A Connotative Space for Supporting Movie Affective Recommendation

Sergio Benini, Luca Canini, and Riccardo Leonardi

***Abstract*—The problem of relating media content to users' affective responses is here addressed. Previous work suggests that a direct mapping of audio-visual properties into emotion categories elicited by films is rather difficult, due to the high variability of individual reactions. To reduce the gap between the objective level of video features and the subjective sphere of emotions, we propose to shift the representation towards the connotative properties of movies, in a space inter-subjectively shared among users. Consequently, the connotative space allows to define, relate and compare affective descriptions of film videos on equal footing. An extensive test involving a significant number of users watching famous movie scenes, suggests that the connotative space can be related to affective categories of a single user. We apply this finding to reach high performance in meeting user's emotional preferences.**

***Index Terms*—Affective meaning, video analysis, connotation, movie recommendation, famous movie scenes**

I. Introduction

PEOPLE finishing work and returning home after a long day, often crave something that can improve or stimulate their emotional state. In these situations, even if the recent proliferation of social media communities and the wide access to digital media enable the possibility of generating automatic suggestions of relevant media experience, the effort of actively searching suitable affective content is still considerable.

Since we are addressing media affective analysis [1] and its use to support content recommendation, the fundamental matter is: to what extent can we trust emotional labels assigned by other individuals to the content they watched? Our opinion is: not much, since emotions are personal, and everyone reacts to events or, in this case, to media content in a way that depends on cultural, personal, and other, even short term, subjective factors. An extreme example is given by the different reactions people have when watching horror movies, which range from extreme fear to amused laugh, even depending on the audience composition (i.e. whether they are watching the movie alone or in group).

Therefore the first question we raise in this work is: can we find some other way of providing an affective content description which is more agreed among users? Or, is there any more effective way to recommend emotional content than relying on other people's affective responses? In our view the answer is: yes, there is, and it relies on "connotation".

Connotation is essential in cinematography, as in any other art discipline. It is given by the set of conventions (such as editing, music, *mise-en-scene* elements, color, sound, lighting, etc.) that influences how the meanings conveyed by the director are transmitted to persuade, convince, anger, inspire, or soothe the audience. While the affective response is on a totally subjective level, connotation is usually considered to be on an inter-subjective level, i.e. shared by the subjective states of more individuals. Using again the example of the horror movie, if we have two people reacting differently to the same film (e.g. one laughing and one crying), they would anyway share a similar view on what the movie is "suggesting", independently from their individual affective responses. For example they would likely both agree in saying that that horror movie atmosphere is grim, the music gripping, and so on, even if these produce different reactions in each of them.

Now, how would connotation be helpful for emotional recommendation? That is, is connotation linked to emotions? Our answer is: yes it is. If we know the emotional reactions of a single user, meaning that he/she has already emotionally tagged some items in the past, then representing media in the proposed connotative space is very helpful to target that single user's emotional desires.

A. Paper Aims and Organization

In this work we develop a space for affective description of movies through their connotative properties. It will be shown how this space can be the basis to establish when filmic scenes lead to consistent emotional responses in a single user.

We prove in fact that movie scenes sharing similar connotation are likely to elicit, in the same user, the same affective reactions, meaning that, when recommending affective items, using connotation properties can be more reliable than exploiting emotional annotations by other users. In our introductory example, the user frightened by the horror movie and wanting more of the same material, would probably be happier if we recommended to him/her other filmic items characterised by similar connotative properties rather than using other people's emotional annotations, since emotional reactions can be very different. As a possible use, these findings will be used to support movie recommendation.

As advantage with respect to the state of art, the proposed solution enables to provide affective descriptions of filmic products which are more objective than those provided by existing emotional models, therefore constituting an inter-subjective platform for analysis and comparison of different feature films on an equal footing. Second, on the basis of the proposed space, further research on computable affective understanding [2] may lead to the definition of more objective methods to assess human emotions elicited by films.

Authors are with the Department of Information Engineering, University of Brescia, via Branze 38, 25123 Brescia, Italy, e-mail: {firstname.lastname@ing.unibs.it}.

The paper is organised as follows. Section II explores filmic emotional theories and recent advances in affective video analysis. Section III provides the overall methodology. Section IV first reviews some concepts derived from psychology, such as *denotative* meaning, *connotation* and *affective response*. Then it explains how to measure connotative meanings, and describes ad-hoc semantic spaces for emotions, art and design objects. The description of the proposed connotative space for movies is then developed in Section V. To assess its validity in terms of users' inter-rater agreement, results of an extensive test on more than 200 users are presented in Section VI. Potentialities of this approach in emotion-based video recommendation are illustrated in Section VII, where we show how the connotative space better meets single users' emotional preferences than using other users' affective annotations. Concluding remarks are gathered in Section VIII.

II. Previous Work

A. Filmic Emotional Theories

Emerging theories of filmic emotions [3][4] give some insight into the elicitation mechanisms that could inform the mapping between video features and emotional models. Tan [3] suggests that emotion is triggered by the perception of "change", but mostly he emphasises the role of realism of the film environment in the elicitation of emotion.

Smith [4] instead attempts to relate emotions to the narrative structure of films. He describes filmic emotions as less character-oriented or goal-oriented, giving a greater prominence to style. He sees emotions as preparatory states to gather information and, more specifically, argues that moods generate expectations about particular emotional cues. According to this view, the emotional loop should be made of multiple mood-inducing cues, which in return makes the viewer more prone to interpret further cues according to his/her current mood. Smith's conclusion that "emotional associations provided by music, mise-en-scene elements, color, sound, and lighting are crucial to filmic emotions", should encourage attempts to relate video features to emotional responses.

Additionally, there are conflicting views on the extent to which emotional responses to films depend on the individual. Soleymani et al. [5] investigate physiological responses to films, exploring a wide range of physiological signals and correlations between users' self reports and the affective dimensions accessible through physiological measurements. Their study emphasises individual differences in affective responses with an in-depth analysis of the correlation between dimensional variables and video features for each subject.

Conversely, Smith [4] and Tan [3] agree on the extent to which emotional responses to traditional films depend on the individual, by confirming that a relatively uniform type of emotional responses are generated across a range of audiences, despite individual variations.

B. Previous Work on Affective Video Analysis

Even if intriguing possibilities could be offered by an emotion-based approach to currently investigated multimedia applications, related works in affective analysis of video content are few, sparse and recent. This limited interest is mainly caused by the apparent impossibility to define an objective method to assess emotions elicited by film videos, unless directly registering individual reactions by recording physiological responses to the video observation.

An alternative and practical way to assess the affective dimension of media is given by the use of the "expected mood", proposed by Hanjalic in [1], i.e. the set of emotions the film-maker intends to communicate when he/she produces the movie for a particular audience with a common cultural background, since this also appears consistent with Smith's conclusions as reported previously.

In a work co-authored with Xu [6], Hanjalic pioneers the analysis of affective video content, through an approach based on direct mapping of specific video features onto the Arousal and Pleasure dimensions of the Pleasure-Arousal-Dominance (PAD) emotional model [7]. They describe motion intensity, cut density and sound energy as arousal primitives, defining an analytic time-dependent function for aggregating these properties and using video frames for the time dimension.

Though the mapping of video properties on a model intended for describing emotions is inspired from previous literature, it has not been thoroughly validated by psychological questionnaires or physiological measurements, which would be proper methods to assess a time-dependent model. Furthermore, the examples of arousal mapping given in [1] refer to live sports events (football matches videos), whose properties may not transfer entirely to the case of other videos and feature films, which have different editing and whose soundtracks are of a different nature (since the latter do not necessarily include spontaneous audience reaction).

To date, emotional characterization has been mainly used to study a narrow set of situations, like specific sport events as in [8] or movies that belong to a particular genre, for example horror movies, as in [9].

Extending this approach, Xu et al. [10] describe emotional clustering of films for different genres, using averaged values of arousal and valence, deduced from video parameters. One inherent limitation of this clustering approach may be the use of a categorical description of target user emotions, with no clear indication that these would be elicited by the viewing of traditional film genres. Such proposed framework performs better for action and horror films than for drama or comedy, fact which authors attribute to the prominence of specific features in the first two genres. This could also be analysed as a more efficient detection of arousal-related features, which tend to characterise these two genres, over valence-related ones, as reflected by the selected video descriptors (e.g., brightness and colour energy as valence features).

De Kok [11] extends some aspects of this work by refining the modelling of colours, in an attempt to achieve a better mapping onto the valence dimension, while Kang [12] describes instead the recognition of high-level affective events from low-level features using HMM, a method also used by Sun and Yu [13]. Performance obtained by Kang in affective classification of movie scenes are outperformed in the work by Wang and Cheong [14]. They propose 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. In the same work [14], authors corroborate the view that audio cues are often more informative than visual ones with respect to affective content.

Bags of affective audio-visual words are recently proposed in [15] for affective scene classification; here authors also introduce an attempt for an intermediate representation, where emotions and events are linked by the use of "topics".

Recently affective descriptions of multimedia items also started to be applied to traditional recommender systems [16]. Tkalcic et al. in [17] propose the usage of metadata fields containing emotional parameters to increase the precision rate of content-based recommenders; by demonstrating that affective tags are more closely related to the user's experience than generic descriptors, they improve the quality of recommendation by using metadata related to the aesthetic emotions of users, and not the intrinsic emotions contained in items.

