Content-Based Prediction of Movie Style, Aesthetics, and Affect: Data Set and Baseline Experiments

Jussi Tarvainen, *Student Member, IEEE*, Mats Sjöberg, *Graduate Student Member, IEEE*, Stina Westman, Jorma Laaksonen, *Senior Member, IEEE*, and Pirkko Oittinen

***Abstract*—The affective content of a movie is often considered to be largely determined by its style and aesthetics. Recently, studies have attempted to estimate affective movie content with computational features, but results have been mixed, one of the main reasons being a lack of data on perceptual stylistic and aesthetic attributes of film, which would provide a ground truth for the features. The distinctions between energetic and tense arousal as well as perceived and felt affect are also often neglected. In this study, we present a data set of ratings by 73 viewers of 83 stylistic, aesthetic, and affective attributes for a selection of movie clips containing complete scenes taken from mainstream movies. The affective attributes include the temporal progression of perceived and felt valence and arousal within the clips. The data set is aimed to be used to train algorithms that predict viewer assessments based on low-level computational features. With this data set, we performed a baseline study modeling the relation between a large selection of low-level computational features (i.e., visual, auditory, and temporal) and perceptual stylistic, aesthetic, and affective attributes of movie clips. Two algorithms were compared in a realistic prediction scenario: linear regression and the neural-network-based Extreme Learning Machine (ELM). Felt and perceived affect as well as stylistic attributes were shown to be equally easy to predict, whereas the prediction of aesthetic attributes failed. The performance of the ELM predictor was overall found to be slightly better than the linear regression. A feature selection experiment illustrated that features from all low-level computational modalities, visual, auditory and temporal, contribute to the prediction of the affect assessments. We have made our assessment data and extracted computational features publicly available.**

***Index Terms*—Aesthetics, content-based analysis, felt affect, film, machine learning, modeling, perceived affect, style.**

I. Introduction

COMPUTATIONAL content-based movie analysis provides a way to describe, summarize and recommend movies automatically without subjective interpretations [1] or genre classification [2], which often fail to describe the stylistic, aesthetic and affective topology of movies accurately. These elements of film expression are interrelated [3], and in particular, the affective content of a movie is often considered to be largely determined by its style and aesthetics [4], [5]. It has been claimed that style, aesthetics and affect communicate more strongly with viewers than high-level semantic concepts such as plot or theme [6]. As such, they are crucial in terms of the overall artistic impression made by a movie [7].

Recently, several studies (e.g. [8], [9]) have attempted to estimate affective movie content with computational features based on attributes of film style. Computational models of affective content would be helpful in movie classification and recommendation systems, even the most sophisticated of which still rely on user-generated metadata such as star ratings [10]. Results have, however, been mixed, and three possible reasons can be suggested: first, a lack of data on perceptual stylistic and aesthetic movie content; second, simplistic modeling of affective content; and third, direct operationalization of perceptual stylistic attributes as computational features.

The purpose of this study is to compare predictions of stylistic, aesthetic and affective movie content by low-level computational features that can be automatically extracted from a movie clip. Addressing the aforementioned shortcomings of previous studies, the prediction is based on ground truth data obtained in a subjective assessment of movie clips representing a wide range of stylistic, aesthetic and affective content. Previous studies have typically modeled affect with a two-dimensional valence–arousal affect space [11], where valence describes the pleasurability and arousal the alertness associated with an affect. In contrast, the current study utilizes a three-dimensional space [12] containing, in addition to the valence dimension, two distinct arousal dimensions, energetic arousal (from tired to awake) and tense arousal (from calm to tense), to obtain more detailed affect data. The study also takes into account the distinction between perceived and felt affect; that is, between the affect expressed in a movie and the viewer's affective response [13], which have been shown [14] to differ both in terms of inter-rater agreement and their relation to stylistic attributes. Lastly, the study avoids making *a priori* assumptions about the computational equivalents of specific perceptual stylistic attributes (e.g. by imposing a single computational measure for the attribute colorfulness). Two prediction methods are compared: multiple linear regression and the recent neural-network-based Extreme Learning Machine (ELM) [15] algorithm.

Manuscript received January 08, 2014; revised June 10, 2014; accepted September 09, 2014. Date of publication September 12, 2014; date of current version November 13, 2014. This work was supported in part by the *Next Media* Tivit SHOK project under Grants 255745 and 251170 of the Academy of Finland and the Doctoral Program in User-Centered Information Technology. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shin'ichi Satoh.

J. Tarvainen, S. Westman, and P. Oittinen are with the Department of Media Technology, Aalto University School of Science, Espoo 02150, Finland (e-mail: jussi.tarvainen@aalto.fi; stina.westman@aalto.fi; pirkko.oittinen@aalto.fi).

M. Sjöberg and J. Laaksonen are with the Department of Information and Computer Science, Aalto University School of Science, Espoo 02150, Finland (e-mail: mats.sjoberg@aalto.fi; jorma.laaksonen@aalto.fi).

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

Digital Object Identifier 10.1109/TMM.2014.2357688

The study aims to answer five research questions. First, which types of movie attributes–stylistic, aesthetic or affective–can be most effectively predicted by computational features? Second, is the prediction more effective for perceived or felt affect? Third, how do the predictions obtained with ELM compare to those obtained with linear regression? Fourth, how does the affective prediction performance change if one targets temporally shorter segments? And fifth, does feature selection affect the accuracy of affective content prediction?

The article is organized as follows. First, the theoretical basis for the collection of the ground truth data is discussed; style and aesthetics are covered in Section II and affect in Section III. Then, an overview of previous studies on affective movie content prediction, which provides the basis for the selection of the low-level features, is given in Section IV. The current study's assessment of the stylistic, aesthetic and affective content of movie clips in an experimental setup is presented in Section V, and the prediction of these assessments by computational features in Section VI. Results are discussed in Section VII and conclusions are drawn in Section VIII.

II. Style and Aesthetics

Style and aesthetics are related concepts in that the aesthetic composition of a movie is understood to be largely determined by its style [7], [16]. In this sense, style is an aspect of aesthetics, and indeed, some studies (e.g. [3]) conflate the terms by including stylistic impressions in their definition of aesthetics. In the context of the current study, however, the two concepts are distinguished by their degree of abstraction. Here, film style is defined as a set of audiovisual means of narration and elicitation of emotions based on specific techniques such as camerawork and the use of color. Aesthetics, on the other hand, covers more abstract impressions (e.g. beauty) that are irreducible to a specific technique. In essence, aesthetics is here used to describe, following [17], the aesthetic effects of movies, and style the means by which they are achieved.

A. Style

Film style can be split into three modalities: visual, auditory and temporal [18]. Since film is an evolving art form, an exhaustive list of its stylistic attributes cannot be given, but certain attributes can be considered central due to their repeated exposure in film theory textbooks (e.g. [4], [18]). For visual style, which concerns properties of the image, these include lighting, contrast, color and framing. For auditory style, which concerns properties of the soundtrack, they include loudness as well as the use of music, dialogue and sound effects. Lastly, for temporal style, which concerns variations of the visual and auditory attributes in time, they include, in addition to temporal manipulation of the aforementioned features, shot duration, shot motion and rhythm. These are discussed, along with their possible computational estimates, in Section IV-A.

Whereas much has been written about the relation between film emotion and narrative structure, theme and characters [1], [19], there is a notable lack of studies on the influence of specific stylistic attributes on affect in film. Still, relations between some of the aforementioned attributes and affect have been shown for other media. For example, valence has been shown to increase with image brightness and saturation [20] and with music tempo [21], and also to be affected in various ways by the rhythm and pitch of human speech [22]. Arousal has been shown to decrease with image brightness and increase with saturation [20], with the loudness and rate of speech [22], with the loudness of music (including an increase in tense arousal) [23] and with the intensity of video motion [24].

Previous studies indicate that both the visual and auditory modalities contribute to the elicitation of emotions in film. Music has been shown to be a particularly strong affect cue, having been found to influence the connotative meaning of a movie clip [25] and predict its perceived emotion ratings [26].

B. Aesthetics

Like style, film aesthetics defies exhaustive definition, partly because of the abstract nature of the concept itself, and partly because of the scarcity of studies on film-specific aesthetic terms. For practical purposes, however, the most commonly-used attributes in studies across various art forms can be identified. Chief among them is beauty, whose prevalence in studies on aesthetic terms has been described "the primacy of beauty" [27]. Other common attributes include complexity, comprehensibility (also called understandability), interestingness, naturalness, novelty (encompassing familiarity, predictability and suddenness) and pleasantness [3], [28]–[30].

The difficulty of defining a set of aesthetic attributes is compounded by domain-specificity in aesthetic word usage. However, according to [31], word usage in aesthetic descriptions of visual art can be generalized to film, with film-specific words mostly related to affect. Since aesthetics is kept distinct from affect (Section III) in the current study, it is assumed that the aforementioned attributes will suffice for its purposes.

Though aesthetics and affect are here studied separately, it is recognized that they are related concepts. Aesthetic appraisal of an artwork is thought to influence the emotions elicited by it [30], and some definitions of aesthetic impression encompass affective experience (e.g. [29], [32]). Lastly, according to [16], the overall aesthetic composition of a movie is largely determined by its mood (that is, by its affective content), which is in turn created with various "cinematic-aesthetic devices".

