# Affective Recommendation of Movies Based on Selected Connotative Features

Luca Canini, Sergio Benini, and Riccardo Leonardi

***Abstract*—The apparent difficulty in assessing emotions elicited by movies and the undeniable high variability in subjects' emotional responses to film content have been recently tackled by exploring film connotative properties: the set of shooting and editing conventions that help in transmitting meaning to the audience. Connotation provides an intermediate representation that exploits the objectivity of audiovisual descriptors to predict the subjective emotional reaction of single users. This is done without the need of registering users' physiological signals. It is not done by employing other people's highly variable emotional rates, but by relying on the intersubjectivity of connotative concepts and on the knowledge of user's reactions to similar stimuli. This paper extends previous work by extracting audiovisual and film grammar descriptors and, driven by users' rates on connotative properties, creates a shared framework where movie scenes are placed, compared, and recommended according to connotation. We evaluate the potential of the proposed system by asking users to assess the ability of connotation in suggesting film content able to target their affective requests.**

***Index Terms*—Affective recommendation, video analysis.**

## I. Introduction

DURING the last few years, the technological evolution and the fast growth of social networks have been shaping a new generation of media consumers. Today, it is extremely easy to access private or shared repositories of multimedia content; as a consequence, the way people enjoy movies, music clips, or home-made videos has dramatically changed, thanks also to the introduction of video on-demand technologies.

In this scenario, a person that feels like watching a movie may rely on the suggestions of his or her group of friends, or on the opinions of a virtual community that shares the same interests. Alternatively, this person could also benefit from the help of a media recommender system with the ability to suggest video content on the basis of his or her user profile, social experience, relationships, and current affective state. The ability of tuning automatic systems according to the emotional state or wishes of users is receiving growing attention, due to the intriguing new possibilities that could be offered by applying affective computing techniques to multimedia systems [1].

Manuscript received January 30, 2012; revised June 4, 2012; accepted July 11, 2012. This paper was recommended by Associate Editor T. Zhang.

The authors are with the Department of Information Engineering, University of Brescia, Brescia 25123, Italy (e-mail: luca.canini@ing.unibs.it; sergio.benini@ing.unibs.it; riccardo.leonardi@ing.unibs.it).

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

Digital Object Identifier 10.1109/TCSVT.2012.2211935

Psychologists have already investigated the emotion-eliciting properties of film media, both in terms of empathy with characters and situations, and in terms of the director's use of established film-making techniques that provide emotional cues. Regarding viewers' empathy, Tan [2] explained a universal affective response in terms of a witness effect in classical Hollywood films, wherein the viewer experiences the real emotions of being a part of the depicted events. Provided they are engaged with the media, viewers' responses are, therefore, a genuine reflection of the affective characterization of a scene.

According to Smith [3], it is not merely empathy with characters that provides the affective cues within film media. Indeed, film-makers make use of techniques of editing, musical scores, lighting, and other aspects of *mise-en-scene* to emphasize a particular emotional interpretation by the viewer. These aesthetic arcs within a film, referred to as *connotation*, plot a continuous path of affective communication, regardless of narrative or plot details, which influences how the meanings conveyed by the director are transmitted to persuade, convince, anger, inspire, or soothe the audience. In cinema as in the literature, we do not merely "read what we see," but connotation brings to our interpretation a range of pre-existing expectations, knowledge and shared experiences that shape the emotional meaning we take from what we see.

### A. Paper Aims and Organization

The severe entanglement between connotation and emotions inspired authors to develop in [4] a space for affective description of movies through their connotative properties. In that work, authors tackled two main research questions.

1) To what extent can we trust emotions registered by other individuals and the content they recommend? The answer was: not much, since emotions are personal, and everyone reacts to any event or to media content in a way that depends on cultural, personal, past experiences and other, even short term, subjective factors. As a possible alternative, perceived connotative properties prove to be more intersubjectively shared than emotions [4].
2) Are connotative rates assigned by users more effective for recommending content than provided affective annotations? The outcome was that movie scenes different in content but similar in connotation likely elicit, in the same user, similar affective reactions. Therefore, using scene similarity based on connotative properties to recommend similar affective content to a single user

is more reliable than exploiting other users' emotional annotations [4].

In this paper, extending the work in [4], we target automatic recommendation of affective content based on audiovisual features. In particular, we investigate the following questions.

1) Can we predict connotative values from audiovisual features? By modeling the relationship between connotative rates assigned by users and selected audiovisual features, we are able to automatically predict connotative values as perceived by users, thus positioning scenes in the connotative space defined in [4].
2) Can we recommend affective content based on predicted values of connotation? Performed tests confirm that recommending movie scenes that are at minimum distance in the connotative space from a query one is an effective strategy for proposing similar emotional content. This verifies that the connotative space constitutes a valid intersubjective platform for affective comparison and recommendation of films.

As a first advantage with respect to the state of the art, the recommendation method here proposed reduces the problem of subjectivity of emotions connected to the use of other people's affective annotations. Since connotative properties are more agreed among people than their emotional reactions, connotation provides a more accurate recommendation method for targeting single users' affective requests.

Second, the proposed learning method that models how to translate low and mid-level properties of video into an intersubjective space for affective analysis of films constitutes a valid nonobtrusive alternative to established methods for performing research on emotions, such as users' self-reporting, monitoring of user's behavior, and neurophysiological signal recording (cited as in [5], on ascending scale of obtrusiveness).

This paper is organized as follows. Section II explores recent advances in affective video analysis and recaps previous findings and experiments in [4] preparatory to this paper. Section III presents the overall methodology, while Section IV describes the audiovisual features extracted to build models for the connotative dimensions. Section V sketches the algorithm used to select features among extracted candidates, which are then mapped onto connotative dimensions by means of the learning methods described in Section VI. Section VII first introduces a validation of the employed model by evaluating the ranking ability of the proposed method on recommendation lists against a ground truth; a user test then assesses performance and potentialities of the proposed framework for affective recommendation of movie scenes. Conclusions and future work are provided in Section VIII.

## II. Previous Work

Recent progress made in the development of affective systems, a well-detailed review of emotion theories, and methods for studying emotions in information science, information retrieval, and human–computer interaction, can be found in the notable work by Lopatovska and Arapakis [5]. Concerning multimedia affective content analysis, this research topic was not popular until a few years ago due to the difficulty in defining objective methods for assessing the affective value of a video and for relating audiovisual descriptors with the emotional dimension of the audience. In this sense, the intuition of Hanjalic represents a breakthrough [6]; the affective dimension of media can be explored because of the expected mood, i.e., the set of emotions the film-maker intends to communicate when he or she produces the movie for a particular audience with a common cultural background. In a work co-authored with Xu [7], Hanjalic pioneers the affective analysis of video content through an approach based on direct mapping of specific video features onto the PA dimensions of the pleasure-arousal-dominance (PAD) emotional model [8]. They describe motion intensity, cut density, and sound energy as arousal primitives, defining an analytic time-dependent function for aggregating these properties along video frames. Though the mapping of video properties on a model intended for describing emotions (PAD) is inspired from the previous literature, it has not yet been thoroughly validated by psychological questionnaires or physiological measurements, which would be proper methods for assessing a time-dependent model.

To date, emotional characterization of videos has been mainly used to study a narrow set of situations, such as specific sporting events as in [9] or, most frequently, movies that belong to a particular genre such as horror movies, as in [10]. Extending this approach, Xu *et al.* [11] described emotional clustering of films for different genres, using averaged values of arousal and valence deduced from video parameters. Such a proposed framework performs better for action and horror films than for drama or comedy, a fact that authors attribute to the prominence of specific features in the first two genres.

Regarding movie scenes, Wang and Cheong [12] proposed to fuse audio and visual low-level features in a heterarchical manner in a high-dimensional space, and to extract from such a representation meaningful patterns by an inference SVM engine. They employed such an approach for probabilistic classification of Hollywood movie scenes into a finite set of affective categories. They also corroborated the view that audio cues are often more informative than visual ones with respect to affective content. In a later work [13], they proposed a motion-based approach combined with an inference engine to recognize different classes of film directing semantics, such as establishing shot, stationary shot and focus-in or focus-out, employed by directors to emotionally emphasize their work.