III. Overall Methodology

As mentioned in the introduction, a spatial representation is proposed to facilitate the derivation of the affective impact of movies thanks to the description of their connotative properties. To this end, Figure 1 explains the adopted workflow:

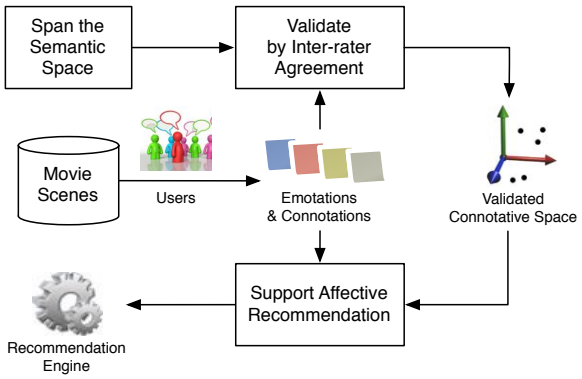


Fig. 1. Diagram describing the paper workflow for modelling the connotative space for movie scene analysis and recommendation.

a) ***Span the Semantic Space****:* Inspired by previous work in the industrial design domain [18], we propose to shape the affective identity of a movie thanks to a representation of its connotative properties according to the theory of "semantic differentials" [19].

b) ***Validate by Inter-rater Agreement****:* To validate the model, on the one hand we ask users to provide emotional annotations on the movie (called *emotations*), and on the other hand to rate some connotative properties; we then measure the level of inter-rater agreement for both (emotations vs connotations). The outcome is that connotative properties are more inter-subjectively shared among users than emotations.

c) ***Support Affective Recommendation****:* Once the proposed connotatitve space has been validated, we further show how to use it as a support for affective recommendation. In the specific, we prove that using connotation properties for recommending items to a user with a known emotional profile, is more reliable than directly exploiting emotional annotations gathered by the entire user community.

The main advantages of the proposed approach are described in the following. As shown in Section II, most of the previous work on affective video analysis tries to directly map low-level representations of content to a defined set of well-known human emotions often using affective models for emotions (such as the PAD model or the Russell's circumplex [20]). Our aim is to develop the first ad-hoc connotative space specifically built for affective movie description and not to directly compare the PAD model (nor Russell's) with the connotative space. In fact, while a point in the PAD describes *one emotion* in terms of pleasure, arousal and dominance, a point in the connotative space describes *one movie segment* in terms of its connotative properties derived from cinematography.

Establishing a direct correspondence between physical representations of video signals and high level users' emotional responses often produces results which are somewhat inaccurate and difficult to validate [14]. This is likely due to the undeniable distance between the measurable properties of filmed objects and the inscrutable and personal nature of emotions. The proposed connotative space is able instead to fill the need for an intermediate semantic level of representation between low-level features and human emotions.

By grounding in the related fields of cinematography and psychology, this solution is helpful in closing the distance between users' reactions and video features from both directions. By adopting semantic differentials on connotation properties, on the one hand it enables the representation of the affective properties of a movie in a more inter-subjective way than directly using emotations. On the other hand it envisages an easier translation process of video low-level properties into intermediate semantic concepts mostly agreeable among individuals, thus avoiding the "bridging at once" process.

IV. Measuring the Affective Meaning

Besides a *denotative* meaning, every concept has an *affective* meaning, or connotation. Denotation, also known as cognitive meaning, refers to the direct relationship between a term and the object, idea, or action it designates. Connotation, also known as affective meaning, refers to the emotive or associational aspect of a term.

For example concepts such as *summertime* and *love* arouse unique assemblages of positive emotional connotations. *Homeless* and *cancer* summon clouds of negative emotional connotations. Other concepts, such as *boxing*, call up mixed positive and negative connotations. Again, a *stubborn* person may be described as being either *strong-willed* or *pig-headed*. Although these have the same literal meaning (i.e. stubborn), *strong-willed* connotes admiration, while *pig-headed* connotes frustration in dealing with someone.

In literature, no author can write with color, force, and persuasiveness without control over connotation of terms [21]. In the same way using the emotional appeal of connotation is essential for any concept, event, or object in disciplines such as design, art, and most interestingly for us, cinematography.

A film is made up of various elements, both denotative (e.g. the purely narrative part) and connotative (such as editing, music, mise-en-scene, color, sound, lighting). A set of conventions, known as film grammar [22], governs the relationships

between these elements and influences how the meanings conveyed by the director are inferred by the audience.

In this sense the intuition by Hanjalic in formulating the "expected mood" concept [1] is brilliant: the affective meaning of a video does not necessarily correspond to the affective response of a particular individual. The affective meaning results from the actions of a movie director, who for example adopts all conventional techniques for giving connotation to his/her latest horror movie. Opposed to this, the actual affective response by individuals is very subjective and context dependent, so that it can be very different.

Therefore, there are at least three possible levels of description for a given object, a video in our case: the **denotative** meaning (what is the described concept), the **connotative** one (by which terms the concept is described) and the **affective response** (how the concept is perceived by a person). Whereas the denotative meaning (resp. the affective response) is on a totally objective (resp. subjective) level, the connotative meaning is usually considered to be on a inter-subjective level, i.e. shared by the subjective states of more individuals.

A. Semantic differential

In the 1950s, Osgood has been credited with the breakthrough of being able to measure the connotative meaning of any concept [19]. By constructing bipolar scales based on semantic opposites (the "semantic differential" scales) such as "good-bad", "soft-hard", "fast-slow", "clean-dirty", "valuable-worthless", "fair-unfair", "warm-cold", and so on, he was able to differentiate attitudinal intensity of persons towards the connotative meanings of words.

The outcome was Osgood's discovery of "semantic space" - the existence of three measurable underlying attitudinal dimensions that everyone uses to evaluate everything in his/her social environment, regardless of language or culture. In the semantic space every concept has an affective meaning that varies along three dimensions: *Evaluation* - goodness versus badness, *Potency* - powerfulness versus powerlessness, and *Activity* - liveliness versus torpidity (EPA) [19].

This structure makes intuitive sense. When our ancestors encountered a person, the initial perception had to be whether that person represents a danger. Is the person good or bad? Next, is the person strong or weak? Our reactions to a person markedly differ if perceived as good and strong, good and weak, bad and weak, or bad and strong.

Subsequent experimentation by many investigators confirmed the validity of Osgood's semantic space and its cross-cultural identity, making the Evaluation, Potency, and Activity (EPA) structure one of the best documented facts in social science (a full bibliography of research in this area is provided by Heise in [23])

B. Semantic spaces for emotions

After Osgood's studies, affect control theory has been used in research on different concepts: emotions, genders, social structure, politics, deviance and law, business, design and art. Contextually, the original EPA space has been declined into spaces which are peculiar to several specific application fields.

Concerning emotions for example, two studies by Russell and Mehrabian [24] provided initial evidence that three independent and bipolar dimensions, Pleasure-displeasure, degree of Arousal, and Dominance-submissiveness (PAD), are both necessary and sufficient to adequately define emotional states. Some years later, evidences that these affective dimensions are inter-related in a highly systematic fashion led Russell to develop his circumplex model of affect [20] (see Figure 2-a), and, more recently, to the Geneva emotion wheel [25] and Plutchik's colour wheel [26] (in Figure 2-b).

This model, made of eight basic emotion categories - joy, sadness, anger, fear, surprise, disgust, anticipation, and acceptance - is a quantized version of the Russell's circumplex, but it better allows us to clearly perceive the "closeness" between arbitrary pairs of emotion categories. Later to our days, Fontaine et al. criticise the fact that many researchers focus exclusively on two-dimensional models mostly involving valence and arousal [27]. Therefore, adopting a theoretically based approach, they show that at least four dimensions are needed to satisfactorily represent similarities and differences in the meaning of emotion words.

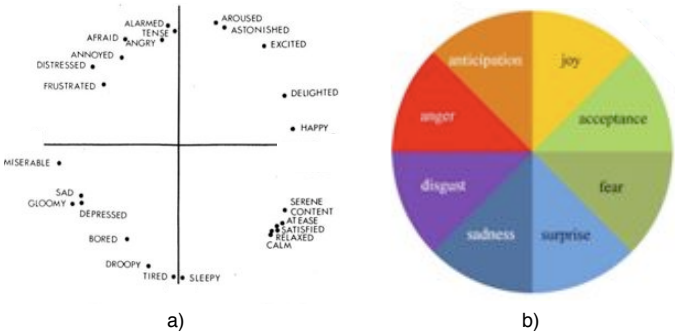


Fig. 2. a) Russell's circumplex model of affection [20] and b) Plutchik's emotion wheel based on eight basic emotions [26], which allows to clearly perceive the "closeness" between arbitrary pairs of emotion categories.

All these emotional models allow for representing emotions as positions in a semantic space. In the years, using such a dimensional approach stimulated many researchers in multimedia, from Hanjalic onwards, to propose primitive low-level audiovisual features as suitable axes for the PAD. Since then, video properties have been commonly (but not properly) mapped in such spaces originally built to represent emotions.

Though very popular, these approaches need now to be advanced by investigating novel semantic spaces proper for movie affective description, instead of unsuitably changing dimensions to spaces originally thought for emotions.

V. Span the semantic space

Osgood asserts that every concept, object, artefact can be positioned in a three dimensional vector space defined by semantic expressions. Spanning the semantic space on a new concept implies first to collect a large number of words describing the domain - typically adjectives - from different sources: pertinent literature, experts, etc. A large number of subjects are then asked to express their judgements on words

and their relation to the concept, by rating on bipolar measurement scales, typically ranging from 1 to 5. After that, collected data are processed using, for example, factor analysis and PCA for dimensionality reduction to discover how different words are related to each other, and in which way they affect concept understanding. Finally, relevant words representing the different dimensions are selected in order to link them to concept properties. Fontaine et al. for example adopt this theoretically based approach in [27] to span the emotion semantic space and represent similarities and differences in the meaning of emotion words.

A. Semantic spaces for products and design objects

When applied to engineering or technical sciences, a correct example of quest for the affective meaning of items is provided by *Kansei* engineering [28], which investigates the links between *feelings* and *properties* of industrial products to understand how the choice of product attributes may affect the customers' emotional perception of the whole product.

With a similar approach, Castelli [18] spans the semantic space by focusing his attention to the sensorial experience in relation to design objects. Through the proposed "qualistic" approach, he identifies the qualities of a product according to inter-subjective evaluations, apart from parameters that can be measured under qualitative and quantitative terms. The proposed qualistic diagram (see Figure 3), which accounts for three axes - *natural, temporal* and *energetic* - is able to support the many variations of a design product allowing to define, develop and manage prototypes that can be compared over time and discussed on equal footing in a wide range of sectors (research and development, marketing, communication).

