surprise*, and *Neutral*. This set obeys the following four criteria and is appropriate for scene-level content in film:

1) Universality: Each emotion can be universally realized and experienced.
2) Distinctiveness: Each emotion is clearly discriminated from the other.
3) Utility: Each emotion should have significant relevance in the film context.
4) Comprehensiveness: This emotion set should be adequate to describe almost all emotions in films.

4.2 Mixed Media Graph (MMG)

After extracting audiovisual features, we use Mixed Media Graph [29] to discover the emotion for the query scene. MMG is a graph-based approach used to find correlations across the media in a collection of multimedia objects. Through MMG, the association between audiovisual features and emotions in scenes can be discovered.

The problem and assumption of MMG are defined as follows:

PROBLEM 4.1: Given a set S of n multimedia objects $S = \{O_1, O_2, \ldots, O_n\}$, each with m multimedia attributes, find patterns/correlations among the objects and attributes.

DEFINITION 4.1: The domain D_i of (set-valued) attribute i is the collection of atomic values that attribute i can choose from. The values of domain D_i will be referred to as the domain tokens of D_i. *Di* can consist of categorical values, numerical values, or numerical vectors.

ASSUMPTION 4.1: For each domain D_i ($i = 1, \ldots, m$), we are given a similarity function $s_i(*, *)$ which assigns a score to each pair of domain tokens.

As shown in Figure 4.4, there are two types of vertices in MMG: object nodes and attribute value nodes. Object nodes stand for multimedia object (e.g. O_1, O_2, O_3), and attribute value nodes represent associated attribute values of object nodes. For example, there are two types of attributes value node: r_i (i = 1, 2, … , 8) and e_j (j = 1, 2, 3, 4). It is permitted that object nodes have missing attribute value nodes. For object nodes with m types of attributes, MMG will be an $(m+1)$-layer graph with m types of nodes and one more type of nodes for the objects.

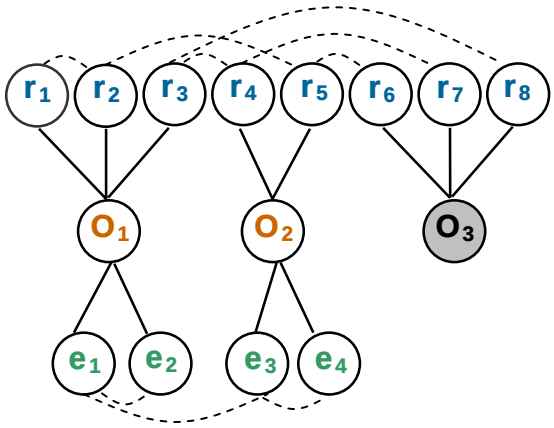

Figure 4.4 Mixed Media Graph with two attribute type: *r* and *e*

The edges in MMG are treated as un-directional. There are two types of links in MMG: *object-attribute-value link* (*OAV-link*), and *nearest neighbor link* (*NN-link*). *OAV-link* is the link between an object node and an attribute value node, represented by solid arc. *NN-link* is the link between two similar attribute value nodes that are in the same type, represented by dashed arc. Each attribute value node has k *NN-link* linking to k most similar attribute value nodes. Note that some attribute value nodes have degree greater than k because the edges are un-directional and nearest neighbor relationship is not symmetric.

After constructing MMG, the “random work with restart” mechanism is used for estimating the affinity of attribute value node to the query node. In detail, take Figure 4.4 for example, to compute

the affinity of attribute node e_2 with respect to query node O_3 - $A_{O3}(e_2)$ - consider a random walker starting from node O_3 selects randomly among the available edges every time, except returning to O_3 with probability c. $A_{O3}(e_2)$ is the steady-state probability that the random walker will reach e_2 from O_3.

In this way, the affinities of e_j (j = 1, 2, 3, 4) with respect to O_3 can be computed. The attribute nodes of *e* type having higher affinity can be regard as the attribute value nodes of O_3.

4.3 Scene Affinity Graph

In this report, we exploit MMG as our classification algorithm and denote our graph as Scene Affinity Graph (SAG). The SAG represents the relationship between audiovisual features and emotions of scenes. We propose two topologies for the Scene Affinity Graph and modify MMG as weighted graph according to different types of attribute. In the MMG, the "random work with restart" mechanism is used to evaluate the affinity of attribute value node to the query node. A random walker randomly chooses one edge of the available edges to move to the next node. But in the Scene Affinity Graph, each type of attributes can be given a weight to characterize its influence.

The two topologies are SAG with seven-attribute and SAG with four-attribute. There are seven type of attribute value node in SAG with seven-attribute - *color*, *light*, *tempo*, *close-up*, *audio*, *text*, and *emotion*. SAG with four-attribute is revised from SAG with seven-attribute by combining the *color*, *light*, *tempo* and *close-up* attributes into *vision* attribute.

4.3.1 Scene Affinity Graph with seven-attribute

The problem and assumption of the SAG with seven- attribute are defined as follows:

PROBLEM 4.2: Given a set *S* of *n* scene objects $S = \{SE_1, SE_2, \ldots, SE_n, Q\}$, each with seven attributes – *color*, *light*, *tempo*, *close-up*, *audio*, *text*, and *emotion* features, to find the association among the query scene object (*Q*) and *emotion* attribute.

DEFINITION 4.2: The domain D_{color} (D_{light}, D_{tempo}, $D_{close\text{-}up}$, D_{audio}) of attribute *color* (*light*, *tempo*, *close-up*, *audio*) is the collection of atomic values that attribute *color* (*light*, *tempo*, *close-up*, *audio*) can choose from. The values of domain D_{color} (D_{light}, D_{tempo}, $D_{close\text{-}up}$, D_{audio}) will be referred to as the domain tokens of D_{color} (D_{light}, D_{tempo}, $D_{close\text{-}up}$, D_{audio}). D_{color} (D_{light}, D_{tempo}, $D_{close\text{-}up}$, D_{audio}) consist of numerical vectors.

DEFINITION 4.3: The domain D_{text} ($D_{emotion}$) of attribute *text* (*emotion*) is the collection of atomic values that attribute *text* (*emotion*) can choose from. The values of domain D_{text} ($D_{emotion}$) will be referred to as the domain tokens of D_{text} ($D_{emotion}$). D_{text} ($D_{emotion}$) consist of categorical values.

ASSUMPTION 4.2: For each domain D_i (*i = color, light, tempo, close-up, audio, text* and *emotion*), we are given a similarity function $s_i(*, *)$ (*i = color, light, tempo, close-up, audio, text* and *emotion*) which assigns a score to each pair of domain tokens.

Graph Construction

1. Object-Attribute-Value link

In scene affinity graph, each object node stands for one scene (video clip). For each object node in SAG with seven-attribute, there are seven types of attribute value nodes – *color*, *light*, *tempo*, *close-up*, *audio*, *text* and *emotion*. Given one object node (i.e. one scene), the number of attribute value nodes in each attribute is given as follows:

- For *emotion* attribute: because each scene belongs to one emotion, the object node has one *emotion* attribute value node.
- For *color* (*light*, *tempo*, *close-up*) attribute: the number of *color* (*light*, *tempo*, *close-up*) attribute value nodes depends on the number of shots in the scene.
- For *audio* attribute: the number of *audio* attribute value nodes is based on the number of shots that is no less than one second in the scene.
- For *textual* attribute: the number of *textual* attribute value nodes is based on the number of feeling words in the scene. One *textual* attribute value node stands for one feeling word and the number of feeling words in a scene could be zero.

2. Nearest Neighbor link

- For *emotion* attribute: the edge between two *emotion* attribute value nodes are linked only when they are of the same emotion.
- For *color* (*light*, *tempo*, *close-up*, *audio*) attribute: the edge between two *color* (*light*, *tempo*, *close-up*, *audio*) attribute value nodes are constructed based on *k*-nearest neighbors.
- For *textual* attribute: the edge between two *textual* attribute value nodes is linked when these two nodes belong to the same class (positive or negative feeling).

