Hernández-García, A., Fernández-Martínez, F., Díaz-de-María, F. (2016). Comparing visual descriptors and automatic rating strategies for video aesthetics prediction. *Signal Processing: Image Communication*, v. 47, pp. 280-288.

DOI: https://doi.org/10.1016/j.image.2016.07.004

Comparing visual descriptors and automatic rating strategies for video aesthetics prediction

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

Universidad Carlos III de Madrid, Leganés

Abstract

Automatic aesthetics prediction of multimedia content is bound to be a powerful tool for information retrieval due to the wide range of applications where it could be used. With this paper we contribute to the research in the field of video aesthetics assessment by carrying out a comparative study of one, the performance of eight families of visual descriptors in accounting for the general aesthetics perception of videos and second, the suitability of different YouTube metadata for providing successful strategies for automatic annotation of a data set. Regarding the descriptors, some families, tested on their own, have provided significant classification rates (62.3% with only two features), which is increased when the best families are combined (65% accuracy). With respect to the metadata, we have created strategies for automatic annotation and found out that using the number of *likes* and *dislikes* (quality-based metadata) provides successful ways of annotating the corpus, whereas the number of *views* (quantity) is not useful for deriving a rate related to aesthetics perception.

Keywords: automatic aesthetics prediction, image descriptors, video descriptors, YouTube, automatic annotation

1. Introduction

Social networks based on audiovisual content like *YouTube*, *Flickr*, *Instagram* or *Vimeo* are currently changing the way we communicate: multimedia resources are be-

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

coming increasingly more important. Nowadays, not only do we find audiovisual content in cinemas and on television, but also millions of hours of videos are available for everyone on the Internet. Statistics like the ones provided by YouTube in [23] claim that more than 6 billion hours of video are watched each month on YouTube, which is an example of the current importance of this kind of content on the web. This is being exploited by many agents, including companies which want to advertise their products, because one big difference between videos available on the Internet and, for instance, television, is that the Internet provides easy tools for sharing and tracking the impact of contents.

Then, with such an amount of audiovisual data available on the web, it is essential to have tools that facilitate their management. In the past and still today, most tools to automatically organize, retrieve or analyze multimedia content were based on text-like information, such as tags or metadata. However, these procedures are being gradually replaced by approaches based on content, a method which offers a range of advantages: information from content is much deeper than simple tags and obviously reflects more accurately the essence of the items. The counterpart is, of course, that dealing with the content is far more difficult than processing text.

In particular, one application motivated by the advantages of using multimedia content to extract information which has gained much interest in recent years is the prediction of the aesthetic value of an image or a video by means of their audiovisual properties, a field which can be referred to as *aesthetics assessment* or *prediction*. The word *aesthetics* has many philosophical connotations, but looking at its Greek etymology one can find that it originally refers to *sensation* or *perception* and these are specifically the meanings we will attribute to that word along this paper, because the potential of inferring information related to aesthetics from an audiovisual element lies in the fact that it allows obtaining an idea of how users or consumers of an audiovisual piece perceive it and feel it.

Aesthetics prediction in multimedia information is a challenging problem because it involves not only dealing with the extraction of information from content, but also inferring objective conclusions from subjective opinions. However, it has gained great interest recently because of the wide range of possibilities it potentially offers. Being

capable of predicting the perception of viewers of a particular picture or video can be of great application in different contexts. For example, it could be used in recommendation systems for better retrieving multimedia information or it could be used to assess the aesthetic quality of an audiovisual production before publishing it, being the latter the application we have tried to exploit in this work.

1.1. Previous works

One of the main applications of aesthetics assessment is in the field of recommendation systems, which is extensively surveyed by Adomavicius and Tuzhilin in [1], also proposing possible improvements and tendencies in the future. Looking at the particular case of YouTube, which is the source of data for our work, a study of its recommendation system was done in [6]. In this paper we found some evidence that recommendation is still based on users' activity, without incorporating elements related to the aesthetics of the videos or other content-based features. Closely related to recommendation systems, automatic aesthetics prediction can also be applied to image and video classification and retrieval with the aim of improving the systems by incorporating elements related to the perception of users or the aesthetic value of multimedia content. A survey on the literature of this field was carried out in [4].