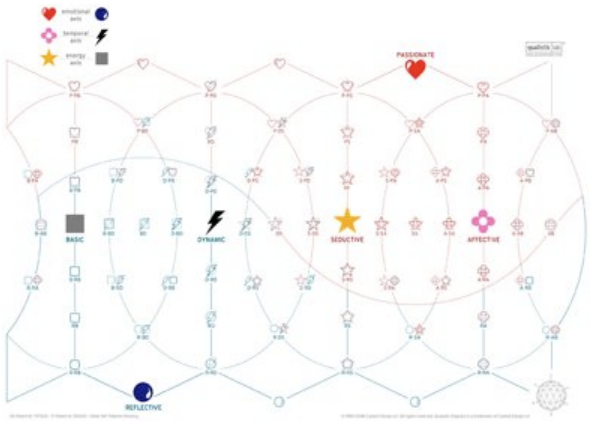


Fig. 3. Qualistic diagram for affective description of design objects [18].

B. Semantic space for movie connotation

In order to shape a connotative space for movies, we do not set our focus on performing the psychologists' work and span a large set of expressions to discover the most relevant word for each dimension, but we extend the notion of piece of art or design to movies as suggested by Castelli [18]. As it happens for literature, art and design, even in cinematography it is essential to keep control over connotation: undoubtedly, in the last century the process of filmmaking has evolved into an art form, including concepts ranging from marketing aspects to social communications. For this reason, we transpose the movie affective identity into cinematographic terms.

In the semantic *connotative space* we propose, the affective meaning of a movie varies along three axes which account for the *natural, temporal* and *energetic* dimension, respectively. As in [18], the natural dimension splits the space into a passional hemi-space, referred to warm affections, and a reflective hemi-space, that represents offish and cold feelings. The temporal axis characterises the space into two other hemi-spaces, one related to slow dynamics and another describing an intrinsic attitude towards high pace and activity. Finally, the energetic axis identifies films with high impact in terms of affection and, conversely, minimal ones.

Following Osgood's evidences, we construct bipolar scales based on semantic opposites and associate to each axis a couple of adjectives in a dichotomic relationship. To the natural axis we link the couple *warm/cold*. The temporal axis is described in terms of *dynamic/slow*, while the dichotomy *energetic/minimal* is associated to the third axis. These choices allow for representing a movie (or a movie segment) in the related connotative space either as a cloud of points or, considering the time component, as a trajectory that describes the evolution of its affective identity.

An example of a movie scene evolution in the connotative space is shown in Figure 4. The trajectory gives an accurate affective characterisation of the movie, being not restricted to a fixed set of emotions or by the previously discussed limitation of emotional models such as the pleasure-arousal scheme.

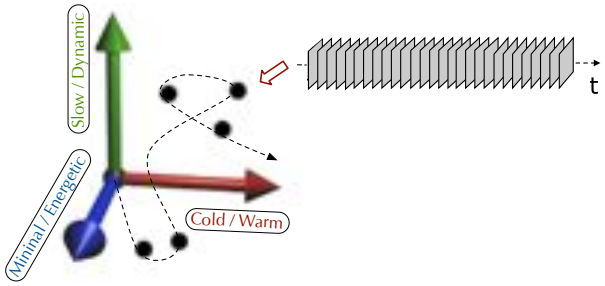


Fig. 4. Connotative space for measuring the affective meaning of movie scenes and a connotative trajectory extracted from a video.

Having decided to borrow the theoretical approach from art and design, we are aware that the proposed space is only one among the possible ones, and probably not the best possible; however, even on the basis of previous studies and experience in video analysis, the chosen dimensions are commonsensical and inherently linked to film properties connected to shooting and/or editing. This choice leaves space for further improvement, in case teams of expert psychologists want to span the semantic space of the related domain with a more rigourous theoretical approach (as those in [20] or [27]), thus optimising the choice of the expressions to associate to each dimension. Nevertheless in the following experimental phase we show how the proposed space, even if directly transposed from the art and design domain, is effective in increasing the inter-rater agreement and in supporting movie affective analysis.

VI. Validate by Inter-rater Agreement

The connotative space aims at being a common ground for movie analysis which is more objectively shared among users than their subjective affective responses. To validate this hypothesis on the model, we ask users to watch some movie scenes and provide annotations on four different semantic concepts: their emotation on the watched scene, and three connotative aspects they perceived while watching the scene. For each movie segment, we measure the level of inter-rater agreement on all four judged concepts, and demonstrate that the agreement among users is higher on provided connotative annotations than on expressed emotations. Using the before mentioned example of the users watching the horror movie (one laughing and one crying), we aim at confirming that they show higher agreement on judging the connotative properties of the movie, than on the experienced affective reactions.

A. Experiment set-up

The experiment is set up as follows. A total number of 240 users are recruited: 195 are students at the University of Brescia, while remaining 45 are chosen among colleagues, family members and friends. Out of these, a total of 140 users fully completed the experiment on all videos, while others performed it only on a subset. The experiment is in the form of a user test and it is performed online, with both support of English and Italian languages.

Data consist of 25 "great movie scenes" [30] representing popular films spanning from 1958 to 2009 chosen from IMDb [31]. They are listed in Table I with their duration, while in Figure 5, a representative key-frame for each scene is shown, so that the reader can recognise at least a few of them.

The choice of "scene" as elementary unit for analysis is supported by the fact that each scene in a movie depicts a self contained high-level concept [32]. Selecting other typical video segments, such as a shot (i.e. an uninterrupted run of a camera take [33]), would not be adequate, since its short average length (a few seconds) does not in general allow to convey a concept and/or induce a well defined emotional state in the user. Conversely, scenes extracted as explained in [34], constitute narrative units mostly autonomous in their meaning [14] even when excerpted from the original context.

From a video property perspective, following the common rule of film production which imposes that a persistent semantics is associated to a long term continuity at least in chromatic composition, lighting and ambient sound [35], most scenes are highly homogenous in terms of low-level features [36] (except for a few of them which are discussed later as counterexamples).

Two criteria guided us in the scene selection. First "great movie" scenes are chosen since we expect that they more easily elicit emotional reactions in the observer, since "they are our memories of segments of films that have achieved a life of their own, compelling us to remember and relive the moment" [30]. Following the definitions in [30] we try to cover all key-ingredients of great movie moments: we have "a striking, cinematically-beautiful image" (2001: A Space Odyssey), "a spectacular action with large crowd sequence" (Ben Hur), "a surprising revelation, or unexpected shock" (Se7en), etc.

TABLE I
Set of scenes excerpted from famous feature films.

Scene	Movie title	Dur
1	*(500) days of summer (2009)*	01:33
2	*2001: A Space Odyssey (1968)*	01:37
3	*Ben Hur (1959)*	01:41
4	*Blade runner (1982)*	01:03
5	*Full Metal Jacket (1987)*	01:41
6	*Ghost (1990)*	01:54
7	*Le Fabuleux Destin d'Amelie Poulain (2001)*	01:35
8	*Le Fabuleux Destin d'Amelie Poulain (2001)*	01:01
9	*Life is Beautiful (1997)*	01:50
10	*Notting Hill (1999)*	00:45
11	*The Matrix (1999)*	01:39
12	*The usual suspects (1995)*	01:17
13	*Se7en (1995)*	02:59
14	*The good the bad and the ugly (1966)*	02:17
15	*Saving private Ryan (1998)*	01:24
16	*Scent of a woman (1992)*	02:22
17	*Vertigo (1958)*	01:30
18	*No country for old men (2007)*	01:41
19	*Dolls (2002)*	00:31
20	*Dolls (2002)*	01:17
21	*Dancer in the dark (2000)*	00:54
22	*Dancer in the dark (2000)*	01:27
23	*The English Patient (1996)*	02:42
24	*Once Upon a Time In The West (1968)*	02:31
25	*The blue lagoon (1980)*	01:40

Second, observe that selected scenes are chosen so as to expectedly cover all categories of elicited basic emotions, while there is no need to cover all content variability of thousands of existing movies. In this sense, selected scenes offer a sufficiently broad spectrum to characterise the limited variability of affective reactions of the audience to movies. Since we are not addressing a statistical study on the variety in emotional video content, but on the users' variability in assigning emotional labels, most important is the cardinality of the users' set. Therefore the number of recruited users is among the largest involved in a test on multimedia affective analysis over the last years.

Fig. 5. Representative key-frames from the movie scene database.

B. Rating concepts on interval scales

To perform the test, every user is asked to watch and listen to $\eta = 10$ randomly extracted movie scenes out of the total $M = 25$, in order to complete the test within 30 minutes. Scenes can be watched as many times as users want, either in English or Italian. After the viewing, users are requested whether they have seen the scene/movie before, and in case they did, they express their guess on the movie title. The whole test can also be interrupted and resumed in different moments. After watching a scene, each user is asked to express his/her annotation on four different concepts.

First the user is asked to annotate the emotional state he/she is inspired with on the *emotation wheel* in Figure 6-a. This model is a quantized version of the Russell's circumplex (Figure 2-a) and presents, as in the Plutchik's wheel, eight basic emotions as four pairs of semantic opposites: "Happiness *(Ha)* vs. Sadness *(Sa)*", "Excitement *(Ex)* vs. Boredom *(Bo)*", "Tension *(Te)* vs. Sleepiness *(Sl)*", "Distress *(Di)* vs. Relaxation *(Re)*". Such a circular model allows us to clearly perceive the "closeness" between arbitrary pairs of emotion categories: relatively close emotions are adjacent to each other so that it is easier to transit to neighbouring emotions than more distant emotions [15]. For self-assessment, the emotation wheel is preferred to other models, such as PAD, since it is simpler for the users to provide a unique emotional label than to express their emotional state by a combination of values of pleasure, arousal and dominance.