III. Affect

Affective science is concerned with the study of affect and emotion. From a psychological perspective, an emotion is a conscious affective state marked by cognitive appraisal, while affect is a broader category encompassing feelings, emotions and moods [33]. Emotion can thus be seen as a subset of affect. However, the two terms are often used interchangeably.

Relevant models of affect are presented in Section III-A. The distinction between perceived and felt affect is discussed in Section III-B.

A. Models of Affect

Several models of affect have been developed to distinguish between emotions and to determine how they relate to each other. These are typically based on either categorical or dimensional emotion theories [34]. Both approaches appear regularly in studies on affective movie content prediction (Section IV-B).

Categorical models of affect are based on a theory of emotions as discrete states governed by distinct neural processes [34]. Categorical models are typically concerned with so-called basic emotions [35], which differ from one another in terms of their physiology, appraisal and behavioral response. Various lists of basic emotions have been proposed, the most influential being Paul Ekman's list consisting of happiness, surprise, fear, anger, disgust–contempt and sadness [35].

A theory based on discrete emotions has been claimed to be ineffective in describing nuances between affective states [34]. As a result, categorical emotion theory has lost ground to dimensional emotion theory, which sees emotions as arising from overlapping neurophysiological systems, thought of as dimensions of a common affective space. The valence–arousal–control model of affect [11] has proven particularly influential. Valence (from negative to positive) describes the pleasurability of an affect, arousal (from calm to aroused) describes the alertness associated with it, and control describes the degree to which the affect is within the subject's control. Control has been shown to play only a minor role in affect [36], leading to the introduction of a simpler two-dimensional circumplex model of affect [37] on which all affective states are expressed as linear combinations of valence and arousal.

The use of a single arousal dimension was challenged in [12], with two distinct arousal dimensions proposed instead: energetic arousal (awake–tired) and tense arousal (calm–tense). This has led to the proposal of a three-dimensional model with one hedonic tone (valence) dimension and two arousal dimensions [38]. It was shown in [39] that energetic and tense arousal are independent and not mixtures of valence and a single arousal dimension, and that more accurate affect modeling can be achieved with two arousal dimensions.

B. Perceived and Felt Affect

An important distinction in affect modeling is that between perceived and felt affect; that is, between the affect expressed in a stimulus (such as a movie) as perceived by the viewer, and the viewer's personal affective response to the stimulus [40]. In this sense, perceived and felt affect are distinguished from one another by the objects of their focus: the former is a property of the *movie*, and the latter is a property of the *viewer*. The perceived affect of a movie can be thought of as its mood, that is, its "affective character" [41] that determines the "aesthetic composition of a cinematic world" [16] and facilitates emotional responses in the viewer [5].

The relation between the two types of affect has mostly been studied in musicology, with the general finding that perceived affect ratings are more objective–and thus more generalizable–than felt affect ratings. In [40], perceived and felt emotion ratings of classical music were found to be mostly similar, although felt emotion was stronger than perceived emotion in connection with valence and weaker in connection with arousal. Perceived affect showed greater inter-rater agreement, indicating that it is less sensitive to individual differences between viewers. On the other hand, in a recent multimedia study [42] perceived and felt basic emotion ratings of non-film video clips were found to be highly consistent.

Studies on the relation between perceived and felt affect in film are scarce, but recent results suggest that perceived affect shows more promise in terms of computational affective content prediction. In [43], an affect space mixing elements of perceived and felt affect was proposed to this end. The authors found greater inter-rater agreement with this space than with a quantized version of the circumplex model of affect [37]. Also, in [14], perceived and felt affect ratings of movie clips were compared on the hedonic tone–energetic arousal–tense arousal space proposed in [38]. Perceived affect ratings occupied a greater portion of the affect space and showed greater inter-rater agreement, thus providing better separation of affect ratings. The study also found stylistic attributes to have a stronger relation with perceived affect, suggesting that style-based computational prediction of affective movie content should focus on perceived affect. Lastly, in [44] it was suggested that perceived affect could be used to estimate an individual viewer's affective response with the help of a viewer profile, which would provide a mapping between the content of a movie and the preferences of the viewer.

IV. Affective Movie Content Prediction

A. Features

The theoretical foundations and performance of visual, auditory and temporal features used recently in studies on affective movie content prediction are discussed below. The reader is referred to the original publications for technical details. In the following, *frame* refers to a single image and *shot* to a sequence of frames between two edit points [18].

1) Visual Features: Visual features can be extracted from a single frame. They are typically related to either color or brightness, although shot size has also been used recently.

Color is considered one of the most expressive devices in visual narration [20]. It is used in film to direct attention, create associations and support narrative development [18]. Consequently, many studies on affective content prediction include color features, such as hue histogram [45], [46], dominant color [9], [47], color variance [46], [48], saturation [2], [45], [49] and color energy [2], [49]. Some of these also appear in the MPEG-7 standard [50]. However, the performance of color features has often been moderate at best. In [51], light source color was found to be poorly suited for valence modeling, and in [9], the removal of color features improved the quality of valence modeling. Also, in [2] color energy performed poorly overall, except when highly saturated colors were present.

Bright images are generally associated with pleasant emotions and dark images with unpleasant ones [18], [20], making brightness (or lightness in the CIE Luv space) an effective predictor of valence. Median lightness and a lightness histogram featured in [46], and in [49] brightness was one of the five visual features used. The proportion of pixels whose brightness falls below a threshold has sometimes been expressed as a shadow proportion measure [2], [46]. The product of the mean and standard deviation of a frame's grey level values can also be used to estimate its lighting key, ranging from bright, low-contrast lighting to dark, high-contrast lighting [48].

Framing, an important attribute of visual style, is determined by camera angle, level and height, as well as shot size [18]. However, only the last of these has so far been successfully modeled computationally [2], [52]. As such, framing remains largely unaccounted for in affective movie content prediction.

2) Aural Features: Since images and sounds engage separate senses, sounds can be used to influence the viewer's emotions independently of images [18]. Five auditory features are presented here: volume, zero-crossing rate, sound energy, frequency and mel-frequency cepstral coefficients.

Volume, which describes the perceived amplitude of sound, is used in movies to direct attention and create affective impressions [18]. Volume has a proven association with arousal [53], [54], but a volume feature is quite rare in computational studies. Still, it can be used to compute other features, such as volume standard deviation and dynamic range [47].

Zero-crossing rate is a simple but common [46], [47], [49] measure of the rate at which the signal changes sign between positive and negative. It has been used in speech recognition and shown to be able to characterize aspects of music, especially percussive sounds [55].

Sound energy is another widely-used [2], [46], [47], [49], [51], [56], [57] feature closely related to volume and often associated with arousal [44]. It is typically defined as the sum of the spectral values of an audio sequence's power spectrum. Thresholding allows it to be used in different ways; in [44], only frequencies above 700 Hz were considered, and in [2], thresholding was used to detect periods of silence.

The short-time Fourier transform of an audio sequence allows the computation of several frequency-domain features [58]. These include the spectral centroid and frequency bandwidth measures, which describe the "center of mass" of the frequency spectrum and the range of the sound's frequencies, respectively [46], [47], [49]. Both of these are related to the sound's timbre, that is, its tonal quality or "brightness" [59].

Mel-frequency cepstral coefficients (MFCC) represent the amplitudes of a sound's short-term power spectrum. They resemble properties of the human auditory system [47] and have proven effective in several studies [2], [56]. In [46], MFCCs were among the best valence and arousal estimates.

3) Temporal Features: In addition to temporal variations in visual and auditory features [2], certain computational features can be considered inherently temporal. Two such features are presented here: shot duration and shot motion, along with a discussion of their use as proxies for rhythm.

The primary form of film expression has been argued to be movement [60], created by editing and shot motion [18]. Editing influences shot duration and the change in space or time across shots [18]. Since detecting the latter requires a semantic interpretation of the content, it is not considered in most studies. Shot duration, however, has been included in practically all recent studies [2], [44]–[49], [56], [61]. It is most commonly expressed as mean duration, but median duration [45] and duration variance [62] have also been used.

Shot motion can result from either camera or object motion [18]. It is commonly computed between adjacent frames, typically with a pixel- or block-level difference measure [2], [9], [47], [48], [62] or motion vectors [44], [46], [51], [56], [63]. The name of the feature varies, with motion activity, intensity and magnitude all appearing regularly. In [9], the removal of the motion feature from the feature set had no significant effect, but in [2], a feature based on weighted pixel-based shot motion was ranked the best visual feature used.

Tempo, which operates on a slow–fast scale [64], is occasionally used as a proxy for the more complex concept of rhythm, which can be slow or fast, linear or non-linear, fluid or fitful [65], [66], and which is yet to be successfully estimated computationally. Both shot duration [44] and motion [67], or a weighted combination of both [51], have been used to model visual tempo. They can also be combined with an auditory tempo feature to model overall tempo [63].

B. Affect Modeling

Both categorical and dimensional models have been used in affective movie content prediction. With categorical models, computational features–or combinations thereof–are mapped directly to specific emotions. Ekman's basic emotions are common [2], [9], [62], [68], though emotions rarely represented in movies are sometimes omitted (e.g. disgust [2]).

