Affective Understanding in Film

Hee Lin Wang and Loong-Fah Cheong

***Abstract*—Affective understanding of film plays an important role in sophisticated movie analysis, ranking and indexing. However, due to the seemingly inscrutable nature of emotions and the broad affective gap from low-level features, this problem is seldom addressed. In this paper, we develop a systematic approach grounded upon psychology and cinematography to address several important issues in affective understanding. An appropriate set of affective categories are identified and steps for their classification developed. A number of effective audiovisual cues are formulated to help bridge the affective gap. In particular, a holistic method of extracting affective information from the multifaceted audio stream has been introduced. Besides classifying every scene in Hollywood domain movies probabilistically into the affective categories, some exciting applications are demonstrated. The experimental results validate the proposed approach and the efficacy of the audiovisual cues.**

***Index Terms*—Affective classification, audiovisual features, emotion, film grammar, movie scene, psychology.**

I. Introduction

WITH the increasingly vast repository of online movies and its attendant demand, there exists a compelling case to empower viewers with the ability to automatically analyze, index and organize these repositories, preferably according to highly personalized requirements and criteria. An eminently suitable criterion for such indexing and organization would be the affective or emotional aspect of movies, given its relevance and everyday familiarity. Endowing an automated system with such an affective understanding capability can lead to exciting applications that enhance existing classification systems such as movie genre. For instance, finer categories such as comedic and violent action movies can be distinguished, which would otherwise have been grouped together in the action category under the present genre classification.

With the ability to estimate the intensity of different emotions in a movie, a host of intriguing possibilities emerges, such as being able to rank just how "sad" or "frightening" a movie scene is. Taken to its logical end, this can lead to personalized affective machine reviewer applications, doing away with the limitations of predefined movie genres. In short, computable affective understanding promises a new emotion-based approach toward currently investigated topics such as automated content summarization, recommendation and highlighting.

Surprisingly, immediately related works in affective classification of general domain multimedia have been few. While many works exist in the wider area of multimedia understanding, ranging from scene segmentation [2], sport structure analysis [3], event detection [4], semantic indexing in documentaries [5], sports highlight extraction [6], audio emotion indexing [7] to program type classification [34], literature in affective classification is sparse and recent. This state of affairs is mainly due to the seemingly inscrutable nature of emotions and the difficulty of bridging the affective gap [8], especially in this case where high-level emotional labels are to be computed from low-level cues.

Of works that deal with affectively-related issues, [9] computed the motion, shot cut density and pitch characteristics along the temporal dimension of movie clips from which emotion profiles known as "affect curves" are obtained in a two-dimensional (2-D) emotional space known as the Valence-Arousal space. Ref. [10] used visual characteristics and camera motion with hidden Markov models (HMMs) separately at both the shot and scene level in an attempt to classify scenes depicting fear, happiness or sadness, while [11] proposed a mean-shift based clustering framework to classify film previews into genres such as action, comedy, horror or drama, according to a set of visual cues grounded in cinematography. Ref. [12] proposed finite state machines (FSMs) with face detection and an audiovisual based activity index to model and distinguish between conversation, suspense and action scenes.

While these works have advanced research in affective classification, their output emotion categories in the affective context are somewhat ad hoc and incomplete [10]–[12] ([9] does not use output emotions). Furthermore, the inputs treated by these works are previews [11] or handpicked scenes [10], which due to the prior manual filtering process, are biased by the aims and methods of the selectors. It remains to show whether these works can be readily extended to treat more emotions as well as to analyze complete movies. Crucially, the following important questions are left unaddressed: How should output emotion categories be chosen? And what should they actually be?

Thus in establishing a successful movie affective understanding system, we put forth, as our first contribution, a complementary approach grounded in the related fields of cinematography and psychology. This approach identifies a set of suitable output emotion categories which are chosen with clear reason, a more complicated task than it seems. The increase in the number and subtlety of these categories results in a more difficult, but also more comprehensive and meaningful classification. In contrast, besides having less complete output emotion categories, previous works are explicitly based on just one of the two fields. In the film affective context, they are thus constricted by the limited information and paradigms at their disposal. Ref. [9] employed only psychology, and [11] cinematography, while [10] mentioned psychology briefly but proceeded solely based on the cinematographic basis.

Manuscript received October 25, 2004; revised October 6, 2005. This paper was recommended by Associate Editor R. Lienhart.

The authors are with the Institute of Infocomm Research, Singapore 119613, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: hlwang@i2r.a-star.edu.sg; eleclf@nus.edu.sg).

Digital Object Identifier 10.1109/TCSVT.2006.873781

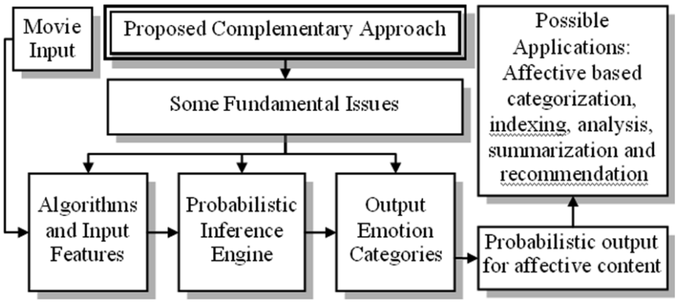

Movie Input
Proposed Complementary Approach
Some Fundamental Issues
Possible Applications: Affective based categorization, indexing, analysis, summarization and recommendation
Algorithms and Input Features
Probabilistic Inference Engine
Output Emotion Categories
Probabilistic output for affective content

Fig. 1. Illustration of scope covered in current work.

For our second contribution, we develop from cinematographic and psychological considerations a set of effective audio-visual cues in the film affective context. Though low-level, some of these features can yield high-level information which helps to bridge the affective gap. For instance, we formulate a visual excitement feature that takes viewer feedback directly into account. Other useful features, which have not been employed in this context, are color energy, chroma difference, music mode and the proportions of music, speech and environ (MSE) audio.

In particular, we propose a probabilistic based approach to extract movie audio affective information from each of the MSE channels in a more suitable and comprehensive manner than other film affective works. Movie audio affective extraction is fraught with many unique difficulties: the challenging nature of speech and music emotion recognition, confusing multiple-speaker presence and MSE mixing. Our approach overcomes these difficulties by splitting the audio analysis units according to cinematographic knowledge and processing each MSE channel differently, before integrating the extracted information by signal duration in a probabilistic manner.

As a broad comparison of our work with others, several main advantages emerge. Foremost among these, we have a set of output emotion categories that are theoretically better founded. Second, we have exploited affective information for audio far more extensively than others, who concentrated on visual cues [9]–[12]. Third, the outputs are expressed probabilistically instead of discretely [11], thus increasing output accuracy. Finally, we have not pre-selected our experimental data; and its size, at about two thousand scenes, is also larger than the next largest video corpus used [10] by about an order of magnitude.

Due to the dominance of the "classical Hollywood cinema" in film [1, p. 89], the scope of this work deals with automatically analyzing and classifying the affective content of Hollywood movie scenes, and in turn the entire movies. The scene, also known as the story or thematic unit, is chosen as the basic unit of analysis, because it conveys semantically coherent content, and is the primary unit of distinct phases of plot progression in film [1, p. 92]. The notion of mise-en-scene, where the design of props and settings revolve around the scene, further enhances its potency [1, p. 176]. Not surprisingly, it is usually the individual scenes that are most sharply etched in the collective memories of the cinema.

The rest of the paper, as illustrated in Fig. 1, is as follows. We introduce the background and explore the fundamental issues of the work in Section II. Section III lays down the proposed complementary approach and discusses how the output emotional categories as well as input features are actually obtained. The design and extraction of the low-level audiovisual cues used for affective classification are elaborated upon in Sections IV and V, while Section VI describes the probabilistic inference engine used in this work. Experimental results are presented in Section VII, followed by the conclusion in Section VIII.

II. Background and Fundamental Issues

Movie affective classification draws upon methodologies from two fields: cinematography and psychology. This section starts off by briefly introducing the necessary foundation of these two fields and the motivation for using them. We also explore various fundamental issues implicit in our approach.

A. *Cinematographic Perspective*

A film is made up of various elements such as editing, sound, mise-en-scene, and narrative. Governing the relationships amongst these elements is a set of informal rules known as film grammar, defined in [14] as "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." The value of film grammar to the present problem lies in the fact that it defines a set of conventions through which the meanings—many of which are affective—of cinematic techniques employed by a director can be inferred.

A quintessential example is that the excitement level 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 offering a computable approach in bridging the difficult transition to high-level semantics such as emotions. Many cues in Sections IV and V are founded on the basis of film grammar.

B. *Psychology Perspective*

Film evokes a wide range of emotions. Hence, a fundamental challenge of movie affective classification lies in the choice of appropriate output emotion representation in film. How do we represent emotions in movies, or relate them to existing emotion studies? These questions mirror some of the most important topics investigated in psychology, which provides emotion paradigms helpful for us in proposing reasonable answers to the questions.

A survey of contemporary theory and research on emotion psychology reveals the most dominant and relevant general theoretical perspectives, respectively known as the Darwinian [38] and cognitive perspectives [39]. The Darwinian perspective postulates that basic emotions are evolved phenomena that confer important survival functions to humans as a species, strongly implying the biological origins and universality of certain human emotions. An impressive body of evidence in human facial expression study by Ekman [16] has identified perhaps the most supported set of proposed basic emotions:

Happy, Surprise, Anger, Sad, Fear and Disgust. This set of emotions, which we call "Ekman's List", are found to be universal among humans, and significantly governs our choice of output emotions and its representation.

On the other hand, the cognitive perspective postulates that *appraisal*, a thought process that evaluates the desirability of circumstances, ultimately gives rise to emotion. Using a dimensional approach to describe emotions under such a paradigm, several sets of primitive appraisal components thought to be suitable as the axes of the emotional space have been proposed [15], so that all emotions can be represented as points in that space. Such a representation is suited for laying out the emotions graphically for deeper analysis. The most popular appraisal axes VAD, proposed by Osgood *et al.* [17] and also Mehrabian and Russell [18], are shown to capture the largest emotion variances, and comprise of Valence (pleasure), Arousal (agitation) and Dominance (control). For this work, we have found a simplified form, the VA space, helpful in visualizing the location, extent and relationships between emotion categories. Dominance is dropped because it is the least understood [33], and its emotional variance accounts for only half that of Valence and Arousal.

Outside psychology, [32] utilized a different set of emotions for machine emotional intelligence. However, that set was chosen for human-computer interaction purposes, and is not suitable for describing affective content in movies.