Irie *et al.*, by proposing a system for affective movie scene classification [14], tackled two main issues: 1) how to extract features that are strongly related to viewers' emotions and 2) how to map the extracted features onto emotion categories. They answered the first question by extracting bags of affective audio-visual words, while for the second one they created a "latent topic driving model" as an attempt for an intermediate representation where topics link emotions to events.

Recently, affective descriptions of multimedia items have also been applied to traditional recommender systems [15]. In [16], Tkalcic *et al.* proposed a framework that describes three stages (entry, consumption, and exit) at which emotions can be used to improve the quality of a recommender system. In a previous work [17], the same research group introduces the usage of metadata fields, containing emotional parameters to increase the precision rate of content-based recommenders for images. By demonstrating that affective tags are more closely related to the user experience than generic descriptors, they improve the quality of recommendation by using metadata related to the aesthetic emotions perceived by users.

Content items can be labeled with affective metadata either explicitly, by asking the user to annotate the observed content with an affective label or, implicitly, by automatically detecting the user's emotional reaction (for a review on implicit human-centered tagging, please refer to [18]). Each of the two approaches has its pros and cons. Again, Tkalcic *et al.* [19] showed that content-based recommendation still works better when explicit labels are used, probably due to the still low accuracy of algorithms that detect affective responses. For this reason, research on improving affective implicit tagging is very active and opening up to a wide range of investigations.

Sicheng *et al.* [20], for example, proposed a video indexing and recommender system based on affective analysis of facial expressions. Users are monitored while watching content and their facial features extracted to infer a probable affective state; on this basis, an affective label is assigned to each movie segment for indexing and recommendation purposes.

Pupillary reflex, gaze distance, and EEG signals are used instead by Soleymani *et al.* in [21] to design an accurate classification protocol for recognizing emotions, attaining comparable performance to users' self-reporting. Although obtained on a fairly limited dataset of 20 video clips and 24 participants, the promising accuracy seems to be easily scalable to a larger population. In a similar fashion, SpudTV [22] within PetaMedia project develops methods for affective implicit tagging of multimedia based on users' EEG signals and peripheral physiological responses.

Recommendation on mobile platforms for providing personalized services that fit users' emotional states was explored by Kim and Choi in [23]. Their EmoSens system maintains affective scoring for various entities in a mobile device, such as applications, multimedia, and contacts. Scoring is based on particular patterns of device usage, which are inferred in a controlled experiment by collecting user feedback.

In the last few years, the problem of tailoring the recommendation experience to user-specific needs has become more evident. Arapakis *et al.* [24] indicated that adapting a general affective model to a specific user introduces a noticeable improvement in the system's ability to discriminate relevant from nonrelevant items. The problem of personal variability in subjects' emotional responses in the case of film content has been recently tackled also in our work in [4], which is summarized in the following paragraphs.

### A. Connotative Space

In [4], we introduced the connotative space as a valid tool for representing the affective identity of a movie segment by those shooting and editing conventions that help in transmitting meaning to the audience. Inspired by similar spaces for industrial design [25], the connotative space accounts for a *natural (N)* dimension that splits the space into a passional hemi-space, referred to as warm affections, and a reflective hemi-space that represents offish and cold feelings (associated

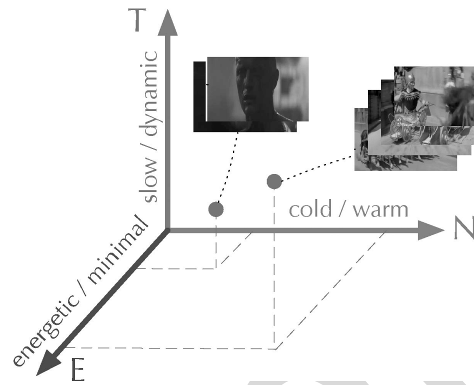



Fig. 1. Connotative space for affective analysis of movie scenes, as in [4].

dichotomy: warm versus cold). The *temporal (T)* axis characterizes the space into two other hemi-spaces, one related to high pace and activity and another describing an intrinsic attitude toward slow dynamics (dynamic versus slow). The *energetic (E)* axis identifies films with high impact in terms of affection and, conversely, minimal ones (energetic versus minimal).

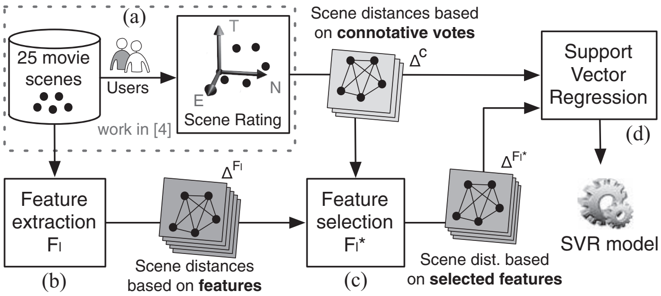
Unlike PAD representation, where each point describes *one emotion* in terms of pleasure, arousal and dominance, in the connotative space, a point (respectively a cloud) describes one (respectively more) *movie segment(s)* in terms of its (their) connotative properties, as shown in Fig. 1.

As a first advantage of using the connotative space, in [4] we showed that the level of agreement among users is higher when rating connotative properties of the movie rather than when they self-report their emotional responses to the same film content. The proposed space seems to fill the need for an intermediate semantic level of representation between low-level features and human emotions, and envisages an easy translation process of video low-level properties into intermediate semantic concepts mostly agreeable among individuals.

The second main outcome provided by analysis in [4] shows how connotation is intrinsically linked to emotions. Specifically, we proved that using connotation for recommending movies to a user whose emotional reactions to the same type of stimuli are known gives better results than exploiting emotional tags by other users. This implies that movie scenes sharing similar connotation are likely to elicit, in the same user, a similar affective reaction. As a consequence, we expect this space to help in reducing the semantic gap between video features and the affective sphere of individuals, thus avoiding the bridging at once process that often inaccurately maps low-level representations to human emotions.

## III. Overall Methodology

While in [4] connotative rates were assigned by users, in this paper we aim to predict connotative values using audiovisual features only. Fig. 2 presents the modelling approach to establish a relation between connotative rates assigned by users and video characteristics. The predicted connotative values are then used for targeting recommendation of affective content in a user test, as described in Fig. 3. The descriptions of the main blocks follow.




AQ:1 Fig. 2. Diagram describing the modeling workflow.

### A. Scene Rating by Users

In the work in [4], we considered a set of 25 "great movie scenes" [26] belonging to popular films from 1958 to 2009 and we asked 240 users to rate each scene on the three connotative dimensions. Following Osgood's evidences [27], rates $Y \in [1, 2, 3, 4, 5]$ were assigned on bipolar Likert scales based on the semantic opposites: warm or cold, dynamic or slow, and energetic or minimal. After rating, the position of a scene $m_i$ in the connotative space is described by the histograms of rates on the three axes $\left(H_i^N, H_i^T, H_i^E\right)$. In this paper, we compute interscene distances between couples $(m_i, m_j)$ by using the Earth mover's distance (EMD) [28] on the rate histograms of each axis $(N, T, E)$ as follows:

$$\Delta_{i,j}^{x} = \mathrm{EMD}\,(H_i^x, H_j^x) \quad x \in \{N, T, E\} \tag{1}$$

which are then combined to obtain the matrix of connotative distances between scenes as $\Delta^C = f\left(\Delta^N, \Delta^T, \Delta^E\right)$ (where function $f$ in [4] is set so as to perform a linear combination of the arguments with equal weights on the three dimensions). In the following, we will refer to these scenes positioned by users' rates as landmarks or training scenes.

### B. Feature Extraction

From movie scenes, we extract features dealing with different aspects of professional content: visual dimension, both color and motion, audio, and film grammar. Since each feature $F_l$ is extracted at its own time scale (frame, shot, and so on), values over a scene $m_i$ are collected in a feature histogram $H_i^{F_l}$ to globally capture its intrinsic variability. For each feature, matrices of interscene distances $\Delta^{F_l}$ are computed as distances between feature histograms.

### C. Feature Selection

To single out those features $F_l^*$ that are the most related to users' connotative rates, we adopt a feature selection criterion based on mutual information.

### D. Regression

A support vector regression (SVR) approach builds a model to relate connotative distances based on users' rates $\Delta^C$ to a function of interscene distances based on selected features $\Delta^{F_l^*}$.