Dimensional models, particularly the valence–arousal model, have also been widely adopted [8], [9], [44], [69], with generally better prediction results than with categorical models. In fact, many studies that use categorical models (e.g. [9], [62]) rely on dimensional models as an intermediate stage in the mapping between low-level features and discrete emotions. The popularity of dimensional models can be attributed to well-established methods for the acquisition of valence and arousal ratings, from self-assessment to psychophysiological measurements [34], and the successful mapping from dimensional data to Ekman's basic emotions (e.g. [2], [69]).

C. Prediction Methods

Basic affective content prediction has typically involved a regression model in which individual feature weights are either determined manually [44] or based on image parameters, such as a saliency map [63]. Recently, though, more advanced methods have become more common, including naïve Bayes classifiers [8], Dynamic Bayesian Networks [62], relevance vector machines [46], Hidden Markov Models (HMM) [9], [45], [70], Gaussian mixture models [69] and generalized state-space models [57]. Neural-network-based methods are more rare, but in [47], a multilayer Perceptron algorithm was applied to genre classification based on low-level features.

The lack of publicly available benchmarking data sets of movie clips makes comparing feature extraction and prediction methods difficult [71]. Since replicating other researchers' work is seldom possible, carrying out reliable comparisons is infeasible without common data sets. The recently-published LIRIS-ACCEDE data set [72] provides felt affect ratings for clips taken from movies shared under Creative Commons licenses, but none for commercially-released movies. As such, its applicability to feature extraction from, for example, mainstream Hollywood movies is questionable. The set also contains no ratings on stylistic or aesthetic attributes. In response to this need, we have made our assessment and computational feature

TABLE I
THE MOVIE CLIPS USED IN THE USER STUDY. TIMECODES ARE TAKEN FROM NTSC DVD RELEASES

#	Movie title	Year	Timecode [h:mm:ss]	Length [m:ss]	Shots
1	*500 Days of Summer*	2009	0:31:20	2:04	23
2	*Amelie*	2001	2:00:35	1:36	15
3	*Army of Shadows*	1969	0:38:40	1:54	15
4	*Before Sunrise*	1995	1:31:57	2:33	31
5	*Blue Velvet*	1986	1:55:32	2:21	16
6	*Children of Men*	2006	0:26:00	2:07	1
7	*Days of Heaven*	1978	0:04:05	1:37	16
8	*E.T.*	1982	1:47:42	1:10	25
9	*Punch-Drunk Love*	2002	1:06:30	1:16	7
10	*Raiders of the Lost Ark*	1981	0:07:45	2:09	59
11	*The Good, the Bad and the Ugly*	1966	2:45:49	2:17	61
12	*The Night of the Hunter*	1955	0:56:30	1:58	15
13	*The Shining*	1980	0:34:59	1:56	19
14	*Vertigo*	1958	0:26:00	1:45	18

data publicly available for other researchers to use in comparing their methods to our results (see Section V-E).

V. DATA COLLECTION

To obtain data on perceptual stylistic, aesthetic and affective content in movies, we conducted a user study whose participants were shown a series of movie clips and asked to assess their stylistic, aesthetic and affective attributes. These ratings are then used to train the algorithms used in the computational prediction experiment (Section VI). For a detailed discussion of the user study, the reader is referred to [14].

A. Participants

In all, 73 participants (44 women, 29 men, $M_{\text{age}} = 27.1$ years, $SD_{\text{age}} = 5.4$ years) took part in the study. Most were university students, from various fields. Thirty-eight percent were experts in film. A participant was considered an expert if he or she had studied film and/or had filmmaking as a hobby. Fluency in English was required of the participants since the movie clips were spoken or subtitled in English.

B. Movie Clips

The sample consisted of 14 movie clips 1–2.5 minutes in length (Table I). They were chosen from a set of 22 candidate clips based on their perceived affect ratings obtained in a pilot study. Due to the large number of participants in the user study, two sets of seven clips with similarly wide affect distributions were chosen, with the sets shown to different groups.

The clips were taken from mainstream movies made between 1955 and 2009 with an average of 180,000 IMDb ratings and an average rating of 8.15/10.[1] They encompassed several genres, such as action, drama, horror and romance. They also varied stylistically in terms of, for example, composition, colors, sound and editing. Nine clips (3–7, 10, 12–14) contained speech; ten (1, 5, 7–14) contained music. All clips were presented in their original language, clip 3 (in French) with subtitles. Each clip contained a complete scene that could be understood without knowledge of the preceding events.

TABLE II
USER STUDY GROUPS, SESSIONS, PARTICIPANTS AND CLIPS

Group	Session	N	Experts [%]	Female [%]	Clips in viewing order
1	Expert	15	67	73	2, 6, 4, 7, 12, 11, 1
	Non-expert	19	5	53	
2	Expert	20	75	55	8, 3, 9, 13, 14, 5, 10
	Non-expert	19	11	63	

C. Procedure

Participants were split into two groups, each viewing a distinct set of seven movie clips. A separate session was held targeting the expert and non-expert participants in either group, resulting in four sessions in all (Table II).

The study was conducted in a movie theater. Participants rated the clips in terms of perceived and felt affect as well as style and aesthetics. Felt affect was rated before perceived affect to prevent the latter, more objective, rating from influencing the former, and because the duration of an emotional response is limited [33].

D. Assessments

Participants used the UWIST Mood Adjective Checklist [38] to produce, for each clip, a rating of the affect expressed in it (perceived affect) and a rating of their personal affective response to it (felt affect), on a discrete scale of [1], [4]. Using the procedure in [38], we transformed the ratings into values in the hedonic tone (HT, corresponding to valence), energetic arousal (EA) and tense arousal (TA) dimensions. These ratings are on a scale of [-1, 1], from negative to positive, tired to energetic and calm to tense, respectively [12].

Using low-level features to detect events within movie scenes was recently shown to be feasible [73]. As these events can be expected to influence the affect ratings [5], we asked the participants to plot the progression of the perceived and felt affect within each clip in time–affect plots. This allowed us to study whether more temporally detailed affect data can improve the prediction of the affective content of the clips.

Each participant drew two curves for both perceived and felt affect: one for hedonic tone (from negative to positive) and one for general arousal (from low to high). We scanned the hand-drawn curves and transformed them to discrete numeric values on a scale of [-1, 1]. Since participants used the scales differently, we normalized each participant's curves to use the full scale, separately for hedonic tone and arousal. An example of the averaged hedonic tone and arousal curves is shown for clip 6 in Fig. 1. The onset of a violent event 25 seconds (20%) into the clip is reflected in the perceived hedonic tone curve [Fig. 1(a)] by a sudden drop, and by a corresponding increase in arousal [Fig. 1(c)]. The effect is similar, but less clearly defined, for felt affect [Figs. 1(b), 1(d)].

For the computational prediction, we split the curves for each clip into four temporal segments of equal duration and computed mean values for each segment across participants. These are used to compare the accuracy of segment-based clip affect prediction to prediction based on movie-wise affect ratings. For this purpose, we also transformed participants' affect ratings into general arousal values, from [-1, 1] [38].

[1][Online]. Available: http://www.imdb.com.

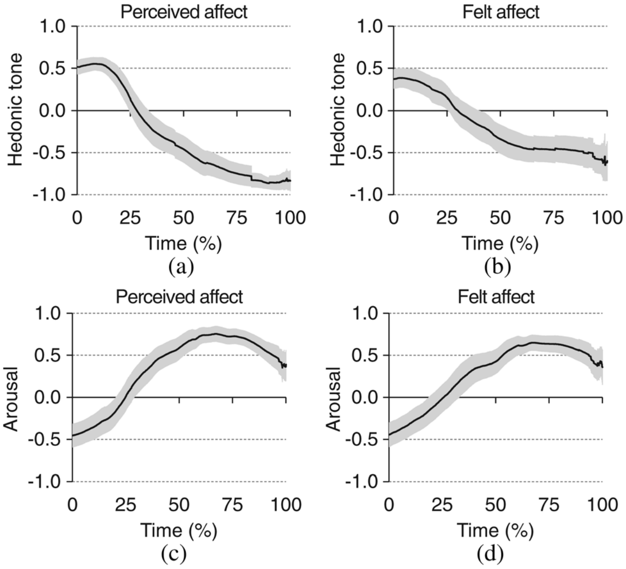

Fig. 1. Averaged affect plots for clip 6. Top row: hedonic tone (from negative to positive) of (a) perceived and (b) felt affect. Bottom row: arousal (from low to high) of (c) perceived and (d) felt affect. The grey areas denote 95% confidence intervals.

Since the curves were drawn after the viewing of each clip, they are limited in their temporal precision, making them unsuitable for determining absolute affect changes at specific moments in time. However, by studying the curves, we found four segments to be sufficient in illustrating their overall trends. Also, in [73], the average duration of a film event was estimated to be around 1 minute, making higher-resolution segmentation unfounded from a narrative standpoint. For these reasons, we limited the number of segments to four per clip, whereby the shortest segment (for clip 8) was 18 seconds long.

In the style and aesthetics assessment, participants rated the applicability of 13 stylistic and 14 aesthetic attributes (Table III) to each clip on a discrete scale of [1], [5]. The attributes were presented without category or modality labels. They contained semantic opposites, such as brightness and darkness. These were rated separately since they are not necessarily mutually exclusive [74]; for example, a clip could be considered both beautiful and ugly for different reasons.