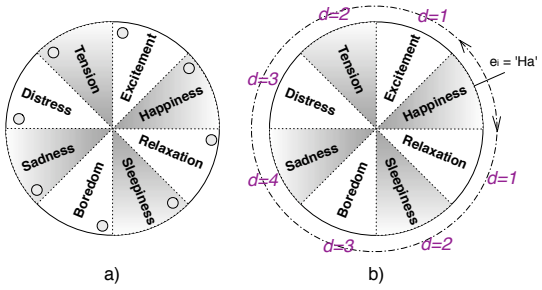


Fig. 6. a) The emotation wheel used by users to annotate emotions; b) on the emotation wheel relatively close emotions are adjacent to each other.

To assess connotation of the movie scenes, users are then asked to rate on a Likert scale from 1 to 5 three concepts accounting for the *natural, temporal* and *energetic* dimensions of the connotative space. Ratings are expressed by using the radio button indicators in Figure 7 on three bipolar scales based on the semantic opposites: *warm/cold* (natural), *dynamic/slow* (temporal), and *energetic/minimal* (energetic), respectively. In particular users are asked to rate:

- the atmosphere of the scene from *cold* to *warm*
- the pace of the scene from *slow* to *dynamic*
- the scene impact on them from *minimal* to *energetic*

since we expect users to be familiar with intuitive concepts such as atmosphere or rhythm, and to be able to evaluate the power of the viewed scene.

Likert scales on 5 levels, commonly used in survey research, belong to the category of "interval" scales [37]: they clearly implies a symmetry of response levels about a middle category and their visual presentation clearly suggests equal spacing among levels and the continuity of the underlying concept.

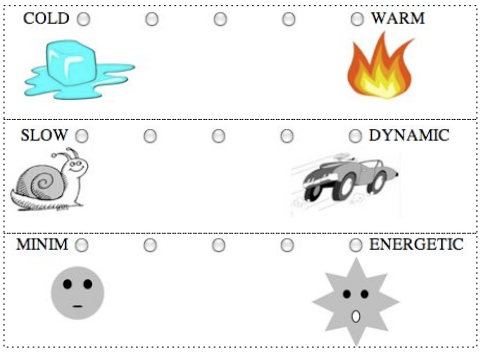


Fig. 7. Bipolar scales based on semantic opposites for rating the connotative properties of movies.

To measure the inter-rater agreement on the four concepts we first need to enable comparison between annotations. For this purpose emotations are converted into 1-to-5 bipolar scales. Observing the emotation wheel in Figure 6-b, we can define the distance d between two emotions e_i and e_j as done by Russell in [20] and recently by Irie et al. in [15], i.e. as the number of steps required to reach emotion e_j from emotion e_i. 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 anticlockwise (see Figure 6-b: if $e_i = Ha$, $e_j = Sa$ then $d(e_i, e_j) = 4$ in both senses). Exploiting distances between emotions, for each scene we then turn emotations into a 1-to-5 bipolar scale by unfolding the wheel only on the five most voted contiguous emotions, as shown in Figure 8.

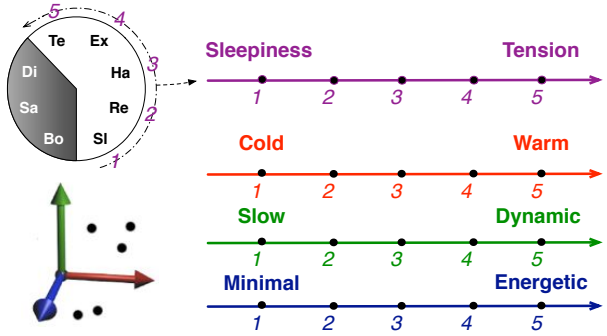


Fig. 8. The most voted 5 contiguous emotions of each scene (in the white sector of the model) are turned into a five-level bipolar scale, thus comparable with the three scales related to the connotative properties of the scene.

The definition of a quantitative distance between elements of the proposed scales is also supported by the theory of the scale of measurements on interval scales [37]. Moreover, the choice of discarding, separately for each scene, the three least voted contiguous emotions, is supported by Osgood, who states that "five possible rates, or five bipolar adjectives, are proven to yield reliable findings" [19]. Also note that the large majority of votes (beyond 95% in average) gathers in few "close" emotions, so that the number of non-counted votes for each

scene is statistically not significant. Finally, even if removed contiguous emotions seem to create a "hole" in the wheel, no discontinuities are actually introduced due to the circular nature of the space: on the circle distances remain linear, and the obtained scale is comparable with other connotative ones.

Inter-rater agreements on each scene can now be computed separately on the four separate concepts, each represented on a 1-to-5 Likert scale, as shown in Figure 8. For each scene i, we have four histograms collecting rates from 1 to 5 on semantic bipolar scales: one histogram H_i^W collects the number of votes of the most rated five contiguous emotations for that scene, while the other three histograms H_i^N, H_i^T, and H_i^E collect the expressed rates for the three connotative axes (natural, temporal, and energetic, respectively).

Examples of normalized histograms are given in Figure 9 for a scene taken from "Ghost" (no. 6) (in purple emotations, in red the natural axis, in green the temporal one, and in blue the energetic one). Observe on the top right corner of the emotation histogram the percentage of votes collected by the five consecutive bins with the largest probability mass, confirming that despite the variability in affective reactions, collected votes tend to gather in few "close" emotions.

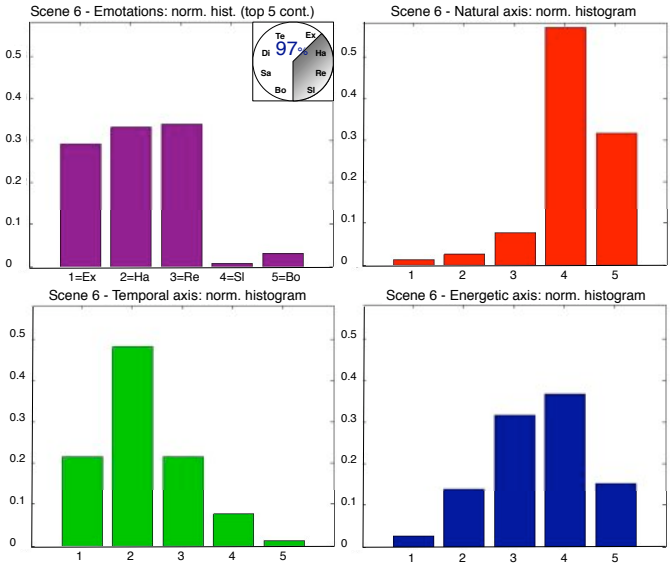


Fig. 9. Top left: normalized histogram H_i^W for the scene taken from"Ghost" of the 5 top rated contiguous emotations, collecting 97% of votes (purple). The other three normalized histograms H_i^N (red), H_i^T (green), and H_i^E (blue) collect the rates for the three semantic differential scales for the same scene.

C. Measuring inter-rater agreement on interval scales

To assess wheter the connotative space is inter-subjectively shared among users, we compare the levels of inter-rater agreement on rates expressed on the three connotative axes with the level of user consensus on the expressed emotations.

There are several techniques for measuring inter-rater agreement, whose suitability depends on the definition of the problem and on the adopted measurement scale. As adopted scales are of the "interval" type [37], one of the most common methods to assess the rater agreement is the measure of the intra-class correlation coefficient (*ICC*) [38]. The intra-class correlation coefficient statistically estimates the degree of consensus among users. It is usually defined in the interval [0,1] where the higher the value the bigger the homogeneity among raters. Since each scene is rated by a different set of users randomly selected from a larger set and we want to measure the overall agreement, the $ICC(1,k)$ perfectly suits our experiment scenario:

$$ICC(1,k) = \frac{BMS - WMS}{BMS} \tag{1}$$

where BMS is the between-scene mean square, and WMS is the within-scene mean square. For a description of the statistical model and the above formula please refer to [38], while details on the computation of BMS and WMS in our specific case are given in Appendix. Values of inter-rater agreement expressed for the three connotative axes and for emotations are given in Table II. Following the recommendation expressed in [38], we assess the statistical significance of our measurements by rejecting the null hypothesis of non-agreement within a confidence interval of 95% (risk level 0.05).

TABLE II
MEASURES OF INTER-RATER AGREEMENT $ICC(1,k)$.

Inter-rater agreement	Emotation	Natural	Temporal	Energetic
$ICC(1,k)$	.7240	**.8773**	**.8987**	**.7503**

The comparison between intra-class correlation coefficients clearly shows that the overall agreement is consistently higher when users are asked to rate connotative concepts of the movies rather than when they have to provide emotional annotations. In particular, the gap with respect to the consensus level on emotations is larger when users are requested to express their impressions on the scene atmosphere and dynamics, while it reduces when users rate the energetic axis, i.e. the energetic impact of the scene. This is not surprising, since the energetic axis is the declination of the third axis of the original EPA space, dimension which, as already observed by Greenwald et al. in [39] and by Hanjalic [6] in the multimedia domain, often plays a limited role in characterizing emotional states with respect to the other two axes. Nevertheless, the consensus even on this third concept remains larger than the agreement on emotional annotations expressed on scenes.

Along with the computation of the ICC expressed on the aggregated scene set, we also perform a scene by scene analysis, by relying on standard deviation of votes (Stevens in [37] states that, beyond *distance*, Likert scales also support concepts such as *mean*, *mode*, and *standard deviation*).

To begin with, we show in Table III the mode of the emotation histogram H_i^W, that is the most rated emotional annotation expressed by users. It gives an indication on the collectively perceived emotion elicited by the scene, stating for example that the "Ben Hur" scene (no. 3) is mostly perceived as tense, while "The blue lagoon" (no. 25) as happy.

Scene-based indicators on inter-rater agreement are instead provided for the single scene i, by the four standard deviations:

- σ_i^W on the emotation histogram H_i^W;
- σ_i^N on the natural histogram H_i^N;
- σ_i^T on the temporal histogram H_i^T;
- σ_i^E on the energetic histogram H_i^E.

TABLE III
MOST RATED EMOTION FOR EACH MOVIE SCENE.