3. Similarity measure for seven attributes

- *color* attribute: the similarity of two nodes is the average of the four features' similarity –FH, MPH, MDH and DCR. The similarity measure for the four features is defined as follows,
  - FH, MPH and MDH: the similarities measures for the three features are the same. It is defined as bin-wise histogram intersection [8]:
    - For two node *x*, *y* in *color* attribute with color histogram H_x, H_y

$$sim(x, y) = \sum_{i=1}^{B} \frac{\min(H_x(i), H_y(i))}{\max(H_x(i), H_y(i))}$$

where i is the bin number, and B is the total number of color bins.

$$B = \begin{cases} 264, & \text{when color histogram is FM.} \\ 12, & \text{when color histogram is MPH.} \\ 144, & \text{when color histogram is MDH.} \end{cases}$$

- DCR: the ratio of smaller value over larger value.
  - For two node *x, y* in *color* attribute with DCR_x, DCR_y

$$sim_{DCR}(x, y) = \frac{\min(DCR_x, DCR_y)}{\max(DCR_x, DCR_y)}$$

- *light* attribute: the ratio of smaller value over larger value.
  - For two node *x, y* in *color* attribute with $lightness_x$, $lightness_y$

$$sim_{light}(x, y) = \frac{\min(lightness_x, lightness_y)}{\max(lightness_x, lightness_y)}$$

- *tempo* attribute: the similarity is defined as the average of similarities for SL and NSPM features.
  - SL: the ratio of smaller value over larger value.
    - For two node *x, y* in *tempo* attribute with SL_x, SL_y

$$sim_{SL}(x, y) = \frac{\min(SL_x, SL_y)}{\max(SL_x, SL_y)}$$

  - NSPM: the ratio of smaller value over larger value.
    - For two node *x, y* in *tempo* attribute with $NSPM_x$, $NSPM_{ys}$

$$sim_{SL}(x, y) = \frac{\min(NSPM_x, NSPM_y)}{\max(NSPM_x, NSPM_y)}$$

- *close-up* attribute: the similarity is defined as the average of similarities for CU and CUR features.
  - CU: the ratio of smaller value over larger value.
    - For two node *x, y* in *close-up* attribute with $NSPM_x$, $NSPM_{ys}$

$$sim_{CU}(x, y) = \frac{\min(CU_x, CU_y)}{\max(CU_x, CU_y)}$$

- CUR: the ratio of smaller value over larger value.
  - For two node *x, y* in *close-up* attribute with CUR_x, CUR_{ys}

$$sim_{CUR}(x, y) = \frac{\min(CUR_x, CUR_y)}{\max(CUR_x, CUR_y)}$$

- *audio* attribute: the similarity of two nodes of *audio* attribute is defined as the average of five audio features' similarities. The similarities of audio features are defined as Euclidian Distance.
- *textual* attribute: for two nodes t_1, t_2 in *textual* attribute (i.e. two feeling words):

$$sim_T(t_1, t_2) = \begin{cases} 1, \text{ if } x \text{ and } y \text{ are the same.} \\ 0.8, \text{ if } x \text{ and } y \text{ are different but belong to the same category.} \\ 0, \text{ otherwise.} \end{cases}$$

- *emotion* attribute: for two node e_1, e_2 in *emotion* attribute (i.e. two emotion words):

$$sim_E(e_1, e_2) = \begin{cases} 1, \text{ if } x, y \text{ are the same.} \\ 0, \text{ if } x, y \text{ are different.} \end{cases}$$

For example, Figure 4.5 illustrates the constructed SAG with seven-attribute with respect to the query scene object *Q*, in which there are two training scene object nodes - {SE_1, SE_2}. Scene SE_1 has two shots and both are more than one second. Scene SE_2 has two shots and only one shot is more than one second. There are one and two feeling words in SE_1 and SE_2 respectively. For scene *Q*, it has two shots and both are more than one second. There is one feeling word in *Q*. In this case,

- SE_1 has *color* attribute nodes - {cr_{11}, cr_{12}}, *light* attribute nodes - {lt_{11}, lt_{12}}, *tempo* attribute nodes - {tp_{11}, tp_{12}}, *close-up* attribute nodes - {cp_{11}, cp_{12}}, *audio* attribute nodes - {a_{11}, a_{12}}, *textual* attribute nodes - {t_{11}, t_{12}}, and *emotion* attribute node - e_{11}
- SE_2 has *color* attribute nodes - {cr_{21}, cr_{22}}, *light* attribute nodes - {lt_{21}, lt_{22}}, *tempo* attribute nodes - {tp_{21}, tp_{22}}, *close-up* attribute nodes - {cp_{21}, cp_{22}}, *audio* attribute node - {a_{21}}, *textual* attribute node - {t_{21}, t_{22}}, and emotion attribute node - e_{21}.
- The query scene object *Q* has *color* attribute nodes - {cr_{q1}, cr_{q2}}, *light* attribute nodes - {lt_{q1}, lt_{q2}}, *tempo* attribute nodes - {tp_{q1}, tp_{q2}}, *close-up* attribute nodes - {cp_{q1}, cp_{q2}}, *audio* attribute node - {a_{q1}, a_{q2}}, and *textual* attribute nodes - {t_{q1}}. In Figure 5.2, the number of nearest-neighbors, *k*, is set to one.

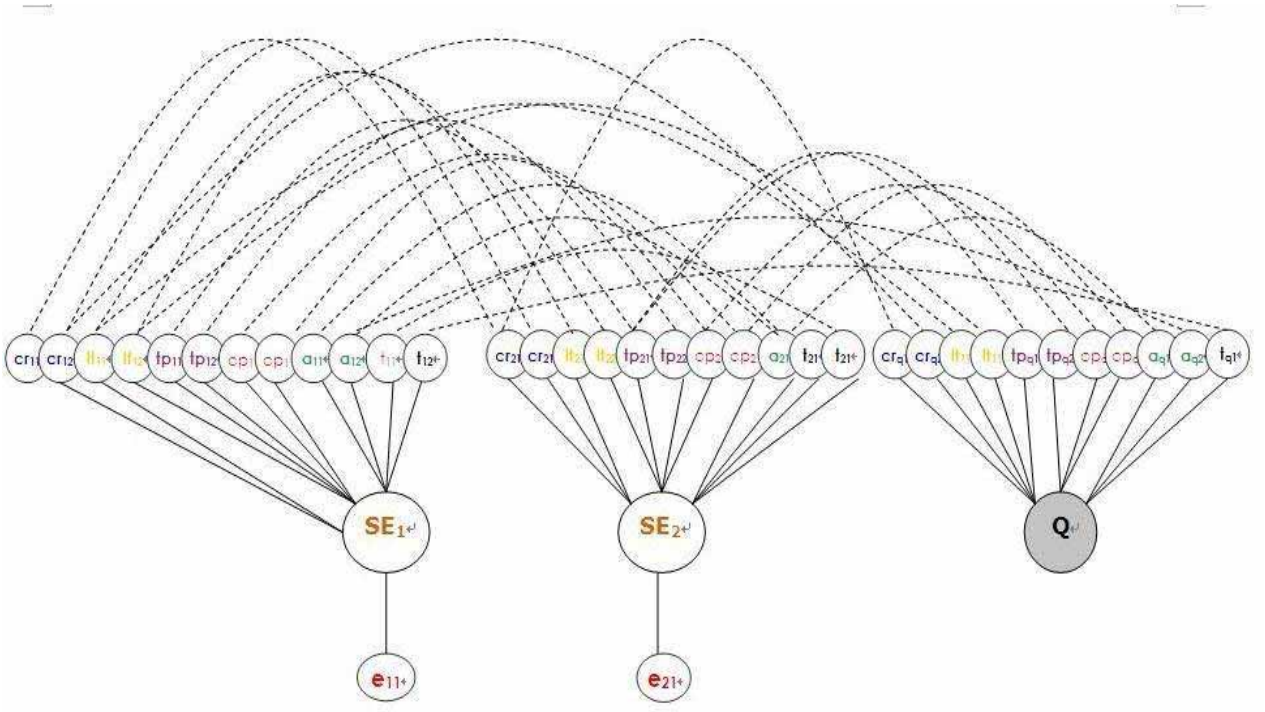

Figure 4.5 Scene Affinity Graph with seven-attribute (*k* = 1)

4.3.2 Scene Affinity Graph with four-feature

The problem and assumption of the SAG with four-attribute are similar with those of SAG with seven-attribute. *Color*, *light*, *tempo*, and *close-up* attributes in SAG with seven-attribute are combined into the *vision* attribute in SAG with four- attribute.

