Expert Systems with Applications (2015). 42(1), 293-305.
DOI: http://dx.doi.org/10.1016/j.eswa.2014.07.033

Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials

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

Universidad Carlos III de Madrid, Leganés

Abstract

In this paper, we present a computational model capable to predict the viewer perception of car advertisements videos by using a set of low-level video descriptors. Our research goal relies on the hypothesis that these descriptors could reflect the aesthetic value of the videos and, in turn, their viewers' perception. To that effect, and as a novel approach to this problem, we automatically annotate our video corpus, downloaded from Youtube, by applying an unsupervised clustering algorithm to the retrieved metadata linked to the viewers' assessments of the videos. In this regard, a regular k-means algorithm is applied as partitioning method with k ranging from 2 to 5 clusters, modeling different satisfaction levels or classes. On the other hand, available metadata is categorized into two different types based on the profile of the viewers of the videos: metadata based on explicit and implicit opinion respectively. These two types of metadata are first individually tested and then combined together resulting in three different models or strategies that are thoroughly analyzed. Typical feature selection techniques are used over the implemented video descriptors as a pre-processing step in the classification of viewer perception, where several different classifiers have been considered as part of the experimental setup. Evaluation results show that the proposed video descriptors are clearly indicative of the subjective perception of viewers regardless of the implemented strategy and the number of classes considered. The strategy based on explicit opinion metadata clearly outperforms the implicit one in terms of classification accuracy. Finally, the combined approach slightly improves the explicit, achieving a top accuracy of 72.18% when distinguishing between 2 classes, and suggesting that better classification results could be obtained by using suitable metrics to model perception derived from all available metadata.

Keywords: automatic video annotation, aesthetic quality assessment, video sentiment analysis, video metadata, YouTube

Email addresses: ffm@tsc.uc3m.es (F. Fernández-Martínez), ahgarcia@tsc.uc3m.es (A. Hernández García), fdiaz@tsc.uc3m.es (F. Díaz de María)

1. Introduction and motivation

The increasing growth of video creation and share, specially over the Internet, and the predictable tendency for the future make the development of techniques and tools to handle videos very necessary. In order to improve the efficiency of searching for videos and offering the users satisfactory results, techniques of video classification [5] and video recommendation [1] have been deeply studied. However, most techniques were based on text, tags or metadata. It has been only in recent years that content-based approaches are being researched. A very challenging and valuable tool for improving searches and user experience would be to develop models that allow recognizing the aesthetic quality of videos, according to what users expect, exclusively relying on video content.

Here, our purpose with this work is demonstrating that it is possible to determine if a video has been positively or negatively perceived by users, building a predicting model based on low-level video descriptors and using as ground truth the labels derived by means of unsupervised learning techniques from *YouTube* metadata inherent to the videos, such as the number of likes or the number of views.

To the best of our knowledge, up until now, automatic aesthetic quality assessment in image and video has been addressed by different approaches, but all of them by using as ground truth explicit scores ranked by users. Although this is not a limiting inconvenient (except for the cost of it), this work suggests to approach the problem of video aesthetics assessment without depending on tags or scores assigned by a group of annotators specifically recruited for such purpose. Instead, we simply rely on real metadata present in YouTube.

Hence, the main idea behind our approach is that we assume these metadata (e.g. the number of *likes* or *views*) to be indicative of the subjective appreciation of a video by its viewers. For example, it is reasonable to think that a video with many likes and a high number of views is more appealing from the user point of view than another video with several *dislikes* and a few number of views. Under this assumption, we use unsupervised clustering techniques to bring together videos with similar metadata, deriving suitable polarity labels and thus, modeling how users have perceived the videos on average. Once we have annotated the set of videos with their corresponding perception labels, we carry out well-known image and video processing techniques for extracting low-level features, some of which can be referred to as novel descriptors. Finally, we employ different supervised classification algorithms to assess how much these features may be indicative of the user appreciation of the video modeled as previously mentioned, taking special notice of how these features can be combined to provide better results. Figure 1 shows a diagram providing a complete overview of the whole process.

The paper is organized as follows: after this introduction, Section 2 presents a literature review of automatic aesthetics assessment techniques applied to both images and videos. Section 3 provides the details of the video corpus acquisition and clustering procedures. Section 4 describes the visual descriptors extracted for the classification task. Section 5 presents the classification results including corresponding discussions and issues. Finally, some conclusions and future work are laid out in Section 6.

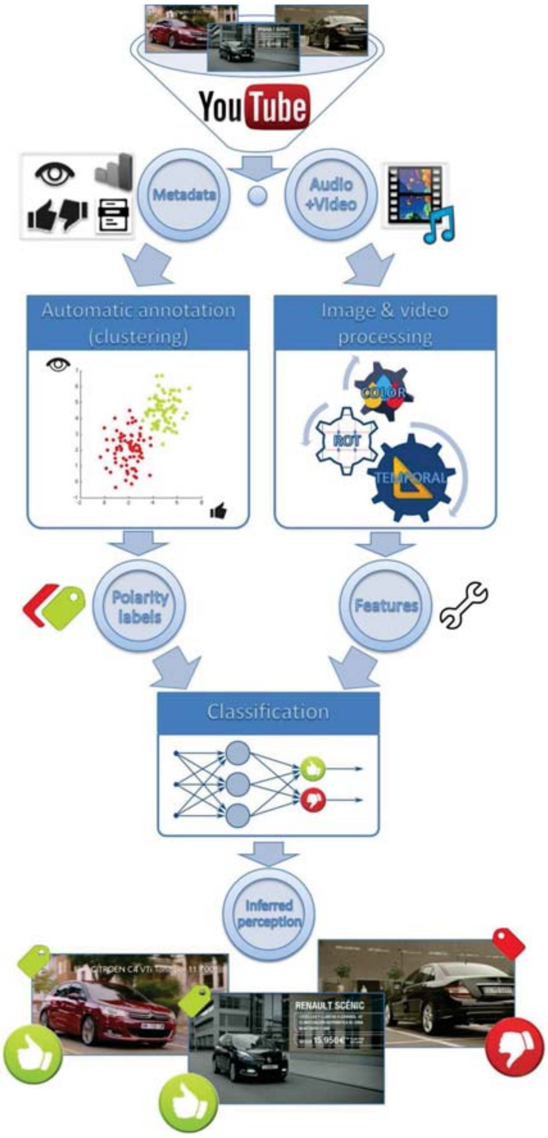


Figure 1: Diagram of the approach overview.

2. Related work

This section is a review of the most relevant research works in the study of subjectivity within multimedia data by means of computational procedures. We will start with an introduction to recommendation and classification systems, as they are the most important domains of applications of this work, and will follow by exposing the latest works in sentiment analysis, which is a field with important relationships with aesthetics assessment. Finally, we will focus on the previous works of aesthetics assessment, both applied to still images and videos.

2.1. Recommendation and Classification Systems

The objectives and applications of this work are closely related to video classification and video recommendation, which are fields of great research interest due to the great amount of available videos today. An important work on video recommendation systems was carried out by Adomavicius and Tuzhilin in 2005 [1], in which they performed a survey of the state of the art at that moment and proposed some improvements. The importance of recommendation systems can be understood by looking at the growth of social networks based on videos and video platforms, such as YouTube. A discussion on the techniques used in the recommendation systems of YouTube is done in [8]. Similarly, video classification techniques have been deeply studied and have still great potential of development. A survey on the literature related to video classification was made in [5] in 2008.

Classification and recommendation systems can be seen as the driving force of other related research works in multimedia applications, such as image and video quality assessment [18, 19], video sentiment analysis or image and video aesthetics assessment.

2.2. Sentiment Analysis

The present work aims to extract subjective information from objective data. Such a purpose is also the goal of sentiment analysis or opinion mining [34], a thoroughly researched field which studies the subjectivity of information through automatic computational procedures. Traditionally, sentiment analysis has focused on extracting sentiment and opinions from text sources of different nature [24, 27]. The first attempt to extend sentiment analysis to audiovisual data was recently carried out by Morency *et al.* [23] in 2011, where they perform a multimodal sentiment analysis of 47 videos from YouTube. Together with the text-based sentiment analysis, they take advantage of the extra information that the audiovisual features add. Their conclusion is that using together text, audio (pauses and pitch) and video (smile and look away) improves the performance with respect to using only one kind of feature. Further research following this study has been made in [29, 35].

