Fernández-Martínez, F., Hernández-García, A., Fernández-Torres, M.A., González-Díaz, I., García-Faura, Á., Díaz de María, F. (2018) Exploiting visual saliency for assessing the impact of car commercials upon viewers. *Multimedia Tools and Applications*, 77(15), pp.: 18903–18933.

DOI: https://doi.org/10.1007/s11042-017-5339-9

The final publication is available at link.springer.com

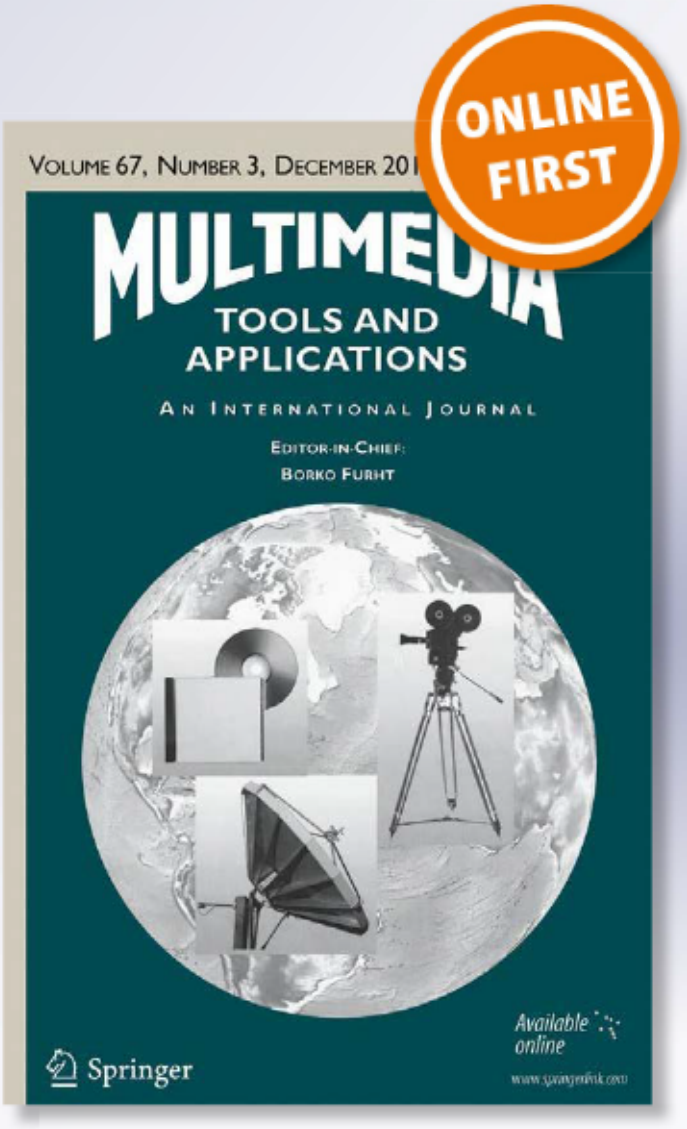
Exploiting visual saliency for assessing the impact of car commercials upon viewers

F. Fernández-Martínez, A. Hernández-García, M. A. Fernández-Torres, I. González-Díaz, Á. García-Faura & F. Díaz de María

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-017-5339-9



Springer

Multimed Tools Appl
https://doi.org/10.1007/s11042-017-5339-9

Exploiting visual saliency for assessing the impact of car commercials upon viewers

F. Fernández-Martínez[1] · A. Hernández-García[2] · M. A. Fernández-Torres[2] · I. González-Díaz[2] · Á. García-Faura[1] · F. Díaz de María[2]

Received: 23 November 2016 / Revised: 19 October 2017 / Accepted: 20 October 2017