E. Benchmark Data

We have made our data publicly available at http://research.ics.aalto.fi/cbir/data/ to allow other researchers to compare their methods with our baseline results. The data consists of the human ratings (Section V-D) and the computational features (Section VI-A). The movie clips can be obtained using the information in Table I.

The assessment data is provided in two forms: in a raw form containing assessments with missing values, marked “NAN” (not-a-number), and in a cleaned form where, for each (movie, participant) pair, all assessments with missing values have been removed. The raw form contains assessments for 42 001 (movie, participant, attribute) triplets and the cleaned form for 38 844 triplets. In our experiment we used the cleaned set.

TABLE III
STYLISTIC AND AESTHETIC ATTRIBUTES ASSESSED IN THE USER STUDY

Category	Modality	Attributes	
Stylistic	Visual	Brightness Colorlessness	Colorfulness Darkness
	Auditory	Dialogue-basedness Music-basedness	Loudness Quietness
	Temporal	Fastness Rhythmicity Smoothness	Fitfulness Slowness
Aesthetic		Beauty Familiarity Pleasantness Simplicity Ugliness Understandability Unpleasantness	Complexity Interestingness Predictability Tiresomeness Unclarity Unfamiliarity Unpredictability

To our knowledge, this is the first publicly available data set of ratings not just of the affective content of movie clips (all from mainstream movies), but also their stylistic and aesthetic attributes. The set also contains hedonic tone and arousal curve data for each clip and participant, useful for studying the effect of temporal changes on affect within movie scenes. The set is expected to facilitate the further development and evaluation of computational descriptors of perceptual attributes.

VI. Assessment Prediction Experiment

A. Computational Features

We automatically extracted a large set of computational features from the clips. They were required to be closely analogous to those described in Section IV-A, easy to implement, and fast to extract automatically. We do not propose any novel features especially designed for the current prediction task. Instead, we use a large number of conventional features, with a total dimensionality of 192, among which the prediction algorithms can make use of the most beneficial ones.

Table IV lists the computational features used. The second column gives the number of components in each feature, i.e. its dimensionality. Visual features were first calculated for each frame and auditory features for each second of audio, and then expressed as the mean of these values averaged over the whole clip. The features marked with an asterisk (*) were also represented by two additional values: their standard deviation and mean-normalized deviation over the whole clip. The mean-normalized deviation was calculated as the ratio between the standard deviation and the mean to compensate for the fact that features with large average values often also exhibit large variation. The features are discussed in more detail below.

Visual Features: We used the average intensity value over all frames as a measure of overall brightness. We also calculated the variation in brightness between different parts of a frame divided into five zones (Fig. 2) to characterize the intra-frame contrast between spatial areas. Colors were represented in either the RGB, CIE Luv or HSV color space depending on the feature. Dominant color was calculated as in [50]. The spatial zones of Fig. 2 were also used to describe spatial color variations. Lastly, we used a five-bin brightness (intensity) histogram as a representation of lighting key. Its first bin can be interpreted as the shadow proportion.

Auditory Features: We used the traditional MFCC and zero-crossing rate features, as well as the audio power spectrum

TABLE IV

COMPUTATIONAL FEATURES USED IN THE PREDICTION, ALONG WITH THEIR BASIC DIMENSIONALITIES AND DESCRIPTIONS. FEATURES MARKED WITH AN ASTERISK (*) WERE REPRESENTED BY TRIPLETS CONSISTING OF THE MEAN, THE STANDARD DEVIATION AND THE MEAN-NORMALIZED DEVIATION OF THE VALUE OVER THE WHOLE CLIP

Name	Dim.	Technical description
Visual features		
Brightness *	1	average intensity value
Brightness variation *	1	standard deviation of intensity over the five zones of Figure 2
Color variation *	1	standard deviation of values over all RGB channels and the five zones of Figure 2
Average colors *	15	average RGB in the five zones of Figure 2
Dominant colors *	6	CIE Luv components of the two dominant colors
Brightness histogram *	5	proportions of intensity values in 20 % bins
Saturation histogram *	5	proportions of saturation values in 20 % bins
Auditory features		
Overall volume *	1	average sound energy
Frequency band energy *	6	energy in octave frequency bands up to 22 kHz
MFCC *	13	Mel-frequency cepstral coefficients
Zero-crossing rate *	1	number of sign changes in the audio signal
Music brightness	1	percentage of energy in frequencies above 1500 Hz
Music event density	1	average frequency of events, i.e. the number of note onsets per second
Music mode	1	estimate of modality (major vs. minor) as a numerical value
Music tempo	1	estimate of tempo (beats per minute) based on periodicities in the onset detection curve
Temporal features		
Shot duration	1	average shot duration
Shot number	1	number of shots
Variation of shot duration	1	standard deviation of shot durations
Motion intensity	1	overall intensity of motion activity
Motion direction	2	dominant direction of motion activity (x and y)
Spatial motion distribution	3	number and size of regions with motion activity
Temporal motion distribution	5	variation of motion activity over shot duration
Between-frame difference *	3	pixel-wise RGB difference, difference in Centrist features, keypoint movement length

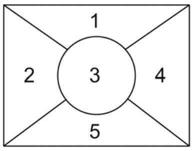

Fig. 2. Spatial image zones used in visual features.

as a representation of both the overall volume and the energy distribution over different octave-wide frequency bands. Also, in addition to our own feature implementations, we extracted four music features–brightness, event density, mode and tempo–using the Music Information Retrieval (MIR) Toolbox [75]. These features were not explicitly targeted for the music sequences in the clips, nor were they tuned to describe the auditory content of the clips in question.

Temporal Features: We used the shot number, average shot duration and variation of shot duration features. As the number and duration of clips in our study was limited, for greater accuracy, we determined the features with manually annotated shot boundaries instead of automatic shot boundary detection, which is often considered a solved technical problem [76]. Also, to express the overall motion in the clip, we used the MPEG-7 Motion Activity feature [50], which has been split into its four components in Table IV: motion intensity, motion direction, and spatial and temporal motion distribution. Lastly, we modeled the difference between adjacent video frames with three measures: the raw average pixel-wise RGB difference, the Euclidean distance between Census Transform (Centrist) [77] features and the average measured movement length of matching SURF keypoints [78] between the frames.

B. Prediction Methods

Our goal is to predict the style, aesthetics and affect ratings of new movie clips based only on human ratings of the training set clips and automatically extracted computational features. To this end, we use two methods, linear regression and the non-linear ELM, that try to learn the relations between low-level features and human ratings. The data collected in the user study (Section V) is used to train and test the predictors.

Both linear regression and ELM are conventional, well-known methods. Their use is motivated by the limited number of clips we have for training the predictors. Being simple models, they can be expected to work better in these conditions than more complex models, such as Support Vector Regression [79], which generally require more training samples. Due to its simplicity and widespread use in machine learning applications, linear regression can be considered the reference method here. ELM, on the other hand, is a recent and popular algorithm that is a notably fast non-linear classification method [15], making it suitable for comparison with linear regression.

Let us assume that we have n training clips, for which we have some computational feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and the corresponding human ratings $y_1, y_2, \ldots, y_n$ of some particular attribute. We then wish to form a prediction model of the relation $y = f(\mathbf{x})$ that can be used to predict the human rating when the computational features are known. Each feature vector $\mathbf{x}_i$ is assumed to consist of p components:

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}] . \quad (1)$$

In our experiments, all the features listed in Table IV were concatenated, leading to feature vectors $\mathbf{x}_i$ with $p = 192$.

Linear Regression: A multiple linear regression model assumes a linear relation; for $i = 1, \ldots, n$

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \gamma + \epsilon_i \tag{2}$$

which can be written more concisely in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{3}$$

where the ratings y_i have been collected in a vector

$$\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$$

and the prediction errors ϵ_i in

$$\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T.$$

Likewise, the feature vectors for each clip have been stacked as rows of the matrix $\mathbf{X}$, with the last column set to ones to accommodate the intercept γ, i.e.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} & 1 \\ x_{21} & \ldots & x_{2p} & 1 \\ \vdots & \ddots & \vdots & 1 \\ x_{n1} & \ldots & x_{np} & 1 \end{pmatrix}, \tag{4}$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_p, \gamma]^T. \tag{5}$$

The least-squares solution $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ minimizes the prediction error $\|\epsilon\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$. For solving the least-squares problem we used the DGELSD routine from the LAPACK library [80] via the Python numpy package for scientific computing[2]. This approach first reduces the matrix $\mathbf{X}$ into a bidiagonal form using Householder transformations. The resulting bidiagonal least-squares problem is solved using a divide and conquer approach, and the solution can be transformed back into the original problem by applying the reverse Householder transformations. The routine uses a regularization that imposes a cut-off ratio for small singular values of $\mathbf{X}$. We set the cut-off ratio to 0.05 times the largest singular value.

