ISCA Archive
http://www.isca-speech.org/archive

2nd Workshop on Speech, Language and Audio in Multimedia (SLAM 2014)
Penang, Malaysia
September 11-12, 2014

Combining audio-visual features for viewers' perception classification of Youtube car commercials

F. Fernández-Martínez, A. Hernández-García, A. Gallardo-Antolín, F. Díaz de María

Universidad Carlos III of Madrid, Leganés
Department of Signal Theory and Communications

ffm@tsc.uc3m.es, ahgarcia@tsc.uc3m.es, agallardo@tsc.uc3m.es, fdiaz@tsc.uc3m.es

Abstract

In this paper, we present a computational model capable of predicting the viewer perception of Youtube car TV commercials by using a set of low-level audio and visual descriptors. Our research goal relies on the hypothesis that these descriptors could reflect to some extent the objective value of the videos and, in turn, the average viewer's perception. To that end, and as a novel approach to this problem, we automatically annotate our video corpus, grouped into 2 classes corresponding to different satisfaction levels, by means of a regular *k*-means algorithm applied to the video metadata related to users feedback. Evaluation results show that simple linear logistic regression models based on the 10 best visual descriptors and on the 10 best audio descriptors individually perform reasonably well, achieving a classification accuracy of roughly 70% and 75%, respectively. Combination of audio and visual descriptors yields better performance, roughly 86% for the top-20 selected from the entire descriptor set, but tipping the balance in favor of the audio ones (i.e. 17 vs 3). Audio content bigger influence in this domain is also evidenced by a side analysis of the video comments.

Index Terms: subjective assessment, video aesthetics, Music Information Retrieval, video metadata

1. Introduction

In a world where new technologies are increasingly more related to multimedia information, the development of tools to make it easier dealing with this type of data becomes essential. One problem that has attracted much research interest in recent years is the development of models to extract subjective information from objective data. Particularly, inferring the perceived value by the potential consumers of multimedia resources (e.g. Youtube commercials) by means of automatic procedures, aimed at analysing both audio and visual content, would be of great application for developing more efficient indexing and recommendation systems.

There are different fields that study computational procedures to extract subjectivity of data, such as sentiment analysis [1, 2, 3]. From a visual content point of view, the one that we are concerned about is aesthetics assessment, which was firstly applied to still images. One of the earliest works on this domain was carried out with the goal of finding out which features correlated better with rankings [4]. More recently, Datta *et al.* [5] proposed 56 low-level image features tested on 3581 pictures with ratings from the web site *Photo.net* and selected the top 15 features that achieved together an accuracy of 70.12% in separating low from high rated photographs. After this successful achievement, several studies followed this line of research adding different contributions [6, 7].

Applied to videos, aesthetics assessment has been only addressed very recently. Nonetheless, low-level visual descriptors have already proven to be indicative of the aesthetic value of the videos and, in turn, of their viewers' perception. To our knowledge, the first attempt to model visual aesthetics in moving pictures was addressed by Moorthy *et al.* [8] in 2010. They collected 160 consumer videos from YouTube and performed a controlled user study to obtain rating labels as ground truth. Then, different frame-level features based on those from [5] and on users reports were extracted from the videos and extended to the temporal level. Finally, they selected the 7 most relevant features and after classification procedures they achieved an accuracy of 73.03%.

On the other hand, background music accompanying commercials has also been shown to be a major component influencing audience responses. For example, Alpert and Alpert early presented an study [9] suggesting that audience moods and purchase intentions may be affected by background music. More recent studies like [10] have clarified that, despite the widespread assumption that virtually any product advertisement is enriched by the mere presence of music, there are much empirical evidence casting doubt on this and suggesting that music can have a neutral or detrimental effect (e.g. audience could perceive the music as innapropriate or unsuitable for the brand message, or they may simply dislike the musician or find the tune annoying or boring) as well as a positive one.

Other related studies have also pursued into the goal of understanding how individuals emotionally respond to common advertisement sounds. For example, [11] suggested a hypotheses based model for predicting the emotional reaction and empirically tested it using data from 153 laboratory participants and 20 different sounds. Results from a survey asking participants about their emotional response towards each particular sound indicated that the emotional response to a sound clip can be predicted by the level of interest generated and how well the sound captured the participant's attention.