Focusing on the relatively new field of aesthetics prediction, within which we can set this work, it is important to remark that before the first attempts with videos, it was firstly studied in still images. One of the earliest approaches towards this domain was carried out by [19] fifteen years ago. In that paper, they aimed to find out which aspects were related to image appeal with a data set of 194 pictures previously ranked by 11 people. They came to the conclusion that image appeal had to be addressed through metrics others than those used for measuring image quality. More recently, Datta *et al.* proposed in [5] 56 low-level image features tested on 3581 pictures with ratings from the site Photo.net and selected the top 15 features related to photographic aspects like the rule of thirds or the depth of field that achieved together an accuracy of 70.12% in separating low from high rated photographs. Several works followed this one by adding different contributions. For instance, [12] carried out a higher-level analysis to assess the aesthetic quality of photographs and Marchesotti *et al.* [14] extended the

study by using a larger and diverse set of features and achieved an accuracy of 89.9%.

Applied to videos, automatic aesthetics prediction has not been addressed until a few years ago. To the best of our knowledge, the first work of this type was performed by [15] in 2010. They collected 160 consumer videos from YouTube and performed a controlled user study to obtain rating labels as ground truth to finally evaluate the usefulness of a set of frame-level features inspired by those of [5] and extended to the temporal dimension, obtaining an accuracy of 73%. [21] used the same data set and extended the work by making a differentiation between semantically independent and dependent features in order to perform a comparative study and [2] proposed a model with features based on psycho-visual statistics. Furthermore, [7] proposed some new features at the video-level based on cinematographic and photographic notions and a model which automatically annotates the video through clustering techniques using YouTube metadata. That paper is the starting point of the present work.

It is remarkable that very recently the research on aesthetics modeling has been extended to incorporate also audio features. To our knowledge, the first works in this regard were [11], in which a wide range of multimodal features is proposed, and [8] which offers a comparative study of the performance between visual and acoustic features.

1.2. Main objectives

The aim of this paper is double: on the one hand, given the advantages of annotating a data set automatically over the common procedure of recruiting people for rating the videos *ad hoc*, we have designed three different strategies for obtaining labels related to how positively or negatively a video is perceived by its users with the aim of finding out which metadata are suitable for that purpose and which are not. One strategy relies on YouTube metadata based on quality, such as the number of *likes* and *dislikes*, another strategy uses the number of *views*, i.e. quantity, and a third strategy combines both. We describe these strategies in more detail in Section 2 and discuss the suitability of them in Section 5. On the other hand we propose a set of video descriptors organized into eight different families, together with a procedure that allows predicting if a YouTube video has been perceived in a positive or negative way, with the objective of performing

a comparative study of the families which enables us to identify appropriate types of features for future research on automatic aesthetics prediction. The visual features are presented in Section 3 and the corresponding discussion in Section 5.

2. Generating viewers' ratings from YouTube metadata

Previous works on aesthetics prediction of videos have used diverse data sets: for instance, [2] tested their features on a database with 1,000 videos of different topics released by NHK in 2013, whose ratings where provided by only 10 people. [15] built a corpus of 160 consumer videos collected from YouTube and annotated also through a survey. On the contrary, we aim to build a system which does not depend on an *ad hoc* procedure for a manually annotation of the videos, but uses instead available data provided by real users and consumers of the videos; for instance metadata collected by YouTube, 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.

2.1. Videos retrieval

The main advantage of annotating our corpus by using YouTube metadata is that they are provided by users as they watch, share and interact with other users and, therefore, these data will be closely related to how viewers actually perceive each video. However, it also has challenging drawbacks and we need to be aware that not every video in YouTube is commensurately assessed. There are some kinds of videos which are more popular than others and cannot be compared in terms of their metadata and the chances are that differences in metadata do not reflect a real difference in the aesthetics assessment. Furthermore, in order to be capable of providing labels for each video according to the users' perception, we need the videos to have a sufficient amount of metadata so that they are representative of the general assessment.

Hence, in order to minimize any possible bias, we have restricted our domain to one single type of videos: car commercials. The choice of this domain is motivated for several reasons: first, advertising videos have similar and limited duration, which is appropriate not only for computational reasons, but also for having certain homogeneity within the corpus. Besides, since the target of every car commercial is to sell the

car it advertises, we reduce the bias and make the metadata be more connected with the users perception of how the commercial is made. 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.

We have chosen YouTube as the source of our videos for two main reasons: it offers a huge amount of available videos (100 hours of video are uploaded every minute according to [23]) and also the high number of users (more than one billion unique users per month according to the same official source) is very convenient for having rich metadata. After a filtering procedure, which is detailed in [7], we make up our data set with a total of 138 videos.