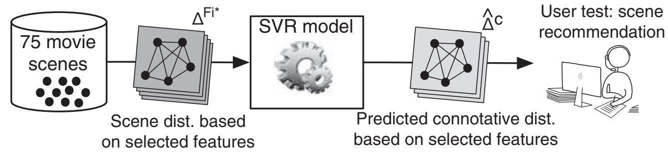



Fig. 3. User test diagram, performed in a recommendation scenario.

TABLE I
LIST OF EXTRACTED FEATURES

| | |
|---|---|
| **Visual** | Dominant color, color layout, scalable color, color structure, color codebook, color energy, lighting key I, lighting key II, saturation⋆, motionDS⋆ |
| **Audio** | Sound energy, low-energy ratio, zero-crossing rate⋆, spectral rolloff⋆, spectral centroid⋆, spectral flux⋆, MFCC⋆, subband distribution⋆, beat histogram, rhythmic strength |
| **Grammar** | Shot length, illuminant color, shot type transition rate |

Descriptors with ⋆ are computed both in terms of average and standard deviation.

### E. Scene Recommendation

Once the model is validated, we are able to predict connotative distances between movie scenes starting from distances based on selected features. As in Fig. 3, which describes the testing scenario, we compute interscene distances on selected features for 75 movie scenes. Then, by the learned SVR model, connotative distances are predicted as $\widehat{\Delta}^C$. The final user test assesses the ability of the connotative space in recommending affective content: users choose a query item and annotate their emotional reactions to recommended scenes that are proposed since at low connotative distance from the query. AQ:2

## IV. Feature Extraction

From movie scenes we extract features to describe professional video content: 12 visual descriptors, 16 audio features, and 3 related to the underlying film grammar, as listed in Table I. For each scene $m_i$ and feature $F_l$, we gather feature values over time in histogram $H_i^{F_l}$. Considering that the system should easily include any new feature, we apply a common quantization strategy to all features by assigning a number of bins that takes the square root of the number of data points in the sample (known as a square-root rule of thumbs).

The selection of the feature set is guided by the following considerations. First, we want our set to include well-known, fast, and effective descriptors. We thus extract MPEG7 visual and motion standard descriptors (dominant color, color layout, scalable color, and others), which are detailed in [29]–[31]. With the same aim, for the audio dimension we include well-studied descriptive features, such as MFCC, subband distribution, and beat histogram, only to cite a few. These have been extensively described and tested in a number of publications, among which [32] and [33] are the most influential.

Second, by scanning recent publications in content-based multimedia affective analysis we select the most promising descriptors, as well as those optimized across several publications (e.g., color energy and lighting key) such as [12] and [34]. From an architectural point of view, since we are aware that a more precise description of the connotative dimensions could be obtained by enlarging the feature set, the proposed system is scalable and open to the insertion of additional features. The considered features are detailed in the following paragraphs.

### A. Visual Features

The visual dimension is perhaps one of the most important ways of communication, which is exploited at its fullest by directors while shaping a film product to convey a specific message. Thus, in our attempt to capture the emotional identity of a movie scene we consider the visual sphere and extract color and motion descriptors, as presented in the following.

We consider MPEG7 color features that proved to be effective in retrieval applications based on visual similarity: dominant color, color layout, scalable color, and color structure [30]. We also extract a codebook constituted by a set of representative colors for a frame, obtained by using a vector quantization approach in the YUV color space [35]. Beyond standard descriptors we employ other visual features believed to have a strong impact on the emotional identity of media content [12]: color energy, lighting key, and saturation.

Color energy is related to the perceptual strength of the color and depends on saturation, brightness, and area occupied by different colors in an image. It also depends on the hue, as in whether it contains more red (energetic) or blue (relaxing) components and the degree of contrast between colors. The result is a scalar indicating for each frame its perceived color energy. For more details, please refer to [12].

Lighting conditions play a key role in scene definition. To capture them we use two descriptors, proposed in [12], referred to as lighting keys. They are related to two major aesthetic lighting techniques: *chiaroscuro*, characterized by high contrast between light and shadow areas, and *flat lighting*, which de-emphasizes the light or dark contrast. Differences between the two illumination techniques lie in the general light intensity and the proportion of shadow area. Thus, for each frame the first descriptor captures the median of the pixels' brightness, while the second, accounting for the proportion of shadow area, uses the proportion of pixels whose lightness falls below the level for which a highly textured surface no longer appears as such [12].

Previous work on affective response to colors proved that saturation and difference in colors are crucial for mood elicitation in subjects [36]. Thus, in addition to the already mentioned features, we adopt two descriptors that account for the average saturation of pixels, as well as their variance.

Finally, motion dynamics are often employed by directors to stress the emotional identity of a scene. To transmit a sensation of speed and dynamism or a feeling of calm and tranquillity, directors often rely on shot pace and type, camera and object motion. The motionDS descriptors introduced in [31] capture the intuitive notion of intensity of action; in particular, we measure the average of motion vector modules and their standard deviation on consecutive frames.

### B. Audio Features

Ambient sound, voices, and music of the soundtrack are forms of expression which play central roles in shaping scene affection and in the process of emotional involvement of the audience [37]. As suggested by a relevant work on audio analysis [32], we decide to describe audio signals in terms of *intensity* (i.e., the energy of the sound, expressed by the amplitude of the associated waveform), *timbre* (related to the spectral shape of the sound and can be seen as the set of qualities that allows us to distinguish two sounds from different instruments), and *rhythm* (related to the repeating sequence of stressed and unstressed beats and divided into measures organized by time signature and tempo indications). In the same work, as well as in other publications (such as [33] and [38]), authors demonstrate that such a description provides high performance for retrieval and classification of audio signals in general, and especially for music.

The choice of privileging features mainly used in musical audio analysis is due to the particular use of audio in movies: scenes that are somehow central to narration are usually stressed due to a particular choice of the soundtrack, e.g., gentle and pleasant music for a romantic moment, loud and aggressive for an action sequence, silences and reprises in a dialogue. In this perspective, audio energy can be seen as a simple but effective clue. In this paper, we consider the energy of an audio signal as the sum of the squared waveform values over 20 ms frames, with 5 ms overlap, as suggested in [32].

Considered timbral features are low-energy ratio and zero-crossing rate in the time domain; spectral rolloff, spectral centroid, spectral flux, MFCC, and subband distribution in the frequency domain. As in [32], except when differently stated, timbral features are initially computed on overlapping frames of 23 ms (analysis windows), so that frequency characteristics of the magnitude spectrum are relatively stable. Actual features are then obtained as average and standard deviation of analysis windows over 1 s, since the sensation of sound "texture" arises following some short-time spectrum pattern in time.

The low-energy ratio is defined as the percentage of analysis windows that have less energy than average within the 1 s window. As an example, vocal music with silences has a high low-energy value, while continuous strings are at a low low-energy value. The zero-crossing rate measures how many times the waveform crosses the zero axis: a periodic and harmonic sound shows a low crossing rate, while a noisy sound is characterized by a high value of this descriptor.

A spectral centroid represents the magnitude spectrum's center of mass of the signal and is interpreted as an index of sound brightness. A limpid sound is usually characterized by a high value of the center of mass, while a dark sound by a low one. Spectral rolloff represents the frequency below which 90% of the energy is concentrated and describes the smoothness of a sound, i.e., the presence of high-frequency harmonics in addition to fundamental tones. Spectral flux, instead, characterizes variations of the frequency spectrum over time. MFCC are perceptually based spectral descriptors widely used for speech and audio classification [32] and are obtained by a linear cosine transform of a log power spectrum on a nonlinear perceptual frequency scale. The last timbral feature is subband distribution, computed as in [33] on overlapping windows of 3 s by decomposing in four subbands using the Daubechies wavelets [39]. Extracted wavelet coefficients from

Fig. 4. Two frames from *A Beautiful Mind*. The left frame evokes a warm sensation, and the other a cold feeling.

each subband provide a compact representation of the energy distribution of the signal in time and frequency.

As for the rhythmic sphere, by using a beat detection algorithm as in [32], which works on chunks of 3 s, 50% overlapping, we derive the beat histogram and its cumulative value as a measure of the rhythmic strength of the audio track.

