Tarvainen, Jussi; Laaksonen, Jorma; Takala, Tapio

Film Mood and Its Quantitative Determinants in Different Types of Scenes

Published in:
IEEE Transactions on Affective Computing

DOI:
10.1109/TAFFC.2018.2791529

Published: 01/04/2020

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Tarvainen, J., Laaksonen, J., & Takala, T. (2020). Film Mood and Its Quantitative Determinants in Different Types of Scenes. *IEEE Transactions on Affective Computing*, *11*(2), 313-326.
https://doi.org/10.1109/TAFFC.2018.2791529

Film mood and its quantitative determinants in different types of scenes

Jussi Tarvainen; Jorma Laaksonen, *Senior Member, IEEE;* and Tapio Takala

***Abstract*—Films elicit emotions in viewers by infusing the story they tell with an affective character or tone – in a word, a mood. Considerable effort has been made recently to develop computational methods to estimate affective content in film. However, these efforts have focused almost exclusively on style-based features while neglecting to consider different scene types separately. In this study, we investigated the quantitative determinants of film mood across scenes classified by their setting and use of sounds. We examined whether viewers could assess film mood directly in terms of hedonic tone, energetic arousal, and tense arousal; whether their mood ratings differed by scene type; and how various narrative and stylistic film attributes as well as low- and high-level computational features related to the ratings. We found that the viewers were adept at assessing film mood, that sound-based scene classification brought out differences in the mood ratings, and that the low- and high-level features related to different mood dimensions. The study showed that computational film mood estimation can benefit from scene type classification and the use of both low- and high-level features. We have made our clip assessment and annotation data as well as the extracted computational features publicly available.**

***Index Terms*—Film, affect, mood, style, content-based analysis.**

I. Introduction

As the amount of digitally available film material grows, so does the need for description methods that can help viewers make informed choices based on their preferences. Currently, viewers rely on such information as genre, actors, and critical reviews. These descriptors, though useful, fail to account for the fact that film is, above all, an affective art form that seeks to elicit emotions in viewers [1]. A description of a film's affective content would therefore be a helpful guide for viewers. It would inform them whether the film is happy, sad, or anything in between; whether it seeks to excite, humor, or startle; whether it is tonally consistent or varied.

Human assessments of the affective content of films are both resource-intensive to collect and prone to subjective bias. For these reasons, the description would ideally be performed automatically, based on computational analysis of the film's audiovisual data. Indeed, recent years have seen a number of studies (see [2] for a summary) that seek to develop computational methods to describe the affective content of film scenes. Most of these studies define affective content in terms of the viewer's affective state or emotional response (e.g. [3]–[5]). Some, however, focus on the perceived affective expression inherent in the film itself (e.g. [6]–[8]). The distinction is important, since viewers' emotional responses are highly personal [5] and therefore difficult to estimate computationally. On the other hand, viewers tend to be in closer agreement about their perceptions of a film's affective intent than their emotional responses to it [9]; for example, they can recognize a horror film's intent to be scary even when they are not personally scared by the film. For this reason, a definition of affective content based on viewers' perceptions of the film, and not their personal emotional responses to it, is better suited for computational analysis. Combined with information about a given viewer's personal preferences, such a description of a film's objective affective content could also be used to estimate the film's subjective effect on the viewer [8]. To avoid confusion, affective content defined as the film's affective character, atmosphere, or tone can simply be called the film's *mood* [10], [11].

In order to be useful in the general case, a computational descriptor of film mood has to work with different types of content. A common way to classify films is by genre, such as action, comedy, and horror. Since human-created genre labels are not always available or reliable, the classification would ideally be based on features detected directly from the film material. However, computational genre classification has proven difficult, one reason being that films, not to mention individual scenes, do not fit neatly into discrete genres. Scene type, on the other hand, is a potentially more useful classification criterion, especially since the analysis is typically carried out on the scene level. Scene type can be expected to affect the optimal selection and weighting of computational features for mood estimation. For example, dialogue features (e.g. [12]) can be expected to be more relevant in dialogue scenes than action scenes. Scene location, time, characters, and dialogue were recently shown to be feasible in video summarization [13], but the use of such information as a prior in computational estimation of affect in film has not been investigated.

In this study, we investigate the *quantitative determinants* of film mood; that is, the various measurable factors – both human-rated attributes and automatically detected computational features – that affect the mood of different types of film scenes. To this end, we carried out two user studies. In the first, we investigated whether viewers could assess film mood directly in terms of various dimensions (e.g. positiveness and tenseness) in order to facilitate the collection of mood assessments for a large set of film clips. We also examined the general influence of two *narrative attribute* groups (events and speech) and two *stylistic attribute* groups (visual and auditory style) on film mood. Based on our findings, we created a more extensive set of 50 film clips representing various typical film

J. Tarvainen, J. Laaksonen and T. Takala are with the Department of Computer Science, Aalto University School of Science, Espoo, Finland. E-mail: {jussi.tarvainen,jorma.laaksonen,tapio.takala}@aalto.fi

scenes manually classified according to four criteria (location, time of day, and the use of music and dialogue). With these clips, we conducted a second user study in which we collected ratings of film mood and individual stylistic attributes for each clip. We then investigated whether the mood ratings differed across scene types, and how the stylistic attributes as well as various computational features correlated with the ratings.

We sought to answer the following research questions:

1) Do indirect and direct assessments of film mood produce similar results?
2) How do different groups of narrative and stylistic film attributes influence mood ratings?
3) Do mood ratings differ across scene types?
4) How do various stylistic attributes and computational features correlate with mood ratings across scene types?

The study contributes to recent work on film affect in affective computing [2] and cognitive film studies [14] in three ways. First, it provides insight into film mood assessment methods. It also illustrates the comparative influence of narrative and style on film mood. Lastly, it shows the benefit of scene type classification and the use of high-level features in computational mood estimation.

The article is organized as follows. The narrative and stylistic attributes of film that inform our selection of scene types and computational features are discussed in Section II. Low- and high-level computational features that can be used in mood estimation are discussed in Section III. Our approach to scene classification is covered in Section IV. Our two user studies are detailed in Section V, and the methods used to process the data by statistical and computational means in Section VI. Results are presented in Section VII and discussed in Section VIII. Conclusions are drawn in Section IX.

II. Narrative and stylistic attributes of film

In their classic film studies textbook *Film Art: An Introduction* [15], Bordwell & Thompson identify two important categories of film attributes: *narrative* and *stylistic*. Narrative attributes are related to the film's story, and stylistic attributes to the use of various film techniques (e.g. camera movement) to tell the story. In other words, narrative attributes involve what happens in the film, while stylistic attributes involve the ways in which the narrative is expressed through film style.

We discuss common narrative and stylistic attributes of film in Sections II-A and II-B. We then discuss their relation to film mood in Section II-C and methods of mood assessment in Section II-D. In accordance with the scope of the study, we limit our discussion to attributes occurring on the scene level.

A. Narrative attributes

In mainstream films [16], "the action will spring primarily from individual characters as causal agents" [15, p. 94]. In other words, characters carry the narrative, supplying story information and driving the plot forward. From this viewpoint, we can distinguish between two groups of character-centric narrative attributes: *events* and *speech*. The former group encompasses various types of action, such as running, sleeping, etc. The latter group involves the communication of narrative information through spoken language, typically dialogue or voice-over narration. Indeed, the traditional film screenplay format provides just this kind of separation of the narrative into visible events and audible speech [17].

The traditional screenplay format also provides a third relevant group of narrative attributes: the *setting*, which describes when and where the action takes place [17]. In a typical screenplay, each scene description begins with single-word indications of the scene's location and time: "int" for an interior scene or "ext" for an exterior scene, and "day" or "night" to indicate the time of day.

B. Stylistic attributes

The stylistic attributes of film are related to specific techniques, such as cinematography or editing [15]. As such, they are often grouped by their modality into *visual* (image-related) and *auditory* (sound-related) attributes. Visual attributes include those related to elements in front of the camera: set design, costume, makeup, position and movement of actors, and lighting. They also include various cinematography-related features, such as camera angle, height, distance, and movement. Auditory attributes, in turn, often simply describe the stylistic use of different sound types – speech, music, and sound effects – or their perceptual properties: loudness, pitch, and timbre. Editing can involve both image and sound and can thus be considered both a visual and an auditory attribute.

The use of attributes related to specific film techniques can be useful in organizing the attributes into groups (e.g. by modality), but is not very intuitive for the average viewer with little or no knowledge of film technique. Thus, when conducting assessments of individual stylistic attributes, it is more feasible to describe them in terms of their perceptual properties [15]. For example, while viewers may not be able to pick apart various photographic techniques in an image, they can describe its brightness and colorfulness, as was shown in [18]. Also, of the three aforementioned perceptual properties of sound, the first one is certainly an intuitive concept to all viewers. Lastly, while viewers may be unable to assess the individual contributions of the music, camera, characters, an editing to a scene's pace, they can still give a general assessment of the scene's overall pace, i.e. its fastness, irrespective of the underlying techniques [19].

C. Relation to film mood

Previous studies have revealed relations between stylistic attributes and emotional responses with various types of media: for example, viewers' arousal levels during viewing of still images have been shown to increase with saturation and decrease with brightness [18], and increase when motion is applied to a still image [20]. Also, valence, energetic arousal, and tense arousal ratings of music and speech have been found to be greater in the case of loud samples than quiet ones [21]. The results of our earlier preliminary study on the determinants of film mood echo these findings by indicating that stylistic attributes in particular are related to film mood [22].