2.2. Annotation

It has been already briefly discussed that employing metadata inherent to the videos to annotate the corpus, instead of recruiting a group of people to watch the videos and rate them, offers a series of advantages: on the one hand, the procedure is less expensive and can be replicated at any time with an extended corpus and, on the other hand, labels are more closely linked to the original viewers of the videos; the raters, i.e. users, are not biased by laboratory conditions. Nonetheless, regarding video aesthetics prediction, to our knowledge, first works were carried out by obtaining labels with participants who rated the videos *ad hoc* ([2, 15, 21]). Conversely, [7] adopted for the first time an automatic procedure for annotating the corpus based on unsupervised learning techniques, such as k-means clustering. In the present work, we are aware that the complexity of applying those unsupervised machine learning techniques over different metadata involves also some risk of introducing certain noise in the labels. For that reason, in this work we follow a simpler method with the aim of comparing the performance of different types of visual features in a more fairly way. The solution adopted has been similar to the ones in [14, 5] when assessing the aesthetics of still pictures, where they used rates offered by users of the *Photo.net* platform.

For the sole sake of simplicity, from the whole set of metadata collected by YouTube for every video, we decide to use only two of them, having, thus, a simple and transpar-

ent method that enables us to derive clear conclusions about both the visual descriptors and the metadata themselves. On the one hand, we have the number of views or *viewsCount*, which refers simply to the number of times a video has been played. On the other hand, we have the likes-dislikes ratio or *ldRatio*, which is actually a combination of two raw metadata provided by YouTube, the number of *likes* and *dislikes*, built according to the following formula:

$$ldRatio = \frac{numLikes}{numLikes + numDislikes} \tag{1}$$

which, obviously, only applies when $numLikes + numDislikes \geq 0$ and otherwise is set to 0. These two metadata have been chosen because they can be easily identified with two commonly used concepts when assessing something: quality (*ldRatio*) and quantity (*viewsCount*). Then, we define three annotation strategies based on the sample median of the metadata as threshold:

- Quality: videos with an *ldRatio* above the median are assigned to one class and vice versa.
- Quantity: videos with a *viewsCount* above the median are assigned to one class and vice versa.
- Combination: four classes are created with the four possible combinations taking the median of *ldRatio* and the median of *viewsCount* as thresholds.

By using this method for getting the classification labels we assume the hypothesis that videos with many views and high *ldRatio* should correspond to videos more positively perceived than those with fewer views and lower rate. Table 1 shows a summary of the strategies and some statistics about the metadata and the yielded classes.

3. Visual descriptors

For constructing the model of automatic aesthetics prediction we have defined 8 families of descriptors. The choice of some of them has been motivated by previous works like [5, 15, 7] which have already proved the convenience of certain types of

	Quality	Quantity	Combination
Metadata	*ldRatio*	*viewsCount*	*ldRatio* & *viewsCount*
Median	0.93	6,917	–
Standard deviation	0.17	116,640	–
In-class mean	0.76/0.99	2,897/93,601	0.71/0.78/0.98/1.0 & 2,707/107,990/66,784/3,005
Videos per class	69/69	69/69	25/44/44/25

Table 1: Statistics of classes and strategies

features for the task of predicting aesthetics both in pictures and videos, but, in addition to that, we have proposed some novel features inspired by photographic and cinematographic rules of thumb, given that it is a common practice to follow those in film-making so as to create aesthetically appealing videos.

It is important to mention that some slight pre-processing has been applied on the videos in order to avoid undesirable distortions in the feature values. In particular, the pre-processing consists of two parts: removal of black frames at the beginning and the end of the videos and removal of black bands around the frames. Both operations are convenient due to the fact that black frames and black bands introduce a big amount of dark pixels that can influence considerably the values of certain features. The procedure for removing such elements was automatic and based on the energy of borders and statistics about the intensity.

The 8 families with the total of 26 features that make them up are described next.

3.1. Intensity

In photography and film-making, intensity is also commonly referred to as brightness. Although it is usually controlled to capture *correctly* exposed images, regarding the useful exposure range of the film or sensor, it can also be used to create many effects by under- and overexposing the image. Besides, on the one hand, the exposure is not the same under day light conditions than indoors, for instance. And on the other hand, the exposure does not have to be necessarily the *correct* one. Hence, this is a

feature that can potentially have some influence on aesthetics.