### C. Film Grammar Features

When watching movies, the feeling is that some film directors have sharply different styles that are easily recognizable. These individual styles can be identified not only in the content, but also from the formal aspects of the films, known as film grammar [40], which encompasses the set of rules followed by a director to convey a certain message.

As proposed in [41], the obvious approach to searching for individual characteristics in the formal side of a director's grammar is to consider those variables that are most directly under the director's control. Among these, shot length (meant as duration), shot type in terms of camera distance to subjects (closeups, medium shots, or long shots), camera movement (such as panning, tilting, or zooming), shot transitions (cuts, fades, dissolves, wipes), and lighting conditions are grammar aspects that can be automatically investigated. In this paper, we consider as a first set of film grammar features, meaning that they are directly under the director's control, the shot length, the color of the illumination source, and the pattern of shot type.

Shot length greatly affects how a scene is perceived by the audience. Longer durations connotate a scene as more relaxed and slower paced, whereas shorter shots give the impression of a faster paced scene [42]. Thus, we extract the average shot length as an effective scene descriptor.

The second feature is related to the spectral composition of the light source, which is often exploited by directors to give a connotative signature to movies. Light used in the shooting process, called illuminant, influences the appearance of every element in the scene: objects do not have their own colors, which are instead due to the interaction with the incident electromagnetic radiation. In Fig. 4, the frame on the left shows a scene with a yellow polarized illuminant which evokes a pleasant sensation, while the one on the right suggests a colder feeling because of the grayish illuminant. Here, for each frame we estimate the illuminant color by improving a white patch algorithm [43] with the procedure we propose in [44].

The third descriptor accounts for the change of employed shot types. Varying camera distance is a common directing rule used to subtly adjust the relative emphasis between the filmed subject and the surrounding scene [13]. This affects the emotional involvement of the audience [40] and the process of identification of viewers with the movie characters. There are, in fact, evident correspondences between the film-maker's choice of shot type and the proxemic patterns [45], i.e., the subjective dimensions that surround each of us and the physical distances one keeps from other people in social life. Although the gradation of distances is infinite, in practical cases the categories of definable shot types can be re-conducted to three fundamental ones: long shots, medium shots, and closeups (see [40] for a complete taxonomy). First, for each scene we estimate the type of employed shots by the algorithm presented in [46]. Then, we define the shot type transition rate as the number of type changes across consecutive shots in a scene, normalized to the total number of shots. As shown in [47], this rate is in fact part of the complex mechanism responsible for triggering audience's emotional involvement, with strong evidences especially on the arousal dimension.

## V. Feature Selection

A feature selection method is applied to disclose the relationships between scene coordinates in the connotative space assigned by users and the related audiovisual features. This step aims at unveiling which are the audiovisual descriptors that mostly affect user's perception of connotative properties to be employed in the regressive models adopted in Section VI.

Feature selection algorithms are very popular in several disciplines [48], such as gene expression, array analysis, combinatorial chemistry, and multimedia analysis. Given a number of descriptors, they aim at discriminating between those relevant for a certain goal from those that are not, allowing the learning step which usually follows to work with a compact set of significant features. The main advantages are reduction of the number of features to be processed, exclusion of redundant or inefficient ones, and a better understanding of the problem.

The definition of the right selection algorithm for a specific problem depends on several aspects. One possible choice is to integrate the feature selection within the subsequent regression algorithm (e.g., to use a support vector approach for feature selection embedded in an SVR), as suggested, for example, in [49]. However, instead of applying such a procedure, called *wrapping*, we prefer to apply a filtering method, i.e., to keep separated selection and prediction. Filtering methods, apart from being in general computationally less expensive than wrappers [50], usually provide an easier understanding of the selection problem. In addition to this, they are independent of the ensuing learning method, thus allowing the study of the effectiveness of the features with different regression approaches, as we perform in Section VI.

For our specific goal of discovering audiovisual features relevant to connotation, a potential issue is redundancy; it is, in fact, likely that if a particular descriptor is relevant, other descriptors that are correlated to the first one result relevant too. For this reason, we employ an information theory-based filter that selects the most relevant features in terms of mutual information with user votes, while avoiding redundant ones: the minimum-redundancy maximum-relevance (mRMR) scheme introduced in [51].

Given the set of $L$ features $\{F_l\}_{l=1,\ldots,L}$ and the user votes $Y$ on each connotative axis, both interpreted as random variables, consider *relevance* ($V$) and *redundancy* ($W$) defined as

$$V = \sum_{F_l \in S} \frac{I(F_l, Y)}{|S|} \quad W = \sum_{(F_l, F_j) \in S \times S} \frac{I(F_l, F_j)}{|S|^2} \tag{2}$$

where $I$ indicates the mutual information and $S$ is the set of selected descriptors. The goal is to select a subset of $M$ features ($M = |S|$, $M < L$) as informative as possible with respect to users' votes ($\max(V)$) and, at the same time, as uncorrelated as possible among themselves ($\min(W)$). A possible criterion (exposed in [51]) to jointly optimize both conditions treating them as equally important is to maximize the difference between quantities in (2): $\max(V - W)$.

To solve this optimization problem, a heuristic called mutual information difference criterion (MID) is used, as in [52]. According to it, the first selected feature $F_f$ is the most relevant $\left(I(F_f, Y) \geq I(F_l, Y), l = 1, \ldots, L\right)$, while other features are added in an incremental way; for each candidate feature $F_l$ not yet in $S$, the quantities in (2) are recomputed as follows:

$$\widehat{V}_l = I(F_l, Y) \quad \widehat{W}_l = \sum_{F_j \in S} \frac{I(F_l, F_j)}{|S|} \tag{3}$$

and the newly selected feature is the one so that

$$\arg\max_{F_l \notin S} \left(\widehat{V}_l - \widehat{W}_l\right). \tag{4}$$

### A. Sample Probabilities on Distances

To compute mutual information $I(.,.)$ it is necessary to sample probabilities of features and votes. However, when dealing with multidimensional feature histograms $H^{F_l}$, the direct application of such a procedure is impractical. This is due to the number of scenes that would be required if we wanted to compute reliable statistics, both marginal and joint, on all possible combinations of feature values and users' votes.

To overcome this issue, for the selection and regression steps we do not take into account actual histograms, but distances between them. Therefore, we do not employ the absolute position of scenes, but the knowledge of how they are placed with respect to all others, both in connotative and in feature spaces. Such a scheme naturally fits our aim, which is, in fact, to recommend movie scenes according to their proximity in the connotative space. Approaches based on distances between items are also closer to the human mechanism of perceiving emotions, which works in a comparative way rather than using an absolute positioning, as shown in [53] for music items.

For our aims we then consider interscene distances based on users' rates $\Delta^x$ as expressed in (1), and distances based on feature histograms $\Delta^{F_l}$. In the specific, for each descriptor $F_l$, the element of $\Delta^{F_l}$ in position $i, j$ is given by

$$\Delta_{i,j}^{F_l} = \text{EMD}\left(H_i^{F_l}, H_j^{F_l}\right) \quad i, j = 1, \ldots, 25. \tag{5}$$