Abstract Content based video indexing and retrieval (CBVIR) is a lively area of research which focuses on automating the indexing, retrieval and management of videos. This area has a wide spectrum of promising applications where assessing the impact of audiovisual productions emerges as a particularly interesting and motivating one. In this paper we present a computational model capable to predict the impact (i.e. positive or negative) upon viewers of car advertisements videos by using a set of visual saliency descriptors. Visual saliency provides information about parts of the image perceived as most important, which are instinctively targeted by humans when looking at a picture or watching a video. For this reason we propose to exploit visual information, introducing it as a new feature which reflects high-level semantics objectively, to improve the video impact categorization results. The suggested salience descriptors are inspired by the mechanisms that underlie the attentional abilities of the human visual system and organized into seven distinct families

✉ F. Fernández-Martínez
fernando.fernandezm@upm.es

A. Hernández-García
ahgarcia@tsc.uc3m.es

M. A. Fernández-Torres
matorres@tsc.uc3m.es

I. González-Díaz
igonzalez@tsc.uc3m.es

Á. García-Faura
agfaura@die.upm.es

F. Díaz de María
fdiaz@tsc.uc3m.es

[1] Information Processing and Telecommunications Center, Universidad Politécnica de Madrid, Madrid, Spain

[2] Universidad Carlos III de Madrid, Getafe, Spain

Published online: 02 November 2017

according to different measurements over the identified salient areas in the video frames, namely population, size, location, geometry, orientation, movement and photographic composition. Proposed approach starts by computing saliency maps for all the video frames, where two different visual saliency detection frameworks have been considered and evaluated: the popular graph based visual saliency (GBVS) algorithm, and a state-of-the-art DNN-based approach. Then, frame-level salience descriptors are extracted from these maps. Next, pooled statistics are used to collapse the obtained frame-level values into video-level descriptors. Finally, a Logistic regression classifier is built upon the subset of video-level features resulting from a feature selection stage. Experimental validation, conducted on a publicly available corpus of 138 commercials collected from YouTube, shows that the proposed salience descriptors are indicative of the impact upon viewers and achieve a similar performance when compared to a method purely based on aesthetics. Besides, the combined approach, exploiting both saliency and aesthetics together, ultimately results in better performance than what can be achieved individually. In addition, the seven families of salience descriptors defined are also compared in terms of classification performance. Finally, a similar study is also performed targeting the distinct pooling techniques used in the video-level feature computation.

Keywords Visual attention · Saliency · Scene analysis · Aesthetics assessment · Feature extraction · Video impact assessment

1 Introduction

The explosion of Internet video and the integration of TV and the Internet have brought new opportunities for advertising-based services. In this regard, growing faster than even search, video ads are the hottest segment of Internet advertising. Most advertisers have already incorporated Internet into their strategies developing on-line videos meant to capture the clicks and the eyeballs of web consumers.

One of the most interesting technological challenges opened up by such services is the development of computational models for the automatic inference of the affective response of the viewer by exclusively relying on the content of the video. It has been only in recent years that content-based approaches for video classification [8] and recommendation [2] are being researched as an alternative to classical text, tags or metadata based techniques.

Typical CBVIR methods usually decribe the content of the whole image in a uniform way by means of low-level visual features. For instance, a reasonably effective computational model that allow recognizing the aesthetic quality of videos was proposed in [20]. This work demonstrated that it is possible to predict the impact of a video on *YouTube* users, thus determining if it has been positively or negatively perceived, by building a predicting model based on low-level visual descriptors, such as color or texture. Similarly, Fernández-Martínez et al. [19] combined the suggested visual features, respectively related to color, textures, composition, montage, etc. with some additional features extracted from the audio content accompanying the video clips, such as rhythm, tonality, timbre, and roughness, finally improving the performance of the inference system. However, a major drawback of these methods is that low-level contents often fail to describe the high-level semantic concepts viewers use to assess a video [49].