Intensity in a picture or frame is the average value of the pixels of the gray-scale version of the image. This image-level feature can be extended to the video level by computing the following statistics:

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

3.2. Hue

The use of hue as a feature for automatically evaluating aesthetics was already introduced by [5] in the case of still images. David Bordwell points out the importance of color on the *mise en scène* in [3, pp. 148–157,186–189] as one of the most effective resources in film-making, and one of the simplest ways of characterizing color is through the hue, which is indeed one of the channels of the well-known color space HSV [20]. Roughly speaking, hue allows identifying colors by an angle from 0 to 360 degrees. As in the case of the intensity, we compute the following statistics:

- *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.

3.3. Saturation

This feature is a neighbor of hue, as it is another channel of the HSV color space. Saturation can be thought as a parameter that measures the purity of the color, i.e. how close to gray a color is. It is expressed as a percentage, being 100% fully saturation and 0% a gray tone. Again, we obtain the average saturation and the standard deviation along the whole video:

- *mean-saturation*: average of the pixel values of the saturation channel of every frame in a video.
- *std-saturation*: standard deviation of the saturation channel.

3.4. Entropy

Since entropy is a statistical measure that refers to the randomness of a variable, applied to images it can describe texture. Textures can be important as it gives an idea of the complexity of an image, which can be exploited to produce a particular effect on the viewers. Four features related to entropy are computed:

- *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. This feature is designed to capture those commercials that insert some extra frames in the video, among the filmed scenes or at the end, to show the brand logo, a car description, and/or the conditions of an offer. Due to the typical monochromatic background these frames usually have very low entropy compared to others.
- *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.

3.5. Temporal segmentation (cuts)

The main characteristic of videos with respect to images is the temporal dimension, thus, features describing this aspect are of great interest in our analysis and, to our knowledge, previous works have not used this kind of descriptors. Temporal segmentation is in film-making and publicity the basis of montage, the editing technique that allows the creation of most effects cinema produces. For example, an action scene has usually many more number of cuts than a calm, descriptive scene [3, 17].

In order to extract features related to this aspect, it is necessary to determine the abrupt transitions between subsequent shots, which are the most common ones. We

have followed the procedure described [22], which uses the sum of absolute differences (SAD) of the gray intensity, I, which is defined for each frame n as follows:

$$D(n) = \frac{1}{H \cdot W} \sum_{x=1}^{W} \sum_{y=1}^{H} |I_n(x, y) - I_{n-1}(x, y)| \tag{2}$$

where H and W denote the frame height and width respectively. The detection performance can be improved by using a discrete version of its second derivative. This offers additional robustness at high speed movements as it detects abrupt transitions of the first derivative:

$$M(n) = -D''(n + 1) = -(D'(n + 1) - D'(n)) \tag{3}$$

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 validation with previously labeled videos) is set to locate cuts. Then, with this information we define some features:

- *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.

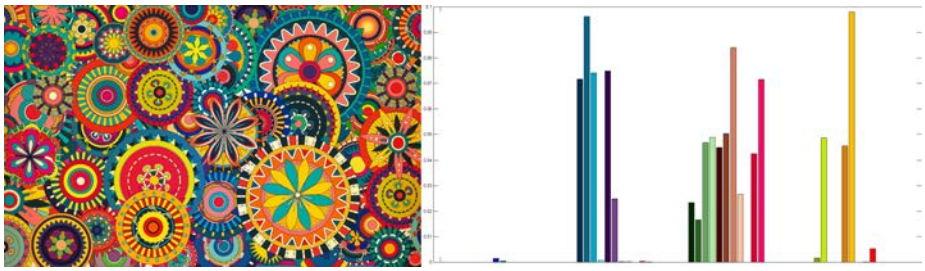
3.6. Frame-level colorfulness

With this visual characteristic, rather than measuring the intensity or vividness of colors, which is described by previous features, we aim to measure the degree of variation of colors. A picture is referred as colorful when it has richly varied colors, in

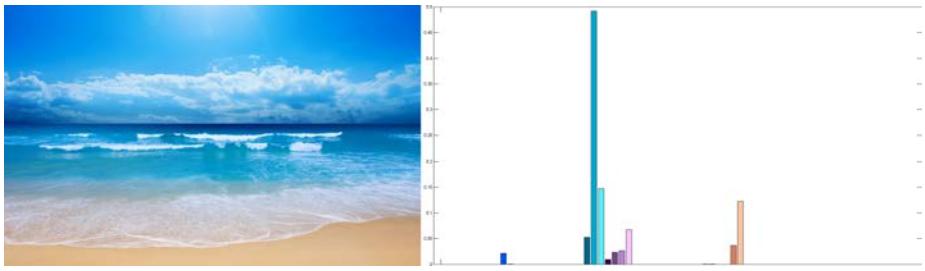
contrast to monochromatic or poorly colored images. From the point of view of analyzing car commercial videos, colorfulness could be of interest to learn whether using colors in the frames, or the absence of them, may attract people.