Still, more empirical evidence on the topic is needed, and the aforementioned categories of narrative and style may

be helpful in this regard. Information on their comparative influence on film mood can function as a guide by identifying areas of filmmaking to focus on when estimating film mood with computational features. If, for example, it turns out that a scene's valence is generally influenced more by narrative events than visual style, it may be feasible to focus one's efforts on developing computational features relating to the former instead of the latter. Also, comparing the influence of narrative and stylistic attributes of film in different types of scenes (e.g. dialogue and non-dialogue scenes) may illustrate worthwhile ways of classifying scenes for mood estimation.

D. Assessment of film mood

The assessment of mood is typically carried out using so-called dimensional models of affect [23]. These models represent moods, emotions, and other types of affect as arising from overlapping dimensions of a common affective space. This space often involves a valence (or *hedonic tone*, HT) and an arousal dimension, where valence describes the pleasurability and arousal the alertness associated with an affect. The use of two arousal dimensions – *energetic arousal* (EA), from "tired" to "awake", and *tense arousal* (TA), from "calm" to "tense" – was proposed in [24]. Since then, several studies have supported the dual-arousal approach (see [25] for a summary).

Though dimensional affect models are most commonly used to characterize human affects, their applicability to the assessment of the affective content of artworks has been shown with music [21]. However, since this concept of an "art mood" has only recently gained popularity in film studies [10], [11], there is not much precedent in the use of dimensional affect models in the assessment of film mood. One exception is our earlier study [22], which used the UWIST Mood Adjective Checklist [26]. In this method, mood is assessed indirectly by way of 24 descriptive terms such as "happy" and "depressed", and the item ratings can then be transformed to mood ratings in the HT, EA, and TA dimensions. However, such indirect assessments are time-consuming to carry out since a large number of ratings need to be collected for each mood rating. This limitation has prompted us to consider whether film mood could also be assessed directly in terms of HT, EA, and TA.

III. Computational features for mood estimation

The computational features that can be used to estimate affective content in film can be split into two types, low- and high-level features, based on the aspects of the film they approximate. These two feature types are introduced below in Sections III-A and III-B. The specific features we used in the current study are detailed in Sections VI-A and VI-B.

A. Low-level features

Low-level features are estimates of stylistic attributes. They approximate "any physical, quantitative aspect [of the film] that occurs regardless of the narrative" [14, p. 149]; examples of such aspects are brightness, color, and movement. Typical visual features include average brightness, shadow proportion (the proportion of brightness values below a given threshold), saturation, color energy (a feature combining brightness, saturation, color contrast, hue, and the spatial distribution of colors), shot duration, and motion (e.g. the average pixel motion across consecutive frames). Typical audio features include volume, pitch (the "highness" or "lowness" of a sound), tempo (the "fastness" of a sound), zero-crossing rate the rate at which the value of an audio signal changes between positive and negative), and mel-frequency cepstral coefficients (a representation of a signal's short-term power spectrum). A recent summary of commonly-used low-level features is provided in [2]; the reader is referred there for details.

The computational estimation of affective content in film is typically based on low-level features because of their relative computational simplicity. However, these features are limited in their ability to estimate affective content fully since they cannot detect many of the higher-level aspects of storytelling and emotional expression in film, such as acting and use of dialogue, that contribute to the viewer's overall impression of film mood [10]. For this reason, we also considered high-level computational features in the current study.

B. High-level features

High-level features are estimates of more complex attributes of film that cannot be defined in strictly technical terms, such as characters' emotional expressions and dialogue contents. However, their relation to specific narrative attributes of film, such as acting, distinguishes them from features describing a film's general aesthetics (e.g. beauty).

High-level attributes are difficult to estimate computationally since their relation to concrete, quantifiable properties of film is not straightforward. Still, recent years have seen the development of features that estimate the emotional expression in human faces, dialogue, and music, all of which are relevant in terms of a film's affective content [27]–[29].

To mention a few publicly available state-of-the-art features in these domains: firstly, the Microsoft Cognitive Services Emotion API[1] is a facial expression recognition tool that assigns each recognized face in an input image a numerical score from 0 to 1 in terms of eight emotions: neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise. In dialogue sentiment analysis, which estimates the attitude of emotion expressed in written text (usually in terms of valence), the Stanford CoreNLP sentiment analysis[2] is a tool based on recursive deep neural network models and a sentiment treebank [30] that places each sentence in the input text into one of five sentiment classes: very negative, negative, neutral, positive, and very positive. The IBM Watson Tone Analyzer[3] is a somewhat similar tool based on the Linguistic Inquiry and Word Count application [31] that assigns each sentence, and the text as a whole, a numerical score from 0 to 1 in terms of five emotions (anger, disgust, fear, joy, and sadness) as well as social tendencies and writing style. Lastly, the Music Information Retrieval (MIR) Toolbox[4] [32]

[1] https://www.microsoft.com/cognitive-services/en-us/emotion-api/
[2] http://nlp.stanford.edu/sentiment/
[3] https://www.ibm.com/watson/developercloud/doc/tone-analyzer/
[4] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/

TABLE I
THE FILM SCENE CLASSIFICATIONS USED IN THE STUDY

Basis	Criterion	Scene type	Definition
Setting	Location	Interior scene	Takes place indoors
		Exterior scene	Takes place outdoors
		Mixed-loc. scene	Takes place in- and outdoors
	Time	Daytime scene	Sunlight / scene context
		Nighttime scene	No sunlight / scene context
Sound	Dialogue	Dialogue scene	Narrative advanced mainly through dialogue
		Non-dial. scene	Narrative advanced mainly by other means
	Music	Music scene	Music clearly audible during main events
		Non-music scene	Music not clearly audible during main events

for MATLAB contains a function for estimating the emotion expressed in a piece of music in terms of the HT, EA, and TA dimensions [33].

IV. Classifying film scenes

In this study, we sought to investigate whether mood ratings of film scenes, as well as their determinants, vary across scene types. To this end, in addition to analyzing all the scenes together, we also manually classified the scenes into types according to four criteria and analyzed each scene type separately.

We looked for classification criteria that are commonly used in film studies, clearly defined in terms of specific aspects of a film, and related to both narrative and stylistic attributes. Based on these aims, we classified the scenes in terms of their setting (Section IV-A) and use of sounds (Section IV-B). The classification principles are summarized in Table I.

A. Setting-based classification

Of the three narrative attribute groups discussed in Section II-A, we used the setting as a scene classification criterion since it separates scenes into discrete types based on location (interior/exterior) and time (day/night). The other two groups, events and speech, do not allow similarly intuitive classifications since the number of possible events or speech contents in a scene is practically infinite.

We also used the label of a mixed-location scene to denote scenes that take place both in- and outdoors (e.g. when the camera follows a character outside). If the scene contained a view of the exterior, we used the presence of sunlight to determine whether it was a day- or nighttime scene. Otherwise we determined the time of day from the scene's context by using commonsense knowledge.

B. Sound-based classification

We also sought to classify scenes based on their visual and auditory style. However, the range of visual expression in an average scene is too wide for classification by visual style to be feasible. For example, a given scene is likely to contain a number of different colors, shot sizes, types of movement, etc., and so cannot easily be classified by these attributes.

Sound, however, is a more fruitful classification criterion. While Bordwell & Thompson's three sound types – speech, music, and sound effects [15] – feature in almost all film scenes, the use of the first two, speech and music, is selective enough to allow scenes to be classified by their prominence on the soundtrack. We therefore classified the scenes based on the prominence of dialogue, i.e. whether the narrative was advanced mainly through dialogue (dialogue scenes) or other types of action (non-dialogue scenes); and the prominence of music, i.e. whether music was clearly audible during the scene's main events (music scenes) or not (non-music scenes). The music could be either diegetic (present within the story world) or non-diegetic (overlaid soundtrack music).

Although dialogue also has a narrative function (as discussed in Section II-A), we consider it a style-based classification criterion here since our classification is not based on the *contents* of the dialogue but only on its prominence in a scene. Also, although Bordwell & Thompson's definition of the "speech" sound type includes voice-over narration, we based the classification on dialogue only since dialogue unavoidably affects the use of other stylistic attributes (e.g. shot composition to keep speaking characters visible, sound editing to keep their mouths in sync with the dialogue) in a way that voice-over does not. Lastly, the concept of a "dialogue scene" is widely used in both film studies (e.g. [15], [17]) and computer science (e.g. [34], [35]), while "speech scene" is not.

V. Data collection

We conducted two user studies (Sections V-A and V-B). We have made all the film style and mood data collected in the second user study publicly available (Section V-C).

A. Study 1

We first carried out a user study whose participants viewed a series of film clips and assessed their mood using indirect and direct methods, as well as the influence of various groups of narrative and stylistic attributes on the mood ratings.

1) Film clips: The stimulus set consisted of 12 film clips between 1–2.5 minutes in duration (Table II), for a total duration of 22 minutes. They were taken from our earlier set of 14 clips, whose selection process is detailed in [22]. The clips were from mainstream films made between 1958 and 2009. Each clip contained a complete scene that could be understood without knowledge of the preceding events.

Clip 7 contained French dialogue and was presented with English subtitles; the other clips were in English and presented unsubtitled. The clips were obtained from DVD discs. Dialogue volume levels were normalized across the clips.

2) Participants: Seven participants (two women, five men) took part in the study. They were recruited personally and given a verbal description of the study. They reported to be fluent in English and to have adequate visual acuity for the task. We had previously found [22] that film expertise (defined as having studied film and/or having filmmaking experience)

TABLE II
THE FILM CLIPS USED IN STUDY 1

#	Film title	Location	Time	Dial. scene	Music scene
1	*Amélie*	int	day	false	false
2	*Children of Men*	int	day	false	false
3	*Before Sunrise*	ext	day	true	false
4	*Days of Heaven*	ext	day	false	true
5	*500 Days of Summer*	mixed	day	false	true
6	*E.T.* [a]	ext	day	false	true
7	*Army of Shadows*	int	night	false	false
8	*Punch-Drunk Love*	ext	day	false	true
9	*The Shining*	int	night	false	true
10	*Vertigo*	int	day	false	true
11	*Blue Velvet*	mixed	day	true	true
12	*Raiders of the Lost Ark*	mixed	day	false	true

int = interior, ext = exterior. [a] 20th Anniversary version.

had no effect on film style and mood ratings. We therefore did not determine the participants' level of film expertise.