2.3. Aesthetics Assessment

Another field that studies subjectivity is known as aesthetics assessment, which was firstly studied in still images. One of the earliest approaches towards this domain was carried out by Savakis *et al.* [31] more than a decade ago. In that paper, they aimed to find out which aspects were related to image appeal through a ground truth experiment in which 11 participants had to rank 194 pictures belonging to 30 different groups. It was found that image appeal was influenced by image quality only regarding objective aspects, so their conclusion was that image appeal had to be addressed through metrics others than those used for measuring image quality. More recently, in 2006, Datta *et al.* [7] 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. The features they selected where all based on photographic aspects or well-established rules of thumb, such as brightness, saturation, hue, metrics of usage of the rule of thirds or depth of field indicators. After this successful work, several studies followed this line of research by adding different contributions. This is the case of [14] or [13], where they carried out a higher-level analysis to assess the aesthetic quality of photographs. In 2011, Luca Marchesotti *et al.* [20] extended the study by using a larger and diverse set of features, including generic image descriptors that added statistics computed from low-level local features. Evaluating their models on images collected from *Photo.net* they achieve an accuracy of 89.9%.

However, aesthetics assessment applied to videos has not been addressed until recent years. A related approach was carried out by [25] in 2012, although it was not strictly aesthetics assessment, but a computational model for automatically separating professional videos from amateur ones. Even though the task is not as challenging as modelling a subjective evaluation, they employed an aesthetics based approach and achieved 91.2% accuracy. To our knowledge, the first attempt of modelling visual aesthetics in moving images was addressed by Moorthy *et al.* [22] 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 [7] and on users' reports were computed from the videos and extended to the temporal dimension through a hierarchical pooling method. Finally, they selected the 7 most relevant features and after classification procedures they achieved an accuracy of 73.03%. This study was extended by [36], using the same set of videos, but differentiating between semantically independent and dependent features in order to perform a comparative study. Finally, in 2013 [3] uses a larger data set of 1,000 videos and proposes a model which relies on features based on psycho-visual statistics.

3. Corpus acquisition and annotation

One of the main aspects of this work in comparison to other related works [22, 36, 3] is that we do not depart from an annotated corpus, but we obtain the labels through unsupervised learning techniques instead. This procedure, as it will be detailed later, consists in deriving or *learning* these labels from video-related metadata, such as the number of likes or the number of views, which we assume to be indicative of the subjective assessment of the videos by viewers.

3.1. Domain selection

Metadata are provided by users, as they watch, interact and share videos. In this regard, when annotating commercials in terms of their aesthetic value by using YouTube metadata, it is very important to define a particular domain from which to collect the videos. For example, it would not be advisable mixing a *Coca-Cola* commercial and a detergent one, mainly because we could not rely on the corresponding metadata to determine any aesthetic difference. Dissemination and public interest of a video are two aspects that may terribly bias related metadata. Therefore, observed metadata differences would then mainly explain the greater dissemination and interest of the *Coca-Cola* video compared to the detergent one, not the actual aesthetic differences between them. Hence, in order to minimize any possible bias, we have restricted our domain to one single type of videos: car commercials.

There are several characteristics in favor of the choice of this particular domain. For instance, the election of advertising videos is appropriate because their duration is limited, which is important not only for computational reasons, but because the variation of the content is limited as well. In addition to this, although there are many different car brands producing video commercials, all of them share the same target: selling a car. Hence, we can reasonably assume related metadata to be more connected to the way the cars are displayed and sold in the commercials and, in turn, to the aesthetic differences resembled by the different spots, rather than to the cars themselves. Thus, our assumption is that polarity differences will depend stronger on the video features than on the content. Nevertheless, despite all these considerations and constraints, we cannot ignore the fact that content dependency cannot be totally avoided (e.g. a particular viewer might be simply *in love* with *Mercedes* cars). Finally, publicity is also a desirable domain because of the marketing applications of the research, which could be of interest for many different agents, such as brands, advertising agencies, consumers or public institutions among others.

3.2. Video filtering

Two main characteristics made YouTube an optimal source of videos for building the corpus: the richness of its metadata and the great amount of available video content (100 hours of video are uploaded every minute [41]). However, because of the huge number of videos, filtering procedures were required to deal with a great diversity.

Corpus collection started with an automatic search of car commercials in YouTube using a list containing only car brands sold in Spain. As another restriction, we also limited our search to results in Spanish language (cross-cultural differences may induce significant bias). Similarly, only videos published after 2010 were retrieved mainly to prevent any temporal bias (changes on the way advertisements are made or on the use of YouTube that people do could seriously affect metadata). Queries were launched containing keywords such as: *advertisement*, *spot* or *campaign*, within the YouTube category *Autos* and retrieving 60 results for each query sorted according to Youtube *relevance* ordering.

Despite the above mentioned search restrictions, the initial corpus (with 2,732 videos) required further filtering procedures. First, we had to delete repeated videos. Even though videos were retrieved through different queries, it happened that the same video was returned by different queries. Second, any video which was not a car commercial should be removed. For this purpose, we included a duration filter so that videos longer than 115 seconds and shorter than 10 seconds were removed. Third, we also checked the inclusion of keywords (same as when performing queries) in the title of retrieved videos, thus, preserving only those including any of them. Fourth, and very important, videos without sufficient metadata (i.e. feedback from viewers) are of no interest for our purposes since no perception label could be derived for them. For this purpose, we removed videos having fewer than 3 raters (value of 3 was adopted to achieve a reasonably good trade-off between the size of our dataset and the reliability of any features derived from related ratings). After these automatic filtering procedures, and to ensure that the collected corpus accomplished due requirements, we finally performed a manual filtering by watching the resulting set of 277 videos and discarding those outdated, without sufficient visual quality, out of our domain (only professional car commercials released in Spain), or with related metadata showing some evidences of the video going viral (i.e. abnormal lifespan or days-to-peak patterns [10]). At the end of this manual filtering, we got a final list with 138 videos [1].

3.3. Available metadata

For a metadatum to be useful for the clustering procedure it must reflect, in some way, the feedback provided by users in terms of how they perceive the video. The metadata that could potentially describe the appeal of a video to users are the following:

- **viewsCount**: number of views or times a video has been played. This metadatum could be indicative of how good or bad a video has been received by users by holding the reasonable assumption that the greater the number of views, the better the viewers' perception.

- **numLikes**: number of likes or times the users have clicked the like button. A clear example of the "more is better" criterion.

- **numDislikes**: number of dislikes or times the users have clicked the dislike button. Opposite of the **numLikes** but equally interesting.

- **favoriteCount**: number of times a video has been selected by a user as a favorite video. Just like **numLikes**, it should also reflect the assessment of a video: the more, the better. However, favoriteCount is most often zero in our domain, so we have discarded this one.

- **rating**: actually the average rating that have been provided by users. This is a special metadatum because it was introduced to reflect the old way YouTube users had to value a video. Until March 2010 [40], instead of a *like* and a *dislike* button, there was a system consisting of five stars from which users could choose from 1 to 5 in halves of star. However, YouTube changed this system because they considered that it did not reflect a real 1-to-5 rating, but just a binary assessment, as it is posted in the official YouTube forum [38, 39, 40]. Hence, the star-based system was replaced by a simpler likes/dislikes system.

- **numRaters**: number of raters or YouTube users who have rated a video or clicked in the *like* or the *dislike* buttons. Ratings can be either positive or negative. Therefore, **numRaters** may be then referred to the impact of a video rather than to its visual aesthetic value. Nonetheless, we find it to be a suitable complement for **rating** (or for the likes-dislikes ratio, that will be introduced next), a quality-like metric which is reasonably well correlated with the actual visual aesthetic value. Particularly, given that the sample size affects the goodness or

[1] The video corpus with the video IDs and related metadata is available at: `http://www.tsc.uc3m.es/~ffm/car-commercials-ids-and-metadata.arff`

reliability of **rating**, we can assume an interaction between them, where the best outcomes occur where large number of raters and high rating values are present together (i.e. a high visual aesthetic value should yield a high rating but also a high number of raters). Hence, their combination would, ideally, allow to distinguish among different strength levels of the underlying visual aesthetic value.

- **numComments**: number of comments or times a video has been commented. Like ratings, comments can also be either positive or negative. Therefore, a similar discussion may be raised concerning **numComments** and the above mentioned quality metrics (or the ratio between positive and negative comments, in case the related comments were annotated in such a way).