For this family of features, we compute the colorfulness of every frame and extend it to the temporal dimension by averaging and computing the standard deviation as usual. In order to compute the colorfulness of a frame we follow a variation of the procedure detailed in [5]. The idea is to compute the 64-bin color histogram (after conversion to the CIE Lab color space [16]) of each frame and compare it through the Earth Mover's Distance [18] with the histogram of an ideal colorful picture, i.e. uniformly distributed.

In order to better illustrate how the histograms reflect the variety of colors and their effect on the colorfulness, a couple of examples of pictures and their colorfulness are presented in Figure 1.



(a) A multicolor image with value of colorfulness $C = 0.67$



(b) An image with predominant blue color and value of colorfulness $C = 0.49$

Figure 1: The image on the top has many different colors, while the image on the bottom is mainly blue. The color histograms show the variety of colors of both pictures.

From frame values, colorfulness can be extended to the video level by computing the following features:

- *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.

3.7. Video-level colorfulness

As an important novelty in this work we have developed a modification of frame-colorfulness. The difference of this family of features with respect to the previous one is that instead of computing the colorfulness of each frame, a value of colorfulness is computed for the set of pixels of the video as a whole. That is, we compute one single color histogram taking into account all the pixels of the video and then compare it to the ideal color histogram as previously explained. Note that this way of defining colorfulness is quite different to the frame-level one. Now a distribution of the feature along the frames is not available, but, instead, we can determine, for instance, the peaks of the histogram, which are indicative of the most predominant colors along the whole video. The particular features derived from this method are the following:

- *video-colorfulness*: colorfulness computed taking into consideration all the pixels of the video at once.
- *first-color*: index (from 1 to 64) of the color with the highest frequency in the histogram.
- *first-color-freq*: relative frequency in the histogram of the first color.
- *second-color*: index of the color with the second highest frequency in the histogram.
- *second-color-freq*: relative frequency in the histogram of the second color.

3.8. Rule of Thirds (ROT)

The rule of thirds (ROT) is a very important rule of thumb in visual arts, such as photography, painting or design. It is the rule for 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. One example of this taken from our data set is shown in Figure 2.

Figure 2: A sample image with the horizontal and vertical third lines

Thirds are used because they approximate 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, something very close to aesthetics. In film-making and photography it is also followed to place the line of the horizon or any other horizontal dividing line within the frame, especially when filming landscapes, being useful to give some priority to the upper or the lower part, depending on where the line is placed.

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 64-bins color histograms, H, corresponding to the two sub-images that the horizontal line generates:

$$D_{ROT} = 32 \cdot \frac{1}{64 \cdot H \cdot W} \sum_{b=1}^{64} |H_{top}(b) - H_{bottom}(b)| \tag{4}$$

where H and W are frame height and width. 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, corresponding to frames from one of our videos, from which the value of the feature applied to the upper third line has been calculated.

(a) $D_{ROT-U} = 0.89$ (b) $D_{ROT-U} = 0.26$

Figure 3: The image on the left follows the rule of thirds, while the image on the right does not. The values of the measure D_{ROT} for the upper third have been computed for both pictures.

This procedure for getting a measure to represent the degree of utilization of the rule of thirds (ROT) is completely novel and, to the best of our knowledge, previous works on aesthetics prediction in videos have not used ROT-based features. 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.

4. Experimental setup

With the aim of briefly summarizing what has been presented up to now, we find important to recall that we have defined three different strategies in terms of labels: quality, quantity and the combination of both. That is, there are 3 different versions of the data set, each with potentially different labels (see Section 2). One of the objectives will be then to compare the different strategies. Besides, we have defined 8 families of descriptors in Section 3 which will be also compared in terms of classification performance.