PROBLEM 5.3: Given a set *S* of *n* scene objects $S = \{SE_1, SE_2, \ldots, SE_n, Q\}$, each with four attributes – *vision*, *audio*, *textual*, and *emotion* features, to find the association among the query scene object (*Q*) and *emotion* attribute.

DEFINITION 5.4: The domain D_{vision} (D_{audio}) of attribute *vision* (*audio*) is the collection of atomic values that attribute *vision* (*audio*) can choose from. The values of domain D_{vision} (D_{audio}) will be referred to as the domain tokens of D_{vision} (D_{audio}). D_{vision} (D_{audio}) consist of numerical vectors.

DEFINITION 5.5: The domain $D_{textual}$ ($D_{emotion}$) of attribute *textual* (*emotion*) is the collection of atomic values that attribute *textual* (*emotion*) can choose from. The values of domain $D_{textual}$ ($D_{emotion}$) will be referred to as the domain tokens of $D_{textual}$ ($D_{emotion}$). $D_{textual}$ ($D_{emotion}$) consist of categorical values.

ASSUMPTION 5.3: For each domain D_i (*i = vision, audio, textual* and *emotion*), we are given a similarity function $s_i(*, *)$ (*i = vision, audio, textual* and *emotion*) which assigns a score to each pair of domain tokens.

Graph Construction

1. Object-Attribute-Value link

In scene affinity graph, each object node stands for one scene (video clip). For each object node in SAG with four-attribute, there are four types of attribute value nodes – *vision*, *audio*, *textual* and *emotion*. Each *vision* attribute value node is composed of *color*, *light*, *tempo*, and *close-up* features. Given one object node (one scene), the number of attribute value nodes in each attribute is:

- For *emotion* attribute: because each scene belongs to one emotion, the object node has one *emotion* attribute value node.
- For *vision* attribute: the number of *vision* attribute value nodes depends on the number of shots in the scene.
- For *audio* attribute: the number of *audio* attribute value nodes is based on the number of shots that are more than or equal to one second in the scene.
- For *textual* attribute: the number of *textual* attribute value nodes is based on the number of feeling words in the scene. One *textual* attribute value node stands for one feeling word and feeling words in a scene could be more than or equal to zero.

2. Nearest Neighbor link

- For *emotion* attribute: the edge between two *emotion* attribute value nodes are linked only when they are the same emotion.
- For *vision* attribute: the edge between two *vision* attribute value nodes are constructed based on *k*-nearest neighbors.
- For *textual* attribute: the edge between two *textual* attribute value nodes is linked when they belong to the same class (positive/negative).

3. Similarity measure for four attributes

The similarity measure for all attributes in SAG with four-attribute is the same with above except *vision* attribute. *Vision* attribute consists of nine features, so the similarity is defined as the average of the nine features' similarity defined above.

Take Figure 4.6 for example, Figure 4.6 illustrates the constructed SAG with four-attribute with respect to the query scene object *Q* revised from Figure 4.5, in which there are two training scene object nodes - $\{SE_1, SE_2\}$ and one testing scene *Q* the same as Figure 4.5. In this case,

- SE_1 has *vision* attribute nodes - $\{v_{11}, v_{12}\}$, *audio* attribute nodes - $\{a_{11}, a_{12}\}$, *textual* attribute nodes - $\{t_{11}, t_{12}\}$, and *emotion* attribute node - e_{11}.

- SE_2 has *vision* attribute nodes - {v_{21}, v_{22}}, *audio* attribute node - {a_{21}}, *textual* attribute node - {t_{21}, t_{22}}, and emotion attribute node - e_{21}.
- The query scene object Q has *vision* attribute nodes - {v_{q1}, v_{q2}}, *audio* attribute node - {a_{q1}, a_{q2}}, and *textual* attribute nodes - {t_{q1}}. In Figure 4.6, the number of nearest-neighbors, k, is set to one.

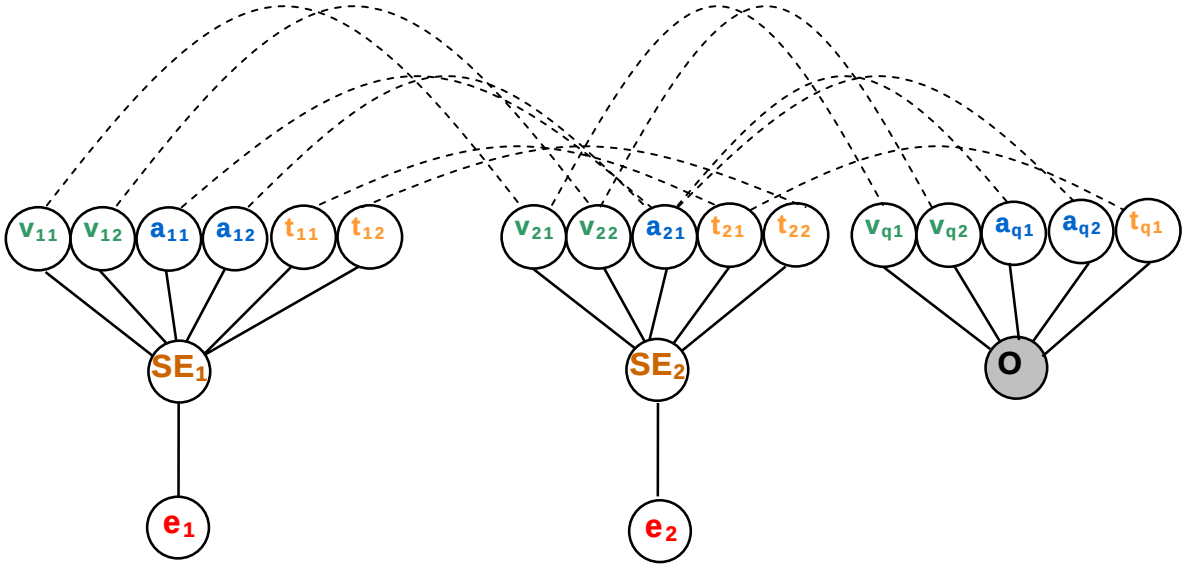

Figure 4.6 Scene Affinity Graph with four-attribute (k = 1)

5. Experiments and Results

5.1 Implementation

Our data consists of ninety scenes which come from twelve movies as shown in Table 5.1. The scenes that were segmented manually according to the criteria adopted in [20]. Fifteen scenes are selected for each emotion category. Every testing scene is classified into one of the six emotions.

To establish the ground truth, the emotions of each scene were manually labeled by two people according to Table 5.2. We took five-fold cross-validation in our experiments. In each test, the Scene Affinity Graph was constructed from the training set and one of the testing scenes.

The steps for scene classification are as below:

Step 1: First, the emotions of the training scenes are labeled manually according to Table 5.2. Each scene has one emotion.

Step 2: Then, the visual, audio and caption stream of each scene are extracted, and all scenes were

segmented into shots.

Step 3: The visual, audio and caption streams are used to compute fifteen features stated in Section 3.

Step 4: After feature extraction, the output features of all scenes and the emotions of the training scenes are used to construct SAG to output the emotion of the testing scene.

Table 5.1 Twelve movies used in our system

Genre	movies
Action	Minority Report
	The Patriot
Drama	Titanic
	In Her Shoes
	Life Is beautiful
	The Departed
	The Pursuit of Happyness
	The Sixth Sense
Romance	Must Love Dogs
	Bruce Almighty
Horror	House Of Wax
	I Know What You Did Last Summer

Table 5.2 Emotion Categories vs. Viewer's feeling [14]

Emotion Categories	Viewer's Feeling
Joyous	exuberance, joyous, enjoyment, happy
Tender Affection	heart-warming, tender, sentimental, relaxed
Sad	depressedsad, bad, hopeless
Fear	scary, fearful, terrified
Anger	exciting, dangerous, aggressive, angry
Surprise	surprised, tense, anticipation

5.1.1 Preprocessing

The visual, audio and caption stream of each scene are processed separately. The visual stream is segmented into shots according to frame difference, and for each shot, one key-frame is extracted to represent the visual detail of the shot. The audio stream is also segmented according to shot boundaries while segments less than one second are being dropped. For caption stream, all words belong to this scene were picked up for later process.

