Affective Analysis on Patterns of Shot Types in Movies

Luca Canini, Sergio Benini, and Riccardo Leonardi
Department of Information Engineering - University of Brescia
{firstname.lastname}@ing.unibs.it

***Abstract*—In film-making, the distance from the camera to the subject greatly affects the narrative power of a shot. By the alternate use of Long shots, Medium and Close-ups the director is able to provide emphasis on key passages of the filmed scene, thus boosting the process of identification of viewers with the film characters. On this basis, we here investigate the use of camera distance in famous movie scenes, highlighting the relations between the employed shot types and the affective responses by a large audience. Results obtained by using statistical classifiers suggest that patterns of shot types constitute a key element in inducing affective reactions in the audience, with strong evidences especially on the arousal dimension. Findings are applicable to support systems for media affective analysis, and to better define emotional models for video content understanding.**

I. Introduction

In the process of film-making, the characteristics of a *shot* (i.e. an uninterrupted run of camera take) are among those variables that are most directly under the director's control, such as *shot length*, intended as shot duration, *shot type* in terms of closeness of the camera to the subject, *camera movement* such as pan, tilt, zooms, *shot transitions* (cut, fades, dissolves, wipes), etc.

While a certain amount of work has been done in investigating most of these characteristics (as in the study in [1]), so far not much attention has been specifically directed towards the analysis of the shot type, that is the distance between camera and the main recorded subject [2].

Varying the camera distance from the subject of interest is a common directing rule used to subtly adjust the relative emphasis between the filmed subject and the surrounding scene [1]. This deeply affects the emotional involvement of the audience [2] and the process of identification of viewers with the movie characters. There are in fact evident correspondences between the filmmakers' choice of shot type and the *proxemic patterns* [3], i.e. the subjective dimensions that surround each of us and the physical distances one tries to keep from other people in social life.

Although the gradation of distances is theoretically infinite, in practical cases the categories of definable shot types can be re-conducted to three fundamental ones: *Long shots* (LS), *Medium shots* (MS), and *Close-ups* (CU).

A Close-up shows a fairly small part of the scene, such as a character's face, in such a detail that it almost fills the screen. This shot abstracts the subject from a context, focusing attention on a person's feelings or reactions, or on important details of the story. Different grades of Close-ups are presented in Figure 1-a.

Fig. 1. Shot types: a) Close-ups, b) Medium and c) Long shots, as in [2].

In a Medium shot, as in the case of the standing actors depicted in the examples of Figure 1-b, the lower frame line passes through the body from the waist down to include the whole body (in this case it is called *Full shot*, FS). In such shots, the actor and its setting occupy roughly equal areas in the frame, while leaving space for hand gestures to be seen. Medium shots are also frequently used for the tight presentation of two actors.

Finally, Long shots show all or most of a fairly large subject (for example, a person) and usually much of the surroundings. This category comprises also *Extreme Long shots* (as shown in Figure 1-c), where the camera is at its furthest distance from the subject, emphasising the background, often used as the opening shot of a sequence to set the scene (also called *Establishing shot*). The reader can refer to [2] for a more detailed taxonomy on shot types.

While making movies, directors choose the shot type which most accurately supports the emotional feeling of that scene: the greater the distance between the camera and the character, the more neutral the expected audience affective response, thus relating the adopted shot type to the corresponding proxemic pattern, as shown in Table I.

In this paper, we investigate the use of shot types in famous movie scenes, highlighting the relations between patterns of camera distances and the affective responses of a large audience. Data gathered by user-self assessments on the Pleasure-Arousal dimensions (PA) of the Pleasure-Arousal-Dominance emotional model [5] are here related to shot type properties by means of two statistical classifiers, that are Markov Models

TABLE I
PROXEMIC PATTERNS AND SHOT TYPES [4].