In order to make fair comparisons between the different families, we have performed identical classification experiments on each of the 8 feature subsets corresponding to the different families. In particular, we have classified with a Logistic Regression model with ridge estimator, based on the well-known method of [13] and using the implementation of the WEKA machine learning software, from the University of Waikato (New Zealand) [10]. As sampling method, we have performed at each experiment 10 random repetitions of 10-fold cross validation ($10 \times$ 10-fold CV), averaging over the folds and hence, getting 10 results for each features family.

5. Discussion of results

The distribution of results provided by the 10 runs per features family and strategy have been shown in Figure 4 as a box plot, which is particularly useful for visualizing the distribution of results of each family in a comparative way: in this case, the edges of each box represent the 25th and 75th percentiles and the median is marked with a solid line within the boxes. The *whiskers* denote the $\pm 2.7\sigma$ limits and data points beyond them are considered outliers and depicted with '+' symbols. There are three boxes per family of features, each for one of the strategies. Note that we have included the *combination* strategy in the same plot for compactness reasons even though the comparison with the others is not fair as it deals with 4 classes instead of 2. The baseline accuracy has been also added to the plot at 50% and at 31.9% for the combination strategy.

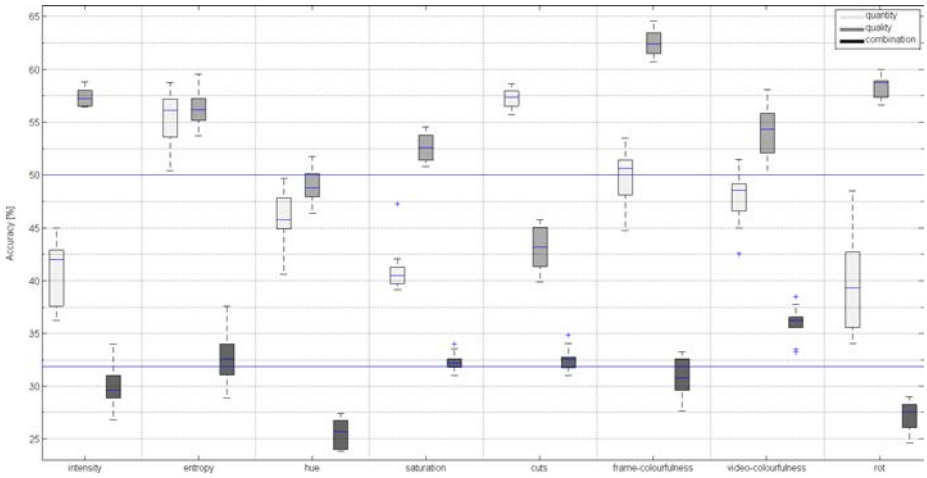


Figure 4: Box plot with the distributions of classification accuracy for each features family and strategy

5.1. Quality vs. quantity

At the first level of analysis we can make a comparison between the performance of each strategy. On the one hand, it seems clear that the combination strategy does not provide promising results in general as most of them are close to the baseline accuracy. In fact, none of the families, except the novel video-colorfulness, has shown to be statistically significant from the baseline scheme, according to the rather conservative corrected paired t-test carried out on every experiment. One possible reason is that the attempt of combining quality and quantity metadata from YouTube to derive definite labels in terms of assessment has not been successful and, together with the conclusions derived in [7] (automatic clustering techniques combining metadata yield to hardly interpretable clusters), it seems that it is not a straight-forward task. On the other hand, the comparison between quantity and quality is more fair and surely more interesting.

At first glance, results from the quality strategy seem to outperform those from quantity. The difference is quite clear in 5 of the families (intensity, saturation, frame and video colorfulness and rule of thirds), while for the entropy family both strategies perform fairly the same, for the hue quality is only slightly higher and for the cuts quantity works exceptionally better than quality. In order to confirm these conclusions we have carried out the Wilcoxon signed-rank test on every family to find out if one

strategy performs better than the other. The Wilcoxon test is typically used to compare paired data as a non parametric alternative to the paired t-test. The results of test, shown in Table 2, confirm the intuitions previously stated.