C. Some Fundamental Issues

We first address a few fundamental issues, beginning from the emotion ground truth labeling stage: should the film affective content be evaluated according to the emotion response of the viewer or what the director intents the viewer to feel? The answer partly hinges on the nature of the currently conceived affective applications. Since they are certainly viewer centric, it is more meaningful to use viewers to calibrate the affective content. This is also consistent with the requirements of future possibilities involving personalized affective applications, which will need viewer emotion response. Not to mention that polling the directors for their intentions rather than the viewers for their emotion responses for numerous movie scenes is far more difficult.

But this raises the question on how the inherent subjectivity of viewer emotion response should be dealt with. Some elements of uncertainty and subjectivity, depending on the unique emotion "makeup" of each individual, are inevitable in the viewer's movie experience. However, the collective mean, or normative emotion response of a statistically large audience is stable and reproducible, especially when dealing with conventional films with a body of accepted "subjective" practices and principles, and thus can be considered objective. Similar assumptions underline the validity of feedback-based psychological studies [18]. For our work, we have thus obtained this normative emotion response to movie scenes in our video corpus from a group of dedicated test subjects.

We emphasize that, though normative emotion response and director intentions broadly concur, they are not equivalent. This is apparent from the difficulties which even highly successful directors have met in conveying their visions. To us, this implies that viewer feedback is an essential element of any viewer-centric film affective system. However, from the standpoint of future works involving personalized affective applications, a potential drawback is the large amount of emotion responses to scenes (of the order of a thousand) required to reliably characterize the unique emotion makeup of an individual viewer, which is too cumbersome for an ordinary user to provide. However, this problem can, we feel, be greatly alleviated by casting the problem of characterizing a viewer as finding the moderately small differences between the individual viewer and normative emotion responses.

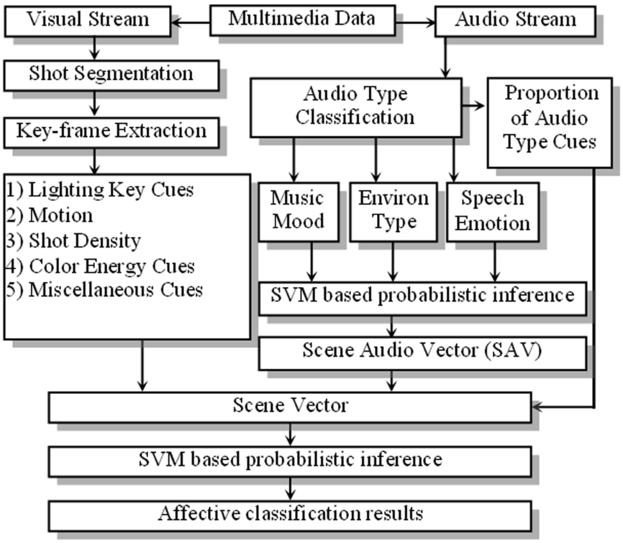


Fig. 2. Flowchart of the system overview.

D. System Overview

We now give a system overview of our affective scene classification system. For consistency, the input to the system comprises of movie scenes manually segmented according to the criteria adopted in [2]. For each scene, the audio and the visual signal are processed separately. The visual signal is segmented into shots and key-frames to facilitate computing visual cues for each scene. The audio signal is then separated according to audio type (music, speech, environ or silence) before being sent into a support vector machine (SVM) based probabilistic inference machine to obtain high-level audio cues at the scene level. The audio and visual cues are finally concatenated to form the scene vectors, which are sent into the same inference machine to obtain probabilistic membership vectors. Fig. 2 illustrates the system overview.

III. Framework Design

As a result of the intended domain of applications and perhaps to simplify matters, all prior related works have relied heavily on just one of three perspectives: Darwinian, cognitive (VA) or cinematographic. In this paper, we demonstrate the advantages of utilizing all three perspectives in affective classification in film domain, and propose a complementary approach that for the first time exploits the information and emotion paradigms methodically from these perspectives to decide on the choice of output emotion categories and low-level input features.

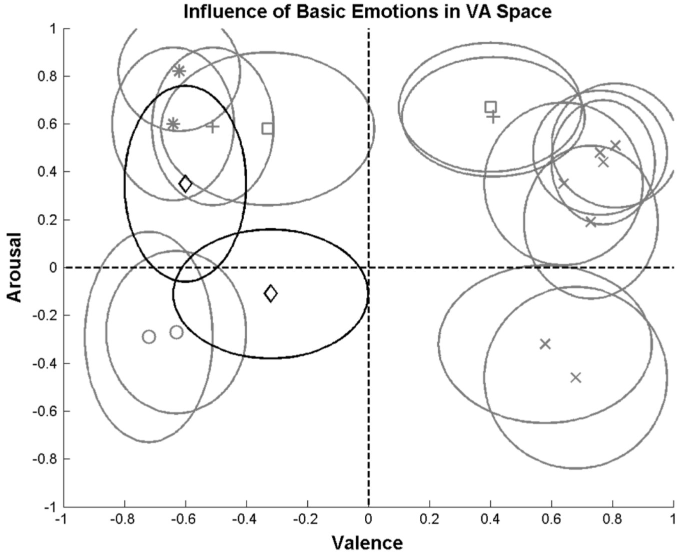


Fig. 3. The VA spaces occupied by basic emotions, Anger (+), Sad (∘), Fear (∗), Happy (x), Surprise (square) and Disgust (diamond), according to manual feedback, centered around mean and bounded by one std. deviation. The exact emotion words used are, in order of increasing arousal in each set: (Angry–anger, aggression), (Happy–relaxed, leisurely, kind, affectionate, enjoyment, joyful, happy), (Sad–sad, depressed), (Fear–fearful, terrified), (Surprise–tense, surprised) and (Disgust–disdainful, disgust).

A. Complementary Approach

1) Characteristics of Each Perspective: The cinematographic perspective provides the advantage of direct insight into film domain production rules, and is eminently suited for formulating new input features. However, its paradigm classifies film according to genre, rather than emotions. Genre is too coarse for emotion categorization, e.g., genres such as drama and romance contain a multiplicity of emotions. Nevertheless there is the possibility of using genre to indirectly gauge the relevance of any proposed emotion categories. The Darwinian perspective provides the theoretical basis on how to categorize emotions meaningfully, but says nothing about other rich information residing in the film domain.

The cognitive (VA) perspective has the advantage of decomposing emotions into its constituent elements. Such representation offers the possibility of visualizing the entire emotion spectrum at a glance in a 2-D feature space, thereby facilitating the analysis of the membership coverage and neighbor relations of different emotion categories. We indeed employ the VA representation when quick visualization is called for. Due to its seeming simplicity, some works have suggested feature-to-VA mapping. But such a proposition is fraught with severe difficulties, especially when applied to the film affective domain. As further explained in the feature selection Section III-D, this is primarily due to the complex distribution of features with respect to emotions.

However, the main reason why we do not adopt the VA as the sole feature space for representing emotions is because some of the output emotions cannot even be sufficiently differentiated therein. We refer the reader to Fig. 3 (details in Section III-B), as well as the dichotomized VAD representations of output emotions (Table II, third column), respectively sourced from the strongest proponents of VAD [18], [19]. By their own accounts, VA space reveals severe to near total overlap between some output emotion categories in the VA space: namely the (Anger, Surprise), (Fear, Anger) and (Disgust, Fear) pairs. These conclusions agree with the assessments of leading emotion theorists who criticized VA for being insufficient to "capture the differences among emotions" [36] and having "little explanatory value, and not much predictive power" [37].

TABLE I
SUMMARY OF COMPLEMENTARY APPROACH

Perspective	Cinematographic	Darwinian	Cognitive (VA)
Area of Strength	Most related to production of film	Organize emotions into families	Represent emotions in VA space well
Tools provided	Film Grammar	Basic Emotions	VA Space
Main Contribution	Input features	Guide to choose initial emotion categories. Input features.	Visualization of emotion membership and neighbor relationship. Input features.

2) Characteristics of Complementary Approach: From the strengths and limitations of being restricted to just one perspective, it is clear that affective understanding in film can benefit from a complementary approach where each perspective offers its tools and paradigms to address facets of the affective problem that it handles well and others are not able to. For our approach to retain the original theoretical bases of these perspectives, we utilize their tools and paradigms in the manner consistent with their purported theoretical strengths and properties, as examined in the preceding sub-section, and summarized in Table I.

Due to a loose underlying consistency amongst the perspectives, features motivated primarily by one perspective may exhibit discernible relationships with others. Thus some may, in attempting to "unify" matters, force features arising naturally from all perspectives (e.g., the underlying physiological basis of speech audio features causes it to map more naturally in the Darwinian perspective) to map to the VA representation before mapping to the output emotions. However, this hierarchical approach introduces information loss, stability and efficiency issues, especially in a complex domain such as affective classification. Instead the complementary approach fuses these features in a heterarchical manner by expressing them directly in a high dimensional space (concatenated into vectors), from which meaningful patterns can be extracted by a powerful inference engine. In this way, complex dependencies amongst features and output emotions can be captured directly and losslessly, thus ensuring greater classification accuracy. The rest of the section will apply this complementary approach to show how the perspectives work together to decide on the choice of output emotions and low-level input features.

B. Output Emotion Categories

A well chosen set of output emotions, besides simplifying complexity, is vital for consistent and principled manual ground truth labeling. In our view, this set should obey the following four criteria:

1) *Universality*: Each emotion can be universally comprehended and experienced.

TABLE II
DESCRIPTOR CORRESPONDENCE BETWEEN DIFFERENT PERSPECTIVES

Output Emotions (Psychological)	Feelings (Viewer)	Dichotomized VA Space	Genre (Cinema)
Anger (Aggression)	Exciting, Dangerous, Aggressive, Angry	-V+A +V+A	Action, Adventure
Sad	Depressed, Sad, Bad, Hopeless	-V-A	Melodrama
Fear	Scary, Fearful, Terrified	-V+A	Horror
Surprise	Surprised, Tense, Anticipation	-V+A +V+A	Suspense, Thriller
Happy	Exuberance, Joyous, Enjoyment, Happy, *Heart-Warming, Tender, Sentimental, Relaxed*	+V-A +V+A	Comedy
Disgust	-	-V+A, -V-A	-
Neutral	Neutral, Boring	(V=0)-A	-

Italicized feelings corresponding to Happy differ from the non-italicized feelings in terms of arousal, a fact that will be used later. The +,- represents the positive and negative half of the Valence(V) and Arousal(A) axes.