Visual (images and scenes) and audio (music and sounds) content in general can help to achieve some cognitive effects on the audience (e.g. attracting attention) and induce some affective responses as well (e.g. creating a particular mood), that can be both considered advertising objectives. However, call for further research on the factors that determine whether both components have a positive, negative or no effect on consumer response to advertising still remains. In this regard, measuring the relative importance of audio content compared to visual on the audience final perception of a commercial seems to be a particularly interesting and worthwhile issue to be investigated.

In this regard, and to the best of our knowledge, all the existing works in inferring the perceived value of videos have used

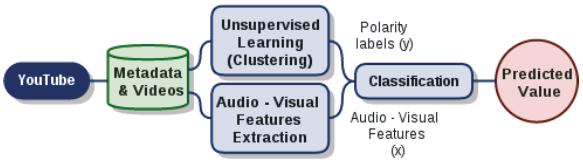

Figure 1: Diagram of the approach overview.

data sets whose videos have been specifically rated for the task through controlled user surveys. This approach has the disadvantage that the ratings do not completely reflect the real effect of the videos on the regular users who watch them in sites like YouTube, since the surveys are performed by a limited number of people who have been given some instructions.

Conversely, in this paper we propose a novel approach to this problem consisting in automatically deriving the ground truth polarity labels by means of an unsupervised learning algorithm applied to the video metadata, which are available at YouTube and have been provided by actual users of the platform as they watch and share the videos, hence, better and more fairly reflecting the actual perception of the videos. This new approach lies on the hypothesis that metadata such as the number of *likes* or the number of views are indicative of the subjective assessment the users give. An overview of the process can be observed in Figure 1.

Therefore, the two main purposes of this paper are:

- to present a novel approach based on Youtube popularity metrics for the automatic annotation of videos (i.e. car commercials) in terms of their expected or potential perceived value.
- to expand upon existing research to investigate how audio-visual content can influence Youtube popularity metrics as common measures of advertising effectiveness.

The automatic analysis of the audio-visual content of an ad allows deriving suitable principles for predicting the above mentioned effects. In this regard, the present work provides some suggestions for the construction of effective computational models of audio and visual content influence on emotions and product orientations. Furthermore, the paper also indicates directions for future investigations of multimodal approaches for analysing the content of the commercials, inferring the advertising effectiveness, and measuring the influence of each tested modality.

The paper is organised as follows: after this introduction Section 2 provides the details of the video corpus acquisition and clustering procedure. Section 3 and 4 repectively describe the audio and visual descriptors extracted for the classification task. Section 5 presents the classification results including corresponding discussions and issues. Finally, some conclussions and future work are laid out in Section 6.

2. Corpus acquisition and clustering

One of the contributions of this work is the automatic annotation of the corpus, instead of using labels obtained after a user survey specifically prepared for the research [8, 12, 13]. For this reason, our first decision regarding the preparation of the corpus was to acquire a suitable one.

2.1. Video domain selection, download and filtering

Selected video domain was basically conditioned by two main requirements. First, it was necessary to have an acceptable amount of metadata so that the clustering algorithm was able to find meaningful clusters. Second, the differences in the metadata between videos should be indicative of the better or worse appreciation of the videos by users. Car commercials reasonably satisfy both.

Once defined the domain, an initial corpus of 2,315 car commercials, and their metadata, was downloaded from YouTube through its API. All the videos were in Spanish language and published after 2010. However, additional filtering procedures were necessary: first, we removed any video that was not a professional car advertisement (i.e. videos with a duration longer than 115 seconds or shorter than 10 seconds). Second, we removed any video without enough metadata and thus, impossible to annotate (i.e. a minimum of 3 raters per video was considered). At the end of the filtering, 138 videos remained and took part of our data set.

2.2. Clustering

In addition to the raw popularity metrics that YouTube provides, we defined two processed metadata in order to simplify the clustering procedure and the interpretation. First, we merged the *likes* and *dislikes* metrics into a single one: the $likes\text{-}dislikes\ ratio$, computed as the proportion of likes from the total number of votes. The other new metric we introduced was the *view-score*, a score in a 1-to-5 scale assigned accordingly to the 20^{th}, 40^{th}, 60^{th}, 80^{th}, and 100^{th} percentile ranks computed for the number of views, respectively.