In addition to the described raw YouTube metadata, some other new metadata have been derived in order to make the clustering procedure more effective.

- **ldRatio**: the likes-dislikes ratio (*ldRatio*), it represents the proportion of likes from the total number of votes (i.e. likes and dislikes) and it is computed as follows:

$$\textbf{ldRatio} = \begin{cases} \frac{numLikes}{numLikes+numDislikes} & numLikes + numDislikes \geq 0 \\ 0 & numLikes + numDislikes = 0 \end{cases} \quad (1)$$

  It combines *numLikes* and *numDislikes* into one single metadatum which replaces them as it enables a joint interpretation, in terms of viewer perception, coherent with the other metadata (i.e. the higher the ratio, the better).

- **viewsCountScore**: a new metadatum that maps the number of views into a score from 1 to 5, according to ranges based on the percentiles of the distribution of data:

$$\textbf{viewsCountScore} = \begin{cases} 1 & 0 \leq viewsCount < 750 \\ 2 & 750 \leq viewsCount < 2,000 \\ 3 & 2,000 \leq viewsCount < 5,000 \\ 4 & 5,000 \leq viewsCount < 15,000 \\ 5 & viewsCount \geq 15,000 \end{cases} \quad (2)$$

  $viewsCountScore$ will be used instead of *viewsCount* as it has been observed a terrible dispersion in the values the latter can take producing a non-linear behavior which does not necessarily reflect real differences regarding the assessment (e.g. a video with 500,000 views is not necessarily ten times better than another one with 50,000 views). Its interpretation is similar to the rest of metadata presented: the higher, the better.

3.4. Corpus annotation

One of the most novel and challenging characteristics of this work is the annotation of the corpus from available metadata through automatic procedures. To the best of our knowledge, previous works on automatic assessment of aesthetic quality and video sentiment analysis made use of either already annotated corpora or carried out an annotation process by recruiting expert annotators. For instance, [20] and [7] made use of the online photo sharing community *Photo.net* as data source, using the average aesthetics score as ground truth for classification purposes. In [3] they used a corpus of videos specifically ranked by 10 individuals for a challenge on the topic. Works [22] and [36] on automatic visual quality assessment used the same corpus of 40 videos annotated by 33 participants. Finally, few existing works on video sentiment analysis, [23] and [29], also manually annotated the opinions expressed on the videos.

Deriving video polarity annotations automatically through unsupervised clustering techniques yields several advantages:

- It is a less expensive procedure. There is no need to recruit experts to annotate the videos.
- It could potentially be more reliable. When under laboratory constraints, annotators might be biased when rating the videos, indirectly making the reliability of the annotations questionable [15]. With the proposed annotation solution, annotations are obtained from metadata provided by actual viewers, hence, reflecting the way potential consumers perceive the commercials.
- It enables a more complex and general assessment model of how viewers perceive the videos. As the procedure relies on several metadata, instead of just a single parameter, a wider definition of the subjective information is implicitly modeled.

3.4.1. Clustering strategies

Given the different metadata described in Section 3.3, it is possible to observe some differences between them according to the way these metadata are provided by users, which, in turn, could suggest that there are two different profiles or types of users watching the videos. On the one hand, there are users who explicitly express their opinion and, on the other hand, users who normally do not explicitly express their opinion, but whose assessment is implicitly provided as they watch videos. Based on these hypotheses, we have defined the following types of metadata:

- **Explicit-opinion metadata**: favoriteCount, rating, numLikes, numDislikes and ldRatio fall within this category. When providing these metadata users do explicitly convey an opinion about the videos, aside from watching them. As we discussed earlier, we regard numComments and numRaters as intensifiers of rating and ldRatio. Hence, they may also be categorized as explicit.
- **Implicit-opinion metadata**: Only viewsCount and viewsCountScore fall within this category. Both metadata are automatically provided by users as they simply watch the videos. Particularly, users are not providing any rate nor any comment, but they are implicitly contributing to the overall video assessment.

Given this categorization, annotation could be approached in three different strategies depending on the type of metadata on which clustering relies: explicit metadata

explicit	**implicit**	**combined**
ldRatio rating numComments numRaters	viewsCountScore	ldRatio rating numComments numRaters viewsCountScore

Table 1: Tested clustering strategies in terms of metadata.

based strategy, implicit metadata based strategy, and the combination of both. Table 1 summarizes the three different annotation schemes and their corresponding metadata.

In view of the visual aesthetic value of a video, we find ldRatio (also rating, which is simply an outdated version of ldRatio) to be a quality ratio that can be considered a reasonably solid and direct evidence of it. Nonetheless, correlation does not need to be necessarily perfect. This may happen to be true, for example, when the hook of a particular commercial is not necessarily connected to its visual aesthetic value but to something else (e.g. the music, viewers could simply be in love with it rather than with the visual content). In such a case, the visual descriptors, and the corresponding elicited aesthetic value, may not suffice to explain the viewers' perception.

By evaluating the explicit-opinion based approach (mostly underpinned by the ldRatio metric) we will basically test whether our suggested visual descriptors are indicative of the quality of the video and, thus, also of its visual aesthetic value.

On the other hand, the number of views (i.e. viewsCountScore) probably has to do more with quantifying the impact of the video, supposedly, regardless of its quality and its corresponding visual appealing. In this regard, it is also important to remark that the number of views of a particular video does not need to be necessarily correlated with its visual aesthetic value. For example, it is possible to find, mostly in domains other than video commercials, certainly disgusting videos that have a considerably high number of views.

Hence, by evaluating the implicit-opinion based approach, we will test whether the proposed visual descriptors are indicative of the mentioned impact of the video instead, which would suggest some dependency between the number of views and the visual aesthetic value as well (i.e. the visual aesthetic value of a video presumably affects, at least to some extent, its number of views).

Finally, the combination of both models, explicit and implicit, will allow us to gain valuable insight about the performance of our visual descriptors when modeling both phenomena, quantity (or impact) and quality, together. In any case, modeling viewers' perception can be reasonably considered quite a complex problem. A problem that we hope to successfully solve, at least partially, by relying on visual descriptors derived from the visual content and its aesthetic value.

The unsupervised clustering algorithm that was chosen for obtaining the annotations is the well-known k-means, one of the most celebrated clustering algorithms first introduced by Lloyd in [17]. For each strategy, the k-means algorithm was run for 4 values of k, producing from $k = 2$ to $k = 5$ independent classes or clusters. Main reason for this was that, given that we are relying on automatic clustering techniques to generate the annotations, it is a good practice to generate and evaluate different number of classes in order to find optimal quality partitionings (i.e. natural grouping of videos is not necessarily achieved with 2 clusters). This makes an important differentiation to previous works, where the usual procedure is to simplify the problem by reducing it to a binary or two-classes classification task. Besides, each of these 12 configurations (i.e. 3 strategies × 4 different number of target clusters) was evaluated with 4 different distance measures: *sqEuclidean* and *cityblock*, both after a Z-score normalization as suggested by [11], and *cosine* and *correlation* (without normalizing, as they already perform an intrinsic normalization). It is important to remark that each clustering strategy could potentially produce a different annotation of the video corpus (up to $3 \times 4 \times 4 = 48$).

4. Visual Features Extraction

Another goal of this work is demonstrating the usefulness of low-level visual features in assessing the user perception of videos. In this regard, we have inspired the decision of which features to test in previous works, such as those from [7] 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.

The decision of which visual features to extract has been motivated and inspired by previous works, such as that from [7, 22], who proved the convenience of some descriptors for assessing the aesthetic value. In this regard we have extracted similar features to previous works such as those related to the intensity, the saturation, the colorfulness or the rule of thirds. We have also motivated the extraction of some novel features (as applied to aesthetics assessment) such as those related to the temporal segmentation or entropy-based features by our knowledge and research in photography and film.

4.1. Temporal Segmentation

In film-making and publicity, temporal segmentation is of great importance, since it is the basis of montage, the editing technique that allows the creation of most effects cinema produces. Montage creates many semantic effects, but quantitatively, the level of segmentation, i.e. the number of cuts, is a good indicator of meaning. For example, an action scene has usually many more number of cuts than a calm, descriptive scene [4, 26]. To our best knowledge, temporal segmentation features have been included here for the first time in a work of aesthetics assessment.