2) *Distinctiveness*: Each emotion is clearly distinguishable from the other.

3) *Utility*: Each emotion should have significant relevance in the film context.

4) *Comprehensiveness*: The emotions in the set should be adequate to describe nearly all emotions in film.

The first two criteria pertain to any general emotion categorization, whereas the last two are relevant for film domain application.

As discussed in the preceding section, the cinematographic perspective offers the notion of genre as possible output emotions but the genre is a movie (and not affective) descriptor, and hence is too blunt and inappropriate for describing scene-level affective content. The Darwinian perspective offers Ekman's List which has been proven through substantial experimental backing to be universally identifiable and distinguishable across cultural borders [40]. Ekman's List has the chief advantages of fulfilling the first two criteria, and is thus used as the principal guide in choosing output emotions.

We now want to adapt Ekman's List to satisfy the other two criteria specific to film. To begin, we investigate the relevance of Ekman's List to cinema viewers. We carried out a survey where nine respondents, each randomly assigned two movies, were asked to propose a word for each individual movie scene that would suitably describe their feelings about it, given a list of 151 emotion words found in [18] as a nonexhaustive guide. We note that because the respondents felt there were many emotionally neutral scenes, we add to the basic emotions in Ekman's List a Neutral emotion category (no emotion). The second column of Table II lists the most commonly suggested emotion words and their correspondences with the emotion categories. From the table, we can see how the more specific emotion words are related to the emotion categories. This, with one exception, attests to the utility of the basic emotions, in the sense that they can be readily associated with the more specific emotion words.

Unsurprisingly, the exception "Disgust" cannot find any correspondence with the genres and viewer feelings. This is primarily due to the lack of scenes that seek to evoke "pure" disgust in the viewer. Furthermore, cinematic scenes with an element of Disgust often contain a strong element of Fear and are thus subsumed under it. Due to its lack of utility in the cinema context, Disgust is henceforth dropped, leaving us with a set of what we term the six output emotions (Happy, Surprise, Anger, Sad, Fear, and Neutral).

For the sake of comparison, Table II also lists the correspondence between the output emotions and their approximate locations in VA space (third column), as well as the rough correspondence with film genres (fourth column). This corroboration serves to strengthen the notion that the output emotions are relevant for describing movie affective content. Note that some of the genres such as drama and romance, which reflect a multiplicity of emotions, cannot be matched uniquely to the output emotions and are thus omitted from Table II.

To check for comprehensiveness, we used VA space as a tool to visualize the extent of coverage of the output emotions. It serves as an approximate test for comprehensiveness, in the VA sense, by showing up any large VA areas neglected by the emotions. To visualize the output emotions, we associate them to the closest related emotion words—drawn from viewer feedback in Table II, second column where possible—as found in the 151 emotion words list. This is because we view each group of emotion words associated to an output emotion as "constituting a family of related affective states, which share commonalities in their expression, physiological activity, and in the types of appraisal that call them forth" [35]. These emotion words are finally mapped to the VA space in Fig. 3, using the mean and standard deviation values of valence and arousal of those emotion words computed from manual feedback [18].

The diagram reflects the extent of coverage of the output emotions. Disgust is mapped for completeness, while Neutral, because it represents absence of emotion, does not have visualization data. The emotion areas which are more densely occupied form a rough U-shape; such distribution reflects the psychological reality that areas of high Arousal-neutral Valence and low Arousal-high Valence characterize uncommon affective states and are thus sparsely occupied. Expectedly, the only significant unoccupied region of VA space is centered on the neutral valence, low arousal area, vindicating the inclusion of the Neutral emotion in the output emotions.

C. *Finer Partitioning of Output Emotions*

As a natural extension of our work, we investigate into the possibility of finer meaningful output emotions. This is motivated by the fact that not all the six output emotion categories have equivalent status in the context of this work. The "Happy" emotion enjoys a privileged position in cinema, as mainstream cinema-goers still prefer to have a positive and enjoyable movie experience. This is attested by the fact that amongst the genres that can be strongly identified with an emotion (Table II, fourth column), comedy is by number of movies the most popular genre (about 75 000), followed by the action genre at a distant second (about 18 000) [45].

The Darwinian perspective provides another motivation to subdivide "Happy", which is observed to contain the most diversity in affective states [35], and hence able to yield sufficiently distinctive finer partitions. This observation is explained by the fact that four of the six basic emotions (Anger, Sad, Fear, Disgust) in Ekman's List have immediate survival functions, and have thus evolved unique reaction patterns [36], including facial expressions, upon which Ekman's List is based. In contrast, "Happy" contains multiple distinctive sub-families of positive feelings under it because due to the lack of immediate survival need, none have evolved its own unique facial expression away from the smiling expression [46]. This is borne out by the cognitive (VA) perspective in Fig. 3, which shows "Happy" affective states to be the most numerous. Hence we feel that it would be useful and cinematically relevant to extend our work by further partitioning the feelings used to describe "Happy".

We now attempt to partition "Happy" with the four criteria of output emotion selection in mind. To retain cinematic utility, comprehensiveness and universality, we confine ourselves to defining the partitions with the feelings used by viewers to describe "Happy" emotion in Table II. Referring to the italicized and nonitalicized words used to describe "Happy" feelings in Table II, there seems to be two sufficiently distinctive sub-families of "Happy". The two reasons supporting this partition are as follows.

Affectively, the nonitalicized feelings encompass enjoyment and exuberance, and such scenes tend to be comedic or merry-making. The italicized feelings embody relaxation, kindness and tenderness, and these scenes are likely to be leisurely or heartwarming. Also in VA space, the two groups of feelings tend to be high and low in arousal respectively (Fig. 3). For the rest of the paper, these partitions are labeled as Joyous and Tender Affections (henceforth abbreviated as TA) respectively. As examples, a scene where a parent comforts a child who has lost a toy falls under TA while comedic situations or boisterous friendly reunions fall under Joyous. Classification tests are carried out for both sets of output emotions before and after this partitioning.

D. Feature Selection

As the summary in Table II shows, depending on the way each perspective views the affective aspect of film, different input features are elicited, which are fused together in accordance with the heterarchical organization of features proposed in our complementary approach. Though features exhibiting law-like or rigorous relationships with the emotion representations of various perspectives are desirable from a classification standpoint, they are incongruous with the vast artistic freedom in the film domain. For instance, the dim lighting characteristic of Fear scenes does not imply TA scenes cannot be similarly lit. Here we obtain features by utilizing guidelines stating how each perspective suggests features with significant affective implications. These guidelines are provided below, while the resultant features are detailed in the next two sections.

1) Cinematographic perspective: There often are film grammar rules with affective implications. Examples include the shot duration and lighting key introduced in Sections V-A and V-C, respectively. The corresponding features can then be computed. Sometimes, these rules suggest features which are mapped to VA space instead of directly to the output emotions. An example is the shot duration in Section V-A.

2) Darwinian Perspective: Many results in categorical perception show that the representation of some entity is critically related to its categorical membership [44], which implies that the underlying representations are different resulting from different categorical membership. The Darwinian perspective, by virtue of furnishing much of the basic emotional categories, clearly influences what features are to be used to represent these emotion categories. For instance, the audio features in Section IV, namely the Audio Type Proportion (ATP) and Scene Affective Vector (SAV), are selected (with the aid of cinematographic rules) in such a way that best separates the set of proposed emotion categories.

3) Cognitive (VA) Perspective: The requirement of VA to stringently reduce features to at most two dimensions renders many complex, multimodal but informative features unsuitable for direct feature-to-VA mapping. This is especially so when features are supposed to model the high-level film domain, thus limiting the usefulness of VA for feature selection. But occasionally, the concept of affective dimensions (in our case valence and arousal) recommends features with a strong connection to those dimensions. Good examples of such features are color energy and visual excitement in Section V.

IV. Audio Features

Hitherto under-exploited, audio cues play an important role in this work. Five channels of information in film have been identified by Metz [13], which are: 1) the visual image, 2) print and other graphics 3) music, 4) speech and finally 5) sound or environmental effects. Interestingly, the majority of them (MSE: music, speech and environ) are auditory rather than visual, implying that the auditory stream is a potentially rich source of information. We show how effective low-level audio cues may be derived based on considerations (particularly in relation to the seven output emotions chosen) discussed in Sections II and III.

A. Audio Type Proportion

Audio type classification refers to the classification of a short audio interval into music, speech, environ or silence, while the audio type proportion (ATP) refers to the relative durations of the respective audio types to the scene duration in any particular scene. Although audio types based on different MSE combinations exist, they are confusing and consequently rarely occur. The exception is speech-music, where the voice follows the affect and tune of the music (like in songs). This combination is classified as music, which is adequate for our classification purposes.

While the audio aspect of film grammar is not as formalized as its visual counterpart, discernable ATP signatures exist for different emotions. This is due to both cinematographic and Darwinian considerations, where different audio types are naturally suited to provoke different emotions. We have illustrated some of these patterns in Fig. 4. As a powerful mood inducing medium, music is suitable for most types of emotions (first set of histograms), as seen from the majority of evenly spread music proportion histograms. This is especially true for Fear scenes, but much less so for Neutral and Joyous scenes. Neutral scenes have downplayed music because of their lack of emotional content and emphasis on dialogue. Interestingly, dialogue is also found to be critical in conveying the intended

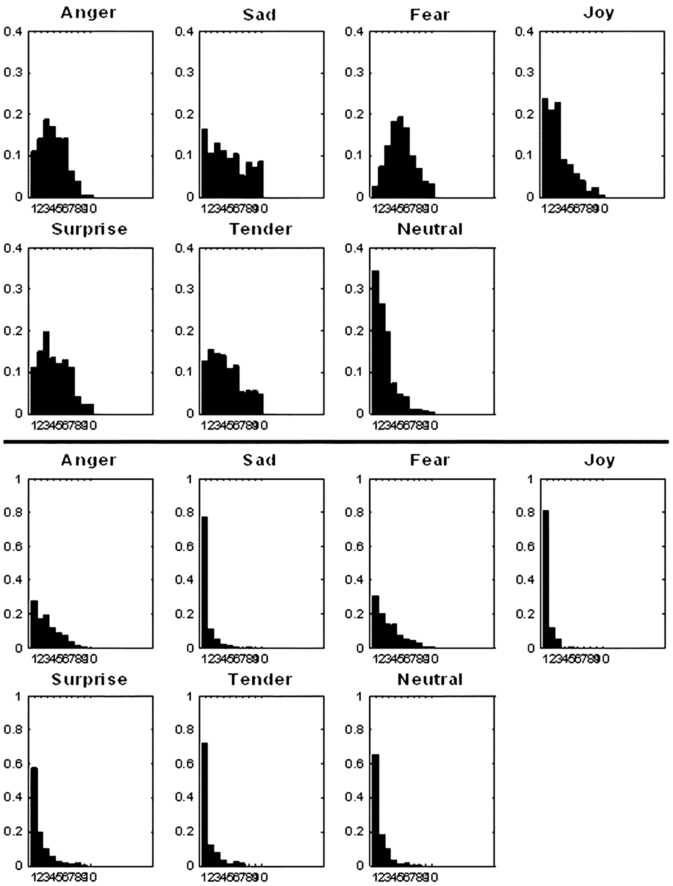


Fig. 4. Music and Environ (top and bottom) ATP histograms, where the x-axis denotes the proportion of the respective audio types.

atmosphere of friendly gatherings or comedic situations, which dominate Joyous scenes.