5.1.2 Visual Feature Extraction

All shots (represented as key-frames), audio segments, and words of each scene are extracted respective features that are defined in Section 4.2.

For each key-frame, we transformed RGB color space into CIELUV space [30]. The CIELUV color space has the advantage of perceptual uniformity, i.e. the same distance between two different points makes equal perceived color difference. The CIELUV color space is divided uniformly into 264 color bins in total. The *color* feature (FM, MPH, MDH and DCR) and *light* feature (lightness) are computed by the color histogram of 264 color bins. The *tempo* and *close-up* features are also calculated. We exploited the face detection function in Open Source Computer Vision Library (OpenCV) [41] to detect the position and the region of a face in a frame. Close-ups are determined by the proportion of detected face size to the frame size. Therefore, totally nine features are extracted to represent each shot. It is noted that MPH, MDH, NSPM, and CUR are features at scene-level, the values of the four features are the same in every shot of one scene. Table 5.3 shows the number of values for each feature in *vision* type. In SAG with four-attribute, each node in *vision* attribute is represented as a vector of 426 dimensions, while in SAG with seven attribute, each node in *color*(*light*, *tempo*, *close-up*) attribute is represented as a vector of 204 (1, 2, 2) dimensions.

Table 5.3 Number of values in *vision* type.

Type	Classes	Feature Representations	Num of values
Vision	Color	Family Histogram	264
		Movie Palette Histogram	12
		Mood Dynamics Histogram	144
		Dominant Color Ratio	1
	Light	Lightness	1
	Tempo	Shot Length	1
		Number of Shots Per Minute	1
	Close-up	Close-Up	1
		Close-Up Ratio	1

5.1.3 Audio Feature Extraction

The audio segments that are more than one second are reserved for feature extraction. Each audio segment is 48kHz sample rate, mono channel and 16bit per sample. In audio feature extraction, five features are extracted from each audio segment. Each audio segment is stands for one *audio* attribute node in SAG. Therefore, each *audio* attribute node in SAG is represented as a vector of 204 dimensions, as shown in Table 5.4.

Table 5.4 Number of values in *audio* type.

Type	Feature Representations	Num of values
Audio	Zero Crossing Rate	12
	Spectral Roll Off	12
	Spectral Centroid	12
	Spectral Flux	12
	Mel Frequency Cepstral Coefficients	156

5.1.4 Textual Feature Extraction

After preprocessing step, the subtitles belong to one scene are collected. In textual feature extraction, these subtitles are compared with the reference feeling word list [42]. If the word (in subtitles) is found in the reference list, it will be picked up from the subtitles of the scene. The reference feeling word list is over 3,000 words. Owing to the limitation of space, we only show the feeling words appearing in our training movies, as shown in Table 5.5. We classified the feeling words of all training scenes into two categories – positive and negative for the similarity measure of two nodes in SAG's *textual* attribute.

5.1.5 Emotion Discovery

The emotion discovery is performed by the Scene Affinity Graph (SAG) mentioned in section 4.3 - SAG with seven-attribute and SAG with four-attribute. In average, the SAG with seven-attribute contains 18300 nodes, and the SAG with four-attribute contains 7400 nodes.

The output emotion of the testing scene is determined by two criteria. The first criterion is the *emotion* node with the maximal probability in SAG, the other is the emotion category with the maximal cumulative probability.

Based on the two proposed topologies, we present two experiments - SAG with seven-attribute and SAG with four-attribute.

Table 5.5 Feeling words appear in out work

A	buried	dear	fired	hollow
afraid	busy	decent	firm	honest
against	**C**	defensive	Fit	honored
alarmed	calm	delivered	Flip	hopeless
alert	careful	depressed	flush	horny
alive	caught	desperate	forced	hot
almighty	charming	dire	forgiven	hurt
alone	cheap	dirty	foul	**I**
amazing	chicken	disgruntled	frank	inappropriate
angry	clean	distant	freaked	incorrigible
appealing	clear	divorced	free	incredible
ardent	clear	down	fresh	innocent
ashamed	close	dropped	fruitful	insane
awake	closed	drunk	fulfilled	intelligent
aware	cold	dull	fun	**J**
awesome	comfortable	**E**	funny	jealous
awful	comfy	embarrassed	**G**	just
awkward	concerned	engaged	gentle	**K**
afraid	confident	enlightened	giving	kind
B	considered	erased	glad	**L**
bad	cool	evil	good	legitimate
beat	correct	excellent	gorgeous	light
beautiful	crap	excited	grand	liked
beloved	crossed	**F**	grave	lonely
better	cruel	fair	great	lost
black	crushed	faithful	grief	love
bliss	cut	fake	**H**	loved
blue	cute	false	happy	lovely
bored	**D**	fantastic	hate	loving
brave	damaged	fast	heavy	lucky
bright	dangerous	fear	helped	**M**
brilliant	dark	filthy	helpful	mad
broken	dead	fine	hilarious	manly

Table 5.5 Feeling words appear in out work (cont.)

mature	poor	sad	Sunny	unlucky
mean	positive	safe	Super	upset
mediocre	powerful	sane	Sure	**V**
messy	precious	satisfied	Surly	violent
mighty	pretty	scared	Sweet	vital
moping	professional	screwed	**T**	vivid
N	promiscuous	secure	talented	voluptuous
naked	promised	serious	Tender	vulnerable
nasty	proper	sharp	Tense	**W**
natural	protected	shocked	Terrible	wacky
negative	proud	sick	Terrific	wanting
nervous	psychotic	simple	terrified	warm
nice	punished	smart	thankful	wasted
normal	pure	smooth	Thrilled	weak
nuts	**R**	soft	Tired	weird
O	ready	solid	Torn	welcome
okay	real	sorrow	Tough	willing
open	reasonable	sorry	true	wise
P	regret	stranded	trusting	wonderful
pain	reliable	strange	typical	worried
panic	rich	strong	**U**	worse
panicked	ridiculous	stubborn	uncomfortable	worthy
passionate	right	stuck	uncovered	wrong
perfect	romantic	stunning	undesirable	
pissed	rough	stupid	unfair	
plain	rude	subtle	unfit	
pleasure	**S**	suffering	unique	

5.2 Experiment on SAG with seven-attribute

This experiment results show that the weights for *color*, *light*, *tempo*, *close-up*, *audio*, *textual*, and *emotion* in SAG are 0:0:1:0:5:0:5 respectively.

As Figure 5.1 shows, firstly we can see there is no clear relationship between the accuracy and the restart probability in the experiment based on the *emotion* node with the maximal probability as the square nodes shows. Second, as the restart probability increase, the accuracy based on the emotion with the maximal cumulative probability will raise as the triangle nodes shows. The maximal accuracy is about 61% with 0.9 restart probability.

As Figure 5.2 shows, generally, based on the two accuracy measures, the accuracy declines as the value of *k* increase.

Because the boundary between *Fear* and *Surprise* emotion is usually confused, we combine them into one emotion. Figure 5.3 and 5.4 show the corresponding modified results. We can see that the maximal accuracy is about 66% when k = 5, 5% better than before.