A temporal segmentation of a video implies to determine the transitions between subsequent shots. Most transitions in video commercials are abrupt cuts and there exist several techniques for detecting the shots boundaries, for example using measures of motion, as presented in [16] or using histograms [37]. For our purposes, in order to detect these transitions, we have made use of the sum of absolute differences (SAD) of the gray intensity [37] which is defined for each frame n as follows:

$$D(n) = \frac{1}{H \cdot W} \sum_{x=1}^{W} \sum_{y=1}^{H} |I_n(x,y) - I_{n-1}(x,y)| \tag{3}$$

The detection performance can be improved by using its second derivative. This offers additional robustness at high speed movements as it detects abrupt transitions of the first derivative:

$$M(n) = -D''(n + 1) = -(D'(n + 1) - D'(n)) \tag{4}$$

with

$$D'(n) = D(n) - D(n - 1)$$

$D''(n)$ is computed for every frame of the video and a threshold (set to 0.18 after several tests with previously labeled videos) is needed to decide if there is a cut at a certain frame or not. Implemented cut related features are:

- **num-cuts**: total number of cuts within a video.
- **longest-shot**: duration in seconds of the longest shot (i.e. a fragment of video between two consecutive cuts).
- **mean-shot-duration**: mean duration of the shots of the video, in seconds.
- **std-shot-duration**: standard deviation of the duration of the shots.
- **mean-cuts-per-min**: mean density of cuts estimated as the absolute number of cuts divided by the duration of the video in minutes.

4.2. Intensity

In photography and film-making, intensity is referred to as brightness. In most situations, under- and overexposure affect the quality of experience of a video, as exposed in [28], so operators usually control it to capture *correctly* exposed images, regarding the useful exposure range of the film or sensor. In other situations intensity might be modulated to create many effects by the image.

Intensity in a still image is referred to as the average value of the pixels of the greyscale version of the image. This image-level feature can be extended to the video level by computing:

- **mean-intensity**: average intensity along all the frames of the video.
- **std-intensity**: standard deviation of the intensity.

Typical black frames (i.e. 0-intensity frames) before and after the content are discarded to not distort the estimated values.

4.3. Entropy

When applied to image processing, entropy measures the randomness of the pixel values which can be used to model and describe textures. Observed textures may give an idea of the complexity of an image, which could help to produce a particular effect on the spectator. Entropy related features are:

- **mean-entropy**: average entropy along all the frames of the video.
- **std-entropy**: standard deviation of the entropy.
- **pct-low-entropy-frames**: percentage of low entropy frames. A frame can be regarded as a low entropy one when its entropy value is below a particular threshold (set to 2.85 after several tests). Most commercial videos tend to insert some extra frames in the video, mainly at the end, to show the brand logo, a car description, and/or the conditions of an offer. These frames, which usually have a monochromatic background or a large portion in a single color (e.g. black or white), and letters or signs in the front, are particularly characterized by having a very low entropy compared to others. Hence, this value will give an idea of the portion of video which is composed by these special frames.
- **low-entropy-end**: a binary feature that states if the end of the video (i.e. last 10% of frames) is mainly formed by low entropy frames, as previously described. For this feature to be instantiated as 1 at least 85% of ending frames must have low entropy. Although most car commercials end with this kind of frames, some others do insert a filmed shot instead, thus, making a difference.

None of these features account for black frames as well.

4.4. Color: Hue and Saturation

Color is a very descriptive characteristic of images and videos which we have translated into computational features following the work of [7]. David Bordwell points out the importance of color on the *mise en scène* in [4, pp. 148–157,186–189] as one of the most effective resources in film-making.

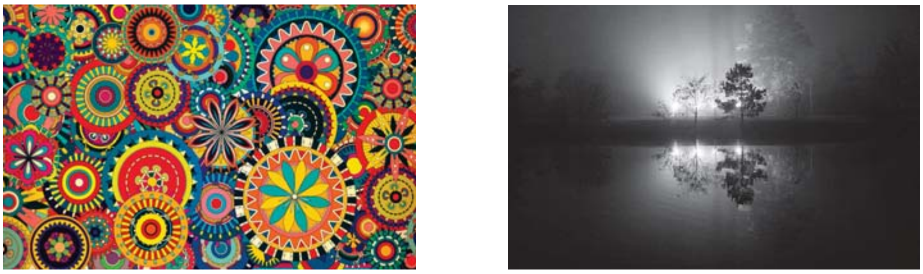
HSV is a well-known and widely used color model [33] which represents color using three intuitive parameters: hue, saturation and value (or brightness). In order to model in a simple way how the predominant colors of a video are, the following features have been implemented:

- **mean-hue**: average of the pixel values of the hue channel of every frame in a video.
- **std-hue**: standard deviation of the hue channel.
- **mean-saturation**: same as mean-hue but referred to the saturation channel.
- **std-saturation**: same as std-hue but referred to the saturation channel.

Black frames are discarded again for their estimation.

4.5. Color: Colorfulness

A picture is referred as colorful when it has richly varied colors. Note that in this case, it is not desired to measure the predominance of certain colors or their saturation as in the previous section, but the color wealth of a frame. From the point of view of analyzing car commercial videos, video colorfulness could be pretty interesting to learn whether the extensive use of multiple colors in the frames, or the absence of them, may attract people. For this purpose we define a feature which expresses the degree of utilization of a great variety of colors of a frame, in contrast to monochromatic or poorly colored images, by comparing each frame with an ideally multi-colored image as in [7]. A couple of examples of pictures and their value of colorfulness are presented in Figure 2.



(a) $C = 63.432$

(b) $C = 122.19$

Figure 2: The image to the left has many different colors, while the image to the right is a black and white image. Colorfulness measures the distance to an ideally multi-colored image, hence, the lower the distance the richer the variety of colors.

From frame values, colorfulness can be extended to the video level by computing the following features (black frames are again discarded):

- **mean-colorfulness**: mean colorfulness along all the frames of a video.
- **std-colorfulness**: standard deviation of the distribution of the colorfulness along all the frames.

4.6. Rule of Thirds

The rule of thirds is one of the most important rules of thumb in visual arts, such as photography, painting or design. It is a rule of image composition that states that the most important subjects in the image should be placed at the horizontal and vertical imaginary lines that divide the image in thirds, giving rise to nine equal parts, or at the intersection of these lines.

The idea behind using thirds is that it approximates the golden ratio, widely present in nature and used already by ancient Greeks in architecture, sculpture and other arts because it gives harmony to the compositions. Apart from being a guide to place the subjects, this rule is also followed to place the line of the horizon or any other horizontal dividing line of the image. If it is placed at the lower third line, it will give more strength and priority to the sky or the upper part and if it is placed at the upper line, it will give

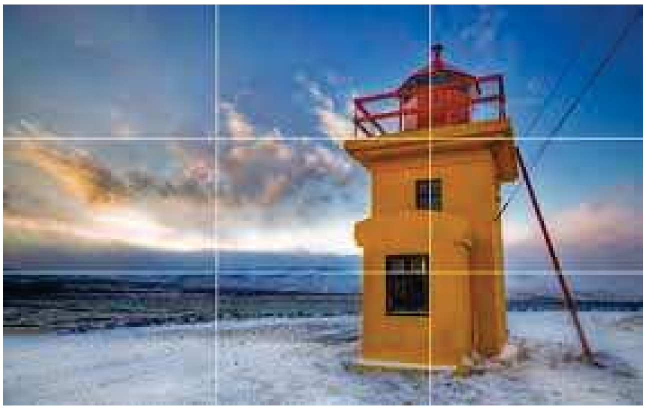


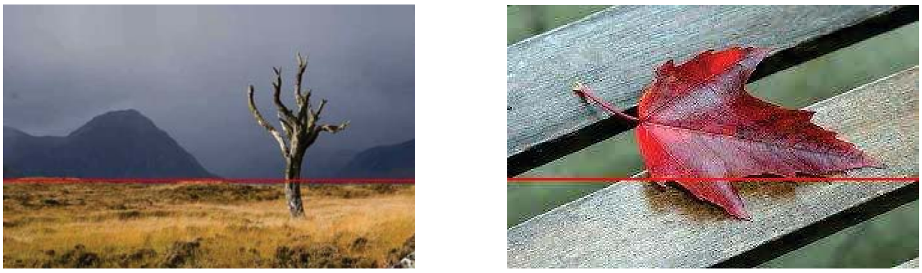
Figure 3: A sample image with the horizontal and vertical third lines

more strength and priority to the ground or the lower part. In video filming this rule is also widely used for placing moving subjects and the horizon, especially when filming landscapes. Therefore, we have developed a technique for measuring the degree of utilization of the rule of thirds for placing the horizon or the important horizontal lines.