For the second set of histograms, scenes with a higher proportion of environ usually feature Aggression (noise from frantic activity, gunfire, etc.) or Fear. However, other emotions have no need for environ noise, hence its presence is minimized. Another salient example is Surprise, mostly generated by silence, which creates unbearable anticipation and tension. Due to the dual connections with the output emotions and cinematography, the ATP audio cues are motivated by the cinematographic and Darwinian perspective.

Having explained the ATP histograms from cinema observation, it is useful to automatically sift out the broad patterns characterizing each audio type histogram series by clustering histograms according to their distributions. To begin the clustering process, we first sort the histograms according to the sum of their own cumulative histograms. The Earth Mover Distance (EMD) [2] is then computed for every pair of consecutively ordered histograms to produce an array $\mathbf{A}_{\text{EMD}}$. In an iterative manner, the ranked histograms are split at an unsplit location of $\mathbf{A}_{\text{EMD}}$ with the largest EMD value, l_{EMD}. This is subject to the maximal EMD possible between any two histograms within any current histogram cluster exceeding l_{EMD}, or unity. For our experiments, exactly two clusters are formed for each audio type histogram series. The results of this ATP clustering are shown in Table III.

TABLE III
RELATIVE AUDIO TYPE PROPORTIONS FOR BASIC EMOTIONS

Emotions	Anger	Sad	Fear	Joyous	Surprise	T.A.	Neutral
Music	+	+	+	-	+	+	-
Speech	-	-	-	+	-	-	+
Environ	+	-	+	-	-	-	-
Silence	-	+	-	-	+	+	-

The symbols [+, -] represent the relatively large and small typical proportions of the respective audio types for each emotion.

To obtain the ATP, the entire audio stream is first divided into scenes based on the manual scene segmentation done beforehand. Then starting and ending with the scene boundaries, type classification is carried out for every two seconds segment for every scene. Silence is first identified by thresholding the average segment energy. Two features, which are chroma difference and low short time energy ratio (LSTER) [24], are then extracted from each segment. Chroma difference is a novel feature effective for differentiating music from environ. Let $\boldsymbol{C}$ be a chroma [28] vector, then the chroma difference is

$$\sum_{i=1}^{11} |C_{i+1} - C_i| \tag{1}$$

where C_i is the ith entry of $\boldsymbol{C}$. Whereas in noise, energy is uniformly distributed in the chroma, in music it is highly concentrated in the frequencies played. This ensures that music tends to have higher chroma difference than environ. Segments are then classified into its MSE type by the two features using a simple SVM.

B. Audio Scene Affect Vector (SAV)

As pointed out, each MSE channel contains significant affective information. Due to the predominance of speech, and strong linkage between speech features and output emotions, speech plays the most important role amongst the triad. Music is similarly informative but due to its modest usage is slightly less influential. Finally, though environ sound is the least distinctive of the triad, it can still distinguish between broad sets of emotions.

This suggests the approach of mapping suitable low-level features for each audio type to output emotions separately, and then integrating such information across a scene to obtain the scene affective composition, also known as the audio scene affect vector (SAV). The SAV is a vector denoting the amounts, or probabilities, of the output emotions existing in a scene. Since the dimension and nature of the SAV is solely dependent on the exact output emotions chosen, it is intimately related to the Darwinian perspective.

Good accuracy for speech and music mood detection has been achieved by the current state of the art. Ref. [26] achieved 78.1% accuracy in classifying speech emotions for Ekman's List of emotions. The dataset comes from twelve speakers, reading standard scripts under laboratory conditions. Ref. [25] obtained 86.3% accuracy in classifying music clips into the four classes of contentment, exuberant, anxious and depression, with data consisting of meticulously prepared clips from the classical and romantic period.

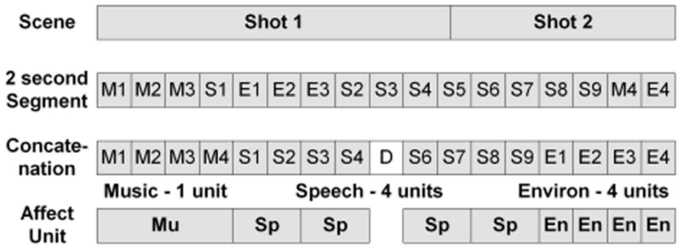


Fig. 5. Illustration of the process of concatenating the segments into affect units to be sent into the probabilistic inference machine. D is for discarded speech segments that straddle shot boundaries.

In contrast, the vast diversity of movies has thrown up for us a very difficult dataset. The speech segments alone contain speech spoken by diverse races, gender, age groups and in multiple English dialects, styles, pitch, speed and volume. Similarly, the music segments feature large diversity of styles from different eras, generated from different instruments. In addition to being subjected to various levels of noise, both music and speech are classified into seven output emotions.

Due to our large database size and the vast variety of audio encountered in each MSE channel, it is impractical to carry out supervised training at the audio segment level. Since there is a need to integrate the information obtained across all channels at the scene level, the proposed algorithm (Fig. 5) is as follows. Conceptually, audio segments belonging to the same type and scene are concatenated together. There is only one constraint: speech segments cannot concatenate across shot boundaries, because they usually denote speaker change, whereas music easily stretches across shots.

These concatenations are then partitioned into equal affect units of 8, 4, and 2 s duration for music, speech and environ sound respectively. These durations are typically the least time required for a confident manual assessment of the affective content for the corresponding audio type. For instance, four seconds is about the least time required to manually discern the emotion of a speech fragment. Affect units are then labeled with the output emotion of the scenes to which they belong. Different sets of features suitable for each audio type [24]–[28], organized into vectors, are then extracted from the affect units.

Here we present the basis behind the features used. More than 50 years of research has yielded speech features affected by psycho-physiological characteristics like air intake, vocal muscle, intonation and pitch characteristics that vary with emotions. Also known as prosodic features, they include speech fundamental frequency, speech power, pitch contour, and frequency sub-band power. We refer the reader to [41], [42] and other excellent references found in [26]. For music mood identification, there is little consensus at present on the exact mechanism whereby music evokes emotions. However, numerous references such as [43] and those found in [25], in accordance with established music knowledge, agree that aspects like music mode, intensity, timbre, and rhythm play important roles in evoking different musical moods.

The more significant features include: 1) energy statistics; 2) LFPC and its delta statistics; 3) low energy ratio; 4) spectral roll-off and centroid; 5) MFCC and its delta statistics; 6) ZCR statistics; 7) spectral flux and normalized version statistics; 8) chroma and its delta statistics; 9) normalized chroma statistics; 10) LSTER; 11) normalized octave energy bands statistics; and 12) music scale. Except for cues 2) and 12), used only for speech and music, respectively, all other cues are extracted from all the audio types. Statistics, where mentioned in conjunction with cues, refers to the mean and variance of the respective cues. These statistics are computed based on equal duration frames (very short audio) within each audio segment. The frame durations used in this work are 100 ms for speech and 200 ms for music and environ.

Here, we introduce the cue that detects music scale, which according to the influential "Doctrine of the Affections" [20], measures the valence of a music piece (minor scale for sad and major scale for happy). In conventional music, it is the music interval relationships between simultaneously played three-notes known as triad chords that determine the music scale. In particular, major and minor scales use four-three and three-four interval patterns respectively between the notes in a triad chord. Hence we have devised an algorithm that uses correlation kernels of the interval patterns on the Fourier transforms of every 200-ms frame of a music piece. The frame is then labeled according to the interval pattern that outputs the highest energy. Finally, the music scale of the music is quantified by the proportion of frames in the music labeled as major.

For each MSE type, the feature vectors are then divided into K groups, with each group being sent into a SVM probabilistic inference machine to obtain the output vectors $V_{\rm au}$ while the remaining $K-1$ groups function as training data. This inference machine, which is mentioned in detail in Section VI, takes an input feature vector and outputs a $N \times 1$ row vector $V_{\rm au}$, where N is the number of affective categories and the entries of $V_{\rm au}$ represent the probabilities of the feature vector belonging to the respective categories. Let the number of affect units, indexed by i, in a scene be Ns. Let their corresponding durations be t_i and output vectors be $V_{{\rm au},i}$, then

$$\text{audio Scene Affect Vector (SAV)} = \text{norm}\left(\sum_{i=1}^{Ns} t_i V_{{\rm au},i}\right). \quad (2)$$

The SAV, which constitutes part of the final scene audio cues, possesses several advantageous qualities. First, it is time weighted to accurately reflect the contribution of every classifying unit. Second, since the output vectors $V_{{\rm au},i}$, contain probabilities, the SAV has a natural probabilistic interpretation; each SAV entry (SAVE) denotes the probability of the scene belonging to the corresponding category like $V_{\rm au}$. Third, due to the integration of information from many affect units, the SAV is far less prone to outlier errors. Finally, using affect units of short durations better models the possibility of affects changing throughout the scene.

Without using SAV, the discrete classification performances for both individual speech and music segments are 45% and 40% respectively. This is due to the vast diversity of audio that are encountered but not sufficiently trained for. However, using the SAV probabilistic framework increases the classification accuracy to 65% and 57%, respectively, at the scene level.