3) Design and procedure: The study was carried out in a one-hour-long session. At the beginning of the session, the participants were given a tutorial on the assessments. They were then shown the set of 12 film clips in the order given in Table II, assessing each clip immediately after viewing. The participants had three minutes to assess each clip. The assessments for each clip were as follows:

1) Indirect assessment of the mood of the clip (24 items)
2) Direct assessment of the mood of the clip (3 items)
3) Assessment of the influence of various attribute groups on mood ratings (12 items)

The indirect assessment of mood was carried out using the UWIST Mood Adjective Checklist [26] comprising 24 items. The direct assessment, in turn, was carried out using three continuous scales, one for each mood dimension (HT, EA, and TA). On each scale, the participants marked the point that best represented the clip's mood in terms of that dimension. The scales spanned a range from "negative" to "positive" (HT), "sleepy" to "energetic" (EA), and "calm" to "tense" (TA) [24]. The midpoint of each scale was marked "neutral". We did not use the more common term "tired" to designate the low end of the EA scale to avoid the term's negative association with boring or old-fashioned material. The participants were instructed to use the full width of the scales and to interpret the dimensions according to their personal criteria.

In the direct mood assessment, the participants were also asked to rate the influence of two narrative attribute groups (events and speech, Section II-A) and two stylistic attribute groups (visual style and use of sounds, Section II-B) on each of the three mood dimensions. The influence of each attribute group was rated on a discrete scale from 1 to 3 (1 = no influence, 2 = some influence, 3 = strong influence). We used a simple three-step scale due to the participants' presumed unfamiliarity with the assessments.

We selected two groups from both attribute categories, narrative and stylistic, to have both categories evenly represented in the study. The included attribute groups and their descriptions in the questionnaire are summarized in Table III. In the questionnaire, the groups were presented without the category labels. Though the description of the stylistic attribute group "sounds" could also include dialogue as a stylistic device alongside the other sound types, we chose not to include it here to avoid confusion with the narrative attribute group "speech", which pertains to the contents of the dialogue and voice-over in the scene.

TABLE III
THE NARRATIVE AND STYLISTIC ATTRIBUTE GROUPS USED IN STUDY 1

Category	Attribute group	Description in questionnaire
Narrative	Events	The events depicted in the clip
	Speech	Dialogue and voice-over
Stylistic	Visual style	Shot composition, cinematography, lighting, colors, editing
	Sounds	Music, sound effects, other sounds

B. Study 2

Based on the results of Study 1, we conducted a second user study with a more extensive set of 50 film clips representing various scene types. The participants assessed the clips according to film mood and stylistic attributes. We then analyzed the mood ratings across scene types, as well as their correlations with stylistic attributes and computational features. A preliminary presentation of Study 2 and its results was previously provided in [36].

1) Film clips: The stimulus set consisted of 50 film clips between 0.5–3 minutes in duration (Table IV), for a total duration of 72 minutes. Each clip again contained a complete scene . The clips were taken from popular mainstream Hollywood films, with an average of 288 817 IMDb[5] ratings and a mean rating of 7.46/10. The films were made between 1980 and 2009, representing the so-called modern blockbuster era of film [16]. Mainstream films of this era have been characterized by two general aspects. The first is an adherence to the so-called classical Hollywood style of filmmaking, whose characteristics include continuity editing, the use of abundant emotional cues, and an emphasis on narrative clarity. The second characteristic, so-called intensified continuity [37], involves the use of techniques such as faster editing, extreme lens lengths, and more mobile camerawork, to control the viewer's attention and generate "a keen moment-to-moment anticipation" [37, p. 24]. These qualities make such films well suited for a study of the constituents of film mood.

To choose the clips, we conducted a pilot study whose participants assessed 50 candidate clips, each from a different film, by their mood. Based on these assessments, we chose 42 of the candidate clips for the actual study. We also included eight additional clips for a final 50-clip set which met the following requirements:

- Even representation of film moods, genres, and production decades 1980–2010
- All the scene types in Table I represented
- A different director and actors for each clip

The clipwise scene type classifications are given in Table IV. The set contained four mixed-location scenes, but we excluded

[5] http://www.imdb.com

TABLE IV
THE FILM CLIPS USED IN STUDY 2

#	Film title	Location	Time	Dial. scene	Music scene
1	*The Assassination of Jesse James ...*	mixed	day	false	true
2	*Death Wish 3*	int	night	false	false
3	*The War of the Roses*	mixed	day	true	false
4	*Ocean's Eleven*	int	night	true	false
5	*V for Vendetta*	int	night	true	true
6	*As Good as It Gets*	int	day	true	false
7	*The Fly*	int	night	false	true
8	*Raising Arizona*	int	day	false	false
9	*The Lord of the Rings: The Fellowship ...*	ext	day	false	true
10	*Red Dawn*	int	night	true	false
11	*Back to the Future Part II*	mixed	night	true	false
12	*Driving Miss Daisy*	ext	night	true	false
13	*Men in Black*	int	night	false	true
14	*Heartbreak Ridge*	int	day	true	false
15	*Pulp Fiction*	ext	day	false	false
16	*Bridget Jones's Diary*	int	day	true	false
17	*Tootsie*	int	day	false	true
18	*Before the Devil Knows You're Dead*	int	day	true	false
19	*Lord of War*	ext	day	false	true
20	*Beverly Hills Cop*	int	day	true	false
21	*The Sixth Sense*	int	night	false	false
22	*500 Days of Summer*	ext	day	false	true
23	*Braveheart*	ext	night	false	true
24	*L.A. Confidential*	int	day	false	true
25	*The Shining*	int	night	false	true
26	*The Right Stuff*	ext	night	false	false
27	*Shakespeare in Love*	int	night	false	true
28	*Meet the Fockers*	int	night	true	true
29	*The Fast and the Furious*	ext	day	false	true
30	*Marie Antoinette*	ext	day	false	false
31	*The Karate Kid*	ext	day	false	true
32	*No Way Out*	ext	day	true	false
33	*The Age of Innocence*	int	night	false	true
34	*Terminator 2: Judgment Day*	int	day	false	true
35	*The Da Vinci Code*	int	night	false	true
36	*The Beach*	int	day	false	true
37	*Before Sunrise*	ext	day	true	false
38	*Police Academy*	int	day	true	false
39	*Pirates of the Caribbean: Dead Man's Chest*	ext	day	false	true
40	*The Thing*	int	night	true	true
41	*Into the Wild*	ext	day	false	true
42	*Robocop*	mixed	day	false	true
43	*A Few Good Men*	int	night	true	false
44	*Desperado*	int	day	true	false
45	*E.T.* [a]	ext	day	false	true
46	*Legends of the Fall*	ext	day	false	true
47	*Hot Shots!*	ext	day	false	true
48	*Wedding Crashers*	ext	night	false	true
49	*Requiem for a Dream*	int	night	true	true
50	*The English Patient*	ext	day	false	true

int = interior, ext = exterior. [a] Original version.

them from the scene-type-specific analysis because of their rarity. As such, each scene classification criterion split the set into two scene types (e.g. interior and exterior scenes).

The clips were obtained from high-definition sources where available and scaled to a maximum width of 960 pixels for presentation. Dialogue volume levels were again normalized.

2) Participants: Forty-two participants (21 women, 21 men, μ_{age} = 27.6 years, σ_{age} = 6.2 years, age range: 23–48) took part in the study. We required self-reported fluency in English of the participants since the film clips were again shown in English without subtitles.

3) Design and procedure: The participants were split into two groups: group A (10 men, 10 women) and group B (11 men, 11 women). To counter any potential rating bias caused by the viewing order of the clips, group A viewed them in the order given in Table IV, and group B viewed them in reverse order.

The study was carried out in a movie theater in a two-hour session. At the beginning of the session, the participants were given a tutorial on the assessments, including a practice assessment. They then viewed and assessed each film clip in turn. The participants had 35 seconds to assess each clip, which was deemed sufficient based on the pilot study.

For each clip, the participants first assessed its mood in terms of HT, EA, and TA. Based on the results of Study 1 (Section VII-A), we used the direct assessment method here.

After the mood assessment, the participants assessed the clip's style in terms of four attributes: brightness (from "dark" to "bright"), colorfulness (from "colorless" to "colorful"), loudness (from "quiet" to "loud"), and fast-pacedness (from "slow" to "fast"). We chose the attributes based on the results of our previous study [22], in which we found the attributes to exhibit good inter-rater agreement and to be related to dimensions of film mood. Each attribute was assessed on a discrete scale from 1 to 5. The participants were again instructed to use the full width of the scales. Since the clip set did not contain any black-and-white or silent material, the participants were asked to use the low ends of the colorfulness and loudness scales to designate modest use of color and sound. In all other respects they were allowed to interpret the attributes freely.

C. Benchmark data

To allow other researchers to compare their methods with our results and further develop computational methods of mood estimation, we have made the data collected in Study 2 publicly available at http://research.ics.aalto.fi/cbir/data/. The data set contains the collected mood and style ratings, 14 700 (film, participant, attribute) triplets in all. For each clip, the data set also contains the scene's timecode, total duration, shot durations, frame-by-frame locations of characters, dialogue contents and timecodes, music timecodes, as well as the values of our computational features (Sections VI-A and VI-B).