The set of metadata that were selected to perform the cluster analysis are: likes-dislikes ratio, view-score, number of comments, number of raters and rating. The cluster analysis was performed through the *k*-means algorithm [14], using the city block (Manhattan) distance measure. As a result of this clustering process, videos were labeled into 2 different classes: 76 "good" or "better" videos on one hand, and 62 "bad" or "worse" videos on another, thus composing the annotated dataset on which classification experiments will be tested.

3. Audio Descriptors

For the extraction of the audio features we have used MIRToolbox [15], a suitable tool for implementing computational approaches in the area of Music Information Retrieval (MIR). MIRToolbox allows the extraction of a large set of musical features from audio files.

Each musical feature is related to the different broad musical dimensions traditionally defined in music theory. We have extracted features related, among others, to tonality, rhythm, timbre, and form. Statistical moments such as centroid, kurtosis, etc., can be applied to either spectra or envelopes, but also to any histogram based on any given feature thus increasing the number of different features up to roughly 400.

In the next subsections we will highlight some of the most interesting features providing potential psychoacoustic evidence of influencing viewer's emotional response.

3.1. Tonality

Generally, musical compositions are organized around a central note, the tonic. Music periodically returning to that central tone exhibits tonality. Specifically, tonality helps to arrange sounds

according to pitch relationships into interdependent spatial and temporal structures, thus characterizing notes, chords, and keys (sets of notes and chords with an specific hierarchy).

The potential for contrast and tension inherent in the chord and key relationships of tonality is well known (e.g. any modulation or movement away from the tonic key creates tensions that may then be resolved by modulation back to the tonic). Hence, related features could be suggested to model the viewers' emotional response.

Examples of tonality features extracted by MIRToolbox are: chromagram (distribution of the signal's energy across a predefined set of pitch classes) and key strength.

3.2. Rhythm

Rhythm, in music, is the placement of sounds in time. In its most general sense rhythm is an ordered alternation of contrasting elements.

Different rhythms or music patterns in time may elicit different reactions and emotional responses from viewers. Rhythm is likely to vary according to the interpretative ideas of the spots' producers who seek to produce some desired effects on the audience. Particularly it could be intentionally deviated for a particular piece of music to better fit or adjust to the visual structure and content of a commercial video.

MIRToolbox enables the estimation of several features related to rhythm: namely tempo, pulse clarity and fluctuation.

3.3. Timbre

Timbre means sound color. In music, timbre is the quality of a musical note or sound or tone that distinguishes different types of sound production, such as voices and musical instruments, even when they have the same pitch and loudness.

As a psychoacoustic hypothesis we expect different timbres (e.g. richness of timbre) or timbre variations (e.g. only music, music and voices, only voices or silence), measured during the spots, could be also indicative of the subjective experience of the viewers while watching them.

The physical characteristics of sound that determine the perception of timbre include spectrum and envelope. MIRToolbox computes, among others, Mel-Frequency Cepstral Coefficients (MFCC), the time envelope in terms of rise, duration, and decay, changes both of spectral envelope and fundamental frequency, as well as many other related basic statistics (e.g. zero-crossing rate, spectral centroid, roll-off, brightness, flatness, etc.).

3.4. Roughness

MIRToolbox provides an estimation of the sensory dissonance, or roughness, related to the beating phenomenon whenever pair of sinusoids are closed in frequency. Particularly, total roughness is measured by computing the peaks of the spectrum, and taking the average of all the dissonance between all possible pairs of peaks. The perceived roughness of a sound is simply how rough it sounds. Assuming rough sounds to be inherently bad or unpleasant, and therefore to be avoided, we can conclude that roughness and annoyance are strongly linked.

4. Visual Descriptors

Regarding the implemented visual features, we have inspired the decision of what features to test in previous works, such as those from [5] and others, who proved the convenience of some descriptors for assessing the aesthetic value, but also in different domain specific characteristics of the videos. We have extracted a total of 21 features, which we present according to the visual aspect they describe.

4.1. Temporal segmentation

In film-making and publicity temporal segmentation is of great importance, since it is the basis of montage, the main source of semantic effects. Quantitatively, the level of segmentation, i.e., the number of cuts, can be an indicator of the type of scene [16]. Transitions between two subsequent shots were determined as in [17] and the following features were extracted: absolute number of cuts, longest-shot, mean-shot-duration, standard deviation of the shot duration, and mean-cuts-per-min.

4.2. Intensity

Intensity is also an important characteristic in film-making and photography, usually referred to as brightness or exposure. It was used in [5] and we extend its meaning to the temporal dimension by computing the average intensity and the standard deviation along all the frames.