V. Visual Features

We describe several visual cues and show their relationships with respect to the perspectives laid out in Section III. As a preliminary, unless otherwise stated, the visual cues are computed exclusively in the hue, lightness, and saturation (HLS) color space. This is justified purely on the psychological evidence [19] that humans perceive the "emotional" influence of colors with respect to its HLS components. HLS histograms, where applicable, are generated by dividing each axis into 20 equal intervals. For reasons of computational efficiency, all visual features (except motion features) are computed from key frames. The first frame of every shot is declared a key-frame and further key-frames from each shot are then selected according to [2]. Then the feature extraction operator Feat[.] is applied to the *Scene* as follows:

$$\mathrm{Feat[Scene]} = \sum_k (t_k \mathrm{Feat[KF}_k]) / \sum_k t_k \qquad (3)$$

where k is the key-frame index of *Scene*, and t_k is the number of frames that key-frame KF_k represents.

A. Shot Duration

From the cinematographic perspective, the perceived passage of time, also known as the pace, is manipulated to great effect by editing effects like cuts, which defines the shot length. As each shot conveys an event, the director can heighten arousal and intensify a scene by increasing the event density via rapid shot changes [22]. 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 [11], [23].

Obtaining the shot boundaries is an essential first step. However, since the system only needs to detect intra-scene shot boundaries, it does not need to consider challenging editing effects like dissolves or fades. We compute the shot boundary profile, **M**, for each pair-wise frames as the sum of L 2-norm between the L histograms for all matching 20×20 blocks between the pair of frames. Shot boundary is declared at frame i only if the frame fulfils these three criteria:

$$C1(i): \left|\frac{\partial \mathbf{M}}{\partial j}\right|_{j=i} > \mathrm{Th}_{b1},$$
$$C2(i): \left|\frac{\partial^2 \mathbf{M}}{\partial j^2}\right|_{j=i} > \mathrm{Th}_{b2} \qquad (4)$$

$C3(i)$: for $i - 15 < j < i + 15, C1(j)$ and $C2(j)$ should both be true only when $j = i$; Th_{b1} and Th_{b2} are experimentally determined thresholds. $C2$ is used to disambiguate true shot boundary from consecutive high peaks observed in the presence of fast or large moving objects (the latter case would have small $C2$). $C3$ encodes the condition that shots must last for a perceivable length of time. Average recall and precision rates are around 94%. The average shot length of a scene is then calculated as (total scene duration)/(no. of shots in scene).

B. Visual Excitement

Motion plays a central role in the cinema experience owing to the intimate correlation between the degree of mental excitement and the perception of motion on screen. This correlation, broadly proven by a psycho-physiological study [21], seems to result from the natural association of fast motion with danger and excitement, as well as new activity or information. From the cognitive (VA) perspective, computing the arousal arising from motion, which we call *visual excitement*, is useful in differentiating between emotions in different halves of the arousal axis (Fig. 3). Hence we explore a method to accurately determine this visual excitement by the motion present in a video sequence.

Existing approaches [9]–[11] have proposed reasonable features to measure visual excitement. However, in the affective context, those features suffer from a somewhat arbitrary mapping to visual excitement. In contrast, our proposed feature is actually obtained from a nonlinear regression of actual psychophysical results obtained for visual excitement, thus reflecting the critical link between the low-level feature and visual excitement.

Our visual excitement measure is based on the average number of pixels that according to human perception are changed between corresponding frames. This change is computed in the perceptually nearly uniform *CIE Luv* space, since visual excitement is intended to model human perception, and frame difference (L 2-norm) calculations are required across the entire spectrum of possible colors. To smooth over noise, the frame difference is calculated over a 20×20 block as

$$x_{\mathrm{fd}} = \sqrt{s_L(L_1 - L_0)^2 + 1/3((u_1 - u_0)^2 + (v_1 - v_0)^2)},$$
$$s_L = \begin{cases} 1/3, & s_{\mathrm{avL}} \geq 1/3 \\ 1/3 + (s_{\mathrm{avL}} - 1/3)^2, & \text{otherwise} \end{cases} \qquad (5)$$

where (L_1, u_1, v_1) and (L_0, u_0, v_0) are the average *CIE Luv* values of corresponding blocks of consecutive frames and a block is declared as changed if x_{fd} is greater than certain threshold $\mathrm{thres}_{\mathrm{fd}}$. Since frames low in the average frame luminance s_{avL} tend to return lower visual excitement values, a scaling factor s_L is used to increase the sensitivity of x_{fd} to luminance differences for dark frames. Let H be the Heaviside step function, N_H the number of blocks in a frame and let k index the blocks of each frame. Then X_{fd} is defined as

$$X_{\mathrm{fd}} = \sum_{k=1}^{N_H} H(x_{\mathrm{fd}}(k) - \mathrm{thres}_{\mathrm{fd}}) / N_H. \qquad (6)$$

Finally, the visual excitement for each scene is computed as

$$(10/N_c) * \sum_{f=1}^{N_c} [X_{\mathrm{fd}} + (X_{\mathrm{fd}})^W]_f \qquad (7)$$

where N_c is the number of frames in the scene and f indexes the frame. In order to prevent bias toward slow motion clips, we add an offset bias $(X_{\mathrm{fd}})^W$ where W is a constant.

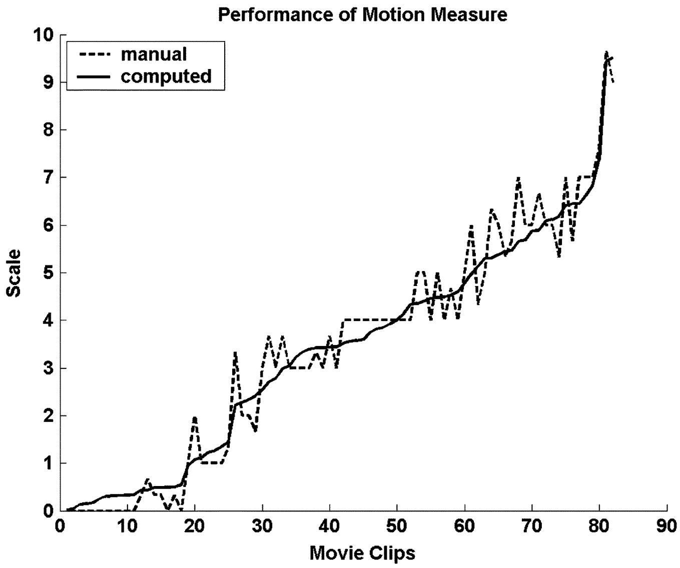


Fig. 6. Graph of the computed visual excitement measure plotted against the manual scale ranking for each movie clip.

To determine the optimal parameters for the visual excitement measure as objectively as possible, a diverse test set comprising of 82 video clips of various types and degrees of motion and lighting conditions are manually selected and segmented from seven movies. These clips feature explosions, large occlusions and special effects averaging around 15 s each and have only one type or degree of motion.

Three test subjects are instructed to give an approximate score to each clip, as far as humanly possible, according to how the motion (not the content) excites them. The clips that each test subject feel to be the most exciting and sedate are assigned 10 and 0, respectively, and used as reference (calibration) clips. Finally all the rest of the clips are scored manually on a linear excitation scale of 0 to 10. Based on the scores, a proposed regression function with suitable parameter value of $W = 0.75$ is obtained, with errors ranging from 0.1% to 13.45% and a mean of 3.91%.

From the results (Fig. 6), it is observed that the measure correlates very closely with the manual scale ranking. As mentioned before, few informative affective cues are suitable for such rigorous regression to even a very limited definition of psychological arousal (visual excitement). However, the results show that the proposed measure is indeed a good indicator of visual excitement.

C. Lighting Key

In the cinematographic perspective, lighting is an extremely powerful tool, used specifically for the purpose of affecting the emotions of the viewer and establishing the mood of a scene. Generally two major aesthetic lighting techniques are frequently employed. Low-key lighting, or *chiaroscuro* lighting, is characterized by a contrast between light and shadow areas, whereas high-key, or *flat* lighting, deemphasizes the light/dark contrast [22], [1]. To generate the light-heartedness and warm atmosphere typical of TA and Joyous scenes, an abundance of bright illumination and a light background, in the form of high-key lighting, is usually employed. In the same vein, film grammar prescribes the use of dim lights, shadow play, and predominantly dark background to recreate the Sadness, Fear, and Surprise for sad, frightening, or suspense scenes [22].

From the above definitions of the two lighting keys, their differences are determined by two factors: 1) the general level of light and 2) the proportion of shadow area. Ref. [11] proposed detecting lighting key using the product of the mean and variance of the brightness of a frame. However, the mean is very sensitive to extreme values and the variance may not be discriminative enough; a high-key lighting frame produced by simply wrapping the brightness values of a low-key lighting frame around the maximum intensity retains the same variance.

We have therefore attempted to formulate two visual features that can accurately quantify the aforementioned components of lighting key in order to better detect it. The median, Med_l, is used as an indicator of the first component, which is the general level of brightness, due to its robustness in the presence of extreme values. The second component, the proportion of shadow area, can be characterized by using the proportion of pixels, Pro_s, whose lightness fall below a certain shadow threshold Th_s. This threshold is experimentally determined to be 0.18, at which an average saturation and highly textured surface no longer appears as textured.

D. Color Energy and Associated Cues

Psychological studies on color have shown that valence is strongly correlated to brightness and to a lesser extent saturation while arousal is strongly correlated to saturation [19]. Thus to capture these affective relationships, we have introduced what we call the *color energy* cue. This cue depends proportionally on the saturation, brightness and area occupied by the colors in a frame [22]. It depends also upon the hue, as in whether it contains more red (energetic) or blue (relaxing) components and the degree of contrast between the colors [22]. From the cognitive (VA) perspective, color energy measures the joint valence-arousal quality of a scene arising from the color composition alone. Thus, the degree of valence or arousal in a scene can be partially inferred by its color energy. For instance, a Joyous effect can be manufactured by setting up a scene with high color energy.

Color Energy is defined as the product of the raw energy and color contrast:

$$\sum_i \sum_j p(c_i) \times p(c_j) \times d(c_i, c_j) \times \sum_k^M E(h_k) s_k v_k \qquad (8)$$

where c is a histogram bin indexed by i, j to iterate over every single bin index in the HLS histogram of an image and $p(\cdot)$ is the histogram probability. $d(c_i, c_j)$ is the L 2-norm in HLS space while M is the total number of pixels, over which index k iterates. s_k, v_k are respectively the saturation and lightness values while $E(h_k)$ is the energy of the hue, assigned a range between [0.75–1.25], depending on its angular distance to blue and red respectively. By the same token, we compute the proportion of pixels with saturation below 20%—an experimentally determined threshold when a color starts to be perceived as gray—to form another cue.