Once $\hat{\boldsymbol{\beta}}$ is determined, one can predict the ratings of new movie clips for which computational features $\mathbf{x}_{\text{new}}$ have been automatically extracted, simply by

$$y_{\text{new}} = f_{\text{linreg}}(\mathbf{x}_{\text{new}}) = \mathbf{x}_{\text{new}}\hat{\boldsymbol{\beta}}. \tag{6}$$

Extreme Learning Machine (ELM): In essence, ELM [15] is a non-linear single-hidden-layer feed-forward network where the parameters of the hidden layer are simply randomized and thus need no training. The model of the input-output dependency of ELM is

$$y_i = \sum_{j=1}^{L} h_j(\mathbf{x}_i)\beta_j + \epsilon_i, \tag{7}$$

or, in matrix form,

$$\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \epsilon, \tag{8}$$

where

$$\mathbf{H} = \begin{pmatrix} h_1(\mathbf{x}_1) & \ldots & h_L(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ h_1(\mathbf{x}_n) & \ldots & h_L(\mathbf{x}_n) \end{pmatrix} \tag{9}$$

and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_L]^T$ are the output weights between the hidden layer of L nodes and the output node. The non-linear hidden node output function is given in the form

$$h_j(\mathbf{x}) = g(\mathbf{x}; \mathbf{a}_j, b_j) \tag{10}$$

where $g(\cdot)$ is typically a sigmoid or a radial basis function (RBF) and the parameters are randomly generated. In the experiments presented here, we have used the Gaussian form

$$g(\mathbf{x}; \mathbf{a}_j, b_j) = \exp\left(-b_j \frac{\|\mathbf{x} - \mathbf{a}_j\|^2}{c_j^2}\right) \tag{11}$$

Finally, one can solve for $\hat{\boldsymbol{\beta}}$ as in the previous section, to get

$$y_{\text{new}} = f_{\text{ELM}}(\mathbf{x}_{\text{new}}) = \sum_{j=1}^{L} h_j(\mathbf{x}_{\text{new}})\hat{\beta}_j \tag{12}$$

$$= \sum_{j=1}^{L} \exp\left(-b_j \frac{\|\mathbf{x}_{\text{new}} - \mathbf{a}_j\|^2}{c_j^2}\right)\hat{\beta}_j. \tag{13}$$

We experimented with different numbers L of hidden nodes. The prediction results improved with an increased number of nodes; however, around $L = 20$ the benefit of adding nodes became relatively small. We opted to use $L = 100$, but in time-critical applications a smaller value can be used, such as $L = 50$, which gives results almost as good in a shorter time.

We used David Lambert's Python implementation of ELM[3]. The values $\mathbf{a}_j$ where randomly uniformly sampled from within the bounding hyperrectangle of the inputs, $c_j = \max(\|x - \mathbf{a}_j\|)/\sqrt{2L}$ and $b_j = 0.01$ was selected by experimentation.

C. Evaluation Procedure and Metrics

Since our sample ($n = 14$) of clips is quite small, splitting it into fixed training and test sets would be infeasible. Instead, we use an n-fold leave-one-out approach where one clip is excluded in turn and the remaining $n - 1$ clips are used for training. The excluded clip is then used to test the predictor.

Our basic evaluation metric is the absolute difference between the predicted value $\hat{y}_i = y_{\text{new}}$, from either (6) or (12), and the ground truth value y_i, i.e. $e_i = |\hat{y}_i - y_i|$. The ground truth is the mean of the human ratings for each clip, which number between 27 and 39, depending on the assessed clip and attribute. The evaluation metric for a single attribute is the average of e_i over all n clips, i.e. over each partition of leaving one clip out for testing and using the others for training.

Comparing the performance of the prediction between different categories of assessed attributes is difficult due to the differing statistics of the human ratings. First, attributes may have different *between-movie deviations* σ_b, i.e. some stylistic, aesthetic or affective attributes are more widely dispersed between

[2] http://www.numpy.org/

[3] https://github.com/dclambert/Python-ELM

movies than others. The between-movie deviation is measured as the standard deviation of the mean ratings of different movies. For example, if perceived arousal is generally rated higher than felt arousal, the prediction errors of the former can similarly appear greater than those of the latter.

Second, the *within-movie deviation* σ_w can also be substantially different for different attributes. The value of σ_w reflects how much the participants disagreed in their ratings. If the ratings vary greatly for a given clip, the computational prediction for it cannot be expected to be very accurate either.

In evaluating the prediction results, we analyze these two deviations and normalize the errors with respect to σ_b. We also compare the absolute prediction errors against the absolute error d of another prediction, which we here call *naïve prediction*, being simply the average of the human-provided attribute values of the $n - 1$ training samples. This is useful as a baseline for comparison: if the prediction method used performs better than the naïve prediction, it can be said to have learned from the low-level computational features something relevant that cannot be inferred simply from the clip ratings.

In addition to the clip-level analysis, we also predict hedonic tone and general arousal ratings based on the four temporal segments acquired from the affect curves (see Section V-D). In this case, each segment of each clip is treated as a separate prediction target and the segments of the 13 other movies are used for training. The other three segments from the same clip are not used for training to avoid including information about the movie whose segment is to be predicted in the training data. In effect, this procedure multiplies the amount of training data fourfold, though the ratings derived from the curves are arguably less reliable than the clip-level numerical ratings.

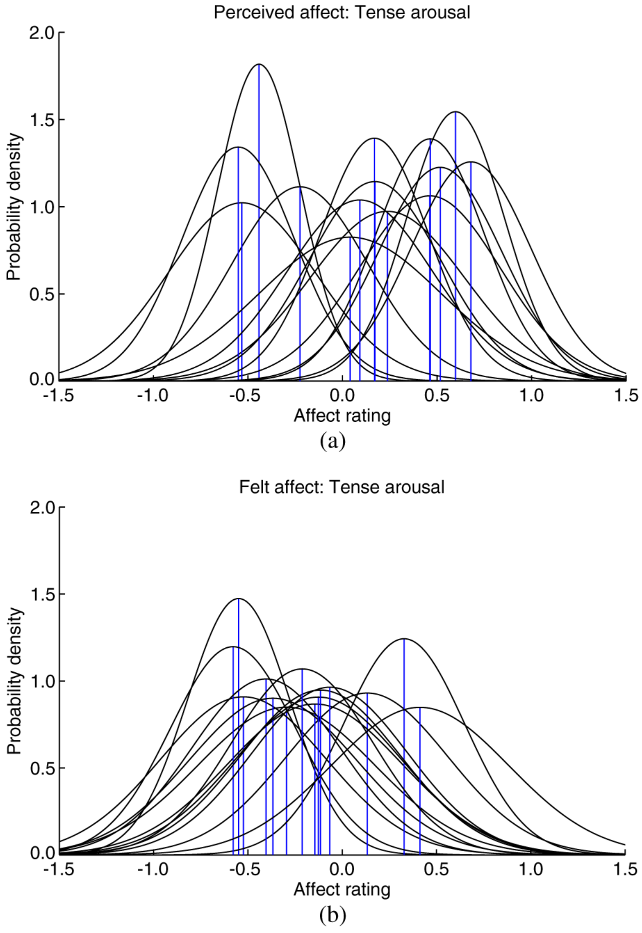

Fig. 3. Probability density functions of the (a) perceived and (b) felt tense arousal ratings of the 14 movie clips. The vertical lines denote average values.

VII. Results

The Gaussian-modeled probability density functions of the tense arousal ratings of all 14 movie clips (Fig. 3) illustrate two general differences between perceived and felt affect. First, average perceived affect ratings were, for each affect dimension, more spread out (i.e. σ_b was larger) than felt affect ratings, indicating better separation of clips by perceived affect. Second, for most clips, the standard deviations of perceived affect ratings were narrower (i.e. σ_w was smaller) than those of the corresponding felt affect ratings, indicating better inter-rater agreement for perceived affect. The greater spread of the vertical lines (representing average ratings), and the narrower probability distributions, in Fig. 3(a) (perceived affect) than Fig. 3(b) (felt affect) illustrate this phenomenon for tense arousal. Still, Fig. 3 also shows considerable overlap between ratings for both affect types; the phenomenon was present to varying degrees in all the attribute categories. Considering the general difficulty of visual pattern recognition tasks, such as optical character recognition or visual concept detection, trivial for humans yet difficult for machines, the overall low level of inter-rater agreement here hints at the even more challenging nature of predicting movie ratings.

Fig. 4 shows the movie-wise felt and perceived affect ratings, as well as their linear regression, ELM and naïve predictions, in a two-dimensional valence–arousal space. Hedonic tone (horizontal axis) corresponds to valence, and the arousal dimension (vertical axis) is represented by energetic arousal.

As in Fig. 3, it can be seen from Fig. 4 that individual ratings for each clip are quite varied, but that standard deviations are slightly smaller for perceived affect. It can also be seen that the average felt and perceived affect ratings for a given clip are generally located in the same quadrant of the valence–arousal space, but that average perceived affect ratings are more spread out, occupying a greater portion of the space. Lastly, the figure indicates that, overall, the content-based predictions perform slightly better than the naïve prediction, which is based only on average user ratings. However, there are also cases where the naïve prediction is better. The results also show that linear regression performs comparably to ELM.

A. Rating Statistics

In order to make a quantitative comparison of the results for the stylistic, aesthetic and affective attribute categories, we calculated the average between- and within-movie deviations (σ_b and σ_w, respectively) over all movie clips. The first block of columns in Table V shows the averages of these statistics across different attribute categories. The values describe the distributions of the human ratings and can be used to analyze their variation and the overlap between clips. In addition, they are useful for assessing the magnitude of the prediction errors.