Proxemic	Social dist.	Emotion	Shot
Intimate	$< .5$m	love, vulnerability	CU
Personal	< 1m	friendship	MS
Social	< 3m	impersonal relat.	MS (FS)
Public	> 3m	formal, detached	LS

[6] and Support Vector Machines (SVM) [7]. The aim is to infer the emotional response of viewers by relying only on patterns of camera distance. We expect that specific sequences of shot types, e.g. an alternated use of different shot types versus the persistent use of a single one, differently affect audience reactions.

This paper is organised as follows. Section II explores recent advances in affective video analysis. Section III presents the overall methodology. Section IV describes the gathering of users' emotional responses to video data. Section V introduces scene modelling and the shot characteristics used for the affective classification of the scene, which is presented in Section VI. Conclusions are finally gathered in Section VII.

II. PREVIOUS WORK

Even if intriguing possibilities could be offered by an emotion-based approach to multimedia applications, the existing related works on affective analysis of video content are few, sparse and recent.

A practical way to assess the affective dimension of media is given by the use of the "expected mood" proposed by Hanjalic in [8], i.e. the set of emotions the film-maker intends to communicate when he/she produces the movie. His approach is based on direct mapping of specific video features onto the arousal and pleasure dimensions. He describes motion intensity, cut density and sound energy as arousal primitives, defining an analytic time-dependent function for aggregating these properties and using video frames as time dimension. However the examples of arousal mapping given in [8] refer to live sports events, whose properties may not transfer entirely to the case of other videos and feature films, which have different editing and whose soundtracks are of a different nature.

To date, emotional characterisation has been mainly used to study a narrow set of situations, like specific sport events as in [9] or movies that belong to a particular genre, for example horror movies, as in [10]. Extending this approach, Xu et al. [11] describe emotional clustering of films for different genres, using averaged values of arousal and valence, deduced from video parameters. De Kok [12] extends some aspects of this work by refining the modelling of colours, in an attempt to achieve a better mapping onto the valence dimension, while Kang [13] describes instead the recognition of high-level affective events from low-level features using HMM, a method also used in [14]. Performance obtained by Kang are outperformed in the work in [15]. It proposes to fuse audio and visual low-level features in a heterarchical manner in a high dimensional space, and to extract from such a representation meaningful patterns by an inference SVM engine. In the same work [15], authors corroborate the view that audio cues are often more informative than visual ones with respect to affective content.

So far, to the best of our knowledge, no studies have been performed regarding the relation between the usage of camera distance and emotional responses of movie viewers.

III. OVERALL METHODOLOGY

The main aim of this work is to relate the usage of shot types in movie scenes with the affective responses by viewers. To this end, Figure 2 explains the adopted workflow:

Gathering Emotional Rating: We first ask users to provide emotional annotations on content by positioning each movie scene in one sector of the Pleasure-Arousal model.

Scene Modelling & Shot Type Features: Then, shot type and other features related to the usage of camera distance are extracted and used to model scenes.

Affective Scene Classification: An attempt to highlight the correspondence between shot type properties and users' emotional responses is done using two statistical classifiers.

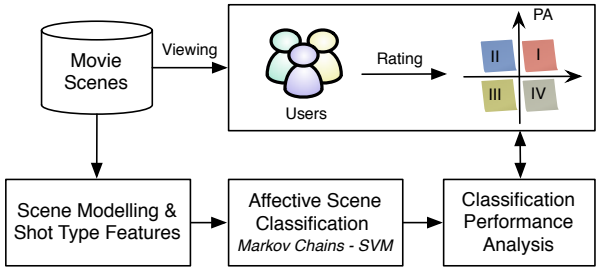

Fig. 2. Paper workflow.

IV. GATHERING EMOTIONAL RATINGS

In order to investigate the relationships between shot types and affective responses, we set up the following experiment. A number of 240 users are recruited: 195 are students at University of Brescia, while remaining 45 are chosen among colleagues and friends. The experiment is in the form of a user test and it is performed online. Data consist of 83 "great movie scenes" [16] chosen to represent popular films spanning from 1958 to 2009 from IMDb [17], for a total duration of more than 3 hours of video and 2311 shots. Complete information on the data set are provided in [18], with scene durations, shot numbers, and movie names.