This measure consists in comparing, by a sum of absolute differences, the color histograms of 64 bins corresponding to the two sub-images that the horizontal line generates:

$$D_{ROT} = 32 \cdot \frac{1}{64 \cdot H \cdot W} \sum_{b=1}^{64} |H_{top}(b) - H_{bottom}(b)| \tag{5}$$

The value of the measure is higher when the difference of the histograms is bigger, hence, the higher the value of this parameter, the higher the degree of utilization of the rule of thirds, as it can be seen in the images in Figure 3, from which the value of the parameter applied to the lower third line has been calculated.



(a) $D_{ROT-L} = 0.930$

(b) $D_{ROT-L} = 0.338$

Figure 4: The image to the left follows the rule of thirds, while the image to the right does not. The values of the measure D_{ROT} for the lower third have been computed for both pictures.

Based on these procedures for computing a measure to represent the degree of utilization of the rule of thirds (ROT), we have defined the following features:

- **mean-hrot-lt**: mean value of the previously described feature along all the frames of a video, applied to the comparison between the sub-images below and above the lower third line.
- **std-hrot-lt**: standard deviation of the distribution of the degree of utilization of ROT along all the frames of a video applied to the comparison between the sub-images below and above the lower third line.
- **mean-hrot-ut**: same as mean-hrot-lt but referred to the upper third line.
- **std-hrot-ut**: same as std-hrot-lt but referred to the upper third line.

5. Results and discussion

After the acquisition of the video dataset, the two main steps of the research process are the clustering analysis performed for its annotation, which has been introduced in Section 3.4.1 and the feature extraction procedure, explained in detail in Section 4. The current section will describe the research methods we have used in the evaluation process to gain an adequate understanding of the actual strengths and limitations of the suggested approaches.

5.1. Experimental setup

Due to the fact that the corpus annotation is based on YouTube metadata and that the nature of these is not uniform, we have adopted a particular experimental setup, which allows us to explore different clustering strategies and combinations in terms of mainly, the metadata that are used for the clustering, but also the distance and the number of classes. These combinations led to multiple data set versions, with different annotations, that must be evaluated and analyzed.

5.1.1. Feature Selection

On each clustering combination, we carried out a feature selection analysis so that we can identify the sets of features that provide better information about the data and their classes. In order to do so, we made use of the well-known WEKA machine learning software, from the University of Waikato in New Zealand [9]. This tool provides a set of feature selection algorithms, from which we pick 6: CfsSubsetEval attribute evaluator with BestFirst search method, SVMAttributeEval with Ranker, ConsistencySubsetEval with GreedyStepwise, InfoGainAttributeEval with Ranker, PrincipalComponents with Ranker and ClassifierSubsetEval with RaceSearch. CfsSubsetEval with Ranker returns the best feature from the whole set, ClassifierSubsetEval with RaceSearch chooses itself the number of features to keep and the rest allow indicating the number of features, *n*, to select. We configured these algorithms so that they provide reduced subsets of features from 1 up to 10 features. At the end of this attribute selection step we generated a total of 44 features subsets for each of the datasets returned by the clustering combinations with potentially different sets of features.

5.1.2. Classification

We used the Experimenter tool of WEKA to test the performance of several classification algorithms over all the features subsets we had previously generated. We tested the following classifiers:

- Rules-based classifiers: ZeroR, used as the baseline scheme, and OneR.
- Bayesian classifiers: Bayes Network and Naive Bayes.
- Function-based classifiers: Logistic, Simple Logistic, SMO (SVM) and Multilayer Perceptron.
- Tree-based classifiers: J48 Tree, ADTree, RandomForest and RandomTree.
- Instance-based classifier: KStar.

Therefore, a classification experiment can be referred to as a combination of: an annotation strategy (together with a number of classes and a particular distance), a feature selection technique and a classification algorithm. The performance of each classification experiment has been measured as the accuracy or the percentage of correctly classified instances. This accuracy is provided by the WEKA Experimenter tool by doing 10 random repetitions of a 10-fold cross-validation on every data set. The WEKA Analyzer provides a corrected Paired T-Tester to check which classification results are statistically better than the baseline scheme (i.e. ZeroR) with a confidence level of 0.95. Among these, the top statistically significant results have been presented for each strategy in Table 2.

Although a two or three-classes scheme could be found to be entirely suitable from a practical implementation point of view (e.g. viewers yielding a lower, an average and a higher perception), we are also interested in analyzing the performance of the suggested approaches for a higher number of classes as it would enable a better granularity for the viewer perception classification. Hence, results for a number of classes ranging from 2 up to 5 are presented. In this regard, it is important to remark that working with a number of classes higher than 5 was practically unfeasible because of sparse data problems (i.e. k-means resulted into underpopulated classes). Details regarding the corresponding clustering distance, the number of features selected by the feature selection algorithm and the used classifier are also specified.

The first important aspect to remark is that we have obtained higher accuracies than those included in the table (e.g. we got up to 57.10% for 3 classes with the combined approach, apparently better than the indicated 50.87%). However, statistical evidence did not suffice to regard such results as significantly better than the corresponding performance references given by the zeroR, so we have not considered them.

From a general point of view, the implicit approach seems to be the worst (except for 3 classes). However, and in spite of its lower performance, the obtained results confirm that the implemented video descriptors are indicative of the number of views a video could potentially receive (viewcount is typically referred to as the index of popularity of a video [6]).

On the other hand, a better performance is achieved by the explicit approach, which suggests that a better partitioning of the video space can be found by relying on explicit

		Accuracy (σ)	**zeroR** (σ)	**Distance**	**# feat**	**Classifier**
2 classes	**EXP**	70.98 (11.20)	55.05 (2.85)	correlation	9	Logistic
	IMP	65.35 (10.86)	55.77 (2.33)	cityblock	4	J48
	COM	72.18 (12.34)	56.48 (1.33)	cosine	4	RandomTree
3 classes	**EXP**	55.52 (10.94)	47.11 (3.47)	cosine	2	Logistic
	IMP	54.29 (11.30)	44.23 (2.33)	cityblock	2	Logistic
	COM	50.87 (12.20)	38.41 (3.11)	cityblock	2	KStar
4 classes	**EXP**	44.54 (12.57)	36.98 (2.25)	cityblock	10	SMO
	IMP	43.38 (10.62)	34.03 (2.94)	cityblock	2	NaiveBayes
	COM	58.79 (9.92)	51.48 (2.43)	cityblock	3	Logistic
5 classes	**EXP**	33.05 (8.18)	23.42 (3.07)	cityblock	10	SMO
	IMP	–	–	–	–	–
	COM	38.37 (11.36)	29.01 (0.88)	cityblock	2	Logistic

Table 2: Summary of top statistically significant results for each strategy and number of classes.

opinion based metadata. As with the implicit strategy, the obtained results again confirm that the implemented video descriptors are also indicative of the better or worse viewer perception levels derived from this type of metadata (e.g. like-dislike ratio).

Finally, the best performance is achieved with the combined approach (e.g. 72.18% for 2 classes or 58.79% for 4 classes), which suggests that the joint use of both types of metadata, explicit and implicit opinion based, could help to better model the viewers' perception.

The resulting performance can be deemed to be satisfactory, particularly if we refer to the combined approach with 4 classes whose error rate is only 13.39% higher than with 2 classes, which suggests an even better predictability by adopting the four classes into two by grouping the classes (e.g. 'high' with 'very high' and 'low' with 'very low').

5.2. Comparison of strategies, features and classifiers

Specific analysis of relevant differences regarding the chosen strategy (i.e. type of metadata), selected features and used classifiers will be carefully addressed in subsequent subsections. In this regard, and although it was not the purpose or the goal of this paper, it is interesting to mention that no statistical evidence has been found about the convenience of using a particular distance for the clustering procedure.Similarly, no relevant differences have been observed neither concerning the tested feature selection techniques nor about the number of features selected for the classification tasks.

The results will be presented as box plots [21]. Box plots are well known to be particularly useful for comparing distributions between several groups or sets of data thus, allowing us to do a fair comparison among the different strategies, features or classifiers while providing an idea of how the corresponding accuracy distributions are. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles and the whiskers extend to the most extreme data points not considered outliers. Every data point lying further than +\- 2.7σ from the median has been considered an outlier (represented with a '+' symbol in the figures).

