Signal Processing: Image Communication 47 (2016) 280–288

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Comparing visual descriptors and automatic rating strategies for video aesthetics prediction

A. Hernández-García [a,*], F. Fernández-Martínez [b], F. Díaz-de-María [a]

[a] *Universidad Carlos III de Madrid, Leganés, Spain*
[b] *Universidad Politécnica de Madrid, Madrid, Spain*

ARTICLE INFO

Article history:
Received 1 February 2015
Received in revised form
11 June 2016
Accepted 13 July 2016
Available online 15 July 2016

Keywords:
Automatic aesthetics prediction
Image descriptors
Video descriptors
YouTube
Automatic annotation

ABSTRACT

Automatic aesthetics prediction of multimedia content is bound to be a powerful tool for artificial intelligence 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 (1) the performance of eight families of visual descriptors in accounting for the general aesthetics perception of videos and (2) 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 YouTube 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 metric related to aesthetics perception.

1. Introduction

Social networks based on audiovisual content like *YouTube*, *Flickr*, *Instagram* or *Vimeo* are currently changing the way we communicate: multimedia resources are becoming 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 [25] 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 devoted 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

* Corresponding author.
E-mail addresses: ahgarcia@tsc.uc3m.es (A. Hernández-García), ffm@die.upm.es (F. Fernández-Martínez), fdiaz@tsc.uc3m.es (F. Díaz-de-María).

http://dx.doi.org/10.1016/j.image.2016.07.004
0923-5965/

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. Main contributions

The contributions of this paper are double: first, 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 3 and discuss the suitability of them in Section 6. The proposed methods for automatic annotation could be replicated to annotate further examples from YouTube. Second, 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 4 and the corresponding discussion in Section 6.

2. Related 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 that 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 addressing the videos domain, still images were firstly studied. One of the earliest works in this regard was carried out by Savakis et al. [20] after the turn of the century. In that paper, they aimed to find out which aspects of images were related to 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 other than those used to measure 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, Khan and Vogel [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%. Recently, image aesthetics evaluation has been addressed by combining local and global structures [27] and aesthetics is also taken into account for automatic photo cropping [28].

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 Moorthy et al. [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%. Yang et al. [23] 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 Bhattacharya et al. [2] proposed a model with features based on psycho-visual statistics. An interesting approach was carried out by Wang et al. [22], consisting in training the algorithms not only with labeled videos, but also still images. Furthermore, Fernández-Martínez et al. [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. Regarding video aesthetics, it is interesting to remark that it is also taken into account in other fields such as video summarization [26,29].

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.

3. Generating viewers' ratings from YouTube metadata

Previous works on aesthetics prediction of videos have used diverse data sets: for instance, the authors of [2,22] tested their features on a database with 1000 videos of different topics released by NHK in 2013, whose ratings were provided by only 10 people and which is not publicly available. Moorthy et al. [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.

3.1. 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.