1	2	3	4	5	6	7	8	9	10
Ha	Te	Te	Sa	Te	Re	Ex	Ha	Sa	Ha
11	12	13	14	15	16	17	18	19	20
Ex	Te	Te	Te	Te	Ha	Te	Te	Re	Sa
21	22	23	24	25					
Sa	Sa	Sa	Bo	Ha					

Standard deviation σ_i^W of histogram H_i^W measures the spread of the emotation distribution around the mean value (remind that emotions are now mapped to a bipolar scale ranging from 1 to 5, as in Figure 8). This value roughly assesses the grade of agreement among users on the emotions elicited by the single scene. Analogously, standard deviations σ_i^N, σ_i^T, and σ_i^E measure the spreads of the distributions of rates on connotative properties around the mean vote assigned to the natural, temporal, and energetic axes, respectively.

In Figure 10 we observe, for each scene, the comparison between standard deviations σ_i^N (red), σ_i^T (green), σ_i^E (blue) and σ_i^W (purple). In general, the standard deviation measured on the emotation histogram H_i^W is evidently larger than the standard deviations computed on rates assigned to the three semantic differential scales of the connotative space, symptom of a lower user consensus on the rated concept.

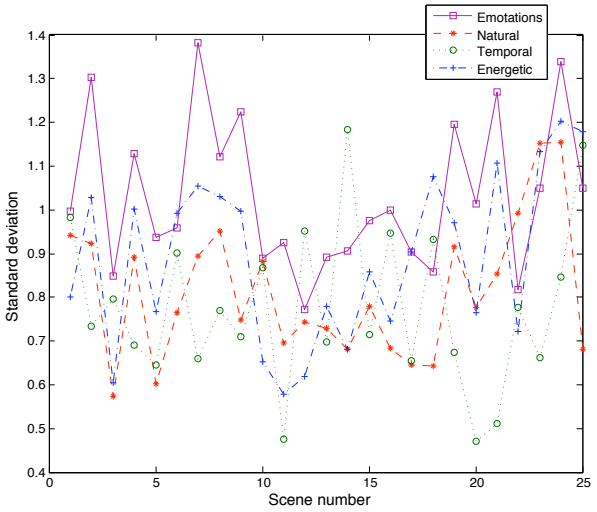


Fig. 10. Comparison between standard deviations (σ_i^N- red, σ_i^T- green, and σ_i^E- blue) obtained by rating scenes on the three axes of the connotative space and the standard deviation of the emotation histogram (σ_i^W- purple).

D. Discussion on scenes

The outcome of the ICC analysis, and the values of standard deviations on single scenes, suggest a higher agreement among persons in judging the connotative properties of a movie scene with respect to sharing similar emotional states in response to the scene. Back to the example of the horror movie, this higher level of inter-rater agreement means that, even if the two viewers might experience different emotional states while watching it, they would likely both agree in judging the movie atmosphere, the scene pace and its emotional impact.

Since the agreement is higher on all three dimensions of the connotative space, this suggests a higher discriminative ability of this space for further analysis and comparison of movies. The broader variation in users' emotations is further endorsed by the fact that the three least rated emotions are discarded from the histogram H_i^W and consequently not considered in the computation of standard deviation σ_i^W, further reducing the actual dispersion around the mean rated value.

By analysing the behaviour of standard deviations in Figure 10, the few scenes which do not follow the general trend are here further discussed as counterexamples.

Scene no. 23 ("The English Patient", Hana reading a book to Laszlo) is very sad, prevailing cold and slow, but interleaved with flashbacks with memories from a happy past and shots with warm atmosphere. Emotionally speaking, the gone happiness of the past reinforces the sadness of the present. However, from the point of view of connotation, it is difficult for humans to univocally tag the scene on the dichotomies *warm/cold* and *energetic/minimal* due to the presence of flashbacks.

Scene no. 14 depicting the final duel from "The good the bad and the ugly" is also worth discussion: the natural and the energetic dimensions are pretty univocally rated, while this is not true for the temporal one. In fact, in this scene the three main characters stare at each other down in the circular center of the cemetery, calculating alliances and dangers in a cinematically famous impasse. Rhythm is therefore very slow at the beginning, but the shotcut rate dramatically increases until the sudden gun drawing which ends the scene. In conclusion, while the atmosphere and energy are well recognisable along the scene duration, its increasing rhythm results in a high standard deviation in rating the temporal dimension.

Finally, scene no. 18 excerpted from "No country for old men" is not univocally interpreted in terms of the impact of the scene (*minimal/energetic*). In this case we notice that the user rate is heavily influenced whether he/she has already seen the movie. In the scene, an elderly rural gas station clerk unawarely saves his life as he calls on a coin flip. In this case, the affective impact of the scene mostly remains at the cognitive level (i.e. the knowledge of the plot) while connotation provided by cinematographic editing is limited. Viewers not knowing the plot and what is actually behind the coin toss, experience difficulties in rating the scene impact, thus assessing for this scene the dominance of the cognitive level over connotation.

From the discussion on previous counterexamples, we conclude that a large variation in the rates assigned to a scene on a specific dimension often reflects a lack of persistency of some low-level features which are likely linked to that dimension. This reinforces our conclusion about the need of developing this intermediate representation and its capabilities to map low-level video properties to user's affective responses.

Based on the analysis on the gathered data, we have strong indications that such a framework constitutes a better inter-subjective common ground for movie analysis and advanced applications than existing emotional models, as we further investigate in the next test.

VII. Support Affective Recommendation

As a possible usage scenario, we investigate hereafter whether connotation can be used for supporting emotional recommendation. In other terms, to meet the emotional wishes of a single user when recommending a movie, is it better to rely on the movie connotative properties (i.e. the connotative space), or to exploit emotations provided by other users (i.e. the emotation wheel)?

In a bit more formal manner, imagine that we know the profile of user u_k, that is the emotations that u_k provided in the past for a small set of movies. Imagine also that u_k wants a precise emotion-evoking movie, for example a relaxing one, and that he/she already emotated at least one movie m_i with the emo-tag *"Re"*. Then, if m_h is the most similar movie to m_i according to the connotative space, while emotations by other users return m_l as the item most similar to m_i in terms of relaxation, will user u_k be happier if we recommend m_h or m_l? From what was shown so far, it is most likely that m_h better fulfils the user's wish.

We will show indeed that the connotative space relates media to single user's affective response better than using emotional annotation by other users, implying that movie scenes sharing similar connotation are likely to elicit, in the same user, the same affective reactions. After showing that rating a movie scene in the connotative space is better agreed among users, if we demonstrate that the connotative properties of the movie are strongly related to a single user's affective response, we reasonably expect this space to be strongly linked with human emotions, thus helping in reducing the semantic gap between video features and the affective sphere of individuals.

A. Top-k lists: a support to recommendation

The envisaged application scenario uses the notion of "top-k list", nowadays ubiquitous in the field of information retrieval, e.g. the list of k items in the "first page" of results by a search or recommendation engine. The idea is that the system returns, on the basis of the user request, a top list of items ranked in the connotative space, which are relevant to the user.

This application scenario cannot be described as a pure recommendation functionality, since to produce the ordered list it employs (at least) one annotated item retrieved from the user profile, as in a *query-by-example* search, such as in [40]. Nevertheless the ranked top lists returned by the system can be used as a valid mechanism for propagating emotional tags of the user profile to "close" items in the database, thus enabling better *filtering of relevant items* from the *non-relevant* as in [17], or to be used as a valid support for integration into traditional recommending methods, both content-based [41] and collaborative filtering ones [16]. Ranking has also the advantage that, since it is based on similarities between items, it is closer to the human mechanism of perceiving emotions which works in a comparative way rather than using an absolute labelling, as shown in [42] for music items.

While in a real application the proposed list would include scenes the user has not seen yet, testing is performed using only those scenes present in the user's profile, as depicted in Figure 11. Once the user expresses an emotional wish, using as a reference those scenes he/she has already emotated as relevant to that emotional wish, we produce two lists of suggested scenes, one in the connotative space and the other based on emotations by all users. Both lists are ordered according to a minimum distance criterium in the corresponding space. To understand which space ranks scenes in a better order, we compare the two lists with a third one, considered as the best target, which is ranked based on the emotations stored in the user's profile.

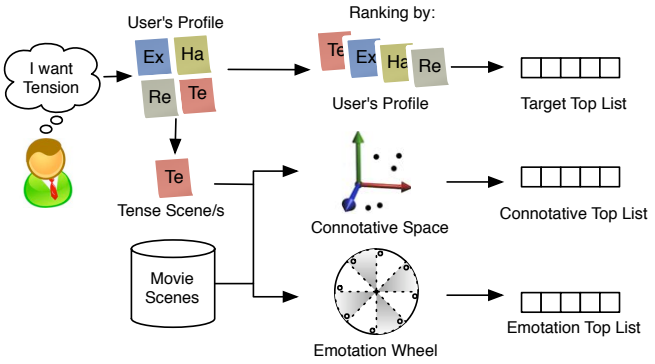


Fig. 11. Given one emotional wish, movie scenes of the user profile are differently ranked in the connotative and emotation space. The two lists are compared with the best target provided by the user's profile.

B. Distances in spaces

Let $\{m_1, m_2, \ldots m_M\}$ be the set of M movie scenes, and $\{u_1, u_2, \ldots, u_U\}$ the set of U users.

Each movie scene m_i is represented in the connotative space by its signature $m_i^C = \{H_i^N, H_i^T, H_i^E\}$, which captures the distributions of votes on the three axes, and in the emotation space by its signature $m_i^W = \{H_i^W\}$, which represents the distribution of affective reactions of all users to the scene. Since each single user u_k voted a number of $\eta < M$ movie scenes on the emotation wheel, on this scene subset we are also able to build a specific user profile, since we gathered an explicit knowledge on the user's personal reactions to them.