4.3. Entropy

When applied to image processing, entropy can describe textures. We have computed the entropy of the gray-scale version of the frames and derived the following features: avg-entropy, std-entropy, pct-low-entropy-frames, which detects the percentage of very simple frames, e.g. with monochromatic background, usually present in car commercials, and a feature that detects if the end of the video has very low entropy.

4.4. Color

Color is a very descriptive characteristic of images and videos which we have translated into computational features following the work of [5]. First of all, we make use of the HSV color model [18] for computing features related to the hue and the saturation (i.e. means and std deviations). Furthermore, colorfulness, a feature that measures how colorful the video is, is computed by extending the implementation of Datta *et al.*

4.5. Rule of thirds

The rule of thirds (ROT) is one of the most important rules of thumb for composition in visual arts, such as photography, painting or design. Among other uses, it is followed for placing important horizontal lines in the image, such as the line of the horizon. Placed in the lower third, it will give more priority to the sky, while placed in the upper third, it will increase the importance of the ground. We have developed a computational method to measure the degree of utilization of ROT. The idea is to compare differences in the color histograms corresponding to the two sub-images that the upper or the lower horizontal lines generate. Hence, the higher the difference, the higher the degree of utilization of ROT, and vice versa.

5. Results and discussion

After annotating the video dataset for 2 different classes and extracting the visual and audio features presented in sections 3 and 4, it was time to perform adequate classification experiments towards evaluating both individual and joint performance of these features when modelling the users' perception.

First of all, a feature selection step is found to be essential, not only to reduce the dimensionality and complexity of the feature-space, but also for a proper and fair comparison to be done between visual and audio features. In this case, we decided to apply the SVMAttributeEval feature selection algorithm provided by the WEKA machine learning software [19]. SVMAttributeEval evaluates attributes using recursive feature elimination with a linear support vector machine. Attributes are selected one by one based on the size of their coefficients, relearning after each one. As a result, a ranked attribute list is generated.

On the other hand, simple linear logistic regression models were used for classification. In this regard, and as part of the adopted experimental setup, each classification result reported has been obtained by performing 10 repetitions of a 10-fold cross-validation scheme.

Finally, the above mentioned classifier is compared to a ZeroR classifier (i.e. predicts the majority class), by means of a corrected paired t-test to check for significance (95% confidence interval). Hence, the evaluation of the results will strictly focus on those which prove to be significantly better than such reference result.

Approach	Top 10 visual	Top 10 audio	Top 20 audio-visual
Feature subset	(3) temporal {(2) shot duration (1) cuts per min (1) intensity {(1) stdev (3) entropy {(2) low entropy (1) average (2) color {(1) hue (1) colorfulness (1) ROT {(1) upper third	(2) tonal {(2) chromagram (1) keyclarity (1) rhythm {(1) tempo (7) spectral (timbre) {(4) mfcc (2) dmfcc (1) flatness	(17) audio {(3) tonal {(2) chromagram (1) keyclarity (1) rhythm {(1) attack time (13) spectral {(6) mfcc (4) dmfcc (1) irregularity (1) zerocross (1) roughness {(1) stdev (3) visual {(1) intensity {(1) stdev (1) ROT {(1) upper third (1) color {(1) hue
Accuracy (stdev)	69.16 (10.59) v	74.79 (11.54) v	85.25 (8.99) v
ZeroR	55.05 (2.85) (results tagged with 'v' when significantly better than this)		

Table 1: Experimental results for each feature subset.

5.1. Individual performance

Given the more reduced set of visual attributes we decided to adopt this as our baseline. Therefore, we started by testing different values for the number of selected attributes. Optimal performance was observed for 10 features. Corresponding accuracy and related top performance features have been detailed in Table 1. According to the evaluation of the visual features, it is important to remark that all the different types of visual features tested, i.e. temporal, entropy or color based, and related to ROT, have attained notable success (there is at least one representative of each in the top-10) thus complementing each other reasonably well.

Then, for a fair comparison between both types of features, we decided to perform an attribute selection process over the entire set of audio features defining a target number of selected features similar to the one showing top performance for visual features (i.e. 10). Resulting performance, as well as related features, have been presented also in Table 1. As it can be observed, top-10 audio features clearly outperform top-10 visual features, hence suggesting a greater influence or impact of audio related features when attempting to model the viewer's satisfaction. We can consider this result our first evidence in that regard. On the other hand, predominance of timbre features (i.e. spectral) is observed among the selected features.