TABLE II

FEATURE RANKING AND RELEVANCE $\widehat{V}$ ACCORDING TO THE mRMR MID SCHEME (SELECTED ONES ARE IN BOLD)

| NATUR. | $\widehat{V}$ | TEMP. | $\widehat{V}$ | ENERG. | $\widehat{V}$ |
|---|---|---|---|---|---|
| **Col.layout** | **0.22** | **Rhyt.str.** | **0.27** | **Sound en.** | **0.18** |
| **Spec.roll.[d]** | **0.14** | **Shot ty.t.r** | **0.15** | **Shot len.** | **0.09** |
| **Light.key II** | **0.11** | **Mot.DS[d]** | **0.25** | Spec.ce.[d] | 0.06 |
| **Illuminant** | **0.19** | Sound en. | 0.10 | Sub.dist.[a] | 0.06 |
| Spec.ce.[d] | 0.07 | Spec.ce.[d] | 0.18 | Satur.[d] | 0.07 |
| Sound en. | 0.06 | Sub.dist.[a] | 0.03 | Col.layout | 0.07 |
| Col.codeb. | 0.21 | Shot len. | 0.07 | Spec.roll.[d] | 0.05 |
| Zero cr.r.[d] | 0.12 | MFCC [d] | 0.08 | Rhyt.str. | 0.12 |
| Shot ty.t.r. | 0.07 | Scal.col. | 0.09 | Spec.flux[d] | 0.04 |
| Col.en. | 0.08 | Satur.[d] | 0.04 | Beat hist. | 0.05 |
| Col.sat.[a] | 0.06 | Spec.cen.[a] | 0.17 | Shot ty.t.r. | 0.05 |
| Sub.dist.[a] | 0.05 | Low en.r. | 0.03 | Col.struc. | 0.04 |
| Mot.DS[a] | 0.07 | Light.key I | 0.05 | Col.en. | 0.04 |
| Shot len. | 0.04 | Spec.flux[a] | 0.04 | MFCC[d] | 0.06 |
| MFCC[a] | 0.06 | Mot.DS[a] | 0.13 | Spec.roll.[a] | 0.04 |
| Col.struct. | 0.08 | Col.en. | 0.04 | Illuminant | 0.04 |
| Scal.col. | 0.16 | Zero cr.r.[a] | 0.04 | Mot.DS[d] | 0.06 |
| Satur.[d] | 0.03 | Zero cr.r.[d] | 0.03 | Low en.r. | 0.03 |
| Low en.r. | 0.04 | Ligh.key II | 0.03 | Sub.dist.[d] | 0.06 |
| Dom.col. | 0.18 | Illuminant | 0.03 | Spec.flux[a] | 0.05 |
| Beat hist. | 0.04 | Spec.flux[a] | 0.04 | Light.key II | 0.04 |
| Spec.flux[d] | 0.03 | Beat hist. | 0.03 | Zero cr.r.[a] | 0.03 |
| Zero cr.r.[a] | 0.12 | MFCC[a] | 0.05 | Light.key I | 0.05 |
| Rhyt.str. | 0.05 | Spec.roll.[d] | 0.02 | Satur.[a] | 0.04 |
| MFCC[d] | 0.06 | Sub.dist.[d] | 0.09 | Scal.col. | 0.07 |
| Sub.dist.[d] | 0.05 | Col.layout | 0.04 | Mot.DS[a] | 0.04 |
| Light.key I | 0.06 | Spec.roll.[a] | 0.02 | MFCC[a] | 0.04 |
| Spec.flux[a] | 0.03 | Col.struct. | 0.03 | Zero cr.r.[d] | 0.03 |
| Spec.roll.[a] | 0.03 | Satur.[a] | 0.02 | Col.codeb. | 0.07 |
| Spec.cen.[a] | 0.05 | Col.codeb. | 0.06 | Spec.cen.[a] | 0.03 |
| Mot.DS[d] | 0.04 | Dom.col. | 0.05 | Dom.col. | 0.03 |

Those computed in average and std dev are indicated with [*a*] and [*d*], respectively.

After normalizing and quantizing EMD distances[1] on five levels as for distances in the connotative space, we compute the mutual information between distances based on feature histograms and connotative distances based on users' rates on a proper number of samples. The MID criterion is then reformulated as follows: for each connotative axis $x \in \{N, T, E\}$, the first selected feature $F_f$ is the one so that

$$I(\Delta^{F_f}, \Delta^x) \geq I(\Delta^{F_l}, \Delta^x) \quad l = 1, \ldots, L \tag{6}$$

while the following features are added as in (4) where:

$$\widehat{V}_l = I\left(\Delta^{F_l}, \Delta^x\right) \quad \widehat{W}_l = \sum_{F_j \in S_x} \frac{I\left(\Delta^{F_l}, \Delta^{F_j}\right)}{|S_x|}. \tag{7}$$

This way, according to the MID criterion, we rank features for each connotative axis, as shown in Table II.

[1]It is worth noticing that the EMD computation is based on the definition of a ground distance, i.e., the distance between two samples of the considered feature. In our work, we use for each feature the *ad hoc* ground distance, as found in the literature: distances as proposed for MPEG7 descriptors in [30], $L^2$ on RBG components for the illuminant color, and so on, while for users' votes expressed on Likert scales we adopt $L^1$ distance.

### B. Relevant Feature Sets

The next crucial aspect is the number of features to select for the regression step; keeping too many descriptors would increase the computational cost of the extraction process, while considering too few descriptors would potentially lead to a poor regression model. Following these considerations we keep, for each connotative axis, only those features that are able to increase the level of mutual information between features and connotative votes above a minimum contribution.

In terms of MID criterion, considering a set $S_x$ of already selected features for the $x$-axis, the next feature in the ranking list $F_l$ is selected if its contribution $\widehat{V}_l - \widehat{W}_l$ [computed as in (7)] satisfies the condition

$$\widehat{V}_l - \widehat{W}_l \geq r \cdot I(\Delta^{F_f}, \Delta^x) \tag{8}$$

where $r \in [0, 1]$ and $I(\Delta^{F_f}, \Delta^x)$ is the mutual information of the first ranked feature for that axis, i.e., the best descriptive one with respect to user's votes. To find the optimal value for $r$, we scan the range of values between 0 and 1 and measure recommendation performance on ranked lists against a ground truth (as described in Section VII-A). In general, we notice that recommendation performance improves when the number of selected features increases, i.e., when $r$ diminishes. However, if $r$ becomes too low, thus including even not so significant or noisy features in terms of mutual information with users' votes, the effectiveness of the system stops increasing. Therefore, during tests in Section VII-A, we determine that the optimal value, in the sense that it maximizes recommendation performance and minimizes complexity in terms of number of descriptors to be extracted, corresponds to $r = 0.15$. By setting this value, we select four features for the natural dimension, three for the temporal one, and two for the energetic one (in bold in Table II). As a reinforcement for the operated choice on $r$, we notice that selected features make intuitive sense for all axes.

As seen in [4], the natural dimension is related to warm or cold affections, and it is voted by users as the scene atmosphere. As expected, selected features for this axis are intuitively involved in the characterization of a scene's atmosphere; they, in fact, describe the color composition (color layout), the variations in smoothness and pleasantness of the sound (spectral rolloff standard deviation) and the lighting conditions in terms of both illumination (illuminant color) and proportion of the shadow area in a frame (one of the lighting key descriptors which is dramatically stressed in the chiaroscuro technique).

The temporal axis has been rated by users in terms of high pace versus slowness. The algorithm returns for this axis the rhythmic strength of the audio signal, which is an index related to the rhythm and the speed sensation evoked by a sound, the pace variation of the employed shot types (shot type transition rate), and the variability of the motion activity (standard deviation on motion vector modules).

User votes on the energetic dimension distinguish items with high affective impact from minimal ones. Selected features are again commonsensical and coherent: the first describes the sound energy, while the second one is the shot length; for example, short shots usually employed by directors in action scenes are generally perceived as very energetic.

TABLE III
APPROXIMATION ERROR ON SCENE DISTANCES BASED ON USERS' VOTES IN TERMS OF RMSE, OBTAINED USING THE REPORTED REGRESSION METHODS

| Regression method | RMSE |
|---|---|
| Polynomial regression | 0.281 |
| Neural network | 0.248 |
| SVR | 0.188 |

Distances are normalized in the range [0, 1].

## VI. REGRESSION

Once features relevant to connotative votes on each axis are picked, we aim at estimating connotative distances $\Delta^x$ based on rates by a function of distances based on selected features

$$\Delta^x \approx \widehat{\Delta}^x = g_x\left(\{\Delta^{F_l}\}_{F_l \in S_x}\right). \tag{9}$$

To define functions $g_x$ that best link the denotative level with the connotative dimensions, we set up a modelling framework using selected features as inputs and connotative votes as desired outputs (not in absolute terms but as distances).

In order to compare different regressive procedures for approximating the desired output starting from the inputs, we test in particular polynomial combination, neural networks (feed-forward neural network trained by a back-propagation algorithm), and SVR models [54] with standard RBF kernel. Modelling functions $g_x$ are then obtained for dimensions $x \in \{N, T, E\}$ by adopting SVR models that are the ones that return the lowest root mean squared error on scene distances based on users' votes, as reported in Table III.