In both spaces, distance matrices between movie scenes can be computed by the Earth Mover's Distance ($\mathcal{EMD}$) [43] between signatures (again supported by [37]), adopting circular distances between emotations and distances between connotative values as ground distances. Therefore in the connotative space, distance matrices of $M \times M$ dimension are computed for each axis:

$$\begin{aligned} \Delta^N &= \delta^N(m_i, m_j) = \mathcal{EMD}\left(H_i^N, H_j^N\right) \\ \Delta^T &= \delta^T(m_i, m_j) = \mathcal{EMD}\left(H_i^T, H_j^T\right) \\ \Delta^E &= \delta^E(m_i, m_j) = \mathcal{EMD}\left(H_i^E, H_j^E\right) \end{aligned} \tag{2}$$

so that the distance matrix Δ^C, accounting for all three dimensions, is a function of Δ^N, Δ^T and Δ^E:

$$\Delta^C = \delta^C(m_i, m_j) = f\left(\Delta^N, \Delta^T, \Delta^E\right) \tag{3}$$

In the emotation space instead, the distance matrix Δ^W is:

$$\Delta^W = \delta^W(m_i, m_j) = \mathcal{EMD}\left(H_i^W, H_j^W\right) \tag{4}$$

On the subset of $\eta < M$ movie scenes voted by a single user u_k, we build the profile of u_k, which is a squared matrix of $\eta \times \eta$ dimension $D^W_{u_k}$ containing the emotional distances between movie scenes expressed by u_k:

$$D^W_{u_k} = d^W_{u_k}(m_i, m_j) \tag{5}$$

where $d^W_{u_k}(m_i, m_j)$ is the circular distance between two emotions on the emotation wheel, that is the number of steps on the emotation wheel between the emotions expressed by u_k for movie scenes m_i and m_j, respectively. Please note that d is not a distance between distributions of votes (as δ is indeed in Equations 2, 3 and 4), but is a distance between emotations given by the same user on different scenes.

C. Ranking lists

Test comparison is not performed on *top-k* lists, as in the envisaged application scenario, but on *full* lists, that is on permutations of all the items in a fixed universe [44]. In fact, since user u_k voted $\eta < M$ movie scenes both on the emotation wheel and in the connotative space, on this restricted scene subset we are able to produce three full lists of scenes for further comparison: the first based on the connotative space ("what the connotative properties suggests"), the second on the emotation space ("what all other users' emotions suggest"), and the last one based on the user's personal profile ("what are the real affective responses of u_k").

Supposing for example that user u_k wants some content eliciting the same emotion he/she already emotated for m_i, the ranked list $v^C_{k,i}$ in the connotative space is computed on the basis of Δ^C (function f in Equation 3 is set so as to perform a linear combination of Δ^N, Δ^T, and Δ^E):

$$v^C_{k,i} = \left(m_{\alpha_1}, m_{\alpha_2}, \dots, m_{\alpha_{\eta-1}}\right), \quad \text{so that}$$
$$\delta^C\left(m_i, m_{\alpha_1}\right) \leq \delta^C\left(m_i, m_{\alpha_2}\right) \leq \ldots \leq \delta^C\left(m_i, m_{\alpha_{\eta-1}}\right)$$

that is in increasing order of distances according to Δ^C.

The full list $v^W_{k,i}$ provided by the emotation wheel shares the same elements of $v^C_{k,i}$, but differently ranked, since distances are here computed on the basis of Δ^W:

$$v^W_{k,i} = \left(m_{\beta_1}, m_{\beta_2}, \dots, m_{\beta_{\eta-1}}\right), \quad \text{so that}$$
$$\delta^W\left(m_i, m_{\beta_1}\right) \leq \delta^W\left(m_i, m_{\beta_2}\right) \leq \ldots \leq \delta^W\left(m_i, m_{\beta_{\eta-1}}\right)$$

With the same approach used to produce the two ranked lists $v^C_{k,i}$ and $v^W_{k,i}$, the target (and optimal) ranked list on the user profile of u_k is built, using distance matrix $D^W_{u_k}$, as a $(\eta - 1)$ dimensional vector:

$$v^{opt}_{k,i} = \left(m_{\gamma_1}, m_{\gamma_2}, \dots, m_{\gamma_{\eta-1}}\right) \tag{6}$$

where distances between m_i and other movie scenes are sorted in ascending order according to $D^W_{u_k}$, that is:

$$d^W_{u_k}\left(m_i, m_{\gamma_1}\right) \leq d^W_{u_k}\left(m_i, m_{\gamma_2}\right) \leq \ldots \leq d^W_{u_k}\left(m_i, m_{\gamma_{\eta-1}}\right)$$

This list is optimal due to the fact that the ranking is built on the user profile, i.e. by the explicit rates expressed by u_k, so that scenes are ordered from the most to the least emotionally similar according to the personal user's emotions. This full list is then used as a target to compare the ranking abilities of the connotative space versus the emotation wheel.

Defining P_k as the set of η scenes rated by user u_k (i.e. his/her profile), full lists in the three spaces are thus computed as $\{v^C_{k,i}\}_{i \in P_k}$, $\{v^W_{k,i}\}_{i \in P_k}$, and $\{v^{opt}_{k,i}\}_{i \in P_k}$. The procedure is then repeated for all users.

The reader should not be misled by the fact that the target lists are computed on votes that the *single user* expressed on the emotation wheel. The wheel collects both the affective reactions of single users, thus determining users' profiles, and the emotations of all other users. Single user's votes are very specific in describing the user's emotional reactions to movie scenes, while the emotations collected on the entire community are likely less accurate in representing the reactions of a single person, as shown in the experiment of Section VI. Our expectation is that connotation works better than emotations by all users to guess the emotional preference of a single user.

D. List ranking comparison: Kendall's tau distance

To compare full lists in the connotative and emotation spaces with respect to the optimal lists, we adopt two different measures for ranking comparison: the well known Kendall's tau distance [45], and an alternative ranking metric we propose, which also accounts for the importance of positioning the most relevant items at the beginning of the list.

Kendall's tau metric on full lists (also adaptable to top-k lists, see [44]) is defined as follows: for each pair $\{m_i, m_j\}$ of movie scenes, if m_i and m_j are in the same order in the two lists v^A and v^B, then $k_{m_i,m_j}(v^A, v^B) = 0$; if m_i and m_j are in the opposite order (such as m_i being ahead of m_j in v^A and m_j being ahead of m_i in v^B), then $k_{m_i,m_j}(v^A, v^B) = 1$. Kendall's tau distance between lists is then

$$K(v^A, v^B) = \sum_{\{m_i,m_j\} \in M \times M} k_{m_i,m_j}(v^A, v^B) \tag{7}$$

and it turns out to be equal to the number of exchanges needed in a bubble sort to convert one permutation to the other. Since we have lists of $(\eta - 1)$ items, Kendall's tau maximum value is $(\eta - 1)(\eta - 2)/2$ which occurs when v^A is the reverse of v^B. Often Kendall's tau is normalised to its maximum value so that a value of 1 indicates maximum disagreement, leading to a normalised Kendall's tau distance in the interval $[0, 1]$.

We compute normalised Kendall's tau distances to compare list rankings in the connotative and emotational spaces with respect to ranking of lists based on the user's profile, that are $K(v^C_{k,i}, v^{opt}_{k,i})$ and $K(v^W_{k,i}, v^{opt}_{k,i})$. Obtained distances can be averaged on a user basis in both spaces (thus measuring how good spaces rank scenes for a given user) as:

$$K(v^C_k, v^{opt}_k) = \frac{1}{\eta} \sum_{i \in P_k} K(v^C_{k,i}, v^{opt}_{k,i}) \tag{8}$$

$$K(v^W_k, v^{opt}_k) = \frac{1}{\eta} \sum_{i \in P_k} K(v^W_{k,i}, v^{opt}_{k,i}) \tag{9}$$

Kendall's tau distances can be also averaged on a scene basis (thus measuring how good spaces are at ranking a specific

scene for all users who rated it) as:

$$K(v_i^C, v_i^{opt}) = \frac{1}{|Q_i|} \sum_{u_k \in Q_i} K(v_{k,i}^C, v_{k,i}^{opt}) \quad (10)$$

$$K(v_i^W, v_i^{opt}) = \frac{1}{|Q_i|} \sum_{u_k \in Q_i} K(v_{k,i}^W, v_{k,i}^{opt}) \quad (11)$$

where Q_i is the subset of users who actually emotated scene m_i. Eventually it is possible to obtain total ranking metrics for the lists on all scenes and for all voting users as:

$$K^C = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|Q_i|} \sum_{u_k \in Q_i} K(v_{k,i}^C, v_{k,i}^{opt}) \quad (12)$$

$$K^W = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|Q_i|} \sum_{u_k \in Q_i} K(v_{k,i}^W, v_{k,i}^{opt}) \quad (13)$$

In Figure 12 the comparison between the two Kendall's tau distances averaged on a scene basis is shown, while distances aggregated on a user basis are compared in Figure 13.

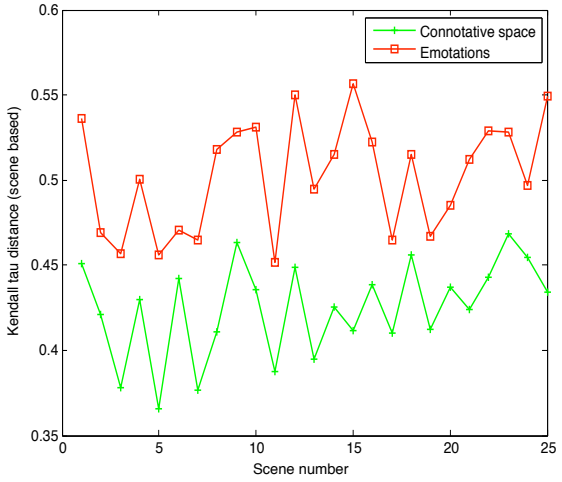


Fig. 12. Distance K computed on a scene basis. Ranking in the connotative space better approximates in all scenes the optimal ranking.