Finally, another weakness of the quantity strategy is that it has not provided significant results with respect to the baseline, whereas quality has proved to be significant in several families (frame-colorfulness, intensity and rule of thirds). An explanation for this difference between quality and quantity is that the latter is a very noisy parameter, in that the sample standard deviation is quite high (116,640 as reported in Table 1) and such a variation might not be indicative of better aesthetic perception by users. On the contrary, quality metadata are intuitevely more connected to the aesthetic value of the video than the number of views, which is affected by many other reasons. Therefore, our conclusion is that the convenience of quantity-related metadata alone for assessing aesthetics is doubtful and its combination with other metadata should be done with extra care.

	int.	entr.	hue	sat.	cuts	f-color.	v-color.	rot
p-value	0.002	0.645	0.065	0.002	0.002	0.002	0.002	0.002
h	1	0	0	1	-1	1	1	1

Table 2: Results of the Wilcoxon signed-rank test for comparing quality and quantity results. $h = 1$ indicates that quality accuracy is significantly higher than quantity, $h = -1$ the opposite situation and $h = 0$ that both are equivalent at the 95% confidence level ($\alpha = 0.05$)

5.2. Comparison of features families

Once we have obtained the first conclusions regarding the strategies, we are in shape to perform the, probably, most interesting analysis: comparing the visual descriptors. In order to simplify the analysis and obtain clearly concluding results we will focus mainly on the quality strategy, which has proved to provide better and more coherent results.

Again, we can repeat the previous procedure of first inferring some conclusions by visual inspection of the box plot and then putting some numbers on the differences. If we observe Figure 4 and look particularly at the boxes of the quality strategy (in

medium gray), we can first mark hue, saturation and cuts families as the ones with worst results, with a median accuracy only slightly better or below the baseline. The cases of hue and saturation are rather normal as their features account for very low-level characteristics of color which are hard to be discriminative of aesthetics on their own. In the case of cuts, better results could be expected, but it turns out to provide particularly low results. Nonetheless, note that, on the contrary, the quantity strategy does perform specially well with this family.

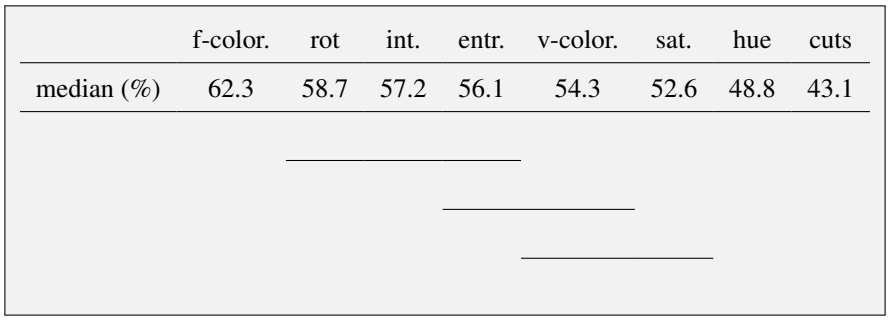


Figure 5: Results of the Mann-Whitney U-test for comparing pairwise the 8 different families of features. Families are sorted from left to right by descending median accuracy. Families underscored by the same line are not significantly different at the 95% overall confidence level ($\alpha = 0.05$), $\alpha = 0.0063$ with Bonferroni correction.

At the opposite extreme we have that frame-colorfulness features have given the highest result (62.3% in median), followed by rule of thirds, intensity, entropy and video-colorfulness. For the sake of rigor we have also performed the statistical analysis of the results. Here, as samples from different families are not paired we have used a Mann-Whitney U-test, a non parametric alternative to the two-sample t-test, to evaluate the significance of every features family pairwise. The conservative Bonferroni approach has been followed, reducing the level of significance by a factor of $1/8$. We report the results in a graphical way in Figure 5, where underlined families denote non significance. We have also included the medians in the figure.

5.3. Merging features families

Lastly, the natural experiment after dealing with the families of features on their own is bringing them together so as to find out if different descriptors can cooperate

and increase the classification accuracy. For that purpose, we have selected the best half of the families, i.e. frame-colorfulness, rule of thirds, intensity and entropy, which make a total of 12 features, and have classified with the same set-up as before: Logistic Regression with 10×10-fold cross-validation. The only difference is that we have let a feature selection algorithm [9] retrieve the 5 to 12 best descriptors, in order not to constrain the classification to the fixed maximum number of them.