Most other public sets of affective video content assessments (e.g. [38], [39]) are based on clips from amateur videos that can be freely distributed online. However, results based on such material do not necessarily generalize to mainstream content. Our data set, being based on mainstream film clips, should be more broadly applicable to popular content. The set is expected to facilitate the development and evaluation of computational descriptors of film style and mood.

VI. METHODS

A. Low-level computational features

In our analysis of the correlations between the computational features and the mood ratings, we tested low-level features related to each of the four assessed stylistic attributes:

brightness and colorfulness (Section VI-A1), loudness (Section VI-A2), and fastness (Sections VI-A3–VI-A5).

1) Brightness and color features: As estimates of the clips' brightness and colorfulness, we computed their average lightness and saturation values using the MATLAB function *rgb2hsl*[6]. We first computed the average value for each frame and then averaged the values across frames. We also computed the average shadow proportion value, defined, following [40], as the proportion of values below 18% of maximum lightness.

2) Audio features: As estimates of the loudness of the clips, we computed their root mean square (RMS) and Loudness Units relative to Full Scale (LUFS) [41] values. We computed the former using the *mirrms* function of the MATLAB MIR Toolbox (v1.6.1) [32] and the latter using Adobe Audition, using each clip's full audio track as input in both cases.

3) Shot duration: As a first estimate of the fastness of the clips, we computed their average shot durations. We first determined the frame positions of the shot boundaries in each clip and then computed the clipwise average of the durations between the boundaries. Though automatic shot boundary detection is often considered a solved technical problem [42], for the purpose of testing the feasibility of the shot duration feature in mood estimation, we annotated the shot boundaries manually for maximal accuracy. We computed both the mean and median shot duration for each clip.

4) Pixel motion: We also computed two other estimates of fastness based on the motion between consecutive frames. In the first of these, pixel motion, we computed the average pixel motion in each clip with the optical flow algorithm. We used the MATLAB function *optic_flow_sand* contained in the "High accuracy optical flow" implementation[7] of the Particle Video method [43].

We computed the optical flow for each frame pair, excluding pairs that crossed a shot boundary (as determined for the shot duration feature). For each frame pair, the algorithm produced two matrices of size $h \times w$, containing the horizontal (u) and vertical (v) components of the pixelwise motion between the two frames. From these, we obtained a scalar optical flow value f_{OF} for each frame pair by computing the mean of the Euclidean distances spanned by the u and v components (i.e. the absolute values of the pixelwise uv motion vectors):

$$f_{\mathrm{OF}} = \frac{\sum_{y=1}^{h} \sum_{x=1}^{w} \sqrt{u_{xy}^2 + v_{xy}^2}}{h \cdot w} \tag{1}$$

We then computed the clipwise average of all the f_{OF} values to obtain a single scalar pixel motion value for each clip. To account for the different aspect ratios between clips, we normalized the clipwise values by dividing them by the length of the clip's diagonal in pixels.

5) Character movement: It has been empirically shown that characters are highly predictive of film viewers' attention [27]. Therefore, as a second motion-based estimate of the clips' fastness, we computed the movement of the characters in them.

We first manually tracked the location of each character's head using Adobe After Effects. Though automatic face detection [44] and recognition [45] algorithms are available, we again sought to use maximally accurate data to test the feature, and thus performed the annotation manually. We included all the characters that were clearly visible in the shot and/or took part in the events depicted, but excluded background extras. We gave a unique identifier to each character that appeared in more than one shot, using "nil" for incidental characters.

We then computed the shotwise movement of each character by adding up the framewise scalar movements of the character's head within the shot, and then averaged out these values across shots and characters to obtain a scalar character movement value for each clip. As with the pixel motion metric, we normalized the values by dividing them by the length of the clip's diagonal in pixels. Since we used scalars, and not vectors, in computing the shotwise mask movement, the values do not reflect the average displacement of the characters, but rather their average total movement within shots, regardless of direction or change in position.

B. High-level computational features

We also tested each of the four high-level features mentioned in Section III-B. These features estimated the emotional content of three aspects of the clips: faces, dialogue, and music (Sections VI-B1, VI-B2, and VI-B3, respectively).

1) Face emotion: We estimated the characters' emotional expressions using the Microsoft Cognitive Services Emotion API. We extracted a still image of each character's head at 12-frame intervals (two images per second), resulting in a set of 14 285 input images from 49 clips (all except clip 50, which had no visible faces).

The Emotion API produced results for 3424 images (24% of the set). Each result was a set of eight emotion scores for each of the analyzed emotions. In cases where there were multiple results for a single character in a given shot, we averaged these out to have at most one set of values for each character in each shot. This process resulted in 0–47 sets of values per clip ($\mu = 14.38$, $\sigma = 10.39$). From these, we computed the clipwise average for each emotion. The resulting features are named according to their respective emotions, e.g. "face anger", "face happiness", etc.

2) Dialogue emotion: We estimated the emotions expressed in the dialogue in the clips using the Stanford CoreNLP sentiment analysis (v3.5.2) and IBM Watson Tone Analyzer tools. We ran the analysis for all 44 clips that contained dialogue, including both dialogue and non-dialogue scenes.

We first manually transcribed the dialogue contents of each clip. Here, too, the process could also be carried out automatically [46], but for our feature-testing purposes we again sought to use the most accurate data available. After transcribing the dialogue contents, we processed them for analysis by expanding all contractions (e.g. "it's" → "it is", "gonna" → "going to") and spelling out all numbers (e.g. "32" → "thirty-two"). We stored each clip's transcribed dialogue in a separate input text file.

With the sentiment analysis tool, we transformed the classifications of each clip's individual sentences into numerical

[6]https://se.mathworks.com/matlabcentral/fileexchange/20292-hsl2rgb-and-rgb2hsl-conversion/

[7]https://se.mathworks.com/matlabcentral/fileexchange/17500-high-accuracy-optical-flow/

values from –2 (very negative) to +2 (very positive) and computed the clipwise averages of these values. The resulting feature is named "dialogue sentiment". With the emotion analysis of the Tone Analyzer tool, we used the whole-text emotion scores directly. The resulting features are named according to their respective emotions, e.g. "dialogue anger", "dialogue joy", etc.

3) Music emotion: We estimated the emotion expressed in the music in the clips using the function *miremotion* [33] of the MIR Toolbox [32]. We used version 1.3 of the toolbox, the recommended version for running the *miremotion* function. We ran the analysis for all 35 clips that contained music, including both music and non-music scenes.

For each clip, we used as input the manually extracted music-containing segments from the clip's audio track. Though automatic music detection [12] could also be used, we again opted for manual classification for maximal accuracy. The algorithm estimated the emotion expressed in each clip's music segments in terms of HT, EA, and TA. The resulting features are named "music-HT", "music-EA", and "music-TA".

C. Analysis

In Study 1, we first compared the similarity of the mood ratings collected in the indirect and direct assessments using Pearson correlation, and their internal consistency using Intraclass Correlation Coefficients (ICC), computed using a two-way random effects, single-measures model (ICC(2,1)) with the absolute agreement criterion [47]. We then analyzed the comparative influence of the narrative and stylistic attribute groups on the mood ratings based on their means across scene types, with confidence intervals (CI) reported at the 95% level.

In Study 2, we first compared the similarity of the mood ratings within each scene type pair (e.g. daytime and nighttime scenes). We assessed the normality of the rating distributions in each scene type using the Shapiro-Wilk test and found all the distributions to be non-normal at $p < 0.05$. We therefore compared the ratings using the non-parametric Mann-Whitney test, which does not require the data to be normally distributed. We ran the Mann-Whitney test using a Monte Carlo method with 10 000 samples.

We then looked for quantitative determinants of the mood ratings. Using Pearson correlation, we studied how the ratings correlated with the stylistic attributes and the computational features across all the clips as well as separately for each scene type. We reported only those correlations that were based on data from at least 10 clips, passed a Bonferroni correction at $\alpha = 0.01$, and had a moderate or large effect size ($|r| \geq 0.30$).

In all these computations, we considered only results significant at $p < 0.01$. We followed Cohen's thresholds for r effect size [48]: small ($|r| \geq 0.10$), moderate ($|r| \geq 0.30$), and large ($|r| \geq 0.50$). We followed Cicchetti's categories for ICC effect size [49]: poor (ICC < 0.40), fair (ICC ≥ 0.40), good (ICC ≥ 0.60), and excellent (ICC ≥ 0.75).

VII. Results

A. Study 1: Indirect and direct mood assessment

The mood ratings given in the indirect and direct assessments were strongly correlated in terms of each dimension: for HT, $r(84) = 0.90$, $p < 0.001$; for EA, $r(84) = 0.89$, $p < 0.001$, and for TA, $r(84) = 0.87$, $p < 0.001$. They also had similarly high levels of internal consistency: for HT, $\text{ICC}_{\text{indirect}} = 0.83$ and $\text{ICC}_{\text{direct}} = 0.85$; for EA, $\text{ICC}_{\text{indirect}} = 0.75$ and $\text{ICC}_{\text{direct}} = 0.69$; and for TA, $\text{ICC}_{\text{indirect}} = 0.78$ and $\text{ICC}_{\text{direct}} = 0.65$. The results therefore answered the first research question in the affirmative and supported the feasibility of direct assessment of film mood.