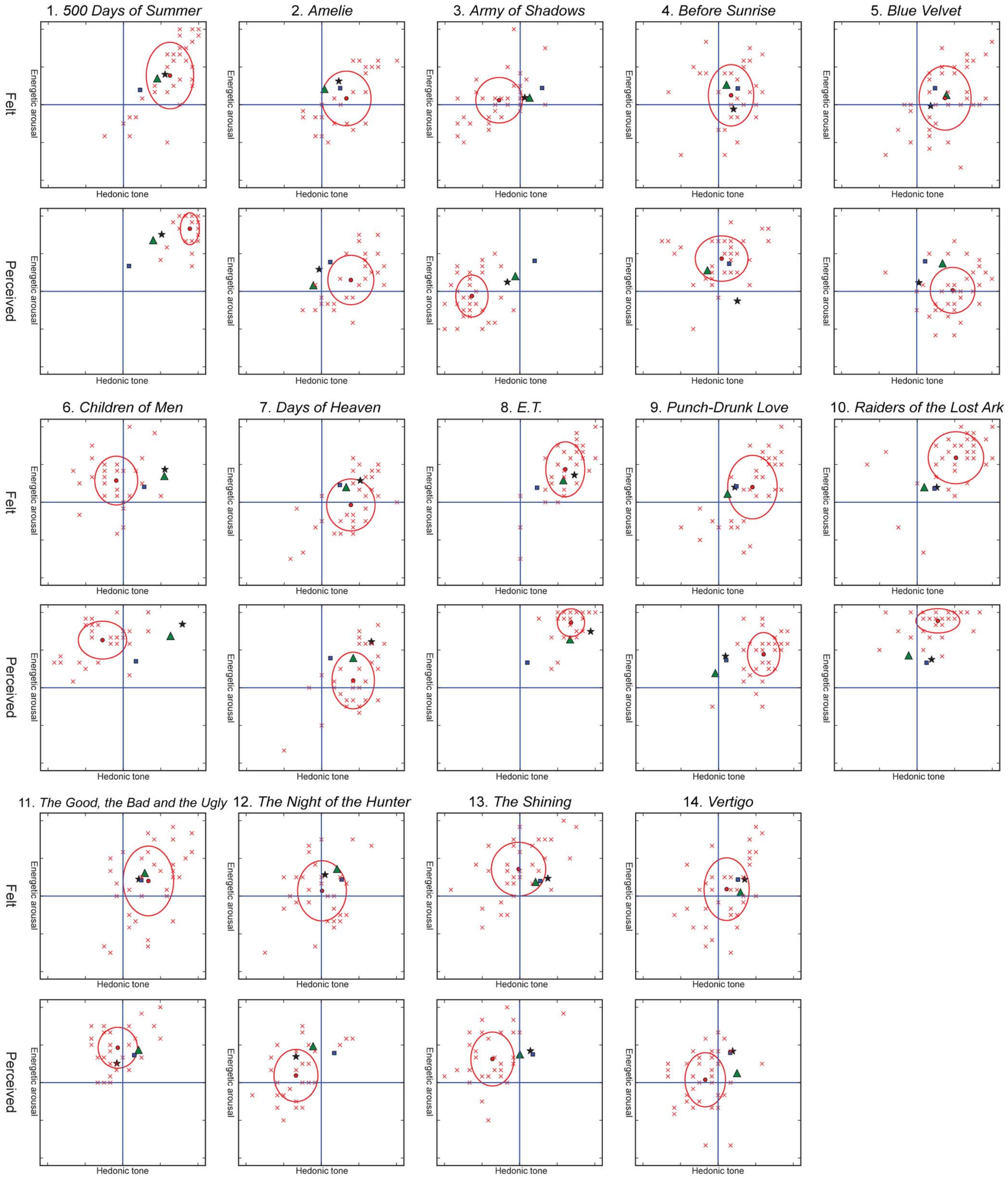

Fig. 4. Movie-wise affect ratings and predictions for the movie clips. In each plot, the horizontal axis represents hedonic tone, from negative (-1) to positive $(+1)$, and the vertical axis represents energetic arousal, from tired (-1) to awake $(+1)$. For each clip, felt affect is shown above perceived affect. Individual human ratings are shown by crosses, and their average and standard deviation by a dot and an oval, respectively. The ELM-predicted rating is shown by a star, the linear regression prediction by a triangle, and the naïve prediction by a square.

The first block of rows contains prediction results for the stylistic and aesthetic attribute categories (see Table III). The second and third blocks give the results for the perceived and felt affect ratings, respectively. These include results for the prediction based on both the movie-wise affect ratings (hedonic tone as well as energetic, tense and general arousal) and the temporal

TABLE V

RATING STATISTICS AND PREDICTION ERRORS ACROSS DIFFERENT ATTRIBUTE CATEGORIES. RATINGS STATISTICS ARE GIVEN IN TERMS OF THE BETWEEN- AND WITHIN-MOVIE DEVIATIONS (σ_b AND σ_w, RESPECTIVELY) AS WELL AS THEIR RATIO (σ_b/σ_w). PREDICTION ERRORS ARE GIVEN WITH RESPECT TO THE WITHIN- AND MOVIE DEVIATIONS AS WELL AS THE NAÏVE PREDICTION ERROR d

	Rating statistics			Prediction errors			
				e/σ_b		e/d	
	σ_b	σ_w	σ_w/σ_b	linreg	ELM	linreg	ELM
Stylistic, visual	0.74	0.93	1.30	0.65	0.68	0.99	1.04
Stylistic, auditory	0.97	0.82	0.87	0.50	0.49	0.75	0.76
Stylistic, temporal	0.61	0.96	2.05	0.65	0.52	1.03	0.86
Stylistic, all	0.76	0.91	1.45	0.60	0.56	0.93	0.88
Aesthetic	0.64	1.00	1.64	0.75	0.82	1.11	1.19
Perceived affect							
Hedonic tone	0.44	0.26	0.58	0.37	0.42	0.44	0.51
Energetic arousal	0.32	0.29	0.87	0.56	0.48	0.76	0.65
Tense arousal	0.40	0.34	0.83	0.74	0.57	1.25	0.97
General arousal	0.29	0.24	0.84	0.85	0.52	1.35	0.82
Perceived affect, all	0.38	0.28	0.72	0.63	0.50	0.95	0.73
Hedonic tone (curve)	-	-	-	-	-	0.78	0.89
General arousal (curve)	-	-	-	-	-	1.13	0.87
Felt affect							
Hedonic tone	0.26	0.32	1.21	0.58	0.54	0.73	0.68
Energetic arousal	0.17	0.38	2.22	0.44	0.60	0.68	0.94
Tense arousal	0.29	0.40	1.35	0.60	0.45	1.15	0.86
General arousal	0.18	0.28	1.47	0.51	0.50	0.98	0.95
Felt affect, all	0.28	0.32	1.28	0.53	0.52	0.89	0.86
Hedonic tone (curve)	-	-	-	-	-	1.10	1.12
General arousal (curve)	-	-	-	-	-	0.84	0.81

segment means computed from the hedonic tone and arousal curves. Since the clip segments originate partially from the same clips, the interpretation of their between- and within-movie deviations is not straightforward. For this reason, σ_b and σ_w are not reported for these assessments.

The between-movie deviation σ_b describes how widely distributed the ratings of different movie clips are. The values illustrate the phenomenon mentioned in Section III-B and exemplified in Fig. 3 that perceived affect ratings have, on average, larger absolute values than felt affect ratings, occupying a greater portion of the valence–arousal space, which in turn means greater variability between clips and larger σ_b values for perceived affect ratings. Style and aesthetics ratings have systematically larger σ_b variability than affect ratings because their rating scale was [1], [5], whereas for the affect dimensions the ratings were scaled to the range $[-1, 1]$.

The within-movie deviation σ_w describes the variation in the human ratings for a given clip, a larger value indicating greater disagreement between participants. Due to the difference in scales, style and aesthetics ratings again have larger values than affect ratings. Perceived affect has smaller variability than felt affect, illustrating the finding, mentioned in Section III-B and seen in Fig. 3, that though viewers' emotional responses (felt affect) to a movie may vary, they tend to agree about what emotions are expressed by it (perceived affect). For example, though all viewers may not find a horror movie scary, they may still agree that its intention is to scare viewers.

The third rating statistics column in Table V shows the average ratio between the within- and between-movie deviations. Note that this ratio is the geometric mean of the movie-wise ratios within a given category, not the direct ratio between the averages in the first and second columns. For style ratings, the values indicate that auditory attributes distinguish between clips relatively well ($\sigma_w/\sigma_b = 0.87$), while for temporal attributes, the ratings have considerably more overlap ($\sigma_w/\sigma_b = 2.05$). The deviation ratio further emphasizes the difference between the perceived and felt affect ratings mentioned in Section III-B. For felt affect, the ratio is so large ($\sigma_w > \sigma_b$ for all dimensions) that ratings for different clips overlap substantially, while perceived affect ratings ($\sigma_w < \sigma_b$ for all dimensions) are more clearly separated.