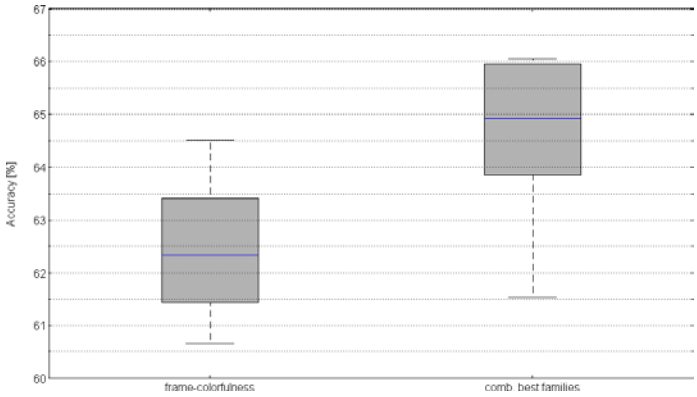


Figure 6: Box plot with the distributions of classification accuracy of the best family (frame-colorfulness) and the combination of the four best families.

With this set-up, the best classification experiment achieves a median accuracy of 64.9%, which turns out to be significantly higher (with 95% confidence) than the results provided by frame-colorfulness features alone, according to a Wilcoxon sign-rank test, with p = 0.0195. For achieving that accuracy it uses 3 rule of thirds descriptors (*mean-hrot-lt*, *mean-hrot-ut* and *std-hrot-ut*), 2 entropy descriptors (*pct-low-entropy* and *end-low-entropy*), 1 frame-colorfulness descriptor (*std-colorfulness*) and 1 intensity descriptor (*std-intensity*). Some important conclusions can be obtained from this experiment: first, the combination of visual descriptors from different families increases statistically significantly the accuracy, although only by about 2.5%. And second, the best combination uses a simple model with *only* 7 features from every selected family, which, on the one hand suggests that combining all features at the same time generates some complexity that might be handled by a more sophisticated classifier and, on the other hand, proves the convenience of the families for assessing the

aesthetics in a complementary manner.

6. Main conclusions and future work

We have presented a comparative study of 8 different families of visual descriptors, most of them based on photographic and cinematographic ideas, for assessing the aesthetic value of 138 car commercials, under the hypothesis that such perception can be modeled by the feedback given by viewers in YouTube by means of *likes* and *dislikes* (quality) and *number of views* (quantity). In this regard, we have proposed a procedure for automatically deriving labels for each video (as an alternatively to methods relying on *ad hoc* surveys) following three different strategies in terms of the employed metadata: by quality, by quantity and combining both.

Two important conclusions can be obtained from this work. First, we have identified some families of descriptors which are suitable for automatically predicting aesthetics in videos, namely frame and video colorfulness, descriptors related to the rule of thirds, intensity and entropy-based descriptors. It is also remarkable that colorfulness and rule of thirds have been used, to our knowledge, for the first time in this work for assessing aesthetics in videos. These findings set a promising path for future research in the field.

Second conclusion is that it has been shown that using quantity-based metadata to model the aesthetic perception is not straight-forward, but, on the contrary, a separation of data based on quality metadata has turned out to be successful. Indeed, with the quality-based annotation three of the features families (frame-colorfulness, intensity and rule of thirds) have provided statistically significant classification rates on their own, reaching an accuracy of 62.3% with only two descriptors from the frame-colorfulness family. Besides, we have achieved a 64.9% rate by combining the best families of descriptors. These conclusions discourage from using quantity metadata in the future, whereas the success of the quality-based annotation seems to point towards the right direction in terms of automatic annotation.

In the future, we aim to add new types of features, like motion-based descriptors and higher-level visual features. A similar comparative study as the one presented

in this work could be done with audio descriptors, with the aim of obtaining refined multimodal predictors in the future. Furthermore, it will be of great interest to collect a larger data set which includes videos from different domains.

[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] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] F. Fernández-Martínez, A. Hernández-García, and F. D. de María. Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials. *Expert Systems with Applications*, pages 293–305, 2015.

[8] F. Fernández-Martínez, A. Hernández-García, A. Gallardo-Antolín, and F. D. de María. Combining audio-visual features for viewers perception classification

of youtube car commercials. In *Proceedinggs of Workshop on Speech, Language and Audio in Multimedia (SLAM)*, 2014.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, Mar. 2002.

[10] 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.

[11] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *Proceedings of The 27th AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[12] 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.

[13] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[14] 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.

[15] 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.

[16] I. C. on Illumination. *Colorimetry: technical report*. CIE technical report. Commission internationale de l'Eclairage, CIE Central Bureau, 2004.

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

[18] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 59–, Washington, DC, USA, 1998. IEEE Computer Society.

[19] A. E. Savakis, S. P. Etz, and A. C. P. Loui. Evaluation of image appeal in consumer photography. *Proc. SPIE*, 3959:111–120, 2000.

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

[21] 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.

[22] 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.

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