This modeling step provides a way to translate video properties into intermediate semantic connotative concepts, which are mostly agreeable among individuals. As a result, the approximated matrix of connotative distances is found as follows:

$$\widehat{\Delta}^C = f\left(\widehat{\Delta}^N, \widehat{\Delta}^T, \widehat{\Delta}^E\right) \tag{10}$$

where function $f$ is set as in (1).

## VII. EXPERIMENTS ON SCENE RECOMMENDATION

The idea of the affective recommendation scenario here proposed is that once a user expresses an implicit emotional wish by selecting a query item (e.g., by choosing a happy scene in his or her opinion), the recommendation algorithm should return a list of candidate movie scenes that are emotionally close to the given query for that user. This kind of query-by-example approach has its roots in information filtering, and goes under the name of content-based recommendation [55] (as opposed to other methods, e.g., collaborative filtering [15]).

Recommendation results are returned as top-k lists, a concept ubiquitous in the field of information retrieval (e.g., the list of $k$ items in the first page of results by a search engine). They are a valid mechanism for propagating emotional tags from the already watched content to close items, thus enabling better filtering of relevant items from the nonrelevant ones as in [17]. The following experiments aim to measure the ability of the connotative space in proposing content relevant to user's emotional preferences.

### *A. Ranking Lists Against a Ground Truth*

To evaluate how good distances based on selected features $\widehat{\Delta}^C$ approximate scene distances computed on users' rates $\Delta^C$, we compare the abilities of the two distance matrices in ranking lists of movie scenes with respect to ground-truth lists built by single users.

This first experiment uses the data gathered by the 240 users on the 25 movie scenes in [4]. The collective users' emotional annotations are expressed in the form of emotional distances $\Delta^W$ between scenes, while the ground-truth lists per each single user $u_k$ are built by the emotional distances $D^W_{u_k}$ between scenes expressed by that specific user. By observing the emotation wheel [4] in Fig. 6 we recall that the distance between two emotions $e_i$ and $e_j$ is the number of steps required to reach emotion $e_j$ from emotion $e_i$, as stated by Russell in [56] and recently adopted by Irie *et al.* in [14] as well as in our test in [4]. As Russell observes, "a score of 1 (is assigned) to the distance between adjacent terms," whereas "a distance 4 is assigned between terms placed opposite on the circle," no matter whether computed clockwise or counterclockwise. Please observe that $D^W_{u_k}$ is not a distance between distributions of votes (as $\Delta^W$ is since it aggregates all users' votes), but is a distance between scene emotions assigned by a single user.

In the proposed test, given a user and a *query* item, all movie scenes are first matched according to how emotionally similar they are to the query item, according to single user's emotional annotations (i.e., the ground truth in $D^W_{u_k}$). Second, this list of scenes is re-ranked based on distances expressed in $\Delta^C$ (i.e., ranking by connotative rates), which expresses the ability of the connotative space in matching the affective preferences of single users. In [4], we have already shown that to recommend movie scenes, connotation ($\Delta^C$) works better than using aggregated emotations by all users ($\Delta^W$) to approximate the ground-truth ranking obtained using $D^W_{u_k}$. Here, we also consider the case when ranking is performed by using the learned models, i.e., how good is the ranking obtained by using the approximated distances $\widehat{\Delta}^C$ provided by the SVR models (i.e., ranking by connotative properties predicted by audiovisual features).

Ranking quality is measured by the Kendall's tau metric $K$ [57], which is equal to the number of exchanges needed in a bubble sort to convert one ranked list to the other one, normalized in the interval [0, 1].

In this process, we apply a five-fold cross validation approach. At each round 20 scenes are used to build the models, and the metric $K$ is measured on the five remainders. Folds are manually arranged using stratification [58], thus ensuring that scenes are balanced as much as possible with respect to the connotative votes assigned by users.

As a result, considering as ground-truth lists those ranked by single users' emotional annotations $D^W_{u_k}$ (for which $K = 0$), the average error performed by using $\Delta^C$ to rank scenes is $K_{\Delta^C} = 0.425$, while the average error performed by using

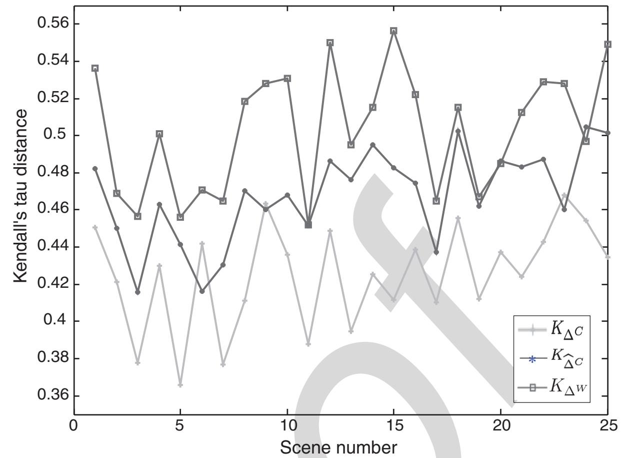



Fig. 5. Kendall's tau metric measuring the quality of list ranking by using connotative distances based on votes ($K_{\Delta^C}$) and by distances approximated with the learning models ($K_{\widehat{\Delta}^C}$) (values gathered on a scene basis and averaged on five-folded models). Since the ground-truth lists are at K = 0, both $\Delta^C$ and $\widehat{\Delta}^C$ perform better than ranking lists by using emotional annotations aggregated by all users ($\Delta^W$).

$\widehat{\Delta}^C$ to rank scenes is just slightly above, $K_{\widehat{\Delta}^C} = 0.467$, however, still inferior than the error performed when using $\Delta^W$ ($K_{\widehat{\Delta}^W} = 0.502$). Inspecting results in Fig. 5 (which shows Kendall's tau scores for each of the 25 scenes, as an average result on the five-folded evaluation) in a comparative way, we can conclude that even if the regression undeniably introduces an error, when the goal is not to replicate exact connotative distances but to obtain a similar ranking, the average ability of the system does not significantly degrade when using $\widehat{\Delta}^C$ instead of $\Delta^C$. More importantly, returned lists using $\widehat{\Delta}^C$ better match the ground-truth lists per each single user than using the aggregated annotations by other users $\Delta^W$, meaning that even connotative properties predicted by audiovisual features are more intersubjectively agreed among people than collective emotional annotations.

### *B. Scene Recommendation: User Test*

The first test that employed a ground truth is here expanded in a larger application scenario for recommending novel movie scenes to users. For this second user test, which is performed online with support of English language, 38 users were recruited. When performing the test, they were not aware of the final aims of the research.

Regarding the scene database, we would like to remark that while ground-truth databases for events and objects analysis in videos are available and relatively easy to build (they can be annotated by one single person and be objectively valid for almost everyone else), large ground-truth video databases where each video scene is emotionally (and subjectively) annotated by a large number of users do not yet exist. In our experiment, in addition to the 25 landmarks, 50 new scenes not previously involved in modelling are adopted as candidates for recommendation for evaluating users' satisfaction with the system, for a total number of 75 scenes. The complete list of employed scenes provided with title, duration, year, IMDb film rank, and (for the new 50 scenes) the available online links for inspection, can be found in [59].


Fig. 6. User test interface: example of a scene to be annotated with an emotional tag chosen among those on the emotation wheel [4] (happiness, excitement, tension, distress, sadness, boredom, sleepiness, relaxation).

While the first test could be evaluated in terms of Kendall's tau metric against relatively short ground-truth lists built by each user, for a database of 75 scenes it is not possible to produce ground-truth lists, since it is unfeasible for each user to rate all new scenes in an limited time without losing focus and attention.

For this reason, each user is asked to query the recommendation system only two times, and rate only three top and three bottom results per each query, for a total number of 14 voted scenes per user (12 rates on scenes plus 2 annotated queries). With the described procedure, 38 users provide more than 500 votes, which allows for gathering reliable statistics on the 75 scenes. To the best of our knowledge, no other work on content affective analysis so far recruited such a large number of users on video recommendation tests (almost 300 users, considering both tests).

To start the test, each user chooses as query items two landmark scenes and tags each with one emotional label chosen among the eight available on the emotation wheel, as shown in Fig. 6. The system returns, for each query, a list of six movie scenes that contains, in a random order, the top-3 close scenes in the connotative space and the three most distant ones from the query (among the 75 total scenes). To verify the ability of the connotative space in recommending similar affective items, we ask users to annotate each proposed scene with one emotional tag, as shown in the interface of Fig. 6.