B. Prediction Results

The right-hand side of Table V contains prediction errors for linear regression and ELM. The prediction errors are shown in two groups of column pairs, the first showing the ratio of the error to the between-movie deviation σ_b, and the second to the average absolute prediction error d of the naïve prediction. The between-movie deviation ratio e/σ_b aims to account for differences in the spread of mean ratings between different attributes. Lastly, naïve prediction deviation ratio e/d values below one indicate that the prediction has benefited from the low-level features. All prediction error ratio values shown in Table V are geometric means of the movie-wise ratios over all clips. We used the geometric mean because it is mathematically more motivated than the arithmetic mean for averaging out ratios.

The between-movie deviation ratio e/σ_b can be used to assess the predictability of the different attribute categories in Table V: stylistic, aesthetic and affective. The values indicate that for ELM, stylistic, perceived affect and felt affect attributes appear roughly equally easy to predict, with average $e/\sigma_b = 0.56, 0.50, 0.52$, respectively. For linear regression, felt affect performs better than the other categories, with $e/\sigma_b = 0.53$, versus 0.63 for perceived affect and 0.60 for stylistic attributes. Among style modalities, auditory attributes appear the easiest to predict for both algorithms. Lastly, aesthetic attributes are clearly the most difficult to predict.

We also investigated, with the e/d metric, which prediction method is more efficient: ELM or linear regression. The results indicate that the relative performance of the two methods varies by category. Overall, ELM performed slightly better for stylistic attributes and perceived affect, and its e/d value is below one for all affect dimensions. Aesthetics is the only attribute category for which both linear regression and ELM have e/d values larger than one, indicating that the computational prediction did not improve on the naïve prediction.

Comparing the segment-based hedonic tone and general arousal curve prediction results to those based on the corresponding movie-wise ratings, it can be seen that the curve data did not improve but worsened the prediction for hedonic tone. For general arousal, on the other hand, the segment-based prediction performed better than the rating-based prediction with linear regression in the case of perceived affect ($e/d = 1.13$ and 1.35, respectively). For felt affect, the segment-based prediction performed better for both linear regression ($e/d = 0.84$ and 0.98) and ELM ($e/d = 0.81$ and 0.95).

C. Feature Selection

Though the number of human-rated movie clips in the data set is not large enough to rank the computational features based on their influence on the prediction accuracy of different attributes, it is nevertheless possible to study the features involved in the prediction of a given attribute.

To investigate which features were the most significant in the prediction of all affect types, we conducted a feature selection experiment, leaving out each of the 14 clips in turn and generating a separate feature selection using cross-validation in the remaining 13-clip subset. Then, the number of times each feature was selected was added up across all 14 experiments for a final result. Within the 13-clip subset, we used a sequential backward selection (SBS) scheme to perform the feature selection. The algorithm terminated when the average absolute prediction error e increased from its previous value by more than a small predefined margin. We set this margin heuristically to 0.005, which is less than the maximum discretization error in the dimensional affect ratings. Due to the computational requirements of the experiment, we conducted it with the linear predictor instead of ELM.

Feature selection did not improve results in terms of e/d. Also, there were no notable differences between the features most often selected for different affect dimensions. The best-performing features for all dimensions included shot duration, dominant colors, motion intensity and direction, spatial and temporal motion distribution, as well as all four MIR features. This shows that all modalities of low-level computational features contribute to the prediction of the affective content.

VIII. Conclusion

We have presented a data set and a setup for automatically predicting stylistic, aesthetic and affective movie content ratings using low-level computational features. We conducted a user study with a large number of subjects to collect human ratings of the style, aesthetics and affect of a limited number of movie clips. Perceived and felt affect were assessed separately, and the clips were mapped into a three-dimensional valence–arousal space with two distinct arousal dimensions, energetic and tense. We have made the data publicly available, along with the low-level features extracted from the clips.

We have also presented a prediction experiment to serve as a baseline for future performance comparisons with our data set. The experiment simulates a situation where the system must predict the average human rating of the stylistic, aesthetic and affective content of a new, previously unknown movie clip. It should be noted that this is different and more difficult than a simple correlation analysis where the ratings are known beforehand. The presented setup imitates a realistic video classification or recommendation system where a new video is classified using only previous ratings of other videos and computational features automatically extracted from the video.

We compared two prediction methods: multiple linear regression and the recent neural-network-based Extreme Learning Machine (ELM) algorithm. Overall, ELM performed slightly better than linear regression. Its performance was also more consistent across the different attribute categories.

Though perceived affect ratings illustrated greater inter-rater agreement and better affect-based separation of clips than felt affect in the user study, in our prediction experiment, both affect types appear equally easy to predict when prediction errors are normalized with respect to the between-movie variation in the human ratings. These results are promising, since for affect-based movie recommendations, the modeling of actual viewer response is more desirable than the modeling of the movie's expressed emotion. However, the result should be verified with a larger data set. Also, since perceived affect, being the more objective of the two measures, has been shown [14] to be better at distinguishing between affect ratings, it remains a worthwhile target in terms of video classification. Perceived affect could also be used to predict an individual viewer's affective response with the help of a viewer profile containing data on their personal preferences [44].

We found stylistic and affective attributes equally easy to predict overall. Among style modalities, auditory attributes were the easiest to predict. Aesthetic attributes were the hardest to predict, and feature-based prediction performed, on average, worse than the naïve baseline prediction. The poor performance of both prediction methods for aesthetic attributes suggests that both style and affect may be more within the grasp of computational methods than aesthetics. The finding is interesting in the sense that though both affect and aesthetics are abstract concepts, the former appears to be more closely linked to low-level features than the latter.

The segment-based affect prediction improved results for general arousal, but worsened results for hedonic tone. This may be because the arousal curves displayed greater within-scene changes than the hedonic tone curves overall. Since the curves were drawn by hand and rescaled in preprocessing, they all contained small-scale temporal fluctuations, which can be interpreted as noise resulting from inaccuracies of freehand plotting. Furthermore, since the hedonic tone curves were, save for a few exceptions (such as clip 6, shown in Fig. 1), in general more flat than the arousal curves, the relative influence of these fluctuations on curve values was greater, resulting in a lower signal-to-noise ratio for the hedonic tone curves, and thereby in more inaccurate prediction. General arousal, on the other hand, benefited considerably from the curve data, suggesting that in terms of arousal prediction, scene-level sampling may be too coarse, since it can neglect the influence of within-scene events on both perceived and felt arousal.

Feature selection did not improve prediction performance. However, it indicated that some of the low-level computational features in our feature set are more important than others and that features from all modalities–visual, auditory and temporal–contribute to the prediction of affect ratings.

Acknowledgment

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*. Hillsdale, NJ, USA: Routledge, 1996.

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

[3] M. D. Augustin, J. Wagemans, and C.-C. Carbon, "All is beautiful? Generality vs. specificity of word usage in visual aesthetics," *Acta Psychologica*, vol. 139, no. 1, pp. 187–201, 2012.

[4] R. Spottiswoode, *A Grammar of the Film*. Berkeley, CA, USA: Univ. California, 1950.

[5] G. M. Smith, *Film Structure and the Emotion System*. Cambridge, U.K.: Cambridge Univ., 2007.

[6] A. Mackendrick, *On Film-making*. London, U.K.: Faber and Faber, 2006.

[7] D. Bordwell, *On the History of Film Style*. Cambridge, MA, USA: Harvard Univ., 1997.

[8] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, "A bayesian framework for video affective representation," in *Proc. ACII*, 2009, pp. 1–7.

[9] R. M. Teixeira, T. Yamasaki, and K. Aizawa, "Comparative analysis of low-level visual features for affective determination of video clips," in *Proc. ICFIT*, 2010, pp. 1–6.

[10] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, pp. 30–37, 2009.

[11] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[12] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford, U.K.: Oxford Univ., 1989.

[13] K. Kallinen, "Towards a comprehensive theory of musical emotions," Ph.D. dissertation, Faculty of Humanities, University of Jyväskylä, , 2006.

[14] J. Tarvainen, S. Westman, and P. Oittinen, "Stylistic features for affect-based movie recommendations," in *Proc. 4th Int. Workshop on Human Behavior Understanding*, 2013, pp. 52–63.

[15] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[16] R. Sinnerbrink, "Stimmung: Exploring the aesthetics of mood," *Screen*, vol. 53, no. 2, pp. 148–163, 2012.

[17] , J. Aumont, A. Bergala, M. Marie, and M. Vernet, Eds., *Aesthetics of Film*. Austin, TX, USA: Univ. Texas, 1983.

[18] D. Bordwell and K. Thompson, *Film Art: An Introduction*. New York, NY, USA: McGraw-Hill, 1990.

[19] N. Carroll, "Film, emotion, and genre," in *Passionate Views: Film, Cognition, and Emotion*. Baltimore, MD, USA: The John Hopkins Univ., 1999, pp. 21–47.

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

[21] A. Gabrielsson and P. Juslin, "Emotional expression in music," in *Handbook of Affective Sciences*. Oxford, U.K.: Oxford Univ., 2003, pp. 503–534.

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

[23] 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.

[24] 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.

[25] S. D. Lipscomb and R. A. Kendall, "Perceptual judgement of the relationship between musical and visual components in film," *Psychomusicology: J. Res. Music Cognition*, vol. 13, no. 1–2, pp. 60–98, 1994.

[26] 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. 1–60, 2007.

[27] T. Jacobsen, K. Buchta, M. Khler, and E. Schrger, "The primacy of beauty in judging the aesthetics of objects," *Psycholog. Rep.*, vol. 94, no. 3, pt. Pt 2, pp. 1253–60, 2004.