Here, the present work aims at contributing to fill the existing gap [43] by evaluating novel impact models augmented with high-level visual saliency features automatically extracted from videos. Visual saliency provides information about the areas of an image

perceived as most important and instinctively targeted by humans when looking at a picture or watching a video [31]. Given that well-known rules and tips are often followed when creating a video, visual saliency can be considered as an additional dimension of the data implicitly embedded in a video by its creator [4], and is commonly accepted as providing a good approximation of what content is intended to be relevant and generating the greatest impact. Intuitively, saliency can play an important role in anticipating the impact of a video, first by accounting for the obvious consideration that not all parts of the image have the same impact from the perceptual viewpoint, and secondly, by providing information, derived from different qualitative and quantitative measures over the identified salient areas in the video images, on how it is grabbing the attention of the viewers. Hence, we will demonstrate that visual saliency can help to effectively improve the accuracy of predicting the impact of a video upon its viewers, and to better understand the actual influence of the visual content on advertising effectiveness in case of car commercials.

Adopting the same annotated set of videos presented in [20], we will start by computing the saliency map, a topographic map that represents conspicuousness of scene locations, for all the video frames. Once these maps are available, a number of different salience features are extracted from the detected salient regions. These frame-level features will be then extended to the temporal dimension by means of different pooling techniques yielding the required features at the video level. Finally, we will employ several well-known classifiers to assess how much these video-level features may be indicative of the viewers' appreciation of the video, taking special notice of how these features can be combined to provide better results. Figure 1 shows an overview of the suggested approach.

The paper is organized as follows: after this introduction, Section 2 presents a literature review of visual attention modeling and automatic aesthetics assessment techniques. Section 3 describes the visual saliency descriptors extracted for the classification task. Section 4 presents the classification results including corresponding discussions and issues. Finally, some conclusions and future work are laid out in Section 5.

2 Related work

In this section we give the reader an insight into the current state of the art in visual attention modeling. First, we cover the fundamental understanding of how visual saliency and

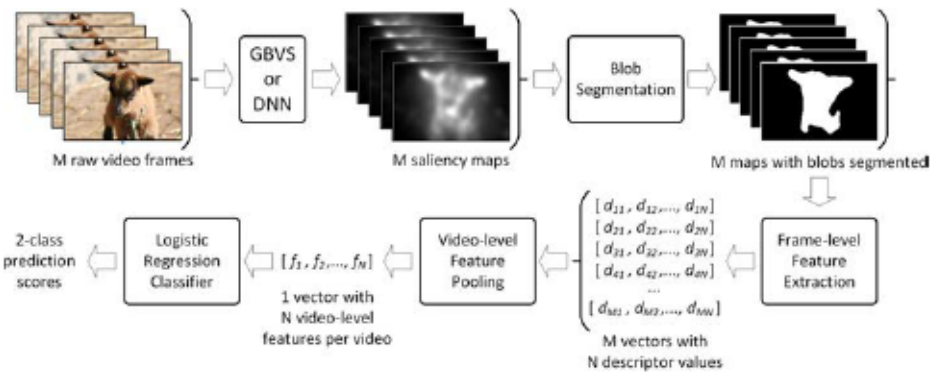


Fig. 1 Overview of the proposed approach

attention work. In this regard, we refer to two of the most important factors typically used to categorize the wide variety of existing saliency models, specifically targeting dynamic models that could be particularly suitable when working with videos. Then, we focus on its practical application providing relevant and useful examples of how visual saliency models can be exploited, not just for visual selective attention purposes, but for other practical applications. Finally, we also review some of the most relevant research works on aesthetics assessment, both applied to still images and videos, and emphasize the novelty of our approach introducing the successful combination of visual saliency and aesthetics together for our video classification task.

2.1 Bottom-up versus top-down models

Visual saliency represents the human visual attention within a visual scene. In 2013, A. Borji, and L. Itti presented a taxonomy of nearly 65 saliency models, which provides a critical comparison of approaches, their capabilities, and shortcomings [5]. According to them, a major distinction among models is whether they rely on bottom-up influences, top-down influences, or a combination of both.

Bottom-up cues are mainly based on the characteristics of a visual scene (stimulus-driven) [59], whereas top-down cues (goal-driven) are determined by cognitive phenomena like knowledge, expectations, rewards, and current goals. Bottom-up attention is fast, involuntary, and most likely feed-forward [17]. On the other hand, top-down attention is slow, task-driven, voluntary, and closed-loop [32].