Since each user, given a query, is required to watch only a limited number of scenes (the top-3 close and the top-3 far items), recommendation results can be evaluated in terms of precision@k: the number of results which are judged to be *relevant* by the user among the first $k = 3$ recommended results.

However, we try to do more than that; instead of stating whether a result is just relevant or not-relevant, the performed test allows us to state to what extent an item is relevant by measuring the scene emotional distances expressed by the user from the query ($d = 0$ "scene with same emotion," $d = 1$ "scene with similar emotion," ..., $d = 4$ "scene with opposite emotion"), which is also closer to the human mechanism of perceiving emotions, which works in a comparative way rather than using absolute terms.

In this sense, if we consider as relevant only those recommendations that are at null emotional distance ($d = 0$), then precision@3 is 0.3 for top-3 close items. However, considering as relevant also items that are emotionally similar ($d \leq 1$), precision@3 raises to a significant 0.68. All precision@3 results are shown in Table IV for both top-3 close and top-3 far scenes at different emotional distances.

TABLE IV
PRECISION@3 RESULTS FOR (LEFT) TOP-3 CLOSE AND (RIGHT) TOP-3 FAR SCENES AT DIFFERENT EMOTIONAL DISTANCES

| top-3 close | precision@3 | top-3 far | precision@3 |
|---|---|---|---|
| d=0 | 0.30 | d=4 | 0.22 |
| d≤1 | 0.68 | d≥3 | 0.59 |
| d≤ 2 | 0.87 | d≥2 | 0.82 |
| d≤ 3 | 0.95 | d≥1 | 0.97 |
| d≤ 4 | 1.00 | d≥0 | 1.00 |

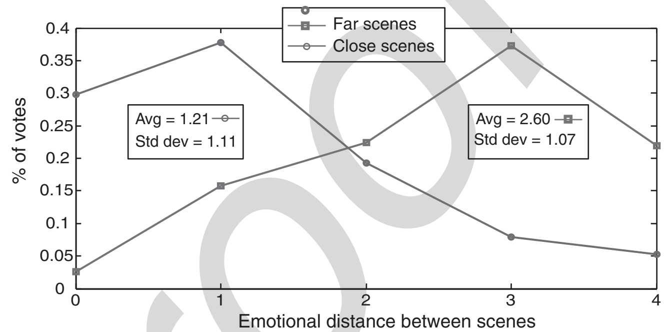



Fig. 7. Histogram (blue) of the emotional distances between the query and the top-3 close recommended items and histogram (red) of the emotional distances between the query scene and the top-3 far scenes. Distances are in the range [0, 4] (0: "same emotion as query" and 4: "opposite emotion").

Fig. 7 summarizes the obtained results by showing the histogram of emotional distances of the top-3 close scenes (blue) and the histogram of distances of the top-3 far scenes (red) from the query. They approximate the probability distribution functions of perceived emotional distances between the query and the recommended items. Since they consider precision computed at different scales of relevance (i.e., at different emotional distances $d$), they can be considered as more informative than a single value of precision@k stating whether a result is just relevant or not.

## C. Discussion

In [4], we have already shown that the connotative judgements are more effective than using people's affective responses in recommending content able to target the emotional request of a single user.

In this paper, we push this result one step further. We, in fact, state that it is possible to automatically position a movie scene in the connotative space by analyzing its audiovisual and grammar features without asking users to express their perception of film connotative properties. By selecting audiovisual descriptors relevant to connotation, we are able to map movie scenes in the connotative space, and to discover similar and dissimilar affective content by computing distances in this space. The first test in Section VII-A demonstrates that using audiovisual properties to derive connotative coordinates introduces a risible drop in performance if compared to connotative rating by users.

In the recommendation scenario, when the user wishes to get some content eliciting a particular emotion, the system automatically proposes content which in the connotative space is close to some items already tagged by the user with the desired emotion. In this final experiment, we checked users' satisfaction with recommendation results by computing the emotional distances between the emotions elicited by the query item and by the suggested scenes.

This way, we have closed the loop: at the beginning the user expresses an emotional wish; to target this, we use already tagged content that elicited in that specific user that emotional reaction. We then look for similar content in the connotative space, and finally ask the user which emotion he or she is inspired with to check the correctness of the emotional recommendation. The outcome of the experiment, as summarized in Table IV and Fig. 7, reveals the effectiveness of the connotative space in proposing content eliciting similar affective reactions.

Notice that content close in the connotative space can be very different in terms of denotative meaning; it happens, for example, that a fight in *Kill Bill II* is recommended on the base of the chariot chase in *Ben Hur*. Even if different in content, both scenes elicit a similar affective reaction in the same user, which is the basic idea of affective recommendation.

A few last considerations on scenes, database dimensions, and future work. Experiments performed in this paper use movie scenes as elementary units since, by definition, each scene in a movie depicts a self-contained high-level concept. We are aware that a recommender system of video scenes has little practical purpose. However, starting from understanding how the system behaves with elementary units of film items is a valid practical approach for future extensions to full movies. To the best of our knowledge, no experiments on affective recommendation on full movies have been attempted so far. Thus, our next research goal is to extend our approach to full movie recommendation.

Of course, working on full movies introduces severe scalability issues to our approach, which are worth discussing. In this paper, each scene is represented as a point in the connotative space. When using full movies instead, the idea is to consider a connotative cloud of scenes or, considering the time dimension, a connotative trajectory that interconnects subsequent scenes in the film.

Even if there is an undeniable technical difficulty in conducting experiments on larger scene databases, we are already tackling this scalability challenge, from both the system and the algorithm time complexity's standpoints. By exploiting the knowledge about the position of a few landmark scenes, it is indeed possible to assign other scenes with absolute positions instead of using distances between scenes. Thus, once a reliable set of landmark scenes is found, new scenes and movies can be added without much complexity, ensuring adequate scalability to the system.

The fact that the system is actually open to the insertion of new scenes and movies, so that users can get more and more recommended items as long as the database increases, is indeed an asset of the system. In fact, while now with 75 scenes it might happen that some scenes have no close neighbors in the connotative space (so that users might be not fully satisfied with the recommended items), the more the database grows, the higher the chances that proper emotional content is found.

## VIII. Conclusion

In this paper, we proposed an affective framework where movie scenes are placed, compared, and recommended by extracting audiovisual and film grammar features. The learning model allowing to link physical features of videos to users' emotional preferences was driven by users' rates on connotative properties, defined as the set of shooting and editing conventions that helped in transmitting meaning to the audience. Connotation here provided an intermediate representation level that exploited the objectivity of audiovisual descriptors to match the emotional queries of single users. To demonstrate the validity of this approach, we conducted a first test of the model against a ground truth to verify the translation process of relevant audiovisual low-level descriptors into connotative properties. Then, a final user test verified the ability of the connotative framework to recommend items matching users' affective requests, thus positively answering to both initial research questions. Further studies on the extension of the current scene-based method to full movies are currently ongoing.

## References


[1] R. W. Picard, "Affective computing: From laughter to IEEE," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 11–17, Jan. 2010.

[2] E. S. H. Tan, "Film-induced affect as a witness emotion," *Poetics*, vol. 23, no. 1, pp. 7–32, 1995.

[3] G. M. Smith, *Film Structure and the Emotion System*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[4] S. Benini, L. Canini, and R. Leonardi, "A connotative space for supporting movie affective recommendation," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1356–1370, Dec. 2011.

[5] I. Lopatovska and I. Arapakis, "Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction," *Inf. Process. Manage.*, vol. 47, no. 4, pp. 575–592, Jul. 2011.

[6] A. Hanjalic, "Extracting moods from pictures and sounds," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006.

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

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

[9] J. Wang, E. Chng, C. Xu, H. Lu, and X. Tong, "Identify sports video shots with 'happy' or 'sad' emotions," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2006. AQ:3

[10] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2005.

[11] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 677–680.

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

[13] H. L. Wang and L.-F. Cheong, "Taxonomy of directing semantics for film shot classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 10, pp. 1529–1542, Oct. 2009.