Both graphical representations suggest the superior ability of the connotative space in ranking movie scenes, since the obtained ranked lists better approximate the optimal rankings suggested by the users' profiles than using emotations. Note that in Figure 13, for the sake of visualisation, users are ordered in ascending order with respect to the connotative ranking distance. Moreover, in order to show coherent values (i.e. obtained for all users on the same number of rated movie scenes), only users that completed the online test by assigning rates to all η movie scenes are taken into account; as a result, the number of considered users in this experiment is lower than the total number of participants.

E. List comparison: ranking metric R

Kendall's tau metric does not differently weight ranking errors performed in the first list positions from those occurring at the end. Thus we propose an alternative ranking metric R which, by assigning positional weights w_j, considers the importance of having the most relevant items at the beginning.

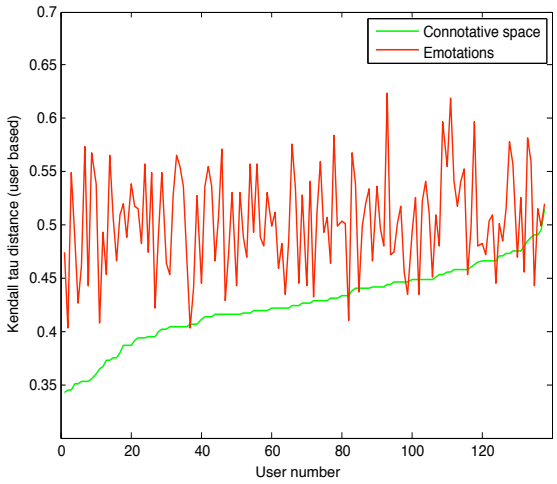


Fig. 13. Distance K computed on a user basis (only who completed the test is considered). For visualisation, users are ordered wrt K in the connotative space: connotative ranking works better than using emotations for most users.

Since ranking based on user profile is optimal (it is based on explicit rates expressed by u_k), we again compare the ranking capabilities of the connotative space versus the emotation space with respect to the optimal ranking metric computed on the user's profile. This optimal bound is defined for user u_k wanting to retrieve items with the same emo-tag as m_i:

$$R_{u_k,m_i}^{opt} = \sum_{j=1}^{\eta-1} w_j \cdot d_{u_k}^W(m_i, v_{k,i}^{opt}(j)) \quad (14)$$

where $w_j \in [0,1]$ are arbitrary positional weights subjected to

$$w_1 \geq w_2 \geq \ldots \geq w_{\eta-1} \quad (15)$$

used to emphasise the importance of ranking in the first positions the most relevant movie scenes (e.g. in the experiment $w = \{0.9, 0.8, \ldots, 0.1\}$ for $\eta = 10$).

Similarly, ranking metrics for lists produced in the connotative and in the emotation space are defined, respectively, as

$$R_{u_k,m_i}^{C} = \sum_{j=1}^{\eta-1} w_j \cdot d_{u_k}^W(m_i, v_{k,i}^{C}(j)) \quad (16)$$

$$R_{u_k,m_i}^{W} = \sum_{j=1}^{\eta-1} w_j \cdot d_{u_k}^W(m_i, v_{k,i}^{W}(j)) \quad (17)$$

where d is the distance between two emotions defined in Section VI, so that we can compare ranking metrics with the optimal one.

All metrics R can be then aggregated, either when ranking items for a specific user, or on the basis of a specific scene:

$$R_{u_k}^{opt} = \frac{1}{\eta} \sum_{i \in P_k} R_{u_k,m_i}^{opt}, \quad R_{m_i}^{opt} = \frac{1}{|Q_i|} \sum_{u_k \in Q_i} R_{u_k,m_i}^{opt} \quad (18)$$

$$R_{u_k}^{C} = \frac{1}{\eta} \sum_{i \in P_k} R_{u_k,m_i}^{C}, \quad R_{m_i}^{C} = \frac{1}{|Q_i|} \sum_{u_k \in Q_i} R_{u_k,m_i}^{C} \quad (19)$$

$$R_{u_k}^{W} = \frac{1}{\eta} \sum_{i \in P_k} R_{u_k,m_i}^{W}, \quad R_{m_i}^{W} = \frac{1}{|Q_i|} \sum_{u_k \in Q_i} R_{u_k,m_i}^{W} \quad (20)$$

where Q_i is the subset of users who actually voted scene m_i. It is again possible to obtain total ranking metrics for the three lists on all scenes and for all voting users as:

$$R^{opt} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|Q_i|}\sum_{u_k \in Q_i} R^{opt}_{u_k,m_i} \quad (21)$$

$$R^{C} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|Q_i|}\sum_{u_k \in Q_i} R^{C}_{u_k,m_i} \quad (22)$$

$$R^{W} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|Q_i|}\sum_{u_k \in Q_i} R^{W}_{u_k,m_i} \quad (23)$$

In Figure 14 the ranking metrics aggregated on a scene basis are shown. The blue curve is the lower bound which represents

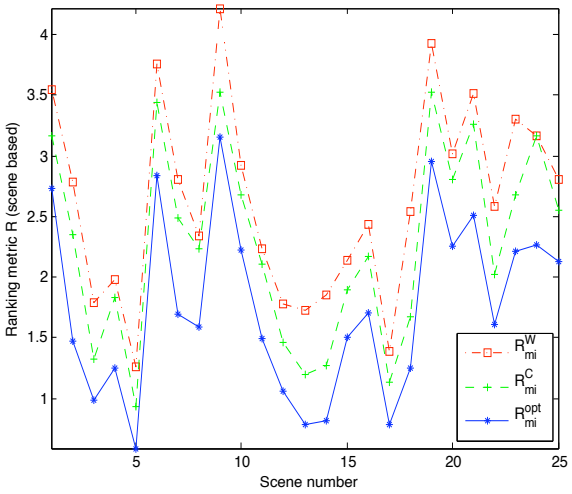


Fig. 14. Ranking metric R computed on a scene-base. Ranking in the connotative space better approximates the optimal ranking for all scenes.

the optimal ranking metric $R^{opt}_{m_i}$ for all movie scenes. The red curve describes the ranking metric $R^{W}_{m_i}$ obtained by using the emotation wheel, while the green one $R^{C}_{m_i}$ is the one obtained by ranking items in the connotative space. For all scenes, the ranking metric $R^{C}_{m_i}$ better approximates the optimal bound $R^{opt}_{m_i}$, thus outperforming the ranking scheme which employs the emotations by other users.

Figure 15 instead shows the comparison between ranking metrics when intended on a user basis. Note that, again in Figure 15, for the sake of visualisation, voting users are ordered in ascending order with respect to the optimal ranking metric $R^{opt}_{u_k}$. In this case, again, the analysis of performance reveals that ranking lists in the connotative space outperforms the ranking scheme obtained by using emotations for a very large majority of users.

F. Discussion

Both metrics aggregated for all voting users and on all scenes are presented and compared in Table IV, which summarises the superiority of the ranking scheme based on the connotative space with respect to the use of emotations, by returning a value closer to the optimal bound provided by ranking based on single users' profiles.

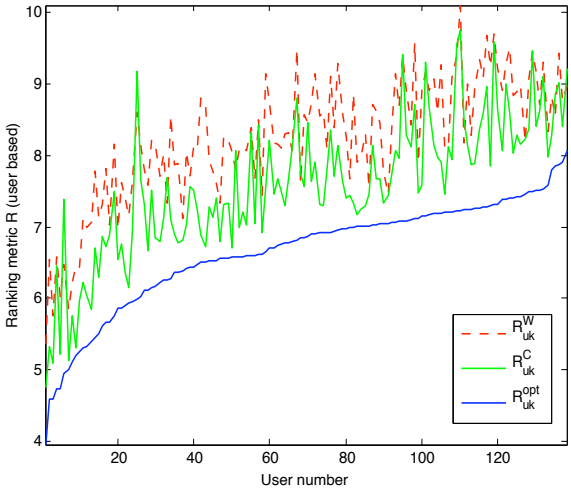


Fig. 15. Ranking metric R on users (only users who completed the online test are considered). For visualisation, users are ordered in ascending order wrt the optimal ranking metric: the connotative ranking scheme works better than using emotations.

TABLE IV
RANKING METRICS FOR ALL VOTING USERS AND ALL SCENES.

Ranking distance	Optimal	Connotative	Emotative
Kendalls' tau	$K^{opt} = 0$	$K^{C} = 0.425$	$K^{W} = 0.503$
R-metric	$R^{opt} = 1.754$	$R^{C} = 2.275$	$R^{W} = 2.631$

The outcome of this test is that the connotative space, when associated to a single user, is more advantageous than using affective tags (emotations) by other users. When using the connotative space in fact, beyond obtaining a stronger agreement in ratings among different users, we are able to better target the emotional wishes of single individuals. Going back for the last time to our introductory example, if the user experiences "relaxation" when watching horror content, he/she will probably feel again relaxed while watching other scenes which are similarly connotated, while if the system recommends commonly considered "relaxing" scenes (thus emotated by other users), he/she will be less satisfied.

As a consequence of the fact that connotative elements in movies strongly influence individual reactions, the proposed space relates more robustly to single users' emotions than using emotional models built on collective affective responses, since for these an agreement in judgement is more difficult to establish. Therefore this work confirms the utility of the connotative space for which it may be easier to achieve a direct correspondence between video physical properties and human emotions.

VIII. CONCLUSIONS

It is nowadays still unfeasible to establish a direct mapping of video physical properties into emotional categories. As an alternative to existing affective models for videos, we develop an ad-hoc connotative space for movie description, which aims at linking connotative elements of movie production to human emotions.

First, we demonstrate that this solution provides connotative descriptions of filmic products which are more objectively

agreed among users than tagging by means of emotional labels. Second, as a support in an application scenario for movie affective recommendation, we demonstrate that the connotative space is linked to the affective categories of a single user. In fact searching for similar videos in the connotative space returns content which better targets the single users' emotional preferences with respect to using emotional annotations from other people. The connotative space thus constitutes a valid inter-subjective platform for analysis and comparison of different feature films on an equal footing.