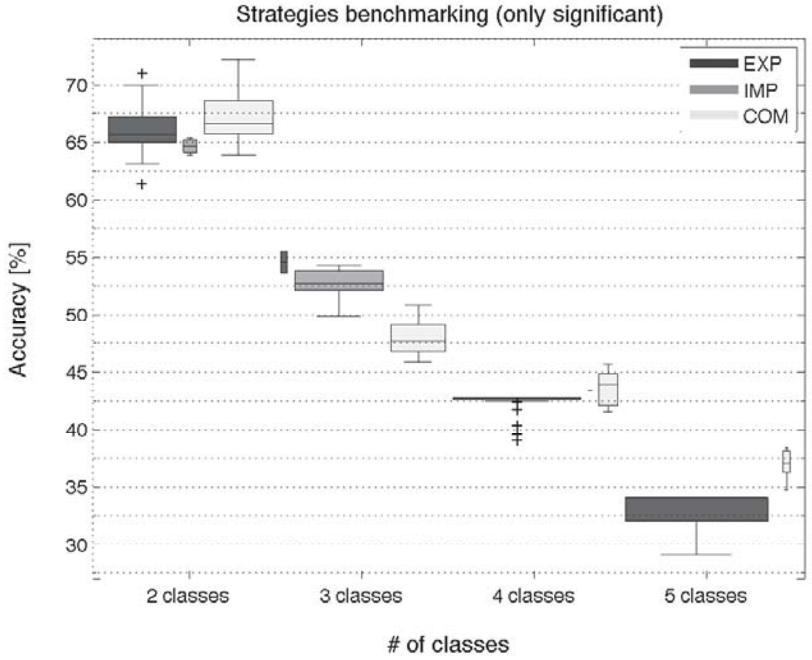


Figure 5: Classification accuracy for each of the strategies and number of classes.

5.2.1. Comparison between strategies

As previously introduced, our experiments covered the comparison between the proposed annotation strategies, namely: based on explicit-opinion metadata, based on implicit-opinion metadata and based on their combination. Figure 5 presents a box plot with 12 different boxes resulting from the combination of the 3 different strategies together with the 4 different numbers of classes tested (actually only 10 boxes are clearly visible in the Figure, given that the performance of the implicit drastically dropped for 4 and 5 classes). Hence, every box in the box plot includes all the statistically relevant classification results obtained for a specific strategy and a particular number of classes regardless of the used clustering distances and the feature selection (including the number of selected features) and classification algorithms, which have been assumed to be irrelevant for the comparison given that the same 'distance-feature selection technique-number of features targeted-classifier' combinations were all explored in all the cases. Variable-width box plots have been used to illustrate the size of each group whose data is being plotted. Particularly, we have made the width of the box proportional to the size of the group (e.g. from all the relevant results obtained for 2 classes, 45% were obtained with the explicit, 10% with the implicit and another 45% with the combined approach respectively).

Finally, in order to test for differences among strategies, a Kruskal-Wallis test [32] has been chosen since normality was questionable and sample sizes within each group (i.e. a particular number of classes) are small. This Kruskal-Wallis test for comparison of strategies indicates that there is a statistically significant difference in the distribution of classification accuracy between the strategies regardless of the number of classes (χ^2 (2) > 16 and $p < 0.001$ for 2, 3 and 5 classes, and χ^2 (2) = 11.19 and $p < 0.005$ for 4 classes). Given that we have obtained a significant Kruskal-Wallis test, we will use multiple Mann-Whitney tests [2] to examine pairwise differences.

Overall, and according to medians depicted in Figure 5, the implicit approach is confirmed as showing the worst performance. This result is not only evident in the measured accuracies, for instance implicit was significantly worse than explicit and combined for 2 classes (Mann-Whitney $p < 0.001$ for both comparisons), but also in terms of the amount of relevant results. Most of them, with the above mentioned exception of 3 classes, were achieved by explicit and combined strategies (implicit was particularly less successful for 4 and 5 classes where only one single relevant result was obtained).

To better understand the observed differences it is important to remark that the implicit approach just relies on one single metadatum (i.e. views score) while the other two rely on a higher number of them, thus, enabling a better partitioning of the video data set and, in turn, better classification results.

On the other hand, as previously introduced in Section 5 and according to top performance results presented in Table 2, the combined approach could be expected to show the best performance. Figure 5 confirms combined approach as showing better performance than explicit and implicit for any number of classes but 3 (Mann-Whitney $p < 0.05$ for corresponding comparisons). In addition to this exception, we can also observe that, despite better than explicit, the relative importance of combined accuracies tends to decrease with the number of classes (i.e. corresponding width becomes narrower), thus favoring the former. Both results, therefore, need further analysis for a better understanding of the actual performance of both strategies.

Work by Chatzopoulou, Sheng and Faloutsos [6], conducted an in depth study of fundamental properties of video popularity in YouTube. After collecting a data set of roughly 37 million YouTube videos, they studied the relationships of the same metrics (among others) that we are making use of.Particularly, they found that viewcount is highly correlated with #comments and #raters, while it exhibits very little correlation with the average rating.

The combined approach makes use of all the metadata. Therefore, every related partitioning is potentially conditioned by such a poor correlation among these metrics. In order to determine whether this could be the reason explaining both observed accuracy deviations we decided to evaluate the resulting clustering partitions from which the classification experiments were performed.

This poor correlation is evident in the example presented in Table 3, that corresponds to the clustering result from which top relevant performance has been obtained for the combined approach with 3 classes (i.e. 50.87% as presented in Table 2).

The table basically presents the centroids representing each cluster. As it can be observed, the resulting clusters could be regarded as natural, given that a natural ordering of the videos has emerged, for every item except for the likes-dislikes-ratio. The observation of these noisy or disordered clusters (i.e. cluster 1 could be tagged either as “HIGH”, according to view score, or as “MED”, according to likes-dislikes ratio),

	Cluster 1	**Cluster 2**	**Cluster 3**
# instances (videos)	53(38.41%)	50(36.23%)	35(25.36%)
AVG # comments	18.96	3.14	2.88
AVG # raters	46.06	8.70	7.71
AVG Likes-dislikes ratio	**0.70**	**0.93**	0.01
AVG Viewscore	**4.81**	**3.12**	2.63
Annotated tag	**HIGH/MED**	**MED/HIGH**	LOW

Table 3: "Combined approach": cluster centroids example for a 3 classes partioning.

		Strategy	
		EXPLICIT	COMBINED
# of classes	2	0.544	0.465
	3	0.563	0.502
	4	0.568	0.486
	5	0.531	0.468

Table 4: Clustering quality assessment: average Silhouette metrics for EXP and COM approaches.

aside from driving into not easily interpretable results from a practical point of view, suggests the importance of further studying the relation between the above mentioned metrics.

As an objective measure to compare the quality of the resulting clusters, we decided to evaluate both strategies using a *silhouette* criterion [12][30]. This criterion computes a silhouette value for each point as a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the i-th point, S_i, is defined as:

$$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

where a_i is the average distance from the i-th point to the other points in the same cluster as i, and b_i is the minimum average distance from the i-th point to points in a different cluster, minimized over clusters. The silhouette value ranges from -1 to $+1$. Hence, a high silhouette value indicates that i is well-matched to its own cluster, and poorly-matched to neighboring clusters. A clustering solution is typically found to be appropriate when most points have a high silhouette value. On the contrary, if many points have a low or negative silhouette value, then the clustering solution may be considered to have either too many or too few clusters. The silhouette clustering evaluation criterion can be used with any distance metric. In our study we have used Squared Euclidean distance.

Table 4 presents, for each strategy (i.e. for each metadata set) and number of classes, the corresponding average silhouette value computed for all the resulting clusters obtained by using the different tested distances to produce the annotations. As it can be observed, measured clustering quality was roughly 10% better when relying only on explicit metadata.