Finally, it is noted that the viewer emotional distance to a scene varies loosely with camera distance, which in turn varies inversely with visual detail. For instance, directors often use close-up shots of characters to facilitate viewer empathy [1]. Thus the visual detail, approximated by the average grey level co-occurrence matrix (GLCM) of a scene and implemented using two-pixel-distant straight and diagonal kernels, is used as a visual cue to characterize the emotional distance of the scene.

VI. Classification and Inference

The features as described by Sections IV and V are extracted and concatenated into row vectors to form the data points characterizing every scene. Due to the highly irregular nature of their probability densities, the classification method needs to be selected with care. In particular, no artificial constraint should be foisted on the data. This excludes parametric based methods or ad hoc rule based methods from consideration. In view of the requirements, we use a specially adapted variant of SVM, which has proven highly successful for classification.

The SVM, which can map vectors from an input space into a possibly infinite dimensional feature space and find the best separating hyperplane therein, has only two adjustable parameters and tends to be less susceptible to the curse of dimensionality [29]. This scheme does not make unwarranted independence assumptions regarding the interaction between audio and visual cues; any such interaction is left to the SVM to learn and exploit. As a kernel based method, it is also able to model extremely complicated class boundaries. Finally, the variant used allows flexibility in the features chosen and more importantly, outputs *a posteriori* probabilities for every category, permitting more refined characterization than binary outputs and allowing the presence of multiple emotions.

To begin with, the data are normalized by shifting the centroid to the origin before dividing it by the mean of the absolute magnitudes. Then K-fold cross validation is used with grid search to obtain the optimal penalty and margin parameters. Subsequently, radial basis kernel SVMs are individually trained for each class pair, so that only features with discriminative value are used. This tailors the features for each class-pair for optimal performance and speed. In line with the ambiguous nature of those training data with dual labels, these data are included in the training set of both labels, and excluded only if the SVM class-pair coincides with the dual labels. *A posteriori* sigmoidals fitted to the decision values of the SVMs are then learnt for each class-pair [30], where the sigmoidals are of the form, with a, b as adjustable parameters,

$$p_i = \frac{1}{1 + \exp(af_i + b)} \tag{9}$$

where p_i is an *a posteriori* and f_i a decision value. These sigmoidals are shown to be very good in modeling the *a posteriori*. A test vector v_t is then processed by each class pair SVM to obtain the decision values, which are in turn fed into the respective sigmoidals to produce class-pair probabilities. These probabilities are finally combined together to compute the *a posteriori* of v_t for every category [31].

TABLE IV
Movies Used for Affective Classification

Action	Horror
The Fifth Element	Ghostship
Speed	Queen of the Damned
Lord of the Rings I	The Haunting
James Bond (Golden Eye)	What Lies Beneath
True Lies	The Others
Men In Black	Dream Catcher
Saving Private Ryan	Ring
Starship Troopers	Gothika
Star Wars I	Legend of the Mummy
Waterworld	
Jumanji	
Drama/Melodrama (D/M)	**Romance/Comedy (R/C)**
Forrest Gump	There's Something About Mary
Magnolia	My Best Friend's Wedding
Ghost	Up Close and Personal
Life is beautiful	Bedazzled
City of Angels	50 First Dates
Artificial Intelligence	Maid in Manhattan
The Sixth Sense	Love Actually
	Bruce Almighty
	Notting Hill

VII. Experimental Results

Our training data consists of 36 full-length and mostly recent mainstream Hollywood movies chosen to represent the more popular films. This translates into 2040 scenes, whose percentage distribution by output emotions are Neutral (24%), Fear (8%), Joyous (13%), Surprise (16%), TA (11%), Anger (17%) and Sad (12%). There is also a diversity of director styles so that the training scenes are likely to be unbiased. Table IV divides these films according to the major genres.

A. Manual Scene Labeling

To obtain the ground truth for experimentation, we attempt to manually match the affective content of a scene to one of the output emotions. If ambiguities arise, we resort to the VA diagram (see text and Fig. 8 in the Appendix). Three persons are employed to independently label each scene. To prevent fatigue and systematic bias, an individual labels only one random movie daily, of a genre different from the previously labeled movie. Except for unanimous decisions that stand, all scenes with dissenting views are reviewed using Fig. 8 as a guide, which usually result in common agreement. Scenes where no agreement can be reached have dual labels; the main label that received two votes, and an alternate label that received one vote. Dual label scenes comprise of 14.08% of all scenes; there are no cases with three differing votes.

B. Discussion

Using a take-one-movie-out approach, we reserve the scenes of one movie for testing while using the rest for training. This approach is repeated for every movie in Table IV, where every testing scene is classified into one of the output emotions. The final aggregated result for all movies is then presented in the form of a confusion matrix given in Table V, and in Table VI for the extended framework. A comparison between Table V and VI reveals that the former can be computed from the latter. The

TABLE V
CONFUSION MATRIX FOR AFFECTIVE CLASSIFICATION (%)

	Anger	Sad	Fear	Joy/TA	Surprise	Neutral
Anger	69.10	4.50	5.34	8.99	7.30	4.78
Sad	2.66	61.13	3.32	15.95	10.96	5.98
Fear	7.03	1.62	84.32	0.00	6.49	0.54
Joy/TA	3.56	9.01	0.21	78.83	2.73	5.66
Surprise	6.16	10.87	10.15	5.073	64.50	3.26
Neutral	4.72	6.74	1.57	10.11	5.62	71.24

TABLE VI
CONFUSION MATRIX FOR EXTENDED FRAMEWORK (%)

	Anger	Sad	Fear	Joyous	Surprise	TA	Neutral
Anger	69.38	3.93	5.62	8.15	7.30	0.84	4.78
Sad	2.66	61.13	3.65	3.99	10.96	11.96	5.65
Fear	7.57	1.08	83.78	0.54	6.49	0.00	0.54
Joyous	6.59	1.55	0.388	80.62	1.94	2.71	6.20
Surprise	6.16	11.23	9.42	1.81	65.22	3.26	2.90
TA	0.00	17.35	0.00	3.20	3.20	71.69	4.57
Neutral	4.50	7.19	1.57	5.84	5.40	4.05	71.46

TABLE VII
CONFUSION MATRIX FOR PAIRWISE AFFECTIVE CLASSIFICATION (%)

	Anger	Sad	Fear	Joyous	Surprise	TA	Neutral
Anger	0	3.76	**6.61**	**7.57**	**6.56**	0.52	4.44
Sad	0	0	3.19	3.19	**11.68**	**15.81**	**6.75**
Fear	0	0	0	0.47	**7.97**	0	1.22
Joyous	0	0	0	0	1.75	2.85	5.61
Surprise	0	0	0	0	0	2.97	4.02
TA	0	0	0	0	0	0	3.91
Neutral	0	0	0	0	0	0	0

clear separation of the Joyous and TA categories confirm that they are distinct categories, and also attests to the discriminating power of the SVM framework used. Henceforth, the rest of the discussion will concentrate on analyzing the more difficult and informative extended framework case. For a clearer analysis of the algorithm performance, the confusion rates between every pair-wise emotion (computable from Table VI) are presented in Table VII.

Of the 21 possible pairs of emotions, Sad-TA, Sad-Surprise, Fear-Surprise, Anger-Joyous, Anger-Surprise, Sad-Neutral, and Anger-Fear are, in descending order, the seven pairs most culpable for errors. The confusion arises due to the frequent co-existence of these emotion pairs, attested by the dual labels that are attached to most of these scenes. For example, tenderness is often portrayed when someone comforts a despondent loved one, thus posing a difficult problem in the identification of the Sad-TA pair. Similarly, because all unpleasant low arousal scenes are classified as Sad, and a major source of unpleasantness is usually due to tension/suspense, therefore, the Sad-Surprise pair also shows high errors.

As for the third pair, Fear-Surprise, these emotions are so intertwined in cinema that it is sometimes hard to even manually differentiate between them. Regarding the fourth pair, Anger-Joyous, the difficulty arises because Joyous comedic scenes are usually slapstick in nature, and contain a fair amount of action elements inside, hence the confusion with Anger. Anger-Surprise and Anger-Fear also tend to co-exist and our experience with the manual labeling process shows that for borderline cases, deciding for one label over another depends to a large extent on one's personal tolerance for fear. A person with low tolerance will be more acutely aware of the fear element, and tend to choose the Fear or Surprise labels. Finally, neutral scenes are dominated by short scenes with dialogue in dull or subdued tones, which are not uncommon in Sad scenes, thus complicating the discrimination of the Sad-Neutral pair.

TABLE VIII
OVERALL CLASSIFICATION RATE (%)

	Correct	Alternate Selected	Incorrect
Visual	42.86	9.87	49.27
Audio	61.39	10.34	28.27
Audio/Visual	74.69	11.13	14.18

TABLE IX
RANKING OF AFFECTIVE CUES

CUES	%	CUES	%
Silence Proportion	38.1	Visual Excitement	20.2
Joyous SAVE	32.1	Speech Proportion	20.2
Fear SAVE	28.7	Median Lighting	19.7
Surprise SAVE	27.5	Shadow Proportion	19.4
TA SAVE	27.4	GLCM	18.3
Anger SAVE	27.4	Average Shot Length	17.6
Sad SAVE	26.2	Color Energy	17.4
Neutral SAVE	26.1	Music Proportion	16.7
Environ Proportion	23.7	Saturation Proportion	16.0

Having discussed the difficult cases, it is nevertheless well to note that the confusion matrix indicates that most scenes are classified correctly. Despite the challenge of classifying every scene of the entire movies, and the significantly larger number and increased subtlety of emotional categories compared to existing works, the overall correct classification rate is 74.69%, or 85.82% if "alternate selected" scenes are included (last row of Table VIII). The second column under 'Alternate Selected' refers to those cases of dual label scenes whose dominant and alternate labels have received the second highest and highest probabilities respectively. Given the intrinsic ambiguity of these scenes and such a stringent criterion imposed, we believe that these so-called "alternate selected" scenes have been adequately classified and fully deserve a separate result category. Empirically speaking, the promising results suggest the classification of the affective categories is well-posed and separable using low-level cues.