To perform the test, every user is asked to watch and listen to 10 randomly extracted movie scenes out of the total 83, in order to complete the test within a total time of 30 minutes. Scenes can be watched as many times as users want, either in English or Italian, and the whole test can also be interrupted and resumed in different moments. Users are requested whether they have seen the scene and/or the movie before, and in case, they are asked for the movie title.

After watching each scene, the user is asked to annotate the emotional state he/she is inspired with on the *emotion wheel* in Figure 3-a. This model is a quantized version of the Russell's

circumplex [19] and presents, as in the Plutchik's wheel [20], eight basic emotions as four pairs of semantic opposites: "Happiness vs. Sadness", "Excitement vs. Boredom", "Tension vs. Sleepiness", "Distress vs. Relaxation".

On the basis of the most rated emotion, each scene is finally placed in one among the four sectors (I, II, III, IV) of the Pleasure-Arousal model (as in Figure 3-b). For this self-assessment phase, the emotion wheel in Figure 3-a is preferred to a direct rating on the PA model, since it is simpler for the users to provide one or more emotional labels than to express their emotional state by a combination of values of pleasure and arousal.

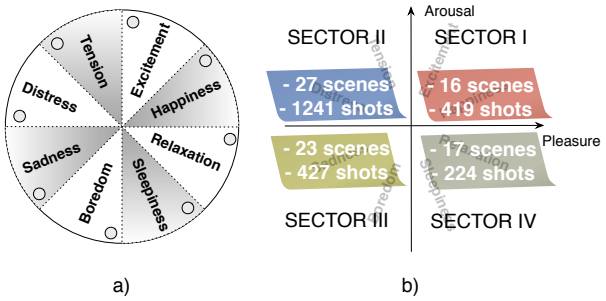

Fig. 3. a) The emotion wheel used by users to annotate emotions; b) Scene distribution among the four sectors of the PA model.

The choice of a "scene" as a elementary unit for analysis is supported by the fact that each scene in a movie depicts a self contained high-level concept, mostly autonomous in its meaning [15] even when excerpted from the movie. *Great scenes* are chosen since they more easily elicit emotional reactions in the viewer since "they are our memories of segments of films that have achieved a life of their own, compelling us to remember and relive the moment" [16].

Selected scenes are chosen so as to expectedly cover all categories of elicited basic emotions, while it is not our goal to cover all content variability of thousands of existing movies. In this sense, selected scenes offer a sufficiently broad spectrum to characterise the limited variability of affective reactions of the audience to movies.

V. Scene Modelling & Shot Type Features

Given the video to investigate, for each shot we compute the corresponding shot type (LS, MS, CU) first by applying any existing technique for shot boundary detection, and then by using an improved version of the method for shot type identification we describe in [21]. As an outcome each movie scene of the database is described as a sequence of shot types; on this basis, the following models and features are proposed to characterise different sectors of the Pleasure-Arousal plane.

A. Markov Chain Modelling

While shooting a scene, a director not only chooses a particular set of shot types, but pays also attention to their sequence, that is the temporal order in which they appear on screen. The presence of a particular pattern or, more basically the passage from a shot type to another one, is part of a group of rules known as film grammar [2]. Being able to model this temporal dependency would be crucial for a good understanding of the filmic product, but the choice of the proper mathematical model can be challenging, since cinematographic rules are sometimes loose and act as general guidelines, also to allow a director to fully express his/her own creativity.

As an attempt to capture these temporal relations we adopt Markov chains that are discrete state stochastic models that work well for temporally correlated data streams [6]. Each Markov chain is described by three *states* $S = \{LS, MS, CU\}$, one for each shot type, the *transition probabilities distribution* $T = \{t_{ij}\}$ that models transitions between consecutive states, and the *initial state distribution* $\pi = \{p_1, p_2, p_3\}$.