[28] K. R. Scherer, "Appraisal considered as a process of multi-level sequential checking," in *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford, U.K.: Oxford Univ., 2000, pp. 92–120.

[29] H. Leder, B. Belke, A. Oeberst, and D. Augustin, "A model of aesthetic appreciation and aesthetic judgments," *Brit. J. Psychol.*, vol. 95, pt. Pt 4, pp. 489–508, 2004.

[30] P. J. Silvia and C. Berg, "Finding movies interesting: How appraisals and expertise influence the aesthetic experience of film," *Empirical Studies of the Arts*, vol. 29, no. 1, pp. 73–88, 2011.

[31] M. Augustin, C. Carbon, and J. Wagemans, "Artful terms: A study on aesthetic word usage for visual art versus film and music," *Perception*, vol. 3, no. 5, pp. 319–37, 2012.

[32] H. Hagtvedt, R. Hagtvedt, and V. M. Patrick, "The perception and evaluation of visual art," *Empirical Studies of the Arts*, vol. 26, no. 2, pp. 197–218, 2008.

[33] M. Lewis and J. Haviland-Jones, *Handbook of Emotions*, 2nd ed. New York, NY, USA: Guilford, 2000.

[34] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development Psychopathol.*, vol. 17, no. 3, pp. 715–734, 2005.

[35] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[36] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgement and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *J. Psychophysiol.*, vol. 3, pp. 51–64, 1989.

[37] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, pp. 1161–1178, 1980.

[38] G. Matthews, D. M. Jones, and A. G. Chamberlain, "Refining the measurement of mood: The UWIST mood adjective checklist," *Brit. J. Psychol.*, vol. 81, no. 1, pp. 17–42, 1990.

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

[40] K. Kallinen and N. Rajava, "Emotion perceived and emotion felt: Same and different," *Musicae Scientiae*, vol. 10, pp. 191–213, 2006.

[41] C. Plantinga, "Art moods and human moods in narrative cinema," *New Literary History*, vol. 43, no. 3, pp. 455–475, 2012.

[42] K. Knautz and W. G. Stock, "Collective indexing of emotions in videos," *J. Documentation*, vol. 67, no. 6, pp. 975–994, 2011.

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

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

[45] H.-B. Kang, "Affective content detection using HMMs," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 259–262.

[46] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proc. ACM Workshop Multimedia Semantics*, 2008, pp. 32–39.

[47] S. Jain and R. Jadon, "Movies genres classifier using neural network," in *Proc. IEEE ISCIS*, 2009, pp. 575–580.

[48] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 11, pp. 1–11, Nov. 2003.

[49] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, "Utilizing affective analysis for efficient movie browsing," in *Proc. IEEE ICIP*, 2009, no. 49, pp. 1853–1856.

[50] , B. S. Manjunath, P. Salembier, and T. Sikora, Eds., *Introduction to MPEG-7: Multimedia Content Description Interface*. New York, NY, USA: Wiley, 2002.

[51] L. Canini, S. Benini, P. Migliorati, and R. Leonardi, "Emotional identity of movies," in *Proc. IEEE ICIP*, 2009, pp. 1821–1824.

[52] M. Xu, J. Wang, M. A. Hasan, X. He, C. Xu, H. Lu, and J. S. Jin, "Using context saliency for movie shot classification," in *Proc. IEEE ICIP*, 2011, pp. 3653–3656.

[53] P. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *J. New Music Res.*, vol. 33, no. 3, pp. 217–238, 2004.

[54] R. Picard*, Affective Computing*. Cambridge, MA, USA: MIT, 1997.

[55] F. Gouyon and F. Pachet, "On the use of zero-crossing rate for an application of classification of percussive sounds," in *Proc. COST G-6 Conf. Digital Audio Effects*, 2000, pp. 3–8.

[56] M. Xu, J. S. Jin, and S. Luo, "Personalized video adaptation based on video content analysis," in *Proc. ACM Int. Workshop on Multimedia Data Mining*, 2008, pp. 26–35.

[57] G. Irie, K. Hidaka, T. Satou, T. Yamasaki, and K. Aizawa, "Affective video segment retrieval for consumer generated videos based on correlation between emotions and emotional audio events," in *Proc. IEEE ICME*, 2009, pp. 522–525.

[58] D. Li, I. K. Sethi, N. Dimitrova, and T. Mcgee, "Classification of general audio data for content-based retrieval," *Pattern Recognit. Lett,*, vol. 22, pp. 533–544, 2001.

[59] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *J. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1493–1500, 1978.

[60] J. Mitry*, The Aesthetics and Psychology of the Cinema*. Bloomington, IN, USA: Indiana Univ., 1997.

[61] X. Lin, X. Wen, Z. Lu, and W. Zheng, "Video affective content recognition based on film grammars and fuzzy evaluation," in *Proc. IEEE MMIT*, 2008, pp. 264–267.

[62] S. Arifin and P. Y. Cheung, "A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 68–77.

[63] S. Arifin and P. Y. Cheung, "Towards affective level video applications: A novel FPGA-based video arousal content modeling system," *Programmable Logic and Applications*, pp. 2–5, 2006.

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

[65] E. A. Sonnenschein, S. Jones, and E. Macleod*, What Is Rhythm?*. Oxford, U.K.: B. Blackwell, 1925.

[66] M. Ghyka*, The Geometry of Art and Life*. : Dover Publications, Inc., 1977.

[67] B. Adams, C. Dorai, and S. Venkatesh, "Automated film rhythm extraction for scene analysis," in *Proc. ICME*, 2001, pp. 1056–1059, IEEE.

[68] S. Zhao, H. Yao, and X. Sun, "Affective video classification based on spatio-temporal feature fusion," in *Proc. IEEE ICIG*, 2011, pp. 795–800.

[69] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence–arousal emotion space for video affective content representation and recognition," in *Proc. IEEE ICME*, 2009, pp. 566–569.

[70] M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," in *Proc. IEEE ICME*, 2005, vol. 61, no. 2, pp. 2–5.

[71] M. Soleymani, M. Larson, and T. Pun, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, Jun. 2014.

[72] Y. Baveye, J.-N. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret, "A large video database for computational models of induced emotion," in *Proc. IEEE ACII*, 2013, pp. 13–18.

[73] J. E. Cutting, K. L. Brunick, and A. Candan, "Perceiving event dynamics and parsing Hollywood films," *J. Experimental Psychol.: Human Perception and Performance*, vol. 38, no. 6, pp. 1476–90, 2012.

[74] U. Schimmack, "Pleasure, displeasure, and mixed feelings: Are semantic opposites mutually exclusive?," *Cognition & Emotion*, vol. 15, no. 1, pp. 1476–1490, 2001.

[75] O. Lartillot and P. Toiviainen, "A matlab toolbox for music feature extraction from audio," in *Proc. DAFx*, 2007, pp. 81–97.

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

[77] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[78] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[79] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Adv. Neural Inf. Process. Syst.*, vol. 9, pp. 155–161, 1997.

[80] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen*, LAPACK Users' Guide*, 3rd. ed. Philadelphia, PA, USA: SIAM, 1999.

Jussi Tarvainen (S'11) received the M.Sc.(Tech.) degree in media technology from Aalto University, Espoo, Finland, in 2011. He is currently a D.Sc.(Tech.) candidate studying computational modeling of film affect at the Department of Media Technology, Aalto University School of Science, Espoo, Finland. His research interests include content-based multimedia analysis and machine learning.

Mats Sjöberg (S'08) received the M.Sc.(Tech.) degree in engineering physics and mathematics from the Helsinki University of Technology, Finland, in 2006. He did his doctoral research on concept detection in multimedia retrieval at the Department of Information and Computer Science, Aalto University School of Science. He currently works at the Department of Computer Science at the University of Helsinki in a project which aims to revolutionize knowledge work supported by novel interactive user interfaces and machine learning methods.

Stina Westman received the M.Sc.(Tech.) degree in automation and systems engineering from the Helsinki University of Technology, Helsinki, Finland, in 2004 and the D.Sc.(Tech.) degree in media technology from the Aalto University School of Science, Espoo, Finland, in 2011.

Since 2014, she has been with CSC – IT Center for Science Ltd, Espoo, Finland, where she is currently an Application Specialist. Her research is focused on information seeking and retrieval in multimedia contexts.

Jorma Laaksonen (S'96–A'97–SM'02) received the D.Sc.(Tech.) degree from the Helsinki University of Technology, Finland, in 1997.

He is presently a Teaching Research Scientist at the Department of Information and Computer Science, Aalto University School of Science, Espoo, Finland. He is an Associate Editor of Pattern Recognition Letters and a founding member of the SOM and LVQ Programming Teams and the PicSOM Development Group. His research interests are in content-based multimodal information retrieval, machine learning and computer vision.

Pirkko Oittinen received the D.Sc.(Tech.) degree.

She is a Full Professor with the Department of Media Technology, Aalto University School of Science, Espoo, Finland. Her research has the mission of advancing visual technologies and raising the quality of visual information to create enhanced user experiences in different usage contexts. The research approach is constructive and seeks to cross disciplinary boundaries. Her current research focuses on interrelations between computational characteristics of still and moving images and visual experience.