In the future, further studies should be conducted to establish a correspondence between the connotative space dimensions and audio and video primitives.

APPENDIX
COMPUTATION OF WMS AND BMS FOR $ICC(1,k)$

Let M be the number of targets (i.e. the number of scenes), k the number of raters for each target and x_{ij} the rating that user j assigns to scene i. If $\bar{x}_i$ indicates the average vote for scene i and $\bar{x}$ is the average of all rates on all scenes, then the intraclass correlation coefficient in its declination $ICC(1,k)$ is computed as in Equation 1. In this formula, WMS is the within-scene mean square, obtained from WSS, which refers to the within-scene sum of squares, that is:

$$WMS = \frac{WSS}{M(k-1)} \tag{24}$$

$$WSS = \sum_{i=1}^{M}\sum_{j=1}^{k}(x_{ij} - \bar{x}_i)^2 \tag{25}$$

The between-scene mean square (BMS) instead, is defined as a normalisation of BSS, the between-scene sum of squares, that is:

$$BMS = \frac{BSS}{M-1} \tag{26}$$

$$BSS = k\sum_{i=1}^{M}(\bar{x}_i - \bar{x})^2 \tag{27}$$

While it is straightforward to compute $ICC(1,k)$ for the three types of connotative measures, where possible rates are $x_{ij} \in \{1,2,3,4,5\}$, we here detail the procedure of calculation for the emotations, due to the circular nature of the adopted metric.

Calculating WMS is not problematic on a scale from 1 to 5, since it is the average of variance terms computed separately on single scenes. Conversely, for what concerns BMS, since the term $\bar{x}$ is the average of all rates on all scenes, the 5-bin histograms H_i^W need to be realigned so that bins in corresponding positions are referring to the same emotions, as depicted in the examples of Figure 16 (where histograms are realigned so that Sa=1 for all of them).

Observe that, to enable realignement, the three discarded contiguous emotions of each scene are assigned zero-valued bins and terms and operations in Equation 27 are computed adopting circular statistics [46]. Notice that distances used in the computation of Equation 27 always take values from 1 to 4 and as such allow for a fair comparison with respect to the $ICC(1,k)$ obtained from the connotative measures.

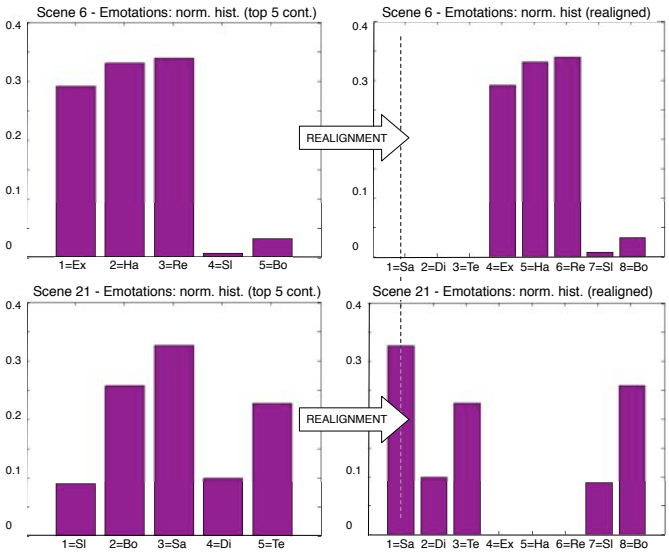


Fig. 16. Example of realignment of emotation histograms on two scenes. This procedure allows for a correct computation of BMS.

REFERENCES

[1] A. Hanjalic, "Extracting moods from pictures and sounds," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, March 2006.

[2] R. W. Picard, "Affective computing: From laughter to ieee," *IEEE Transactions on Affective Computing*, vol. 1, pp. 11–17, 2010.

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

[4] G. M. Smith, *Film Structure and the Emotion System*. Cambridge: Cambridge University Press, 2003.

[5] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *ACM workshop on Multimedia semantics Proceedings*, Vancouver, Canada, October 2008, pp. 32–39.

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

[7] A. Mehrabian, "Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament," *Current Psychology: Developmental, Learning, Personality, Social*, vol. 14, pp. 261–292, 1996.

[8] J. Wang, E. Chng, C. Xu, H. Lu, and X. Tong, "Identify sports video shots with "happy" or "sad" emotions," in *Proceedings International Conference on Multimedia and Expo (ICME)*, Toronto, Canada, July 2006.

[9] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Proceedings of International Conference on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, July 2005.

[10] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *ACM international conference on Multimedia Proceedings*, Vancouver, Canada, 2008, pp. 677–680.

[11] I. de Kok, "A model for valence using a color component in affective video content analysis," in *The 4th Twente Student Conference on IT Proceedings*, Enschede, January 2006.

[12] H.-B. Kang, "Affective content detection using HMMs," in *ACM international conference on Multimedia Proceedings*, Berkeley, CA, USA, November 2003.

[13] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, Berlin, Heidelberg, 2007, pp. 594–605.

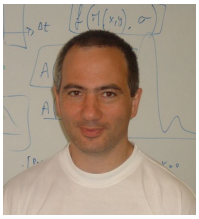
[14] H. L. Wang and L. F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, June 2006.

[15] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie af-

fective scene classification," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 523 –535, October 2010.
[16] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.
[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] C. T. Castelli, "Trini diagram: imaging emotional identity 3D positioning tool," *Internet Imaging*, vol. 3964, no. 1, pp. 224–233, 1999.
[19] C. Osgood, G. Suci, and P. Tannenbaum, *The Measurement of Meaning*. Urbana, IL: University of Illinois Press, 1957.
[20] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, pp. 1161–1178, 1980.
[21] R. M. Weaver, *A Rhetoric and Composition Handbook*. New York, NY: William Morrow & Co., 1974.
[22] D. Arijon, *Grammar of the Film Language*. Silman-James Press, September 1991.
[23] D. Heise, *Expressive Order: Confirming Sentiments in Social Actions*. New York: Springer, 2007.
[24] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, pp. 273–294, September 1977.
[25] K. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, Jan. 2005. [Online]. Available: http://ssi.sagepub.com/cgi/content/abstract/44/4/695
[26] R. Plutchik, "The Nature of Emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.
[27] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.
[28] J. J. Dahlgaard, S. Schtte, E. Ayas, and S. M. Dahlgaard-Park, "Kansei/affective engineering design: A methodology for profound affection and attractive quality creation," *Special Edition on Affective Engineering, The TQM Journal*, vol. 20, no. 4, 2008.
[29] "User test," www.ing.unibs.it/∼luca.canini/tests/index.php, require credentials to luca.canini@ing.unibs.it.
[30] "What is a "Great Film Scene" or "Great Film Moment"? An introduction to the topic," http://www.filmsite.org/scenes.html.
[31] "Internet movie database," *www.imdb.com*. [Online]. Available: www.imdb.com
[32] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video retrieval systems," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 4, Jun 1999.
[33] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot detection and condensed representation," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 28–27, March 2006.
[34] S. Benini, L.-Q. Xu, and R. Leonardi, "Identifying video content consistency by vector quantization," in *Proc. of WIAMIS'05*. Montreux, Switzerland, 13-15 April 2005.
[35] H. Sundaram and S.-F. Chang, "Computable scenes and structures in films," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 482–491, December 2002.
[36] J. A. Coan and J. J. B. Allen, *Handbook of Emotion Elicitation and Assessment*. New York: Oxford University Press, 2007.
[37] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, pp. 677–680, 1983.
[38] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.
[39] M. K. Greenwald, E. W. Cook, and P. Lang, "Dimensional covariation in the evaluation of pictorial stimuli," *Journal of Psychophysiology*, vol. 3, pp. 51–64, 1989.
[40] J. J. M. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ser. ICME'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1436–1439.
[41] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web: methods and strategies of web personalization*. Springer-Verlag, 2007, pp. 325–341.
[42] Y.-H. Yang and H. H. Chen, "Music emotion ranking," in *ICASSP*. IEEE, 2009, pp. 1657–1660.
[43] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proceedings of the 1998 IEEE International Conference on Computer Vision*, Bombay, India, January 1998, pp. 59–66.
[44] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, Philadelphia, PA, USA, 2003, pp. 28–36.
[45] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*. Edward Arnold, 1990.
[46] P. Berens, "Circstat: A matlab toolbox for circular statistics," *Journal of Statistical Software*, vol. 31, no. 10, pp. 1–21, 2009.

Sergio Benini received his MSc degree in Electronic Engineering (cum laude) at the University of Brescia with a thesis granted by Italian Academy of Science (2000). Between 2001-03 he has been working in Siemens Mobile Communication R&D. He received his Ph.D. in Information Engineering from the University of Brescia (2006), working on video content analysis. During his Ph.D. he conducted a one year placement in British Telecom Research, U.K. working in the "Content & Coding Lab". He is currently Assistant Professor at the University of Brescia.

Luca Canini received his MSc in Telecommunications Engineering (cum laude) at the University of Brescia with a thesis which won a prize granted by the Italian Marconi Foundation. He is currently a PhD candidate in the same university. During his PhD studies he has been a visiting student at the IVE Lab, University of Teesside (UK) and at the DVMM Lab, Columbia University (USA).



Riccardo Leonardi has obtained his Diploma (1984) and Ph.D. (1987) degrees in Electrical Engineering from the Swiss Federal Institute of Technology in Lausanne. He spent one year (1987-88) as a post-doctoral fellow with the Information Research Laboratory at UCSB (USA). From 1988 to 1991, he was a Member of Technical Staff at AT&T Bell Laboratories. In 1991, he returned briefly to the Swiss Federal Institute of Technology in Lausanne. Since February 1992, he has been appointed at the University of Brescia to lead research and teaching in the field of Telecommunications. His main research interests cover the field of Digital Signal Processing applications, with a specific expertise on visual communications, and content-based analysis of audio-visual information. He has published more than 100 papers on these topics.