Expert Systems with Applications 42 (2015) 293–305

Contents lists available at ScienceDirect

Expert Systems with Applications

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

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, Spain

ARTICLE INFO

Article history:
Available online 8 August 2014

Keywords:
Automatic video annotation
Aesthetic quality assessment
Video sentiment analysis
Video metadata
YouTube

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.

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 (Brezeale & Cook, 2008) and video recommendation (Adomavicius & Tuzhilin, 2005) 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

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

http://dx.doi.org/10.1016/j.eswa.2014.07.033
0957-4174/

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. Fig. 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.

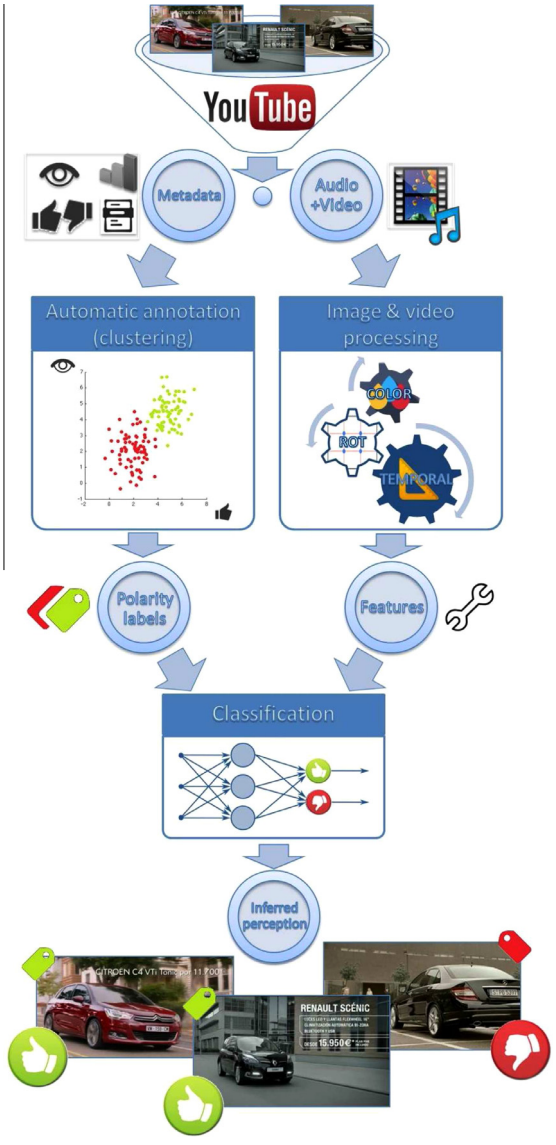

Fig. 1. Diagram of the approach overview.

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 (Adomavicius & Tuzhilin, 2005), 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 Davidson et al. (2010). 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 Brezeale and Cook (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 (Luo, Wang, & Tang, 2011; Luo & Tang, 2008), 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 (Westerski, 2009), 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 (Nasukawa & Yi, 2003; Pang, Lee, & Vaithyanathan, 2002). The first attempt to extend sentiment analysis to audiovisual data was recently carried out by Morency, Mihalcea, and Doshi (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 Rosas, Mihalcea, and Morency (2013) and Wollmer et al. (2013).

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