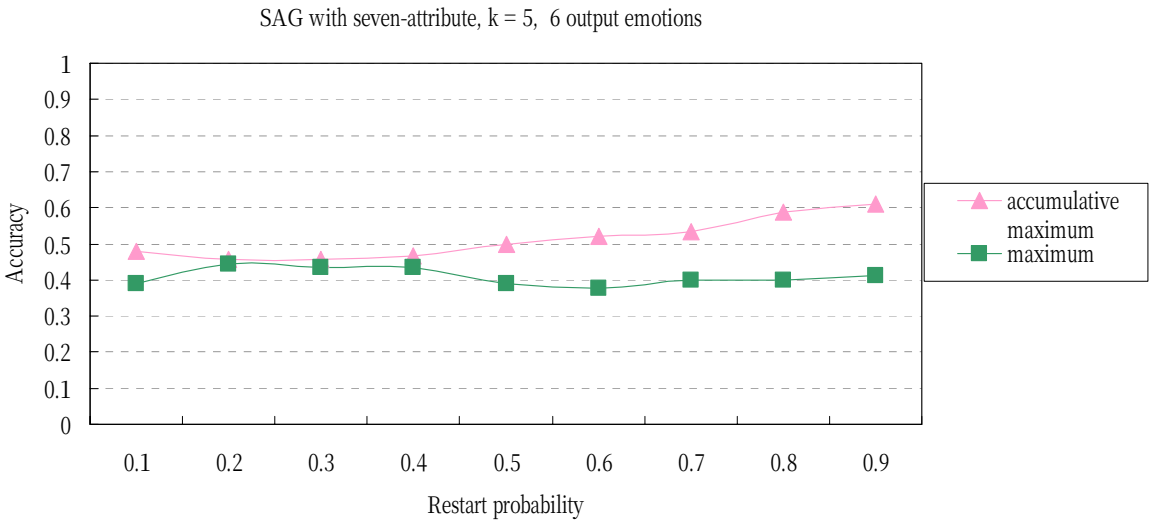

Figure 5.1 Accuracy with different restart probability

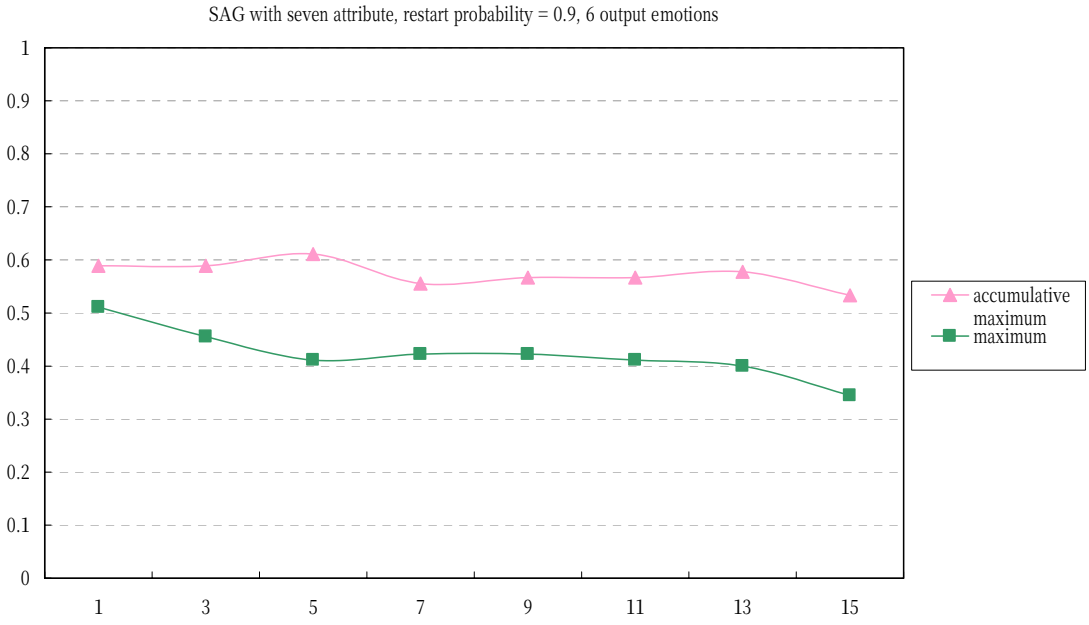

Figure 5.2 Accuracy with different *k* value.

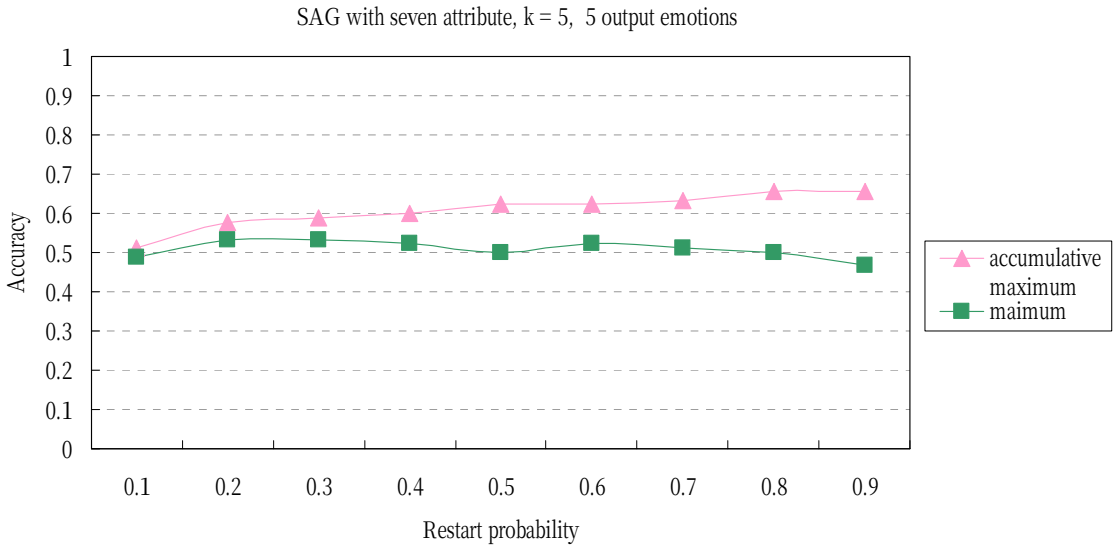

Figure 5.3 Accuracy with different restart probability

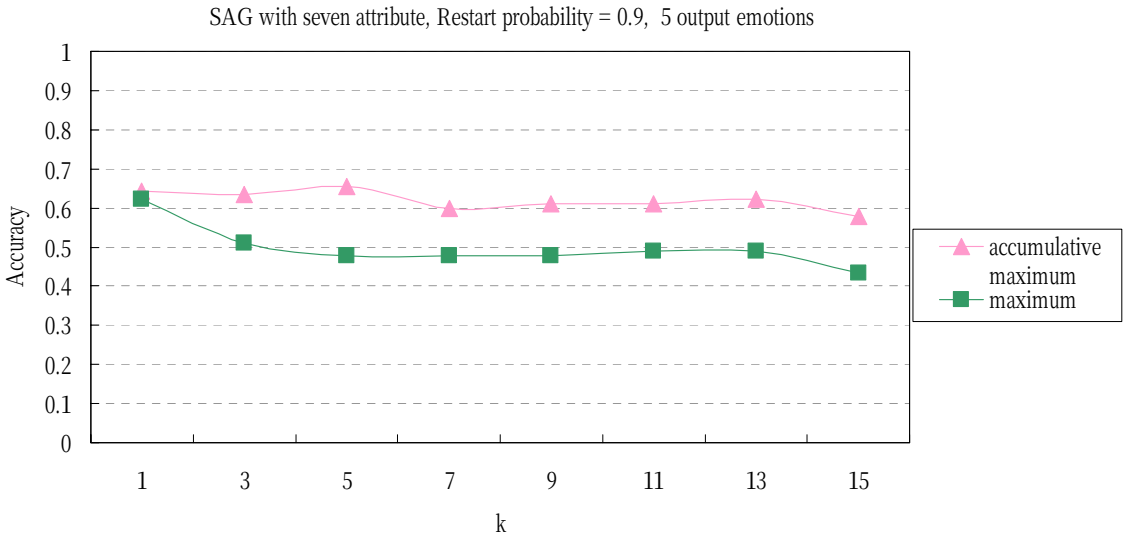

Figure 5.4 Accuracy with different *k* value

5.3 Experiment on SAG with four-attribute

This experiment results show that the weights for *vision*, *audio*, *textual*, and *emotion* in SAG are 0:1:0:1 respectively.

As Figure 5.5 shows, when the restart probability increase, the accuracy raise based on both criteria. The maximal accuracy is about 64% with 0.9 restart probability.