The results also shed light on the relative influence of the audio and visual cues on classification. Rows 2 and 3 of Table VIII present the classification results using the audio and visual cues individually and jointly. First, it is evident that combining both audio and visual cues together for classification significantly outperforms either of the cues individually. The results also corroborate our view that audio cues are far more informative than visual cues with respect to affective content, conforming to our initial expectations. To the extent that the visual cues presented in this paper have captured the visual reality, we observe that there is a general lack of strong correlation between the simple low-level visual cues presented here and the affective content. For instance, except at the extreme regions of the HSL color space, color does not correlate well with the affective content [19]. This lack of correlation is

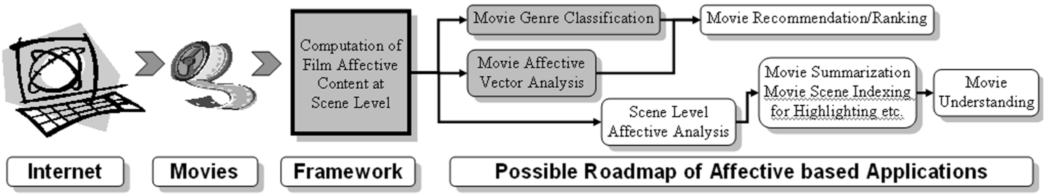

Computation of Film Affective Content at Scene Level
Movie Genre Classification
Movie Affective Vector Analysis
Scene Level Affective Analysis
Movie Recommendation/Ranking
Movie Summarization Movie Scene Indexing for Highlighting etc.
Movie Understanding
Internet
Movies
Framework
Possible Roadmap of Affective based Applications

Fig. 7. Illustration of possible roadmap for applications based on affective understanding in film. Shaded areas denote completed tasks.

compounded by the fact that the director is constrained by the plot and general settings in the amount of freedom to set up visual environments that will evoke the desired moods.

However, these difficulties do not seem to arise as severely for low-level audio cues, where sound in film seems to be more immediately purposeful than visual details. In fact, the correlation between the low-level audio cues and the scene affect is in general so strong that unless there is a good reason for them to contradict (e.g., for comedic effect), the scene itself can easily be misinterpreted or appear jarring. Table IX ranks each cue according to the average rate of correct SVM classification between every pair of categories using only that cue (the take-one-out method). This table corroborates the finding that at the low-level, audio is more informative than visual cues. In particular, if one views absence of sound as an audio cue (a negative kind of audio cue), the top eight cues are all audio cues, with silence proportion as the most effective one. The sheer presence of absolute silence can be most dramatic and unsettling at times.

The performance of our algorithm compares favorably with the 78.7% reported by Kang [10], the only work we are aware of that has performed affective classification on Hollywood scenes. However, with regards to Kang's results, there are several important caveats. The test and training sets contain only selected scenes that are unambiguously manually labeled as one of only three classes: happy, sad or fear. The scenes were also selected from only six movie segments each lasting half an hour, as opposed to all the scenes in a movie.

C. Application

Machine understanding of the affective aspect of Hollywood multimedia can enhance and complement existing classification systems at several levels of resolution. Here we demonstrate applications at two levels: the more generalized movie genre level, and the more refined movie affective vector level (Fig. 7). Other possible applications include using scene-level affective results for story unit extraction.

1) Movie Genre Level: As it is, movie genres are sometimes too blunt to reflect the true character of a movie. For instance, genre labels seldom differentiate between comedic action and *film noir* action movies, or between tender drama and melodrama movies, although these differences substantially impact the movie experience. An obvious application of our work is to offer a more refined classification of any given movie and to detect dual genre movies, thus complementing existing genre classifications. In general, the genre of a movie can be largely determined by the proportion of time occupied by each of the affects. For example, a movie that has a significant amount of Fear and Surprise scenes is likely to belong to the horror genre. Therefore, we let every movie be characterized by a movie affective vector (MAV), or $\boldsymbol{V}_i$

TABLE X
MOVIE GENRE CLASSIFICATION BASED ON SCENES

Movie Title	Manual	1st Label	2nd Label
Jumanji	A	H	A
Lord of the Rings I	A	H	A
Saving Private Ryan	A	D/M	A
Up Close and Personal	R/C	D/M	R/C
Notting Hill	R/C	D/M	R/C
Life is Beautiful	D/M	R/C	D/M
Artificial Intelligence	D/M	A	D/M

The labels refer to the genres: Action(A), Horror(H), Romance/Comedy(R/C) and Drama/Melodrama(D/M).

$$\mathbf{V}_i = \mathrm{Norm}\left(\sum_{s=1}^{N_i} \mathbf{\Lambda}_{s,i}\tau_{s,i}\right) \tag{10}$$

where i, s, and N_i are the movie index, scene index and total number of scenes respectively. $\mathbf{\Lambda}_{s,i}$ is the vector of probabilities of each affective category and $\tau_{s,i}$ is the duration of sth scene in the ith movie. Effectively, $\boldsymbol{V}_i$ captures the affective content of a movie in terms of both the duration and confidence level of each affect.

The notion of affective vector then can be readily extended to the genre, $\boldsymbol{V}_g$, where the summation and normalization is carried out over each genre rather than a movie. The genre of a movie can then be determined by the distance measured between an MAV and that of a genre. We adopted the symmetrical Kullback-Leibler distance measure. Let the individual entries of the query and genre affective vectors be $V_{i,m}$ and $V_{g,m}$, indexed by m which runs over every affective category, and i, g which are the movie and genre indices respectively. The measure $M_a(i,g)$ is then defined as

$$M_a(i,g) = \sum_{m=1}^{N_i}\left(V_{i,m}\log\left(\frac{V_{i,m}}{V_{g,m}}\right) + V_{g,m}\log\left(\frac{V_{g,m}}{V_{i,m}}\right)\right). \tag{11}$$

A movie i is then assigned the genre g that returns the lowest $M_a(i,g)$. Using take-one-out testing, 80.6% of all the movies are assigned the correct genre. The rest of the movies are listed in Table X, where the third and the fourth columns list the labels having the highest and second highest confidence respectively. As can be seen, the second labels of all these movies correspond to the manual labels.

Further inspection reveals that these results indeed reflect the dual nature of the majority of the movies in Table X. For example, *Life is Beautiful* is actually a romance/comedy in the

TABLE XI
MOVIE LEVEL AFFECTIVE VECTOR

Action	**Agr**	**Sad**	**Fear**	**Joy**	**Sur**	**TA**	**Neu**	**Horror**	**Agr**	**Sad**	**Fear**	**Joy**	**Sur**	**TA**	**Neu**
The Fifth Element	31	10	1	9	32	5	12	Ghostship	9	8	34	3	20	5	22
Speed	69	8	8	3	4	3	6	Queen of the Damned	11	16	29	1	33	3	7
Lord of the Rings I	19	11	25	9	28	4	4	The Haunting	4	6	23	0	39	3	24
James Bond (Golden Eye)	45	11	3	3	18	6	15	What Lies Beneath	4	30	14	2	22	10	17
True Lies	41	8	2	7	14	5	23	The Others	9	26	11	2	43	6	3
Men In Black	33	3	1	2	42	1	17	Dream Catcher	19	7	42	6	6	3	17
Saving Private Ryan	24	26	3	5	10	19	13	Ring	2	7	77	1	4	3	6
Starship Troopers	60	7	2	10	4	3	13	Gothika	5	19	55	1	9	5	6
Star Wars I	36	9	4	18	7	3	24	Legend of the Mummy	8	6	19	1	28	2	37
Waterworld	36	8	1	10	24	9	12								
Jumanji	13	26	13	3	29	7	9								
Drama/Melodrama (D/M)								**Romance/Comedy (R/C)**							
Forrest Gump	17	19	2	10	4	21	28	There's Something About Mary	12	3	1	65	2	4	12
Magnolia	21	23	4	8	8	15	21	My Best Friend's Wedding	8	5	0	28	2	52	4
Ghost	23	12	9	3	6	22	24	Up Close and Personal	13	14	2	7	5	38	21
Life is beautiful	21	17	2	37	6	9	7	Bedazzled	12	2	2	68	4	4	8
City of Angels	11	33	3	5	8	26	15	50 First Dates	7	5	2	61	2	17	6
Artificial Intelligence	18	32	8	2	17	9	15	Maid in Manhattan	8	5	0	22	1	47	16
The Sixth Sense	4	41	4	3	18	14	17	Love Actually	3	8	0	32	2	46	9
								Bruce Almighty	12	7	2	44	2	18	16
								Notting Hill	12	16	1	16	5	21	29

The abbreviations for the emotions are respectively: Aggression (Agr), Sad (Sad), Fear (Fear), Surprise (Sur), Tender Affections (TA) and Neutral (Neu).

first half and a melodrama in the second half, while *Saving Private Ryan* is accurately described as a melodrama with a strong action element. On the other hand, *Artificial Intelligence* is correctly classified as an adventure (action) movie with much melodrama. Interestingly, although *Lord of the Rings I* is billed as an action movie, it has far more than its expected share of scary scenes. Classified by the algorithm as being most similar to the horror genre, the movie would warrant a caution for young children viewing it.

2) Movie Affective Vector (MAV) Level: At the next finer level of analysis, the MAV offers a more detailed picture of a movie. Due to the probabilistic inference framework adopted, the relative amounts of each affective component within a movie can now be estimated, as shown in Table XI. This facilitates ranking of the movies according to a very useful and hitherto unimplemented aspect: its affective content. For instance, the affective vector of a movie can rank just how "happy" or "aggressive" etc. a movie is.

We survey Table XI for broad quantifiable trends by genre, noting that the existence of characteristic genre MAV patterns underlies the consistency of the MAV. Aggression and Surprise feature prominently for the action genre while the D/M genre is dominated by TA and Sad elements. The R/C movies are marked by strong Joyous/TA affects while horror movies tend toward Fear/Surprise inducing scenes.

At the very top of the Aggression ranking list are *Speed* and *Starship Troopers*; the former is unrelentingly fast paced while the latter is extremely violent. Unsurprisingly, most other movies from the action genre followed closely, with the exceptions being *Saving Private Ryan*, which has strong melodramatic elements, and *Jumanji*, which, being of the type "family entertainment," abstained from overt violence. Popular horror cinema can generally be differentiated by the main directing technique employed to induce fear: creating overtly threatening situations (Fear) or the more subtle tension (Surprise), which shows up clearly in their MAVs. *The Ring* and *Gothika* are outright frightening, while *The Others* depends far more on Surprise; the rest of horror movies possess a rather even mix of both elements.

As a rule of thumb, the summation of Fear and Surprise is a good indicator of the "scariness" of a movie. According to this indicator, *The Ring* and *What Lies Beneath* would correctly be the scariest and mildest movies, respectively. The same pattern appears for R/C movies, which generally belong to two groups: comedy/slapstick (Joyous) or sentimental (Tender Affection). Similarly, MAV can be used to classify these two groups and to rank them according to the summation of Joy and TA: *Bedazzled* is slapstick, *My Best Friend's Wedding* is sentimental, and *Notting Hill* aptly has the MAV of a subtle drama depicting a tortuous romance.