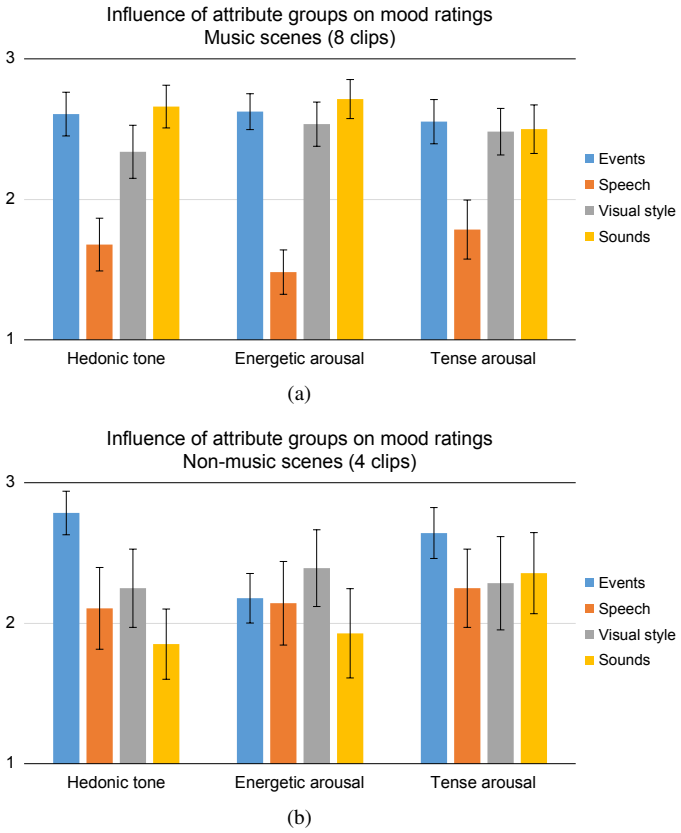

Fig. 1. Study 1: human-rated influence of narrative and stylistic attribute groups on mood ratings: (a) music scenes and (b) non-music scenes. Error bars denote 95% confidence intervals. Influence scale: 1 = no influence, 2 = some influence, 3 = strong influence.

B. Study 1: Narrative and stylistic attributes of film

We examined the influence of the four attribute groups on each mood dimension for all of the clips together, as well as separately for each scene type pair listed in Table I. We found the use of music to differentiate the comparative influence of the attribute groups on each dimension, as shown in Figure 1.

The results for the music scenes (Figure 1(a)) resembled those for all the clips together. Here, all three mood dimensions exhibited a similar order of influence, with speech having a much lesser influence on the mood ratings than the three other attribute groups. However, with the non-music scenes (Figure 1(b)), all four attribute groups had a similar influence on the EA and TA ratings, while events had a much greater influence on the HT ratings than the other attribute groups.

These results suggest, first, that the events depicted in a scene are generally important in terms of mood. Second, they imply that in music scenes, style is as important as the events

TABLE V
STUDY 2: MOOD RATING STATISTICS

Rating	Scenes	n clips	n ratings	μ	σ	M-W r
Hedonic tone	All	50	2024	−0.02	0.55	n/a
	Interior	27	1089	−0.16	0.50	**0.36**
	Exterior	19	775	0.25	0.53	
	Daytime	31	1257	0.02	0.54	0.09
	Nighttime	19	767	−0.08	0.55	
	Dialogue	19	765	−0.14	0.47	0.17
	Non-dialogue	31	1259	0.06	0.58	
	Music	29	1180	0.01	0.60	-
	Non-music	21	844	−0.06	0.47	
Energetic arousal	All clips	50	2022	0.11	0.48	n/a
	Interior	27	1089	0.03	0.43	0.18
	Exterior	19	773	0.18	0.52	
	Daytime	31	1257	0.17	0.50	0.19
	Nighttime	19	765	0.00	0.44	
	Dialogue	19	765	0.00	0.41	0.19
	Non-dialogue	31	1257	0.17	0.51	
	Music	29	1178	0.16	0.51	0.14
	Non-music	21	844	0.03	0.44	
Tense arousal	All clips	50	2023	0.14	0.52	n/a
	Interior	27	1089	0.25	0.48	0.27
	Exterior	19	774	−0.05	0.54	
	Daytime	31	1257	0.08	0.52	0.14
	Nighttime	19	766	0.22	0.52	
	Dialogue	19	765	0.18	0.47	-
	Non-dialogue	31	1258	0.11	0.55	
	Music	29	1179	0.15	0.55	-
	Non-music	21	844	0.11	0.48	

M-W r = Mann-Whitney test effect size. Only results significant at $p < 0.01$ are listed (n/a = not applicable in the case of all clips). Moderate and large effects ($|r| \geq 0.30$) are in boldface. For these pairs, the scene type with the greater mean rating is also in boldface.

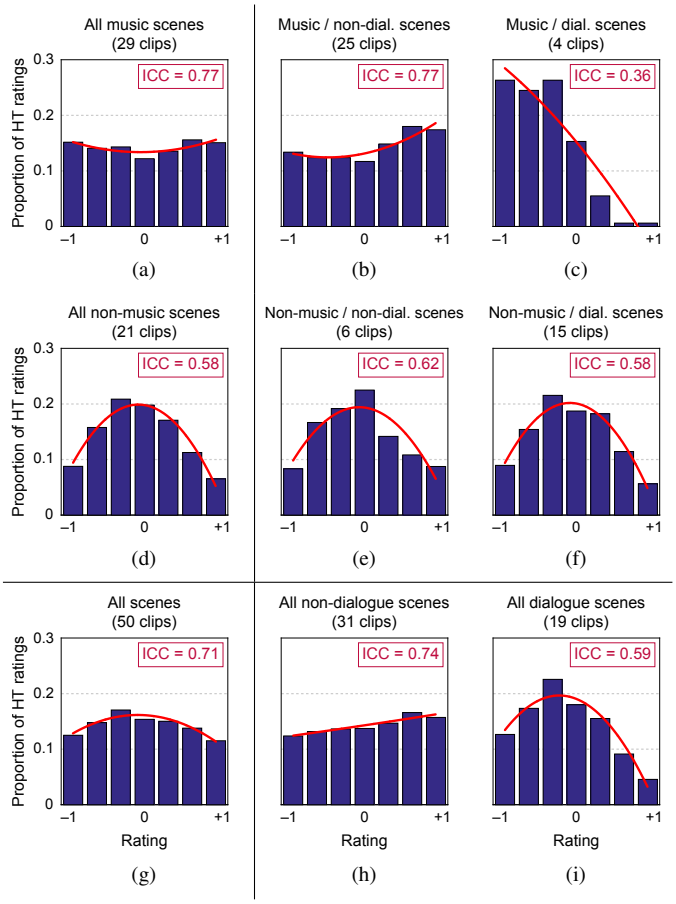

Fig. 2. Study 2: HT rating distributions of (a) music, (b) music / non-dialogue, (c) music / dialogue, (d) non-music, (e) non-music / non-dialogue, (f) non-music / dialogue, (g) all, (h) non-dialogue, and (i) dialogue scenes. The red trendlines represent second-degree polynomial fits to the data. For each scene type, the inter-rater agreement (ICC) value of the HT ratings is given in the top right corner of the corresponding figure.

in terms of mood, while in non-music scenes, this is only the case with the arousal dimensions. Lastly, they indicate that in music scenes, speech is not very important in terms of mood, while in non-music scenes, it is as important as style.

The particularly low influence of speech on mood in the music scenes may be partially explained by the fact that although the narrative was advanced mainly by means other than speech in 75% of both the music and non-music scenes, the music scenes contained on average fewer spoken words (43 words) than the non-music scenes (144 words). Overall, the low number of clips limits the generalizability of the results. Nevertheless, the results provide an answer to the second research question by suggesting that sound-based scene classification can bring out differences in the comparative influence of speech, visual style, and sounds on film mood, and that in music scenes, visual and auditory style can have as great an influence on mood as the scene's events.

C. Study 2: Mood ratings across scene types

To determine whether the participants' mood ratings differed by scene type, we analyzed the difference in the mean ratings within each scene type pair (e.g. the HT ratings of the dialogue and non-dialogue scenes) in terms of the Mann-Whitney test effect size r. Rating statistics for the three mood dimensions and for the four scene classification criteria (location, time, dialogue, music) are shown in Table V.

Table V shows that while several mood rating pairs exhibited statistically significant differences, the effect sizes were small in all but one case: the HT ratings of interior and exterior scenes ($|r| = 0.36$), the latter being rated more positive than the former. We therefore also examined the distributions of the ratings, seeking to determine whether they would reveal differences in mood rating patterns within the scene type pairs.

We found that both of the sound-based classification criteria, dialogue and music, produced divergent HT rating distributions in their respective scene type pairs: the dialogue scenes were rated differently from the non-dialogue scenes, and the music scenes from the non-music scenes. The HT rating distributions of these scene types are presented in Figure 2.

With the music scenes (Figure 2(a)), the ratings covered the entire range of HT values evenly, from extremely negative to extremely positive, while the non-music scenes (Figure 2(d)) had a Gaussian rating distribution, with relatively few extreme ratings. A similar effect occurred with the dialogue and non-

dialogue scenes, but in the opposite direction: the non-dialogue scenes (Figure 2(h)) exhibited a flat HT distribution that covered the entire range of values, while the dialogue scenes (Figure 2(i)) had a Gaussian rating distribution.

However, since there was overlap between the scene types (e.g. some music scenes were also non-dialogue scenes), we investigated the individual contributions of the two sound types, dialogue and music, on the rating distributions in more detail. We split both the music and non-music scenes into two subtypes: those that were also non-dialogue scenes, and those that were also dialogue scenes. The HT rating distributions of these subtypes are shown in Figures 2(b) and 2(c) (top row) for the music scene subtypes, and Figures 2(e) and 2(f) (middle row) for the non-music scene subtypes.