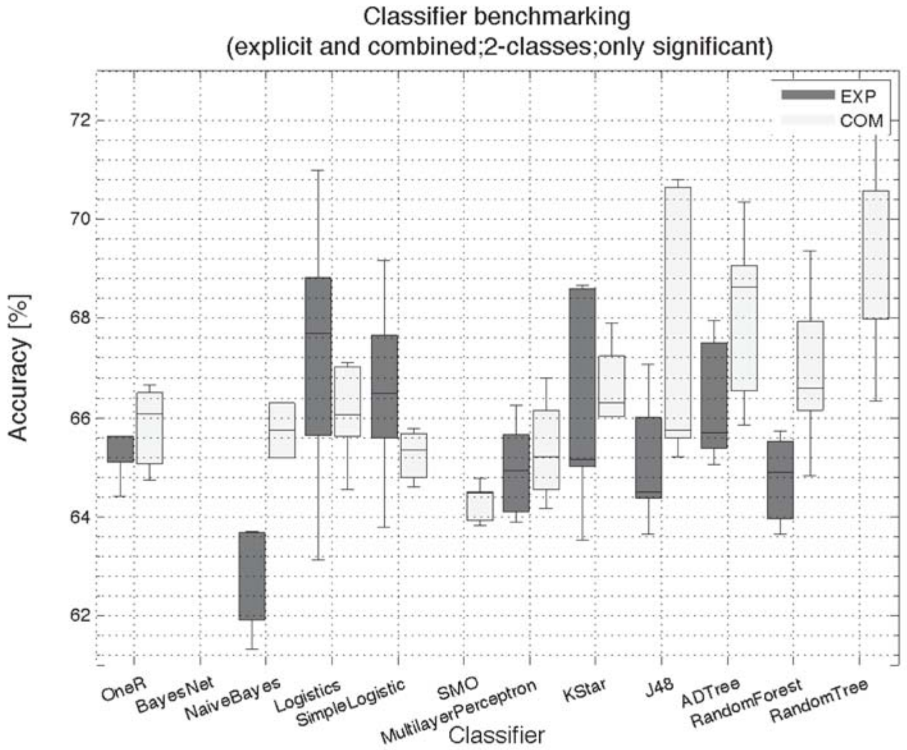


Figure 6: Classification accuracy distribution for each specific classifier and strategy including only relevant results achieved for 2 classes.

Finally, and though the correlation between the number of views and ratings was confirmed to be weak, therefore, inducing the above mentioned partitioning problems, best results have been obtained by exploiting all the metadata together.

Possible applicability issues due to the non inmediate interpretability of the elicited clusters, mostly when using 4 or more, could be tackled, at least to some extent, by simply deciding or weighting which of them, views or ratings, is more relevant for the application domain (i.e. similarly to the popular debate around quality and quantity). In any case, we expect future work to resolve this issue. Particularly, it would be interesting and worth trying to find out some other metrics derived from both the number of views and ratings that could lead to more accurate and interpretable clusters, hence enabling better classification results by making the best of the combination of any available metadata.

5.2.2. Comparison between classifiers

The different quality of the resulting clusters is also evident in the effectiveness of the classifiers we have tested. Corresponding classification results are presented in Figure 6, where box plots of statistically relevant results obtained from the evaluated 2-classes annotations have been included for both strategies, explicit and combined, to allow their comparison. In this regard, it is interesting to confirm that, as it can be derived from the figure, the explicit approach yields best performance mostly with relatively simple methods that do require the data to be linearly separable like Logistic and Simple Logistic (linear classifiers are favored by better quality or more linearly separable clusters). Kruskal-Wallis test for differences across classifiers was significant (χ^2 (12) = 22.61 and $p < 0.005$) although according to Mann-Whitney test there was no significant difference between Logistic and Simple Logistic.

On the other hand, when relying on lower quality clusters derived for the combined approach, fitting linear models becomes a harder problem. Clustered data may be assumed to be less readily fitted for a linear regression so that best performance is shown by mainly non-linear classifiers such as RandomTree and ADTree, which both of them are decision tree-like classifiers well known to be particularly suited to problems that do not have linear decision boundaries in their original feature space. Kruskal-Wallis test for differences across classifiers was also significant in this case (χ^2 (12) = 41.26 and $p < 0.001$). Nonetheless, no significant difference was observed between RandomTree and ADTree according to Mann-Whitney test.

5.2.3. Comparison between features

Finally, as one objective of this work was to find out which features are most indicative of the perception of the videos by users, an additional analysis has been performed to determine those features, if any, that have contributed most to achieve significant classification results. In order to simplify the analysis, again it will be focused only on the results achieved for 2 classes. Besides, considering that the performance gap between both top approaches (i.e. explicit and combined) is reasonably small and mainly to prevent any possible bias in the feature selection process induced by the different characteristics of the corresponding clustering results (as previously discussed in the preceding subsection), all the statistically relevant results obtained for both the explicit and the combined strategy and triggered by a specific feature have been grouped together. Hence, Figure 7 presents a box plot describing the 2 classes accuracy distribution for each specific feature whatever the used strategy, clustering distance, feature selection (including the number of selected features), and classification algorithm.

As it can be observed in Figure 7, every visual feature, except *std-entropy*, *mean-hrot-lt*, and *mean-colorfulness*, has proven to be helpful in the automatic assessment of the user perception of a video. A relevant conclusion is the proof of coherence between the set of features proposed in this work and the useful features proposed in previous works [7, 22]. It should be noted that the classification accuracy distribution for each feature in this figure, except for those cases in which only one single feature was selected, does not reflect the individual prediction performance of the corresponding feature, but mostly in combination with other features.

Applied Kruskal-Wallis test indicated that there is a statistically significant difference in the accuracies achieved by the different features (χ^2 (20) = 54.12 and $p < 0.001$). However, Mann-Whitney tests to examine pairwise differences did not find any significant difference (i.e. $p < 0.05$) between any pair of features ranked in the top 16 out of the 21 tested. Therefore, it may be concluded that all the different types of features tested, i.e. temporal, entropy or color based, and related to ROT, have attained notable and similar success, thus, complementing each other reasonably well.

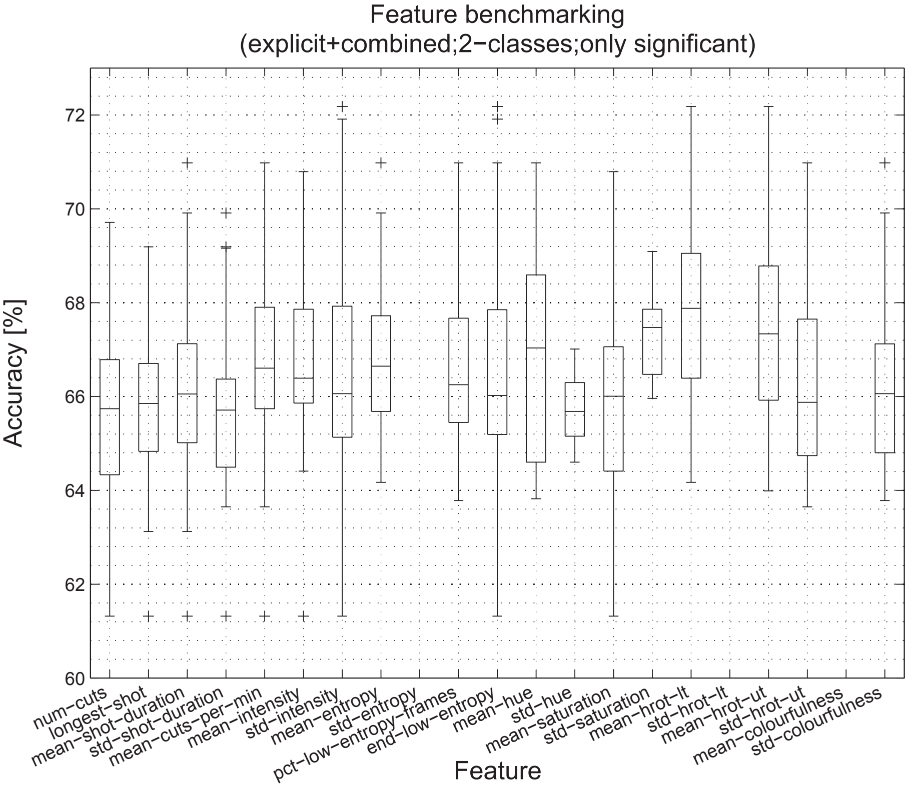


Figure 7: Classification accuracy distribution for each specific feature by grouping together the explicit and combined relevant results achieved for 2 classes.

6. Conclusions and Future Work

In this paper we have presented a computational method for assessing the aesthetic quality of car commercials retrieved from YouTube, extending the research into automatic aesthetic quality assessment, which is still in exploration.