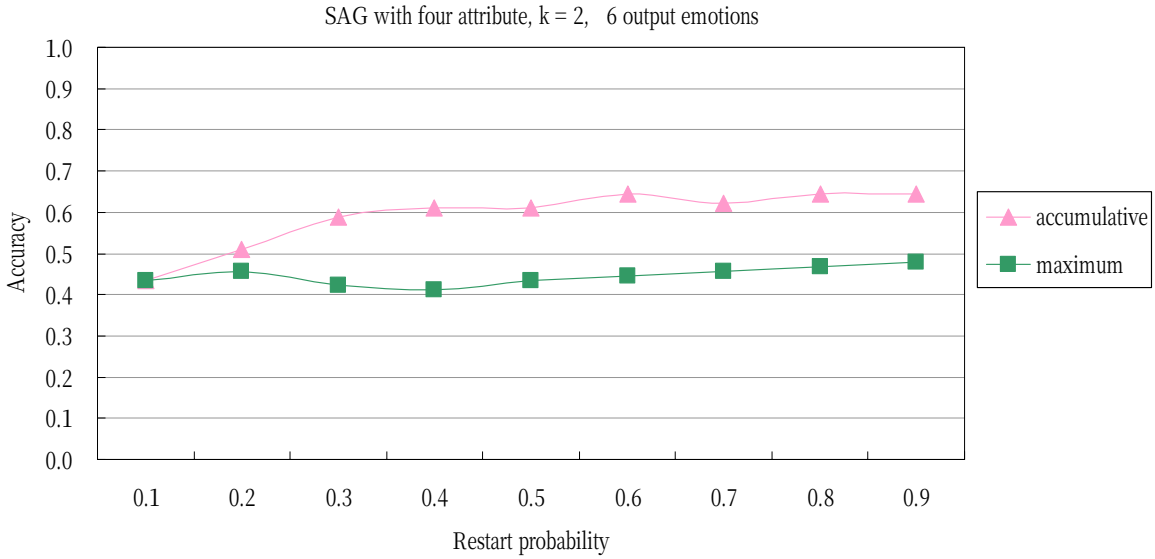

Figure 5.5 Accuracy with different restart probability

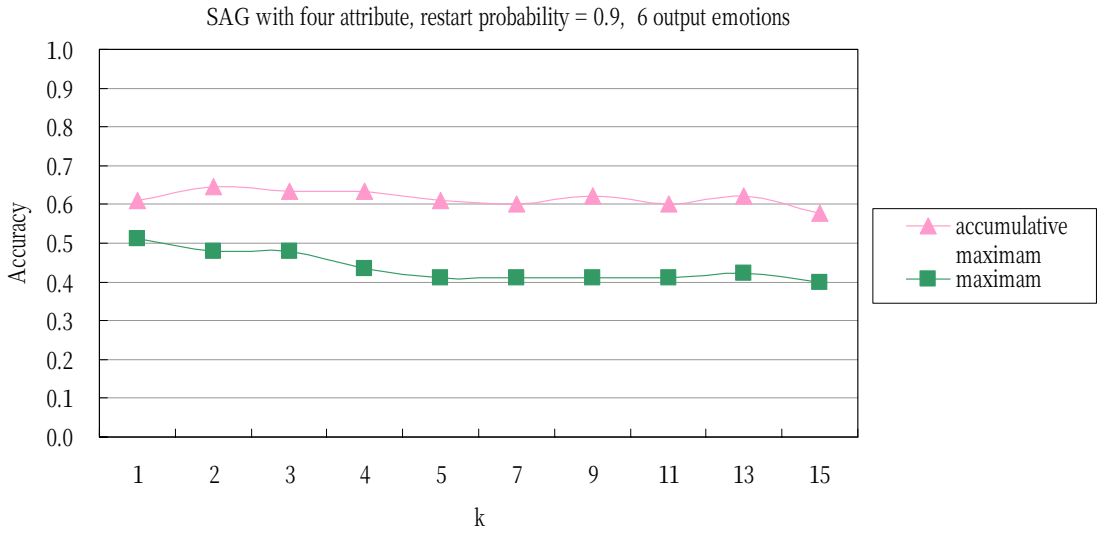

Figure 5.6 Accuracy with different *k* value.

As Figure 5.6 shows, the accuracy declines as the value of *k* increase based on the *emotion* node with the maximal probability. The maximal accuracy is about 64% when *k* = 2.

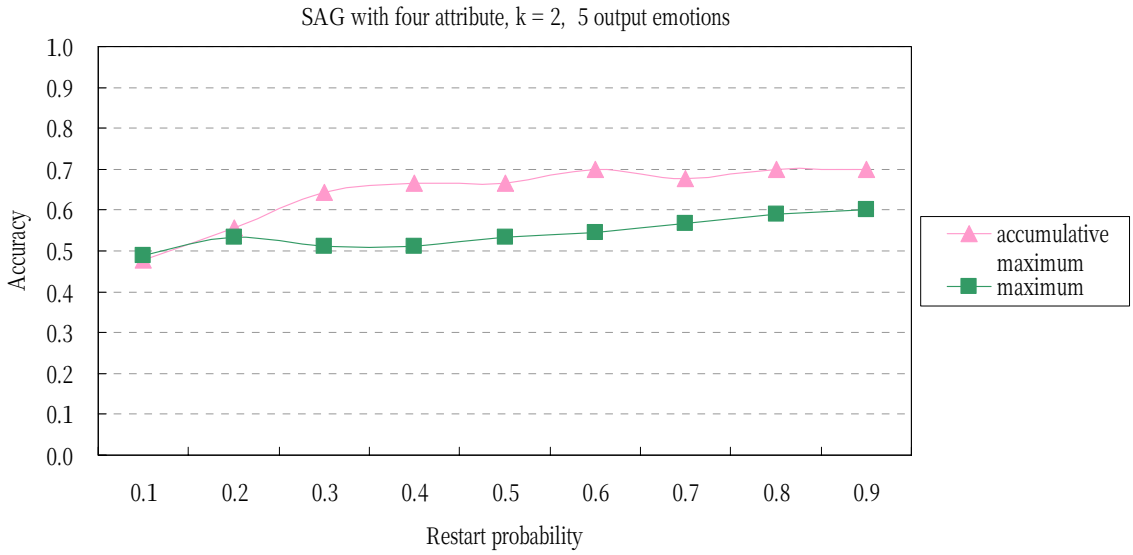

Figure 5.7 Accuracy with different restart probability.

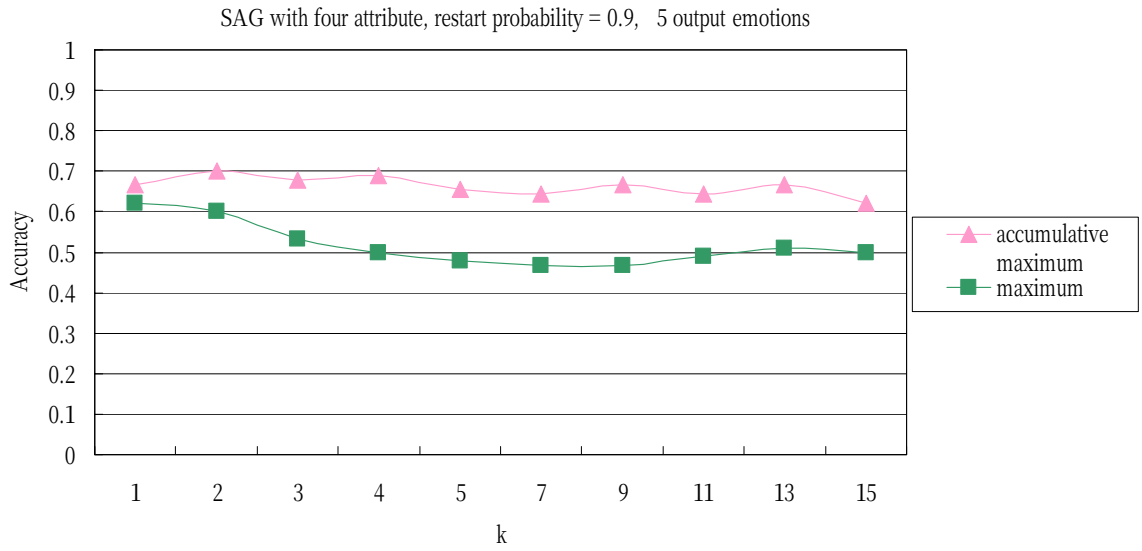

Figure 5.8 Accuracy with different restart probability.