A Markov chain can model each sector of the Pleasure-Arousal plane as in Figure 4, or characterise different PA hemiplanes, e.g. the high/low arousal hemiplanes (sector I $\cup$ II, and sector III $\cup$ IV, respectively) or the high/low pleasure ones (sector I $\cup$ IV and sector II $\cup$ III, respectively).

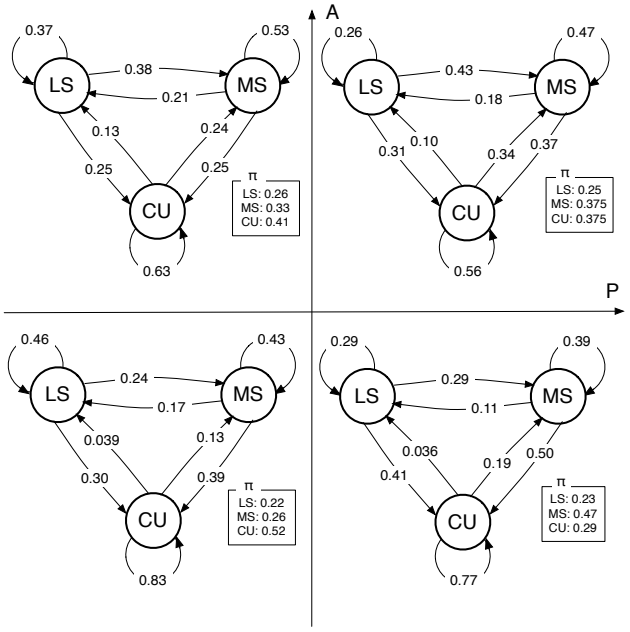

Fig. 4. Markov chains modelling scenes in the four sectors of the PA plane.

Observing the figure[1], some considerations arise. First for all models, LS is the least likely shot for a scene to start with: they are employed only to open particular scenes, describing the surroundings and the atmosphere which characterises an entire section of a movie, or to close a whole narrative chapter. Second, when passing from a LS to another type, the probability to come back is low, especially if the intermediate shot is a CU. Again, this is imputable to the fact that LS are mainly introductory and when the scene evolves they are

[1]The sum of transition probabilities outgoing each state in Figure 4 may differ from 1 due to numerical approximations.

seldom used. Third, once in a CU state there is a considerable probability to remain in that state. This is particularly evident for scenes with low arousal with a significant presence of dialogues, which are central for the development of the plot, especially in non-action movies.

B. Shot Type Features

The Markov models proposed so far to characterise different sectors of the PA plane are here enriched with other features derived from the shot type.

Stationary distribution $\overline{\mu}$: For each PA sector, the presence of each shot type can be computed as the *stationary distribution* $\overline{\mu}$ of the associated Markov chain. For any finite state Markov chain with a unique stationary distribution, $\overline{\mu} = \{\mu_{LS}, \mu_{MS}, \mu_{CU}\}$ is found as solution of

$$\overline{\mu}\, T = \overline{\mu} \tag{1}$$

We present in Figure 5-a the stationary distributions in the four sectors of the PA model, which account for the percentages of different shot types in the four sectors. By inspecting Figure 5-a, we observe that Close-up presence is high in all sectors of the PA model. This might induce to think that CU are adopted for communicating a wide set of emotions. We rather tend to believe that this is due to the diegetic nature of many modern movies, where the plot often unfolds thanks to dialogues between characters who narrate situations and events. Therefore, CU percentage is very high in sectors III and IV which contain scenes - typical of certain dramatic content - characterised by low levels of arousal and with a massive presence of dialogues.