[14] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[15] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[16] M. Tkalcic, A. Kosir, and J. Jurij Tasic, "Affective recommender systems: The role of emotions in recommender systems," in *Proc. RecSys Workshop Hum. Decision Making Recommender Syst.*, 2011.

[17] M. Tkalcic, U. Burnik, and A. Kosir, "Using affective parameters in a content-based recommender system for images," *User Model. User-Adapt. Interact.*, vol. 20, no. 4, pp. 279–311, 2010.

[18] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging," *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 173–180, Nov. 2009.

[19] M. Tkalcic, A. Odic, A. Kosir, and J. F. Tasic, "Impact of implicit and explicit affective labeling on a recommender system's performance," in *Proc. 19th Int. Conf. Advances User Modeling*, 2012, pp. 342–354.

[20] Z. Sicheng, H. Yao, S. Xiaoshuai, P. Xu, R. Ji, and X. Liu, "Video indexing and recommendation based on affective analysis of viewers," in *Proc. ACM Int. Conf. Multimedia*, Dec. 2011.

[21] M. Soleymani, M. Pantic, and T. Pun, "Multi-modal emotion recognition in response to videos," *IEEE Trans. Affective Comput.*, vol. 99, no. 2, pp. 211–223, Apr.–Jun. 2012.

[22] S. Koelstra, "SpudTV," in *Proc. NEM Summit*, Sep. 2011.

[23] H.-J. Kin and Y. S. Choi, "EmoSens: Affective entity scoring, a novel service recommendation framework for mobile platform," in *Proc. 5th ACM Conf. Recommender Syst.*, Oct. 2011.

[24] I. Arapakis, K. Athanasakos, and J. M. Jose, "A comparison of general versus personalized affective models for the prediction of topical relevance," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 2010, pp. 371–378.

[25] C. T. Castelli, "Trini diagram: Imaging emotional identity 3-D positioning tool," *Int. Soc. Opt. Eng.*, vol. 3964, pp. 224–233, Dec. 1999.

[26] *What Is a "Great Film Scene" or "Great Film Moment"? An Introduction to the Topic* [Online]. Available: http://www.filmsite.org/scenes.html AQ:4

[27] C. Osgood, G. Suci, and P. Tannenbaum, *The Measurement of Meaning*. Champaign, IL: Univ. Illinois Press, 1957.

[28] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[29] T. Sikora, "The MPEG-7 visual standard for content description: An overview," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 696–702, Jun. 2001.

[30] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[31] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 720–724, Jun. 2001.

[32] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[33] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in *Proc. Conf. Acoust. Music Theory Applicat.*, 2001.

[34] 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.

[35] S. Benini, A. Bianchetti, R. Leonardi, and P. Migliorati, "Extraction of significant video summaries by dendrogram analysis," in *Proc. Int. Conf. Image Process.*, Apr. 2006.

[36] P. Valdez and A. Mehrabian, "Effects of color on emotions," *J. Exp. Psychol.*, vol. 123, no. 4, pp. 394–409, 1994.

[37] T. Holman, *Sound for Film and Television*. Waltham, MA: Focal Press, 2002.

[38] G. Tzanetakis, "Automatic musical genre classification of audio signals," in *Proc. ISMIR*, 2001.

[39] I. Daubechies, "Orthonormal bases of compactly supported wavelets II: Variations on a theme," *J. Math. Anal.*, vol. 24, no. 2, pp. 499–519, Mar. 1993.

[40] D. Arijon, *Grammar of the Film Language*. Beverly Hills, CA: Silman-James Press, 1976.

[41] B. Salt, *Moving Into Pictures. More on Film History, Style, and Analysis*. London, U.K.: Starword, 2006.

[42] K. Choroś, "Video shot selection and content-based scene detection for automatic classification of TV sports news," in *Internet: Technical Development and Applications* (Advances in Intelligent and Soft Computing, vol. 64). Berlin/Heidelberg, Germany: Springer, 2009, pp. 73–80.

[43] J. V. de Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2207–2214, Sep. 2007.

[44] L. Canini, S. Benini, P. Migliorati, and R. Leonardi, "Emotional identity of movies," in *Proc. 16th IEEE Int. Conf. Image Process.*, Nov. 2009.

AQ:5 [45] E. T. Hall, *The Hidden Dimension*. Anchor, 1990.

[46] S. Benini, L. Canini, and R. Leonardi, "Estimating cinematographic scene depth in movie shots," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010.

[47] L. Canini, S. Benini, and R. Leonardi, "Affective analysis on patterns of shot types in movies," in *Proc. 7th Int. Symp. Image Signal Process. Anal.*, Sep. 2011.

[48] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[49] J.-B. Yang and C.-J. Ong, "Feature selection for support vector regression using probabilistic prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 343–352.

[50] W. Duch, "Studies in fuzziness and soft computing," in *Filter Methods* (Physica-Verlag). Berlin, Germany: Springer, 2006, pp. 89–118.

[51] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[52] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005.

[53] Y.-H. Yang and H. H. Chen, "Music emotion ranking," in *Proc. ICASSP*, 2009, pp. 1657–1660.

[54] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

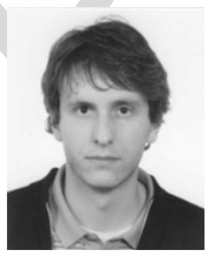
[55] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*. Berlin, Germany: Springer-Verlag, 2007, pp. 325–341.

[56] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[57] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*. London, U.K.: Edward Arnold, 1990.

[58] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross validation," in *Encyclopedia of Database Systems*. 2009. AQ:6

[59] *A Complete List of Adopted Movie Scenes for Affective Recommendation* [Online]. Available: http://www.ing.unibs.it/sbenini/misc/TCSVT-movie-scene-full-list.xlsx

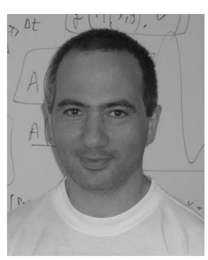
**Luca Canini** received the M.Sc. (*cum laude*) and Ph.D. degrees in telecommunications engineering from the University of Brescia, Brescia, Italy.

He is currently with the Department of Information Engineering, University of Brescia. During his Ph.D. studies, he was a Visiting Student with the IVE Laboratory, University of Teesside, Middlesbrough, U.K., and with the Digital Video/Multimedia Laboratory, Columbia University, New York. In 2012, he cofounded Yonder Labs, an independent company specializing in multimedia content analysis and complex data understanding. AQ:7

Dr. Canini was awarded by the Italian Marconi Foundation for his M.Sc. thesis.

**Sergio Benini** received the M.Sc. degree (*cum laude*) in electronic engineering from the University of Brescia, Brescia, Italy, and the Ph.D. degree in information engineering from the University of Brescia in 2006, specializing in video content analysis.

He is currently an Assistant Professor with the University of Brescia. From 2001 to 2003, he was with Siemens Mobile Communication Research and Development. During his Ph.D. studies, he was with the Content and Coding Laboratory, British Telecom, U.K., for one year. AQ:8



**Riccardo Leonardi** received the Diploma and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 1984 and 1987, respectively.

He has been with the University of Brescia, Brescia, Italy, since 1992, leading research and teaching in the field of telecommunications. He was a Post-Doctoral Fellow with the Information Research Laboratory, University of California, Santa Barbara, in 1987. He was a member of the Technical Staff of AT&T Bell Laboratories from 1988 to 1991. He joined the Swiss Federal Institute of Technology in 1991. His current research interests include digital signal processing, with a specific expertise in visual communications, and content-based media analysis. He has published more than 100 papers on these topics. AQ:9

# AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

AQ:1= Please provide description text for labels (a)–(d) in the caption of Fig. 2.
AQ:2= Please verify the text “since at low connotative distance from the query” for clarity.
AQ:3= Please provide page range in Refs. [9], [10], [16], [20], [22], [23], [33], [35], [38], [44], [46], [47].
AQ:4= Please provide author name and date in Refs. [26], [59].
AQ:5= Please provide publisher location in Ref. [45].
AQ:6= Please provide publisher name and location in Ref. [58].
AQ:7= Please provide location of “Yonder Labs.”
AQ:8= Please provide locations of “Siemens Mobile Communication Research and Development” and “British Telecom.”
AQ:9= Please provide location of “AT&T Bell Laboratories.”

END OF ALL QUERIES