Two of the subtypes, music / dialogue scenes (Figure 2(c)) and non-music / non-dialogue scenes (Figure 2(e)), contained a low number of clips (four and six, respectively). These scene types are understandably rare, since most scenes in mainstream films contain either music or dialogue, and when both are present, the other tends to assume a dominant role so that the two sound types do not compete for the viewers' attention. In any case, because of the low number of clips, we cannot draw definitive conclusions about these subtypes. We can, however, discuss the other two subtypes, music / non-dialogue scenes (Figure 2(b), 25 clips) and non-music / dialogue scenes (Figure 2(f), 15 clips), with more confidence.

The subtype distributions revealed that the music / non-dialogue scenes (Figure 2(b)) – that is, scenes in which the soundtrack was dominated by music – had a similarly flat HT rating distribution as the set of all music scenes (Figure 2(a)). This characteristic distinguished the subtype from the non-music / dialogue scenes (Figure 2(f)); that is, scenes in which the soundtrack was dominated by dialogue. The dialogue-dominated scenes had a Gaussian HT rating distribution, just like the set of all non-music scenes (Figure 2(d)). In other words, the scenes dominated by music seemed to be just as likely to feature extreme HT ratings as neutral ratings, while extreme HT ratings were rare in the scenes dominated by dialogue. The participants were also in greater agreement about their HT ratings in the case of the music-dominated scenes ($\text{ICC}_{\text{HT, music / non-dialogue}} = 0.77$) than the dialogue-dominated scenes ($\text{ICC}_{\text{HT, non-music / dialogue}} = 0.58$). In other words, the music-dominated scenes exhibited more between-clip variance but less between-participant variance in terms of HT.

In an answer to the third research question, the scene types did not notably differentiate the mood ratings in terms of mean ratings, but did differentiate them in terms of their distribution shapes. In particular, the music-dominated scenes and dialogue-dominated scenes had distinctive mood profiles in terms of HT: the mood of the music-dominated scenes was likely to be anything from extremely negative to extremely positive, while the mood of the dialogue-dominated scenes tended to be more muted.

D. Study 2: Determinants of mood ratings

Pearson correlations between the mood ratings and the human-rated stylistic attributes as well as the computational features are shown for all the scene types in Tables VI (setting-based classification) and VII (sound-based classification). Both tables first list the correlations with the stylistic attributes, then the low-level features, and lastly, the high-level features.

The shot duration (Section VI-A3) correlations listed in the tables were computed using the median shot duration, which performed better than the mean. Also, the only face emotion feature (Section VI-B1) included in the tables is the happiness feature ("face happiness") since it was the only one to produce meaningful results. For the same reason, the only included Tone Analyzer dialogue emotion feature (Section VI-B2) is the joy feature ("dialogue joy").

Looking first at the correlations between the human-rated stylistic attributes and the mood ratings in Tables VI and VII, we notice that the HT and TA ratings were not related to the stylistic attributes in the general case (all 50 clips): all the correlations were small except for the one between colorfulness and HT, and even this correlation was not reflected with all the scene types. The EA ratings, however, produced large or moderate correlations with the loudness and fastness attributes in the general case as well as with each individual scene type. We also found a general trend in terms of the scene types: all the EA correlations were stronger with the external and daytime scenes than the interior and nighttime scenes.

The results were similar with the low-level computational features: the HT and TA ratings produced almost exclusively small correlations, whereas the EA ratings produced quite a few moderate and large correlations. Interestingly, the saturation feature completely failed to reflect the aforementioned colorfulness–HT correlation, suggesting that a more sophisticated computational correlate of this attribute is needed. On the other hand, in the case of EA, the large correlations with loudness and fastness were reflected in the correlations with audio RMS and loudness, and pixel motion and character movement, respectively. All these features produced moderate or large correlations with the EA ratings in the general case as well as with each scene type in the sound-based classification. Of the two audio features, audio loudness fared slightly better than audio RMS; in fact, its correlation matched that of the perceptual loudness attribute. Of the two motion features, character movement fared slightly better, though the features' performances were mostly similar. Also, as expected, shorter shot durations (i.e. faster editing tempo) did correlate with greater EA ratings in the general case. Still, the shot duration feature was consistently outperformed by both of the two motion features. Overall, as with the perceptual attributes, almost all the correlations between the low-level features and the EA ratings were stronger with the external and daytime scenes. They were also uniformly stronger with the non-dialogue and music scenes than the dialogue and non-music scenes, respectively.

The high-level features complemented the low-level features exactly, correlating strongly with the HT and TA ratings and poorly with the EA ratings. In the general case, all six high-level features produced moderate correlations with the HT ratings, and four of them (face happiness, dialogue joy, music-EA, and music-TA) also with the TA ratings. Face happiness performed the best overall, producing moderate or large corre-

TABLE VI
PEARSON CORRELATIONS OF STYLISTIC ATTRIBUTES AND COMPUTATIONAL FEATURES WITH MOOD RATINGS
(SETTING-BASED SCENE CLASSIFICATION)

Attribute / feature	Hedonic tone					Energetic arousal					Tense arousal				
	all	int	ext	day	night	all	int	ext	day	night	all	int	ext	day	night
Brightness	0.26	0.21	0.15	**0.32**	0.15	0.17	-	0.25	0.15	-	−0.17	−0.19	-	−0.14	-
Colorfulness	**0.36**	**0.32**	0.27	**0.40**	0.29	0.13	-	0.17	0.22	-	−0.20	−0.28	-	−0.12	−0.29
Loudness	-	-	0.18	0.14	-	**0.51**	**0.38**	**0.56**	**0.57**	**0.34**	0.18	0.13	0.28	0.26	-
Fastness	0.11	-	0.15	0.18	-	**0.63**	**0.51**	**0.67**	**0.66**	**0.52**	0.27	0.18	**0.38**	**0.33**	0.23
Lightness	0.20	-	-	0.28	-	0.28	0.14	**0.43**	**0.32**	-	−0.10	-	-	-	-
Shadow prop.	−0.20	-	-	**−0.34**	-	−0.27	−0.14	**−0.44**	**−0.34**	-	0.11	-	-	-	-
Saturation	-	0.22	−0.20	-	-	**−0.32**	−0.23	**−0.33**	−0.27	**−0.37**	−0.11	−0.21	-	−0.15	-
Audio RMS	0.17	-	0.24	**0.30**	−0.15	**0.47**	**0.31**	**0.56**	**0.52**	**0.32**	0.10	-	0.25	0.12	-
Audio loudness	0.13	-	0.28	0.24	-	**0.52**	**0.33**	**0.62**	**0.58**	**0.36**	0.14	-	0.28	0.21	-
Shot duration	-	-	-	-	-	**−0.33**	−0.29	−0.27	**−0.37**	−0.28	−0.19	-	**−0.37**	**−0.38**	-
Pixel motion	0.20	0.20	-	0.19	0.18	**0.45**	**0.46**	**0.39**	**0.41**	**0.48**	-	-	0.21	0.15	-
Char. movement	0.12	-	-	0.21	-	**0.51**	**0.42**	**0.51**	**0.51**	**0.44**	0.11	0.18	0.24	0.18	-
Face happiness	**0.47**	0.29	**0.43**	**0.54**	**0.32**	-	−0.20	-	-	−0.28	**−0.40**	**−0.32**	**−0.41**	**−0.36**	**−0.51**
Dial. sentiment	**0.39**	0.22	**0.30**	**0.46**	0.26	0.19	-	-	0.22	-	−0.16	-	-	−0.13	−0.23
Dial. joy	**0.36**	**0.36**	0.22	0.28	**0.62**	-	-	−0.26	-	-	**−0.34**	−0.14	**−0.39**	−0.28	**−0.51**
Music-HT	**0.42**	-	**0.53**	**0.58**	-	**0.40**	0.23	**0.49**	**0.42**	0.17	−0.26	−0.18	−0.22	−0.27	-
Music-EA	**0.40**	**0.47**	**0.37**	**0.41**	**0.57**	-	-	0.19	-	-	**−0.38**	**−0.50**	**−0.31**	**−0.40**	**−0.42**
Music-TA	**−0.40**	**−0.52**	**−0.37**	**−0.38**	**−0.62**	-	-	-	-	-	**0.41**	**0.51**	**0.34**	**0.43**	**0.38**

int = interior scenes, ext = exterior scenes, day = daytime scenes, night = nighttime scenes. Only correlations significant at $p < 0.01$ after Bonferroni correction and based on data from at least 10 clips are listed. Moderate and large correlations ($|r| \geq 0.30$) are in boldface.