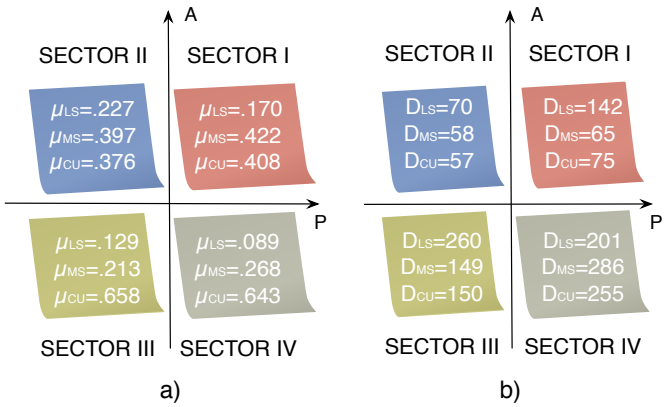

Fig. 5. a) Presence of shot types $\overline{\mu}$ and b) average shot duration $\overline{\mathcal{D}}$ in PA sectors.

The use of Close-ups reduces on the contrary in scenes of action, when there is the need to convey a general impression rather than specific information. In this case, the use of motion (both camera and objects) becomes very important, for example in communicating *excitement* or other positive aroused feelings, so that camera distance needs to increase to include a larger view on happenings. As a further confirmation, we observe that Long shots are mainly gathered in the PA hemiplane expressing high values of arousal.

Finally, Medium shots are probably not specific to a definite set of emotions, thus finding a fair level of employment in all types of filmic material.

Shot Type Duration $\overline{\mathcal{D}}$: Markov chains do not provide information on "how long" the model remains in a state. This is computed aside as the average duration of shots (in frames) for all types: $\overline{\mathcal{D}} = \{\mathcal{D}_{LS}, \mathcal{D}_{MS}, \mathcal{D}_{CU}\}$, which are shown in Figure 5-b for each sector of the PA model.

By inspecting Figure 5-b, we first observe that LS last longer than the other types of shots in most sectors, thus confirming their descriptive or introductory role. Second, notice that sector II is characterised by shots of short duration (about 60 frames on average): frequent shot cuts are typical for these scenes, which are annotated by users with labels such as *tense* or *distress*. In general this is also true for sector I, characterised by high level of arousal, even though the associated content in this case expresses more positive feelings such as *happiness* and *excitement*. We conclude that the permanence of the model in a state is another potential clue able to distinguish video content located in the two opposite arousal hemiplanes.

Entropy rate $\mathcal{H}$: A measure accounting for the chain complexity is given by the *Markov entropy rate*, which is a measure of the difficulty to predict the process evolution. It is defined as in [6] as:

$$\mathcal{H} = -\sum_{i,j} \mu_i \cdot t_{ij} \cdot log(t_{ij}) \tag{2}$$

When applied to the case of a movie scene, entropy rate estimates the difficulty to predict the shot type that will appear after the current one. Measures of Markov entropy rates for different sectors of the PA model are given in Figure 6-a. Observing the figure, there is a strong evidence that high arousal sectors are less predictable in shot type behaviour than low arousal ones.

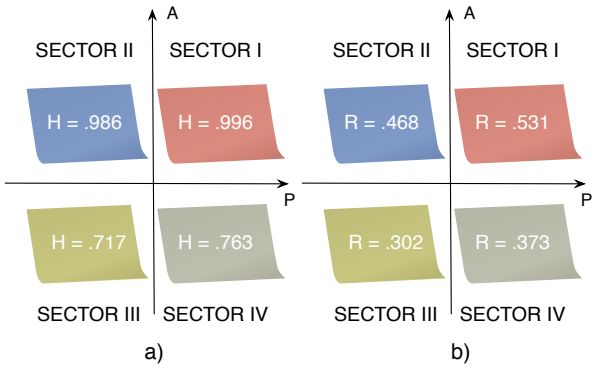

Fig. 6. a) Markov Entropy Rate H and b) Shot Type Change Rate $\mathcal{R}$ for the PA model.