As mentioned in Section 5.2, we combine *Fear* and *Surprise* emotion into one emotion. Figure 5.7 and 5.8 show the corresponding modified results. We can see that the maximal accuracy is about 70% when $k = 2$, 5% better than before.

The experiments above show that the accuracy is closer with the restart probability than the value of k in both topologies. The output emotion of the testing scene is determined based on the emotion category with the maximal cumulative probability always has higher accuracy. And we find out that the *audio* features are the most influential.

6. Conclusions

In this report, we investigate automatic emotion annotation of scenes by the discovery of the relationship between the audiovisual features and the emotions of scene based on film grammar. We exploit fifteen features from six classes - color, light, tempo, close-up, audio, and text. Two topologies of graph were investigated based on Mixed Media Graph approach to find the association between emotions and features.

The experiments show that audio features are the most crucial for affective classification. The experiments show better accuracy about 65% when we put emphasis on audio and emotion attributes. We found that the scenes that evoke Fear or Surprise emotion are not easy to be distinguished. Thus, these two emotions were treated as one emotion and the accuracy is up to 70%. Hence, it is an efficient way to utilize the Mixed Media Graph approach for affective classification.

The experiments above show that the accuracy is closer with the restart probability than the value of k (k-nearest neighbors) in both topologies. The output emotion of the testing scene is determined based on the emotion category with the maximal cumulative probability always has higher accuracy.

Future work includes other audiovisual feature based on film grammars and the exploration of graph-based mining algorithm for emotion discovery.

REFERENCE

[1] 范世鎮與劉志俊，MPEG-4 電影資料之內涵式摘要擷取與角色分析，數位生活與網際網路科技研討會，2004，台南。

[2] B. Adams, C. Dorai, and S.Venkatesh, “Toward Automatic Extraction of Expressive Elements from Motion Pictures: Tempo,” IEEE Transactions on Multimedia, Vol. 4, No. 4, pp. 472–481, Dec. 2002.

[3] D. Arijon, *Grammar of the Film Language*, Los Angeles, CA: Silman-James Press, 1976.

[4] Christopher J. C. Burges, ”A Tutorial on Support Vector Machines for Pattern Recognition,” *Journal of Data Mining and Knowledge Discovery*, pp. 121-167, 1998.

[5] C. Colombo, A. Del Bimbo, and P. Pala, “Semantics in Visual Information Retrieval,” *IEEE Transaction on Multimedia*, Vol. 6, No. 3, pp. 38–53, 1999.

[6] A. R. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace, 1999.

[7] R. Dietz and A. Lang, “Affective Agents: Effects of Agent Affect on Arousal, Attention, Liking and Learning,” *Proceedings of Cognitive Technology Conference*, San Francisco, CA, 1999.

[8] N. Dimitrova, J. Martino, H. Elenbaas, and L. Agnihotri, “Color SuperHistograms for Video Representation,” *IEEE International Conference on Image Processing (ICIP ‘99)*, 1999.

[9] P. Ekman, “Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion,” J. Personality Social Psych., Vol. 54, No. 4, pp. 712–717, Oct. 1987.

[10] L. Giannetti, *Understanding Movies*, Prentice Hall, Englewood Cliffs, NJ, 2005.

[11] A. Hanjalic and L. Q. Xu, “Extracting Moods from Pictures and Sounds: Towards truly personalized TV,” *IEEE Signal Processing Magazine*, Vol.23, No.2, pp. 90-100, Mar. 2006.

[12] A. Hanjalic and L. Q. Xu, “Affective Video Content Representation and Modeling,” *IEEE Transaction on Multimedia*, Vol.7, No.1, pp.143-154, February 2005.

[13] A. Hanjalic and L. Q. Xu, “User-oriented Affective Video Content Analysis,” *Proceedings of IEEE CBAIBL*, Kauai, Hawaii, pp. 50-57, December 2001.

[14] H. L. Wang and L. F. Cheong, “Affective Understanding in Film,” *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.16, No.6, pp.689-704, 2006.

[15] H. B. Kang, “Affective Content Retrieval from Video with Relevance Feedback,” *International Conference on Asian Digital Libraries*, Kuala Lumpur, Malaysia, pp. 243-252, December 2003.

[16] H. B. Kang, “Affective Content Detection using HMMs,” *Proceedings of 11th ACM International Conference on Multimedia*, Berkeley, California, U.S.A, November 2003, pp. 259-262.

[17] G. Kirouac, Les émotions: Monographies de psychologie.. Sillery: Presses de l’Université du Québec, 1992.

[18] F. F. Kuo, M. F. Chiang, M. K. Shan, and S. Y. Lee, “Emotion-based Music Recommendation by Association Discovery from Film Music,” *Proceedings of ACM International Conference on Multimedia*, pp. 507-510, Singapore, November 2005.

[19] P. J. Lang, “The Emotion Probe: Studies of Motivation and Attention,” *American Psychologist*, Vol.50, No.5, pp.372-385, 1995.

[20] Y. Li, S. H. Lee, C. H. Yeh, and C. C. J. Kuo, ”Techniques for Movie Content Analysis and Skimming,” *IEEE Signal Processing Magazine*, Vol. 23, No. 2, pp. 79-89, March 2006

[21] L. Lu, H. Jiang and H. J. Zhang, "A Robust Audio Classification and Segmentation Method,"

Proceedings of ACM International Conference on Multimedia, Ottawa, Ontario, Canada, pp. 203-211, September 2001.

[22] L. Lu, H. J. Zhang, H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Transaction on Speech and Audio Processing*, Vol. 10, No. 7, pp. 504-516, 2002.

[23] F. H. Mahnke, Color, Environmental and Human Response, *Van Nostrand Reinhold*, NY, 1996.

[24] S. Moncrieff, C. Dorai, and S. Venkatesh, "Affect Computing in Film through Sound Energy Dynamics," *Proceedings of ACM International Conference on Multimedia*, pp. 525-527, Ottawa, Ontario, Canada, September 2001.

[25] A. Ortony, G. Clore, and A. Collins, *The Cognitive Structure of Emotions*, Oxford University Press, New York, 1988.

[26] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, The Measurement of Meaning, Urbana, IL: Univ. of Illinois Press, 1957.

[27] J. Y. Pan, H. J. Yang, P. Duygulu, and C. Faloutsos, "Automatic Image Captioning," *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, Taipei, Taiwan, pp.1987-1990, June 2004.

[28] J. Y. Pan, H. J. Yang, C. Faloutsos, and P. Duygulu, "GCap: Graph-Based Automatic Image Captioning," *Proceedings of 4th International Workshop on Multimedia Data and Document Engineering*, Washington, DC, USA, July 2004.

[29] J. Y. Pan, H. J. Yang, C. Faloutsos, and P. Duygulu, "Automatic Multimedia Cross-modal Correlation Discovery," *Proceedings of 10th ACM International Conference on Knowledge Discovery on Database SIGKDD*, Seattle, Washington, pp. 653-658, August 2004.

[30] D. S. Park, J. S. Park and J. H. Han, "Image Indexing Using Color Histogram in the CIELUV Color Space," *Proceedings of 5th Japan-Korea Workshop on Computer Vision*, Korea, pp.126-132, 1999.

[31] G. Peeters, "A Large Set of Audio Features for Sound Description (Similarity and Classification)," in the CUIDADO project. Technical report, Ircam, Paris, France, Apr. 2004.

[32] A. Ramalingam and S. Krishnan, "Gaussian Mixture Modeling using Short-Time Fourier Transform Features for Audio Fingerprinting", *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '05)*, Amsterdam, Netherlands, pp. 1146-1149, July 2005.

[33] Z. Rasheed, Y. Sheikh, and M. Shah, "On The Use of Computable Features for Film

Classification," *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 15, No. 1, pp. 52–64, Jan. 2005.

[34] J. A. Russell and A. Mehrabian, "Evidence for a Three-Factor Theory of Emotions," J. Res. Personality, Vol. 11, pp. 273–294, 1977.

[35] A. Salway and M. Graham, "Extracting Information about Emotions in Films," *Proceedings of 11th ACM International Conference on Multimedia*, Berkeley, California, pp. 299-302, November 2003.