Besides using the MAV on its own for analysis, it can also complement the manually assigned genre of a movie if available, which provides the context for a more refined interpretation of the MAV. For example, whereas a high Neutral score for D/M movies is understandable, it usually signifies boredom for horror movies, as the obscurity of *Legend of the Mummy* can attest. Sadness in all other genres except D/M should also generally be interpreted as the disquieting or oppressive as opposed to the tear inducing variety. Movies from D/M genre tend to represent the whole gamut of human emotions more evenly, being a reflection of life. Here, Aggression reflects the degree of interpersonal conflict while TA and Sad indicates the amount of melodrama.

From the aforementioned paragraphs, MAV analysis is able to yield broadly accurate ranking results according to different affects and even differentiate between different sub-genres, leading to automatic movie recommendation according to personalized affective preferences. To our knowledge, this is a capability not yet available in existing systems. With further investigation, more interesting and subtle patterns from MAV analysis are likely to emerge.

VIII. CONCLUSION

In this paper, a complementary approach has been proposed to study and develop techniques for understanding the affective content of general Hollywood movies. We laid down a set of relevant and theoretically sound emotional categories and employed a number of low-level features from cinematographic and psychological considerations to estimate these emotions. We discussed some of the important issues attendant to automated affective understanding of film. We demonstrated the viability of the emotion categories and audiovisual features by carrying out experiments on large numbers of movies. In particular, we introduced an effective probabilistic audio inference scheme and showed the importance of audio information. Finally, we demonstrated some interesting applications with the resultant affective capabilities.

Much work remains to be done in this largely unexplored field. First, with regards the shortcomings of our work, the small proportion of scenes that are wrongly classified shows up the inherent limitation of low-level cues (especially visual) in bridging the affective gap. Therefore, in the immediate future, we intend to implement more complex intermediate-level cues to further improve present results. Second, the existence of multiple emotions in scenes requires a more refined treatment. Finally, we will also investigate the possibility of finer sub-partitioning of the present affective categories, as well as further scene affective vector level analysis.

APPENDIX
APPENDIX: VA SPACE FOR GROUND TRUTH ARBITRATION

To arbitrate over the output emotion assigned to emotionally ambiguous scenes that defy manual labeling of ground truth, we attempt to lay out the output emotions in the VA space (Fig. 8) so that the entire emotional spectrum can be visualized in a glance and used as a guide. Note that Fig. 8 is meant to conceptually depict the neighboring relationships between categories and approximate spheres of membership rather than literally demarcating crisp boundaries. In cases of ambiguity, the ground truth labeler can use the VA map as a last resort to arrive at a more objective final label. For example, in several scenes of *The Sixth Sense*, the protagonists conversed in worried tones. Although Worry does not clearly belong to any of the categories, thinking in terms of VA axes suggests that the scene falls within the low A-V- part of the VA space; hence these scenes are categorized under Sad (see also [47] for using VA space to categorize images emotionally).

To function as an arbitrating tool, Fig. 8 should avoid the emotion overlap discussed in Section III-A and illustrated by Fig. 3, while preserving cinematic relevance and comprehensive coverage in the VA space. Hence it is necessarily a modification of Fig. 3, and the boundaries of the output emotions have been suitably modified. Surprise is situated in negative valence regions to reflect the fact that Surprise scenes are mostly tense and suspenseful, as opposed to being pleasantly surprisingly. Aggression has also been shifted closer to neutral valence to acknowledge that such scenes are usually meant to excite, and not to provoke extreme infuriation.

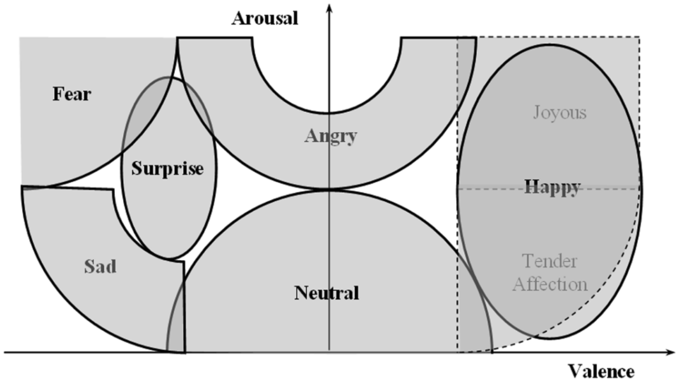


Fig. 8. Conceptual illustration of the approximate areas where the final affective output categories (in bold) occupy in VA space.

REFERENCES

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

[2] B.T. Truong, S. Venkatesh, and C. Dorai, "Automatic scene extraction in motion pictures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 5–15, Jan. 2002.

[3] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues," *IEEE Int.. Conf. Multimedia Expo*, vol. 3, pp. 309–312, 2003.

[4] S. Moncrieff, S. Venkatesh, and C. Dorai, "Horror film genre typing and and scene labeling via audio analysis," in *IEEE Int.. Conf. Multimedia Expo*, 2003, vol. 2, pp. 193–196.

[5] N. Haering, R.J. Qian, and M.I. Sezan, "A semantic event-detection approach and its application to detecting hunts in wildlife video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 6, pp. 857–868, Jun. 2000.

[6] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball program," *ACM Multimedia*, pp. 105–115, 2000.

[7] A. Salway and M. Graham, "Extracting information about emotions in films," *ACM Multimedia*, pp. 299–302, 2003.

[8] A. Mittal and L.F. Cheong, "Framework for synthesizing semantic-level indexes," *Multimedia Tools Appl.*, vol. 20, no. 2, pp. 135–158, 2003.

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

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

[11] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, Jan. 2005.

[12] Y. Zhai, Z. Rasheed, and M. Shah, "A framework for semantic classification of scenes using finite state machines," in *Proc. Conf. Image and Video Retrieval*, 2004, pp. 279–288.

[13] J. Monaco, *How to Read a Film: Movies, Media, Multimedia*, 3rd ed. Oxford, U.K.: Oxford Univ. Press, 2000.

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

[15] R. R. Cornelius, *The Science of Emotion. Research and Tradition in the Psychology of Emotion*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[16] 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.

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

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

[19] P. Valdez and A. Mehrabian, "Effects of color on emotions," *J. Experimental Psych.: General*, vol. 123, no. 4, pp. 394–409, 1994.

[20] J. W. Hill, *Baroque Music: Music in Western Europe*. New York: W. W. Norton, 2005, pp. 1580–1750.

[21] R. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss, "Attention to television: Alpha power and its relationship to image motion and emotional content," *Media Psychol.*, vol. 5, pp. 283–301, 2003.

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

[23] B. Adams, C. Dorai, and S. Venkatesh, "Toward automatic extraction of expressive elements from motion pictures: Tempo," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 472–481, Dec. 2002.

[24] L. Lu, H. Jiang, and H.-J. Zhang, "A robust audio classification and segmentation method," *ACM Multimedia*, pp. 103–122, 2001.

[25] D. Liu, L. Lu, and H.-J. Zhang, "Automatic mood detection from acoustic music data," *Proc. Int. Symp. Music Information Retrieval (ISMIR'03)*, pp. 81–87, 2003.

[26] T. L. New, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, pp. 603–623, 2003.

[27] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. Int.. Conf. Spoken Language Processing*, 1996, pp. 1970–1973.

[28] W. Chai and B. Vercoe, "Structural analysis of musical signals for indexing and thumbnailing," in *Int.. Conf. Digital Libraries*, 2003, pp. 27–34.

[29] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[30] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Microsoft Research* 1999 [Online]. Available: http://research.microsoft.com/~jplatt

[31] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Ann. Statis.*, vol. 26, no. 2, pp. 451–471, 1998.

[32] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[33] H. Schlosberg, "Three dimensions of emotion," *Psychol. Rev.*, vol. 61, pp. 81–88, 1954.

[34] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," *ACM Multimedia*, pp. 295–304, 1995.

[35] P. Ekman, "Facial expression and emotion," *Amer. Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[36] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*. New York: Wiley, 1999, ch. 3.

[37] A. Ortony and T. Turner, "What's basic about basic emotions," *Psychol. Rev.*, vol. 97, no. 3, pp. 315–331, 1990.

[38] C. Darwin, *The Expression of Emotions in Man and Animals*. Chicago, IL: Univ. Chicago Press, 1872/1965.

[39] A. Magda, *Emotion and Personality*. New York: Columbia Univ. Press, 1960.

[40] R. C. Solomon, "Back to basics: On the very idea of basic emotions," *J. Theory of Social Behavior*, vol. 32, no. 2, pp. 315–331, 2002.

[41] I. Fonagy and K. Magdics, "A new method of investigating the perception of prosodic features," *Language and Speech*, vol. 21, pp. 34–49, 1978.

[42] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoustics Soc. Amer.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[43] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research*. Oxford, U.K.: Oxford Univ. Press, 2001.

[44] K. J. Kurtz, "Category-based similarity," in *Proc. 18th Annu. Conf. Cognitive Science Soc.*, 1996, p. 790.

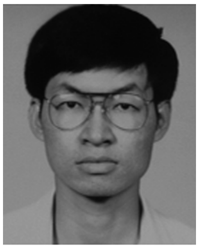
[45] The Internet Movie Database. [Online]. Available: http://www.imdb.com/

[46] P. Ekman, "Are there basic emotions," *Psychol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992.

[47] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. Gainesville, FL: Ctr. for Research in Psychophysiology, Univ. Florida, 1995.

Hee Lin Wang received the B.Eng. degree in electrical and computer engineering from the National University of Singapore in 2001. He is currently a researcher at the Institute of Infocomm Research, ASTAR (Agency of Science, Technology and Research), and is pursuing the Ph.D. degree at the Vision and Image Processing Laboratory, National University of Singapore.

His research interests include biometric processing, affective classification, multimedia indexing, MRF augmented particle-filter tracking, augmented reality, and motion segmentation.



Loong-Fah Cheong was born in Singapore on June 1, 1965. He received the B.Eng. degree from the National University of Singapore and the Ph.D. degree from the Center for Automation Research, University of Maryland at College Park, in 1990 and 1996, respectively.

In 1996, he joined the Department of Electrical and Computer Engineering, National University of Singapore, where he is now an Assistant Professor. His research interests are related to the basic processes in the perception of three-dimensional motion, shape and their relationship, as well as the application of these theoretical findings to specific problems in navigation and in multimedia systems, for instance, in the problems of video indexing in large databases.