Shot type change rate $\mathcal{R}$: Another interesting index is the one related to the changes in shot types across consecutive shots. In the specific it can be computed as:

$$\mathcal{R} = \sum_{i,j \neq i} \mu_i \cdot t_{ij} \tag{3}$$

Notice that while low values of $\mathcal{R}$ imply low values of $\mathcal{H}$, the opposite is not always true. It is interesting to observe in Figure 6-b that higher values of $\mathcal{R}$ are noticeable in sectors I and II of the PA model, suggesting that a frequent change of shot type might be one of the mechanisms responsible for inducing aroused emotional reactions in viewers. These are the same scenes where CU are least employed (i.e. *exciting* or *tense* ones). Conversely, the permanence of the same camera distance across consecutive shots increases (i.e. lower level of $\mathcal{R}$) in low arousal sectors, which are those where we register a massive presence of CU, often focusing on character's reactions and feelings.

VI. Affective Scene Classification

From the previous analysis, we expect that the ways directors use camera distance may be effective in inducing certain emotional states in viewers. In the following we provide proof of this assumption by verifying the ability of the models and shot type features described in Section V to perform affective scene classification, at first using Markov chains, then employing SVM.

A. Markov Chain Classifiers

We here test the ability of the previous Markov models to discriminate between emotional states. At first the classification task is carried out by using the four models in Figure 4, thus considering each sector of the PA plane as a target class. Each scene is classified by inspecting its sequence of shot types and selecting among the four Markov chains the one that most likely generates the sequence in exam. Classification performance against the ground-truth, however, are not satisfying. Although Markov chains offer a statistical model allowing a good understanding of the temporal dependencies between shot types, their classification ability is limited by the restricted set of model parameters $\{S, T, \pi\}$.

A further attempt of classification without employing the *extra* features described in Section V-B tries to focus on pleasure and arousal hemiplanes. To this purpose we model the high and low arousal hemiplanes (sectors I $\cup$ II and sectors III $\cup$ IV, respectively) by two distinct Markov chains, and the same is done for the low and high pleasure ones (sectors II $\cup$ III and sectors I $\cup$ IV, respectively). Performance obtained with respect to the pleasure axis are again not promising, while it is worth analysing those based on the discrimination low/high arousal (reported in Table II).

To obtain them, we perform a five-fold cross-validation [22]: the scene database is divided into five subsets using stratification (thus ensuring that each class is represented with approximately equal proportions in the folds - classes here are low/high arousal) and then five iterations of learning and validation are executed such that at each iteration a different fold of data is used for validation and the remaining four to build the two Markov models. At every run, each scene in the validation fold is mapped on the two learned models, collecting the probabilities associated to each edge it crosses. As before, the scene is said to belong to the class whose model obtains the highest sum of probabilities.

Classification results are averaged on the five runs and reported in Table II in terms of *accuracy*, *precision* and *recall*. Accuracy is the most common way of assessing classication results and it measures the proportion of true results (both true positives and true negatives). Precision indicates how many from those marked as positive (true and false positives) are actually so, while recall considers how many positives are correctly classified.

TABLE II
Classification performance obtained using Markov models on arousal hemiplanes.

Class	**Acc. (%)**	Prec. (%)	Rec. (%)
High arousal	**59.1**	55.2	80.0
Low arousal	**59.1**	68.0	39.5

Results in Table II, even if just above the threshold of classification by coin toss, however suggest that the considered Markov chains provide at least a first level of discrimination between scenes with high and low level of arousal. Considering that also the shot type features presented in Section V-B suggest the same ability in discriminating scenes with respect to arousal, in the next section we discuss a model implemented by using SVM which takes into account the shot type properties introduced in Section V-B.

B. SVM Classifiers

To validate the hypothesis that characteristics related to the shot type are linked to the arousal state conveyed by movie scenes to the viewers, we here build a model based on SVM which is able to perform affective classification on arousal starting from shot type properties discussed above.