5.2. Joint performance

Next, we combined both audio and visual features to evaluate their joint performance. However, rather than simply taking the top-10 visual features and directly combine them with the top-10 audio ones, we decided to re-run the attribute selection process with the whole set of available features, regardless of their audio or visual nature. To that effect, we adopted a number of 20 selected features as our target, mainly for comparison purposes with the previous individual approaches. Assuming such a reference, expected selection should be the simple combination of both top-10 sets in case of a similar relevance for both types of features. On the contrary, resulting top-20 selection was mostly composed of audio features as it can be observed in Table 1, where only 3 out of the 20 top features were visual. This can be considered our second evidence of the better fit of audio related features when modeling the viewer's satisfaction in this particular domain. For completeness, we extended the analysis to higher number of selected features with similar results (i.e. top accuracy 86.31% was achieved with 35, including the same 3 visual features).

5.3. Analysis of comments

Comments can be a powerful resource to identify the sentiment, attitudes, and emotions that viewers attach to the commercials. Hence, we carefully examined the content of all the comments corresponding to the videos in our dataset. Particularly, and in order to find further evidences of the greater influence of audio content on viewers' perception, we manually tagged each comment either as related to audio or not, mainly by identifying references to the music, the sound effects, or the person speaking in the videos. Similarly, visual comments were also tagged by identifying those specifically addressing: objects, places or persons appearing in the spot, explicit references to the montage or the producers, to a particular scene of the spot, etc. As an example of the former, a viewer might comment the following about a particular video: *"Wow! I'm absolutely in love with this song! Can't stop listening to it!"*. Regarding the latter, a possible example would be: *"Amazing landscape!"*.

As a result, Table 2 summarises the corresponding statistics, we measured roughly 40% of comments to be connected to audio related issues while only 16% to visual. This significant imbalance can be considered a third evidence of audio features prevailing over visual ones in our Youtube metrics based perception model.

Type	# comments	Percentage
Audio	328	40.39%
Visual	129	15.89%
Others	355	43.72%
Total	812	100%

Table 2: Analysis of comments.

6. Conclusions and Future Work

In this paper we have presented a computational method for assessing the perceived value of car commercials retrieved from YouTube.

First, the significant results we have obtained successfully validate the use of clustering techniques as an alternative for the automatic annotation of a video corpus in terms of Youtube popularity metrics as a suitable model of viewers' perception.

Second, the performed classification experiments have also demonstrated that both audio and visual content have an important influence on viewer's perception and advertising effectiveness. Indicative automatically extracted features have been identified in both cases. In this regard, the subset of selected visual features is relatively diverse, whereas in the audio one timbre related features seem to be predominant.

Third, we have decomposed both factors measuring their relative influence in modelling viewers' impressions. Suitable experimental setup has been adopted to validate this assessment thus providing a better explanation of the emotional response to commercials in this domain. Particularly, although visual content and related features have proven to be helpful, their role turns to be rather complementary when compared to audio. This result has been found to be coherent with a detailed analysis performed over the comments of the videos.

These results enable further research following the suggested approach to improve the performance of classification and recommendation systems. In the future, apart from increasing the size of the data set, it would be also interesting to explore the possibility of extending the approach by including textual features. In this regard, the clustering based annotation procedure could benefit from the application of natural-language processing (NLP) techniques. For instance, sentiment analysis may be performed to classify the polarity of the related comments, hence enabling their use as relevant metadata to be further included in the annotation process.

Finally, computational models for predicting the audience perception of a commercial should definitely account for the effect of other variables such as [9]: the audience demographics, personality and life-style, cognitive and affective involvement in the communication setting, familiarity with the music, the places shown, the actors, and the interaction of all of these with the product and use-situation stressed in the commercial. Additional research could be conducted in that regard.

7. References

[1] Morency, Louis-Philippe and Mihalcea, Rada and Doshi, Payal, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from The Web", in International Conference on Multimodal Interfaces (ICMI 2011), Nov, 2011, Alicante, Spain.

[2] Martin Wollmer and Felix Weninger and Tobias Knaup and Bjorn Schuller and Congkai Sun and Kenji Sagae and Louis-Philippe Morency, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context", in IEEE Intelligent Systems Journal, IEEE Computer Society, Volume 28, Number 3, ISSN 1541-1672, 2013, pp. 46-53.