The first relevant contribution of the paper is the use of clustering techniques for the automatic annotation of a video corpus which have been successfully validated as a novel alternative by means of the feedback provided by the viewers of the videos. This feedback is extracted from video related metadata, which is assumed to be indicative of the subjective perception of viewers. To the best of our knowledge, this is the very first time that a computational model to automatically assess the viewer perception of videos does not rely on annotations provided by recruited experts or trained annotators, but simply on video related metadata directly retrieved from Youtube. In this regard, annotations have been automatically assigned to videos by adopting three different strategies according to the way these metadata are generated by users.

Second, new video descriptors have been proposed and tested in multiple classification experiments where almost the 21 suggested features demonstrated to be indicative of the subjective perception of the viewers. These features have particularly shown a better performance when predicting perception derived from explicit metadata (i.e. viewers explicitly express their opinions and judgments, e.g. # of likes) than when relying on a simpler model just quantifying perception in terms of the number of views received (i.e. implicit metadata based approach).

On the other hand, the combination of all available metadata has yielded the best classification results. However, it has also produced less accurate and interpretable clusters, particularly when using a high number of them (i.e. mostly with 4 or more), where applicability turns to be less immediate and simple. This result suggests that there is still room for a more complex prediction model, yet to be explored, which could enable even further improvement and also facilitate the applicability of the suggested approach when numerous different video categories are required.

Best classification accuracy achieved is 72.18% for 2 classes, although it is also relevant to mention that satisfactory and significant results have been also obtained with up to 5 classes (e.g. a top accuracy of 58.79% has been measured for 4 classes), the number of classes that the study has been extended to. These results enable further research following the suggested approach to improve, for instance, the performance of classification and recommendation systems based on aesthetics characteristics.

From an applicability point of view, recommendation systems could benefit from this work as it facilitates an alternative to collaborative filtering, which requires available ratings, particularly when dealing with videos without previous considerations or assessments done by any viewers. Similarly, automatic video indexation and retrieval systems may elicit new taxonomies guided by the suggested visual descriptors or enable the retrieval of videos according to some specific visual features (e.g. retrieve only particularly "colorful" videos). Regarding the commercial or wide dissemination video production, the presented approach could be exploited to predict the expected success of the video or to perform a retrospective analysis of existing videos mainly to discover successful keys or tips for an efficient visual language, tips that could be then adopted for subsequent video productions. Moreover, automatic video summarization technology may also be revamped by summarizing video content by focusing on particularly valuable scenes (i.e. those with a high aesthetic value).

In the future, research should be extended to different video domains mainly to test whether the obtained results could be generalized and scaled to different scenarios. Besides, it would be also worth exploring new features. Particularly, the proposed set of video features could be completed, for instance, with typical motion and object detection related feature descriptors such as SIFT or HOG. In this regard, it would be particularly interesting to work towards adopting a multimodal approach by combining not only visual, but also audio and textual features. Domains like this, i.e. car commercials, where music plays a very important role, suggest that audio features could potentially be of great help in the assessment of subjective perception of videos.

Other possibilities to explore could be the detection of viral videos or the application of natural language processing techniques to extract useful information from user comments to enable better partitionings of video data sets.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, Jun 2005.

[2] R. Bergmann, J. Ludbrook, and W. P. J. M. Spooren. Different outcomes of the wilcoxon-mann-whitney test from different statistics packages. *The American Statistician*, 54(1):pp. 72–77, 2000.

[3] S. Bhattacharya, B. Nojavanasghari, D. Liu, T. Chen, S.-F. Chang, and M. Shah. Towards a comprehensive computational model for aesthetic assessment of videos. In *ACM Multimedia*, Grand Challenge, October 2013.

[4] D. Bordwell and K. Thompson. *El arte cinematográfico: una introducción*. Paidós Comunicación 68 Cine, 4 edition, 1995.

[5] D. Brezeale and D. J. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, (3):416–430, Oct 2008.

[6] G. Chatzopoulou, C. Sheng, and M. Faloutsos. A first step towards understanding popularity in youtube. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, pages 1–6, 2010.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III*, ECCV'06, pages 288–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[8] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The youtube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 293–296, New York, NY, USA, 2010. ACM.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, nov 2009.

[10] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann. Viral video style: A closer look at viral videos on youtube. In *International Conference on Multimedia Retrieval*. ACM, 2014.

[11] V. Kathiresan and P. Sumathi. An efficient clustering algorithm based on z-score ranking method. In *Computer Communication and Informatics (ICCCI), 2012 International Conference on*, pages 1–4, 2012.

[12] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York, 1990.

[13] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR (1)*, pages 419–426. IEEE Computer Society, 2006.

[14] S. S. Khan and D. Vogel. Evaluating visual aesthetics in photographic portraiture. In *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, CAe '12, pages 55–62, Aire-la-Ville, Switzerland, Switzerland, 2012. Eurographics Association.

[15] S. Lebai Lutfi, F. FernáNdez-MartíNez, J. M. Lucas-Cuesta, L. LóPez-LebóN, and J. M. Montero. A satisfaction-based model for affect recognition from conversational features in spoken dialog systems. *Speech Commun.*, 55(7-8):825–840, Sept. 2013.

[16] M. Leszczuk and Z. Papir. Accuracy vs. speed trade-off in detecting of shots in video content for abstracting digital video libraries. In F. Boavida, E. Monteiro, and J. Orvalho, editors, *IDMS/PROMS*, volume 2515 of *Lecture Notes in Computer Science*, pages 176–189. Springer, 2002.

[17] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, Sept. 2006.

[18] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, editors, *ICCV*, pages 2206–2213. IEEE, 2011.

[19] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 386–399, Berlin, Heidelberg, 2008. Springer-Verlag.

[20] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, pages 1784–1791, 2011.

[21] J. W. T. McGill, R. and W. A. Larsen. Variations of boxplots. *The American Statistician*, 32(1):12–16, Oct 1978.

[22] A. K. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. In *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pages 1–14, Berlin, Heidelberg, 2010. Springer-Verlag.

[23] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *International Conference on Multimodal Interfaces (ICMI 2011)*, Alicante, Spain, Nov 2011.

[24] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. In J. H. Gennari, B. W. Porter, and Y. Gil, editors, *K-CAP*, pages 70–77. ACM, 2003.

[25] Y. Niu and F. Liu. What makes a professional video? a computational aesthetics approach. *IEEE Trans. Circuits Syst. Video Techn.*, 22(7):1037–1049, 2012.

[26] M. Ondaatje and W. Murch. *El arte del montaje*. Plot Ediciones, 1 edition, 2007.

[27] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *In proceedings of EMNLP*, pages 79–86, 2002.

[28] P. Romaniak, L. Janowski, M. Leszczuk, and Z. Papir. A no reference metric for the quality assessment of videos affected by exposure distortion. In *ICME*, pages 1–6. IEEE, 2011.

[29] V. P. Rosas, R. Mihalcea, and L.-P. Morency. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, 2013.

[30] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.

[31] A. E. Savakis, S. P. Etz, and A. C. P. Loui. Evaluation of image appeal in consumer photography. volume 3959, pages 111–120, 2000.

[32] S. Siegel. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill series in psychology. McGraw-Hill, 1956.

[33] A. R. Smith. Color gamut transform pairs. *SIGGRAPH Comput. Graph.*, 12(3):12–19, Aug. 1978.

[34] A. Westerski. Sentiment analysis: Introduction and the state of the art overview. Technical report, Jun 2009.

[35] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

[36] C.-Y. Yang, H.-H. Yeh, and C.-S. Chen. Video aesthetic quality assessment by combining semantically independent and dependent features. In *ICASSP*, pages 1165–1168. IEEE, 2011.

[37] B.-L. Yeo and B. Liu. Rapid scene analysis on compressed video. *IEEE Trans. Cir. and Sys. for Video Technol.*, 5(6):533–544, Dec. 1995.

[38] YouTube. Five stars dominate ratings. `http://youtube-global.blogspot.com.es/2009/09/five-stars-dominate-ratings.html`, December 2013.

[39] YouTube. The video page gets a makeover. `http://youtube-global.blogspot.com.es/2010/01/video-page-gets-makeover.html`, December 2013.

[40] YouTube. Youtube api v2.0 - ratings. `https://developers.google.com/youtube/2.0/developers_guide_protocol_ratings`, December 2013.

[41] YouTube. Youtube statistics. `http://www.youtube.com/yt/press/statistics.html`, November 2013.