TABLE VII
PEARSON CORRELATIONS OF STYLISTIC ATTRIBUTES AND COMPUTATIONAL FEATURES WITH MOOD RATINGS
(SOUND-BASED SCENE CLASSIFICATION)

Attribute / feature	Hedonic tone					Energetic arousal					Tense arousal				
	all	D	nD	M	nM	all	D	nD	M	nM	all	D	nD	M	nM
Brightness	0.26	**0.32**	0.22	0.27	0.23	0.17	-	0.20	0.22	-	−0.17	−0.23	−0.13	−0.14	−0.23
Colorfulness	**0.36**	**0.45**	**0.31**	**0.39**	**0.30**	0.13	-	0.17	0.13	-	−0.20	**−0.33**	−0.12	−0.21	−0.18
Loudness	-	−0.14	0.16	0.15	−0.16	**0.51**	**0.49**	**0.49**	**0.50**	**0.51**	0.18	0.19	0.19	0.18	0.19
Fastness	0.11	-	0.17	0.14	-	**0.63**	**0.57**	**0.64**	**0.64**	**0.60**	0.27	0.25	**0.30**	0.29	0.22
Lightness	0.20	0.13	0.18	0.24	-	0.28	-	**0.31**	**0.34**	-	−0.10	−0.21	-	-	−0.17
Shadow prop.	−0.20	−0.26	−0.14	−0.21	−0.17	−0.27	-	**−0.30**	**−0.32**	-	0.11	0.23	-	-	0.18
Saturation	-	-	-	-	0.20	**−0.32**	−0.23	**−0.34**	**−0.36**	−0.23	−0.11	-	−0.18	-	-
Audio RMS	0.17	-	0.16	0.25	−0.19	**0.47**	**0.39**	**0.48**	**0.50**	**0.45**	0.10	-	0.15	-	-
Audio loudness	0.13	−0.25	0.20	**0.31**	−0.28	**0.52**	**0.32**	**0.56**	**0.56**	**0.41**	0.14	-	0.20	-	0.16
Shot duration	-	0.23	-	−0.17	0.22	**−0.33**	-	**−0.43**	**−0.44**	−0.25	−0.19	−0.15	−0.22	−0.24	−0.17
Pixel motion	0.20	0.24	0.13	0.24	-	**0.45**	**0.41**	**0.43**	**0.46**	**0.41**	-	-	0.13	-	-
Char. movement	0.12	-	-	0.21	-	**0.51**	**0.47**	**0.49**	**0.54**	**0.48**	0.11	-	0.19	-	0.14
Face happiness	**0.47**	**0.59**	**0.47**	**0.46**	**0.59**	-	-	-	0.19	-	**−0.40**	**−0.48**	**−0.38**	**−0.34**	**−0.50**
Dial. sentiment	**0.39**	**0.42**	**0.31**	**0.37**	**0.44**	0.19	-	-	0.25	-	−0.16	−0.25	-	-	**−0.35**
Dial. joy	**0.36**	0.21	**0.35**	**0.54**	**0.35**	-	−0.17	−0.17	-	−0.21	**−0.34**	**−0.30**	**−0.35**	**−0.33**	**−0.41**
Music-HT	**0.42**	n/a	**0.43**	**0.42**	n/a	**0.40**	n/a	**0.47**	**0.45**	n/a	−0.26	n/a	−0.25	−0.24	n/a
Music-EA	**0.40**	n/a	**0.40**	**0.39**	n/a	-	n/a	0.14	0.14	n/a	**−0.38**	n/a	**−0.39**	**−0.37**	n/a
Music-TA	**−0.40**	n/a	**−0.41**	**−0.40**	n/a	-	n/a	-	-	n/a	**0.41**	n/a	**0.42**	**0.41**	n/a

D = dialogue scenes, nD = non-dialogue scenes, M = music scenes, nM = non-music scenes. Only correlations significant at $p < 0.01$ after Bonferroni correction and based on data from at least 10 clips are listed (n/a = insufficient clips). Moderate and large correlations ($|r| \geq 0.30$) are in boldface.

lations with all mood dimensions and scene types expect one (HT, interior scenes). This is an understandable result since practically all film scenes contain faces, but not all contain dialogue or music. The face happiness feature's correlations were positive with HT and negative with TA, indicating, quite understandably, that happy faces were associated with a positive mood and a lack of happy faces with a tense mood.

In the general case, the dialogue sentiment feature correlated well with the HT ratings and dialogue joy with both the HT and TA ratings. Interestingly, dialogue joy performed better with the non-dialogue scenes than the dialogue scenes. The TA correlations of both features were negative, indicating a relation between negative dialogue contents and a tense mood.

The three music emotion features correlated well with both HT and TA overall, though we could not assess their performance with the dialogue and non-music scenes due to an insufficient number of clips. Interestingly, each feature correlated with two mood dimensions: all three features correlated with the HT ratings, music-HT also correlated with the EA ratings, and both music-EA and music-TA also correlated with the HT

TABLE VIII
PERFORMANCE OF COMPUTATIONAL-FEATURE-BASED LINEAR REGRESSION MODELS OF FILM MOOD

Covariates	Adjusted R^2		
	HT	EA	TA
Low-level features	0.09	0.37	0.07
High-level features	0.44	0.31	0.26
Low-level & high-level features	0.62	0.46	0.35

and TA ratings. This complementary behavior also occurred with scene types: though the correlations between music-HT and the HT ratings were not significant with the interior and nighttime scenes, music-TA produced the strongest of all the HT rating correlations with just these scene types.

In an answer to the fourth research question, the results showed that the EA dimension was associated more strongly with specific stylistic attributes and their corresponding low-level features, whereas HT and TA were associated more strongly with high-level features related to emotional expression in faces, dialogue, and music. Our test of the general performance of the two feature types in linear regression models supported this finding (Table VIII): of models constructed with either low- or high-level features as covariates, the former performed better with EA, and the latter with both HT and TA. Understandably, models containing both types of features performed best with each mood dimension.

VIII. DISCUSSION

The results provide insight into film mood assessment, the influence of narrative and style on mood, and the use of scene classification and high-level features in mood estimation. The implications of the study's main findings are discussed below.

The similarity between the mood ratings produced by the direct and indirect assessment methods indicates that film mood is an intuitive concept to viewers and that indirect assessment is not required for accurate results. This result allows future assessments to be conducted in a more streamlined fashion, which should facilitate the collection of ground-truth mood data for the computational estimation of film affect. With the two arousal dimensions, though, indirect assessment exhibited slightly greater inter-rater agreement. This is understandable since the energeticness and tenseness of mood are less familiar concepts than positiveness. A single arousal scale, from low to high arousal, would probably be more intuitive, but would also result in the loss of the additional information provided by two arousal dimensions, as we have previously shown [22]. In all, the strong correlations between the two assessment methods across each dimension suggest that direct assessment of a three-dimensional representation of film mood is feasible.

Our analysis of the influence of various attribute groups on film mood confirmed a certain primacy of narrative attributes: the events depicted in the scene are always crucial in terms of its mood. Of course, the narrative and stylistic aspects of a film are always intertwined [15] – put simply, style affects how events are perceived. Still, the result does serve as empirical support for the view that in narrative film, the on-screen action itself takes on an affective quality [10]. This phenomenon was further illustrated by the strong correlations we found between the HT and TA ratings and computational features related to such narrative attributes as acting (face happiness) and dialogue (dialogue sentiment, dialogue joy).

In light of our results on the mood rating distributions across scene types, dialogue-dominated scenes do not appear to be as likely to feature extremely negative or positive moods as music-dominated scenes. Dialogue is certainly often used to express and elicit emotions [29] in film, just like music is, but if we consider dialogue only in terms of its prominence in a scene, regardless of its content or tone, it becomes possible to see how it may actually be associated most commonly with muted moods. If film music serves to make the emotional content of a scene more salient [50], then dialogue, insofar as its primary purpose is to communicate information relevant to the narrative [29], can be seen to have an opposite effect, accentuating cognitive aspects of the scene over affective ones.

The increased likelihood of extreme moods in music-dominated scenes is unlikely to be caused by music alone, though. Rather, keeping in mind the tendency for aesthetic and affective coherence in mainstream film, i.e. the simultaneous use of several audiovisual cues to achieve a "unified and coherent" mood [51, p. 157], the effect is likely to be the result of many concurrent stylistic factors. In other words, prominent music is most commonly used in scenes that also utilize other elements of style, such as shot compositions, camera movement, and editing, for aesthetic and emotional expression. Indeed, our result regarding the greater influence of both visual and auditory style on the mood of music scenes supports this very idea. Still, film scenes are easier to classify based on the prominence of dialogue and music than many other stylistic attributes (Section IV-B), and automatic sound type classification has previously been shown to be feasible [12], [52]. Also, in the current study we found high-level features targeting the emotional expression in dialogue and music to be closely related to film mood. Sound-based scene classification may therefore prove useful as a practical way to classify film scenes by their affective expression.

Setting-based classification also showed promise in the current study: the correlations between the EA ratings and the computational features were stronger with the external and daytime scenes. Indeed, it is understandable that the lesser physical constraints of exterior scenes in particular would allow for a wider range of stylistic expression in terms of loudness and fastness, which were most closely associated with the EA ratings here. Overall, the results encourage future studies on computational film mood estimation to consider the use of both setting- and sound-based scene classification.

The correlations indicate that low-level features alone cannot estimate all three mood dimensions. On the other hand, high-level features related to aspects of emotional expression turned out to be most useful, in most cases exceeding even the human-rated stylistic attributes in terms of their correlations with the HT and TA ratings. While we cannot say how closely these features would match with corresponding human ratings (e.g. perceived face happiness), their sheer number of strong correlations with mood ratings suggests that the

emotions expressed in faces, dialogue, and music influence a film scene's overall valence and tenseness.

With regard to faces, the result indicates that the valence of a film scene is often reflected in the valence of its characters. This finding supports the theory that mainstream cinema strives for coherence, wherein the characters, dialogue, and stylistic devices work together to establish a coherent mood [10]. Also, the fact that TA ratings were related to a lack of detected happy faces, but not to an abundance of sad faces, suggests that a lack of expressed happiness is a characteristic of tense scenes, though it could also be explained by smiles being easier to detect computationally. In any case, using the clip's audio track to guide the selection of frames to analyze for facial expressions [53] might improve the performance of the face emotion features. For example, errors caused by mouth movements during speech could be avoided by analyzing the facial expressions only in non-dialogue segments.

The strong performance of the dialogue features indicates that text-based sentiment analysis, which has found many applications in such contexts as social media, news, and commerce [54], can also be effectively used for film affect analysis. A future study could utilize recent deep-learning-based multimodal fusion techniques (e.g. [55]) in order to incorporate accompanying information such as facial expressions, body language, and intonation into the analysis.