[3] Veronica Perez Rosas and Rada Mihalcea and Louis-Philippe Morency, "Multimodal Sentiment Analysis of Spanish Online Videos", in IEEE Intelligent Systems Journal, IEEE Computer Society, Volume 28, Number 3, ISSN 1541-1672, 2013, pp. 38-45.

[4] Savakis, Andreas E. and Etz, Stephen P. and Loui, Alexander C. P., "Evaluation of image appeal in consumer photography", in Proc. SPIE, 2000, Volume 3959, pps. 111-120.

[5] Datta, Ritendra and Joshi, Dhiraj and Li, Jia and Wang, James Z., "Studying Aesthetics in Photographic Images Using a Computational Approach", in Proceedings of the 9th European Conference on Computer Vision - Volume Part III, Springer-Verlag, series ECCV'06, 2006, ISBN 3-540-33836-5, 978-3-540-33836-9, Graz, Austria, pps. 288-301.

[6] Khan, Shehroz S. and Vogel, Daniel, "Evaluating Visual Aesthetics in Photographic Portraiture", in Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging, Eurographics Association, Series CAe '12, 2012, ISBN 978-1-4503-1584-5, Annecy, France, pps. 55-62.

[7] Luca Marchesotti and Florent Perronnin and Diane Larlus and Gabriela Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors", in Proc. ICCV, 2011, pps. 1784-1791.

[8] Moorthy, Anush K. and Obrador, Pere and Oliver, Nuria, "Towards Computational Models of the Visual Aesthetic Appeal of Consumer Videos", in Proceedings of the 11th European Conference on Computer Vision: Part V, Springer-Verlag, Series ECCV'10, 2010, ISBN 3-642-15554-5, 978-3-642-15554-3, Heraklion, Crete, Greece, pps. 1-14.

[9] Judy I. Alpert and Mark I. Alpert, (1989), "Background Music As an Influence in Consumer Mood and Advertising Responses", in NA - Advances in Consumer Research Volume 16, eds. Thomas K. Srull, Provo, UT : Association for Consumer Research, Pages: 485-491.

[10] Lincoln G. Craton, Geoffrey P. Lantos, (2011), "Attitude toward the advertising music: an overlooked potential pitfall in commercials", Journal of Consumer Marketing, Vol. 28 Iss: 6, pp.396 - 411.

[11] Carmen Lewis, Cherie Fretwell, Jim Ryan, (2012), "An Empirical Study of Emotional Response to Sounds in Advertising", Vol. 12, Iss. 1, pp. 80 - 91.

[12] Yang, Chun-Yu and Yeh, Hsin-Ho and Chen, Chu-Song, "Video aesthetic quality assessment by combining semantically independent and dependent features", in ICASSP, 2011, IEEE, ISBN 978-1-4577-0539-7, pps. 1165-1168.

[13] Bhattacharya, Subhabrata and Nojavanasghari, Behnaz and Liu, Dong and Chen, Tao and Chang, Shih-Fu and Shah, Mubarak, "Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos", in ACM Multimedia, Series Grand Challenge, October, 2013.

[14] Lloyd, S., "Least Squares Quantization in PCM", IEEE Trans. Inf. Theor., March 1982, Vol. 28, Number 2, ISSN 0018-9448, pps. 129-137, IEEE Press, Piscataway, NJ, USA.

[15] Olivier Lartillot, Petri Toiviainen, Tuomas Eerola, "A Matlab Toolbox for Music Information Retrieval", in C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, 2008.

[16] Bordwell, David and Thompson, Kristin, "El arte cinematográfico: una introducción", 1995, 4th edition, Chapter 3.7 "La relación entre plano y plano: el montaje", Paidós Comunicación 68 Cine.

[17] Yeo, Boon-Lock and Liu, Bede, "Rapid Scene Analysis on Compressed Video, IEEE Trans. Cir. and Sys. for Video Technol., December 1995, Vol. 5, Number 6, ISSN 1051-8215, pps. 533-544, IEEE Press, Piscataway, NJ, USA.

[18] Smith, Alvy Ray, "Color Gamut Transform Pairs", SIGGRAPH Comput. Graph., August 1978, Vol. 12, Number 3, ISSN 0097-8930, pps. 12-19, ACM, New York, NY, USA.

[19] Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. "Data Mining: Practical Machine Learning Tools and Techniques", (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.