SVM are supervised learning methods used for classication and regression, playing an increasing role in signal processing, pattern recognition and image analysis. The principle is that, given two classes of data which are not separable by a linear function, a SVM projects data into a higher dimensional space (via kernel representation), where the separation problem is solved by building an optimal separating hyperplane which maximises the functional margin.

To feed the learners we use the descriptors presented in Section V-B, thus obtaining the following eight-dimensional vector of features: $\{\overline{\mu}, \overline{\mathcal{D}}, \mathcal{H}, \mathcal{R}\}$. As for the models in Section V-A, classification task is performed by using a five-fold cross-validation scheme, employing stratification to divide data. In each run the penalty term C and ξ of a standard RBF kernel $K(x, y) = \exp(-\xi \left\|x - y\right\|^2)$ are obtained via cross validation for parameter selection via a process of grid search on four folds. The best couple ($\hat{C}$, $\hat{\xi}$) is then used to train the four folds and generate the final model, which will be tested on the fifth fold. Performance are then averaged on the five runs, to obtain the results shown in Table III.

Results reported in Table III indicate that the SVM models built using the shot type characteristics are highly effective in classifying the level of arousal conveyed by a scene and

TABLE III
CLASSIFICATION PERFORMANCE BY SVM.

Class	**Acc. (%)**	Prec. (%)	Rec. (%)
High arousal	**80.7**	77.6	88.3
Low arousal	**80.7**	85.3	72.6

perceived by the viewers. As an outcome of this analysis we conclude that camera distance and the use of specific patterns of shot types play an evident role in the mechanism of arousal elicitation in the audience, while their effect on the pleasure emotional component, if existing, is not yet proven.

VII. CONCLUSIONS

In this work we performed a study in the field of emotional analysis of movies which demonstrates the evident role of camera distance in the mechanism of emotion elicitation. Tests have been carried out so that to highlight the relations between the employed sequences of shot types and the affective responses by a large audience. As an outcome, we reveal that by the alternate use of Long shots, Medium and Close-ups, the director is able to drive the level of perceived arousal of the watched scene. Conversely, it is still unproven the existence of a connection between the usage of shot types and the pleasure dimension of the emotional state. Findings are applicable to support systems for media affective analysis, video content creation, and tools for assisting automated editing.

REFERENCES

[1] H. L. Wang and L.-F. Cheong, "Taxonomy of directing semantics for film shot classification," *IEEE Transactions on Circuits and Systems for Video Technologies*, vol. 19, pp. 1529–1542, October 2009.

[2] D. Arijon, *Grammar of the Film Language*. Silman-James Press, September 1991.

[3] E. T. Hall, *The Hidden Dimension*. Anchor, October 1990.

[4] L. D. Giannetti, *Understanding Movies, 2nd edition*. Prentice-Hall, 1976.

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

[6] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

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

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

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

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

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

[14] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden markov models," in *The 2nd international conference on Affective Computing and Intelligent Interaction*, Berlin, Heidelberg, 2007, pp. 594–605.

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

[16] "What is a "Great Film Scene" or "Great Film Moment"?" http://www.filmsite.org/scenes.html.

[17] "Internet movie database," *www.imdb.com*. [Online]. Available: www.imdb.com

[18] "Database of movie scenes for ispa-2011: a complete list," *www.ing.unibs.it/∼luca.canini/pub/ispa2011.html*. [Online]. Available: http://www.ing.unibs.it/luca.canini/pub/ispa2011.html

[19] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, pp. 1161–1178, 1980.

[20] R. Plutchik, "The Nature of Emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[21] S. Benini, L. Canini, and R. Leonardi, "Estimating cinematographic scene depth in movie shots," in *IEEE International Conference on Multimedia & Expo*, Singapore, 19-23 July 2010.

[22] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross validation," *Encyclopedia of Database Systems*, 2009.