Lastly, the strong correlations of each music emotion feature with two distinct mood dimensions could be a reflection of the emotional coherence of mainstream film music, e.g. in the sense that the music in positive film scenes tends to be not just positive, but also energetic. This explanation is supported by the fact that film music is rarely consciously attended to by viewers and needs simplicity and clarity to come across through the film's other narrative and stylistic attributes [56].

Together, these speculations amount to a call for further study of the perceptual relations between a film scene's individual aspects of emotional expression and its overall mood. At the same time, our results indicate that the computational estimation of film affect can benefit from scene type classification and the use of high-level features. A follow-up study could investigate which combination of low- and high-level features would work best in a mood prediction model, or whether deep learning methods should be favored instead to avoid relying on individual features that may not apply to all types of content. Alternatively, a combination of the two approaches might lead to best results, as suggested recently by [57].

IX. Conclusions

We conducted a study on the quantitative determinants of film mood. We found that direct assessment of mood in terms of three dimensions – hedonic tone (HT), energetic arousal (EA), and tense arousal (TA) – produced similar ratings as the more elaborate indirect assessment, indicating that it is a feasible assessment method. We also found that the film scenes in which music played a more prominent role exhibited more extreme moods in terms of valence, as well as a stronger influence of visual and auditory style on all three mood dimensions. Both setting- and sound-based scene classification brought out differences in the correlations of mood ratings with stylistic attributes and computational features. Lastly, the low- and high-level computational features complemented each other: the style-based low-level features correlated well only with the EA ratings, while the state-of-the-art high-level features correlated well with both the HT and TA ratings.

The results indicate that studies on affective content in film would benefit from taking the distinctions between scene types into account. In a promising finding in terms of computational estimation of affect, features that estimate loudness and motion (in the case of EA) and face happiness (in the case of HT and TA) appear to perform well with various scene types.

Acknowledgment

This work was supported by the Finnish Cultural Foundation, the Media Industry Research Foundation of Finland, and the Academy of Finland's Finnish Centre of Excellence in Computational Inference Research (COIN). The calculations utilized computer resources from the Aalto University School of Science "Science-IT" project.

References

[1] E. S. Tan, *Emotion and the Structure of Narrative Film: Film as an Emotion Machine*. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.
[2] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 410–430, 2015.
[3] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.
[4] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, "Comparative analysis of low-level visual features for affective determination of video clips," in *Proc. IEEE Int. Conf. Future Information Technology*, 2010, pp. 1–6.
[5] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.
[6] R. Parke, E. Chew, and C. Kyriakakis, "Quantitative and visual analysis of the impact of music on perceived emotion of film," *Computers in Entertainment*, vol. 5, no. 3, pp. 1–21, 2007.
[7] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *Proc. IEEE Int. Conf. Image Processing*, 2009, pp. 1853–1856.
[8] 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, 2011.
[9] J. Tarvainen, S. Westman, and P. Oittinen, "Stylistic features for affect-based movie recommendations," in *Proc. Int. Workshop on Human Behavior Understanding*, 2013, pp. 52–63.
[10] C. Plantinga, "Art moods and human moods in narrative cinema," *New Literary History*, vol. 43, no. 3, pp. 455–475, 2012.
[11] R. Sinnerbrink, "Stimmung: Exploring the aesthetics of mood," *Screen*, vol. 53, no. 2, pp. 148–163, 2012.
[12] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.
[13] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf, "Visual movie analytics," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2149–2160, 2016.
[14] K. L. Brunick, J. E. Cutting, and J. E. DeLong, "Low-level features of film: What they are and why we would be lost without them," in *Psychocinematics: Exploring Cognition at the Movies*, A. P. Shimamura, Ed. New York, NY: Oxford University Press, 2013, ch. 7, pp. 133–148.
[15] D. Bordwell and K. Thompson, *Film Art: An Introduction*, 8th ed. New York, NY: McGraw-Hill, 2008.
[16] D. Bordwell, *The Way Hollywood Tells It: Story and Style in Modern Movies*. Berkeley, CA: University of California Press, 2006.
[17] A. Mackendrick, *On Film-Making: An Introduction to the Craft of the Director*. New York, NY: Faber and Faber, 2006.
[18] P. Valdez and A. Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology: General*, vol. 123, no. 4, pp. 394–409, 1994.

[19] K. Pearlman, *Cutting Rhythms: Shaping the Film Edit*. Burlington, MA: Focal Press, 2009.

[20] R. F. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss, "Emotion processing in three systems: The medium and the message," *Psychophysiology*, vol. 36, no. 5, pp. 619–627, 1999.

[21] G. Ilie and W. F. Thompson, "A comparison of acoustic cues in music and speech for three dimensions of affect," *Music Perception*, vol. 23, no. 4, pp. 319–330, 2006.

[22] J. Tarvainen, S. Westman, and P. Oittinen, "The way films feel: Aesthetic features and mood in film," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 9, no. 3, 2015.

[23] L. F. Barrett, M. Lewis, and J. M. Haviland-Jones, Eds., *Handbook of Emotions*, 4th ed. New York, NY: Guilford Press, 2016.

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

[25] U. Schimmack and R. Reisenzein, "Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation," *Emotion*, vol. 2, no. 4, pp. 412–417, 2002.

[26] G. Matthews, D. M. Jones, and A. G. Chamberlain, "Refining the measurement of mood: The UWIST Mood Adjective Checklist," *British Journal of Psychology*, vol. 81, no. 1, pp. 17–42, 1990.

[27] T. J. Smith, "Watching you watch movies: Using eye tracking to inform cognitive film theory," in *Psychocinematics: Exploring Cognition at the Movies*, A. P. Shimamura, Ed. New York, NY: Oxford University Press, 2013, ch. 9, pp. 165–191.

[28] R. J. Ellis and R. F. Simons, "The impact of music on subjective and physiological indices of emotion while viewing films," *Psychomusicology*, vol. 19, pp. 15–40, 2005.

[29] S. Kozloff, *Overhearing Film Dialogue*. Berkeley, CA: University of California Press, 2000.

[30] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. ACL Conf. Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[31] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[32] O. Lartillot and P. Toiviainen, "A MATLAB toolbox for musical feature extraction from audio," in *Proc. Int. Conf. Digital Audio Effects*, 2007, pp. 1–8.

[33] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2009, pp. 621–626.

[34] A. A. Alatan, A. N. Akansu, and W. Wolf, "Comparative analysis of hidden Markov models for multi-modal dialogue scene indexing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 4, 2000, pp. 2401–2404.

[35] A. Muhammad and S. M. Daudpota, "Content based identification of talk show videos using audio visual features," in *Proc. Int. Conf. Machine Learning and Data Mining in Pattern Recognition*, 2016, pp. 267–283.

[36] J. Tarvainen, J. Laaksonen, and T. Takala, "Computational and perceptual determinants of film mood in different types of scenes," in *Proc. IEEE Int. Symp. Multimedia*, 2017, pp. 185–192.

[37] D. Bordwell, "Intensified continuity: Visual style in contemporary American film," *Film Quarterly*, vol. 55, no. 3, pp. 16–28, 2002.

[38] Y. Baveye, J.-N. Bettinelli, E. Dellandra, L. Chen, and C. Chamaret, "A large video database for computational models of induced emotion," in *Proc. IEEE Int. Conf. Affective Computing and Intelligent Interaction*, 2013, pp. 13–18.

[39] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists," *Cognition and Emotion*, vol. 30, no. 5, pp. 827–856, 2016.

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

[41] International Telecommunication Union, *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*, ITU Std. Rec. ITU-R BS.1770-3, 2012.

[42] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[43] P. Sand and S. Teller, "Particle video: Long-range video motion estimation using point trajectories," *Int. Journal of Computer Vision*, vol. 80, no. 72, 2008.

[44] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[45] C. Lu and X. Tang, "Surpassing human-level face verification performance on LFW with GaussianFace," in *Proc. AAAI Conf. Artificial Intelligence*, 2015, pp. 3811–3819.

[46] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. London, UK: Springer, 2015.

[47] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[48] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[49] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment*, vol. 6, no. 4, pp. 284–290, 1994.

[50] N. Carroll, *Theorizing the Moving Image*. New York, NY: Cambridge University Press, 1996.

[51] C. Plantinga, *Moving Viewers: American Film and the Spectator's Experience*. Berkeley, CA: University of California Press, 2009.

[52] M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu, "A three-level framework for affective content analysis and its case studies," *Multimedia Tools and Applications*, vol. 70, no. 2, pp. 757–779, 2014.

[53] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, 2013.

[54] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[55] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[56] A. J. Cohen, "Music as a source of emotion in film," in *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda, Eds. New York, NY: Oxford University Press, 2001, ch. 11, pp. 249–272.

[57] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: performance for emotion prediction in videos," in *Proc. IEEE Int. Conf. Affective Computing and Intelligent Interaction*, 2015, pp. 77–83.

Jussi Tarvainen received the D.Sc. (Tech.) degree in media technology from the Aalto University School of Science, Finland, in 2017. His doctoral thesis examines the perceptual assessment and computational modeling of film mood. His research interests include cognitive film studies, content-based multimedia analysis, and computational aesthetics.

Jorma Laaksonen received the D.Sc. (Tech.) degree in 1997 from the Helsinki University of Technology, Finland, and is presently a senior university lecturer at the Department of Computer Science, Aalto University School of Science. His research interests are in content-based multimodal information retrieval, machine learning and computer vision. He is an Associate Editor of Pattern Recognition Letters, IEEE senior member, and a founding member of the SOM and LVQ Programming Teams and the PicSOM Development Group.

Tapio Takala is a professor of computer science at the Aalto University School of Science. His research interests include virtual reality, affective computing, embodied/enactive interfaces, and computational creativity, with applications in art and entertainment.