[36] J. Saunders, "Real-Time Discrimination of Broadcast Speech/Music," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, Vol. 2, Atlanta, Ga, pp. 993-996, May 1996.

[37] E. Scheirer and M. Slaney, "Construction and Evaluation of A Robust Multifeature Speech/Music Discriminator", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, Munich, 1997.

[38] R. E. Thayer, *The Biopsychology of Mood and Arousal*, New York, Oxford University Press, 1989.

[39] C. Y. Wei, N. Dimitrova, and S.F. Chang, "Color-Mood Analysis of Films on Syntactic and Psychological Models," *Proceedings of IEEE International Conference on Multimedia and Expo. (ICME '04)*, Taipei, Taiwan, pp. 831-834, June 2004.

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

[41] http://www.intel.com/technology/computing/opencv/index.htm

[42] http://eqi.org/fw.htm

計畫成果自評

■ 就研究內容與原計畫相符程度、達成預期目標情況

本研究計畫主要在研究音樂與情緒的關聯探勘。我們分別利用電影中的配樂與字幕、音訊及視訊資訊的搭配關係，研究音樂與情緒的關聯探勘。在原本的計畫中，我們在第一年研究以字幕的文字作情緒偵測、symbolic music 的各種音樂特徵的萃取、情緒與配樂特徵之關聯模型及關連探勘演算法。第二年則研究延伸至 Wave-form Music 的音樂特徵分析。我們擴展 Emotion Detection 的技術，考慮電影中的視覺資訊與 wave-form music 音樂特徵分析。我們利用 Multi-modal Emotion Detection 技術，來提升原本由電影字幕偵測情緒的準確率，進而提升音樂情緒探勘的準確率。

■ 研究成果之學術或應用價值

本計畫的研究成果，在學術價值方面，我們提出利用電影配樂做為訓練資料，用以自動偵測音樂情緒的方法。我們也提出利用電影拍攝手法中的視覺特徵來判斷電影所傳達的情緒，藉以協助音樂情緒的判斷。

在應用價值方面，計畫中音樂情緒探勘的技術及應用，將有助於台灣發展與音樂美感有關的科技。在應用系統方面，如前所述，本計畫所開發的技術將可應用在音樂推薦、音樂檢索、電子寵物上。尤其隨著數位家庭科技的發展，本計畫的應用將可普遍地應用在未來數位家庭中。

■ 是否適合在學術期刊發表或申請專利

本計畫的研究成果除了三篇碩士論文之外，部分研究成果已經發表在部分研究成果已經發表在 2007 年於德國舉行的 ACM International Conference on Multimedia、2008 年於台灣舉行的 Pacific-Rim Conference on Multimedia、2006 年的 TAAI 11th Conference on Artificial Intelligence and Applications。我們也正在整理研究結果，準備投稿到國際學術期刊。

■ 主要發現或其他有關價值

本計畫的研究蒐集了十二部電影的 100 個場景(scene)，切割成 3000 多個鏡頭(shot)。針對每個鏡頭，我們擷取其視覺、聽覺、字幕特徵，利用 Mixed Media Graph 學習分析其情緒。實驗結果顯示準確率可達到 65%，其中音訊資訊對於電影情緒的偵測有顯著的影響。六種情緒中，surprise 與 fear 最不容易區別。如果不區分這兩種情緒的話，準確率將升高到 70%。

附件二

可供推廣之研發成果資料表

□ 可申請專利　☑ 可技術移轉　　日期：97 年 7 月 31 日

國科會補助計畫	計畫名稱：由電影配樂中探勘音樂情緒之研究(一) 計畫主持人：沈錳坤 計畫編號：NSC 95-2221-E-004-009-MY2　學門領域：資訊工程
技術/創作名稱	基於電影拍攝手法之電影場景自動標記情緒系統
發明人/創作人	沈錳坤、廖家慧
技術說明	電影場景自動標記情緒系統主要分為三個模組，鏡頭切割模組、特徵值擷取模組與情緒標記模組。 欲查詢的電影場景透過鏡頭切割模組切割出此場景之特徵值擷取模組產生四類特徵值，這四類特徵值乃根據電影的拍攝手法所決定。擷取出來的特徵值將輸入到情緒標記模組來產生此電影場景的情緒。情緒標記模組我們採用 Mixed Media Graph(MMG)演算法。
可利用之產業及可開發之產品	1. 可利用於文化、娛樂與音樂教育產業 2. 電影搜尋與音樂配樂
技術特點	運用資料探勘技術，對電影視覺內容分析。
推廣及運用的價值	電影情緒的判斷可協助電影配樂情緒的判斷，進而搭配音樂。

出席國際學術會議心得報告

計畫編號	NSC 95－2221－E－004－009-MY2
計畫名稱	由電影配樂中探勘音樂情緒之研究
出國人員姓名 服務機關及職稱	沈錳坤, 政治大學資訊科學系
會議時間地點	日本大阪 (Osaka, Japan), 97 年 5 月 19 日至 5 月 23 日
會議名稱	Pacific-Asia Knowledge Conference on Knowledge Discovery and Data Mining

一、參加會議經過

Pacific-Asia Conference on Knowledge and Discovery and Data Mining (PAKDD)是自 1997 年起由亞太地區資料探勘領域學者所發起舉辦的國際學術會議。PAKDD 在 Data Mining 研究領域中為極具代表性的國際學術會議。

本屆會議共收到 312 篇來自於 34 個國家的論文投稿，所發表的論文包括了 37 篇 long paper，40 篇 regular paper 及 36 篇 short paper。其中 long paper 的 accept rate 為 11.9%，regular paper 的 accept rate 為 12.8%，short paper 的 accept rate 為 11.5%。會議的第一天為四場 workshops。正式的議程從第二天開始。台灣的學者包括台大的陳銘憲教授、成大的曾新穆教授、師大的柯佳伶教授、交大的彭文志教授、雲科大蘇純繒教授、屏科大蔡正發教授等都有論文發表。

正式議程在開幕式後，由 CMU 的 Christos Faloutsos 的 Keynote speech 開始。Prof. Faloutsos 的博士論文研究是 signature files，我的碩士及博士論文部分機制即建立在其博士研究成果上。後來在 1994 年於台灣舉行的 ICDE 也有機會向其請教討論。近年來，Faloutsos 致力於 graph mining, link mining, social network mining 的研究。其所提出的 power law 在 graph mining 領域具有很大的影響力。此次 Faloutsos 的 keynote speech 主題是 Graph Mining: Laws, Generators and Tools，此主題乃延續其發表在 ACM Computing Survey 的論文。其中，Faloustos 提到 social network 中的 key player problem，這也正是目前我正在研究的問題。

當天及接下兩天的 session 主題有 Privacy Preserving Data Mining, Web Mining, Clustering, Network Mining, Feature Extraction and Construction, Frequent Itemset Mining, Sequence Mining, Outlier Detection, SVM and Regression, Rule Discovery, Spatial and Image Mining,

Semi-supervised Mining, Text Mining, Stream Mining, Classification, Applications 等。

二、與會心得

PAKDD 在 Data Mining 領域是權威的會議之一，由每年發表的論文中可以看到 Data Mining 領域的研究發展趨勢。相較於今年 Data Mining 其他重要學術會議，例如 ACM KDD, IEEE ICDM, SIAM SDM，PAKDD 在 social network mining, link mining 的論文數量較少些。但今年 PAKDD 不少與 Data Mining 應用相關的論文。這或許也代表 Data Mining 領域的基本核心技術如 frequent itemset mining、clustering、classification 已趨近成熟。

此外，今年的 PAKDD 台灣領域的學者與前幾年相較，有下降的趨勢。以最佳論文獎為例，兩篇來自於日本，兩篇來自於中國。或許是國內學者這幾年來比較看中學術期刊論文的發表。

Prof. Faloutsos 有關 graph mining 的 keynote speech 對於 graph mining 的相關領域及其特性，都有完整的整理與介紹。尤其，對於 social network mining 與一般 graph mining 的關係，有精闢的分析。他也點出幾個重要研究方向。對於我們正研究的 social network mining 在方向上有很大的幫助。