Top-down factors play an important role in attentional selection. For example, previous studies based on attention experiments [11, 65, 84] have demonstrated that some objects or elements, such as text or human faces, are naturally salient for humans, being totally independent of the way they are shown in the scene. In principle, subjects selectively direct the attention when visualizing a scene depending on both [34], top-down and bottom-up factors, but in practice models have been focusing on each of them separately. In this regard, most models fall into the bottom-up category [9, 29, 69, 85], mainly because bottom-up models provide a generic approximation of attention and deal with aspects that are independent of any internal state of the subject, thus being easier to understand and to measure.

Color, intensity, orientation, and movement are just a few examples of bottom-up visual features that contribute to the selective attention process [40]. Bottom-up attention models [58] estimate a saliency map with the spatial distribution of saliency, where saliency is measured as a scalar quantity at every point in the visual field (i.e. every pixel) by determining how different every given location is from its surround attending to such features. The fundamental model by Itti/Koch [32, 33] has been probably the most frequently used for this purpose. Nonetheless, saliency algorithms are constantly evolving, allowing more accurate saliency maps, for instance, through the identification of salient areas (also known as blobs) by taking into account dissimilarities in pixels neighborhood.

The Graph-Based Visual Saliency (GBVS) framework, proposed by Harel et al. in 2006, is another bottom-up visual saliency model based on graph computations which is able to predict human fixations on the salient regions more reliably than the Itti/Koch algorithm and other tested algorithms [26]. The detection algorithm consists in two simple main steps: first, features from the given image are extracted based on biological fixations, building several feature maps with them. Then, these feature maps are combined and normalized forming the final saliency map. The salient regions found using this method are also found to be more cohesive than with other methods while maintaining high accuracy.

Recently, more graph-based approaches have been proposed. Jiang et al. introduce the discriminative regional feature integration (DRFI), which integrates regional contrast, property and backgroundness descriptor together to formulate the master saliency map [35, 80]. Based on graph-based manifold ranking (MR), the work of Yang et al. [83] utilizes the four boundaries of the input image as background prior to extract foreground queries for the final saliency map. In [44], Li et al. have introduced two major innovation aspects: the erroneous boundary removal process to optimize the image boundary selection and the regularized random walks ranking to improve the foreground saliency estimation. Finally, Qin et al. have recently proposed a novel bottom-up method based on a propagation mechanism dependent on Cellular Automata, an intuitive updating mechanism which exploits the intrinsic relevance of similar regions through interactions with neighbors [66].

At present, deep neural networks (DNNs) have been applied to detect salient objects achieving state-of-the-art performance [46, 79]. These data-driven saliency models aim to directly capture the semantic properties of salient objects in terms of supervised learning from a collection of training data with pixel-wise saliency annotations. For instance, the Deep Contrast Learning model for Salient Object Detection proposed by Li and Yu, is based on an end-to-end deep contrast network that consists of two complementary components, a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. The first stream directly produces a saliency map with pixel-level accuracy from an input image, while the second one extracts segment-wise features very efficiently, and better models saliency discontinuities along object boundaries. Finally, a fully connected CRF model is also incorporated to improve spatial coherence and contour localization in the fused result from these two streams. Recently reported experimental results have demonstrated that this deep model significantly improves the state of the art [42].

2.2 Spatial versus spatio-temporal models

Most of the visual saliency models are only considering spatial information such as contrast. These models are called spatial and were designed at first for still pictures. They are also commonly referred to as static saliency models as they do not consider the temporal dimension. However, in the real world, the visual information we receive is constantly changing due to either egocentric movements or dynamics of the world. The Human Visual System (HVS) is highly sensitive to the relative motion. Consequently, our visual attention is dependent on both current scene saliency as well as the accumulated knowledge from past scenes. Hence, there are also models called spatio-temporal based on the motion present in videos. By applying these dynamic attention models we should be able to capture scene regions that are important in a spatio-temporal manner.

Video saliency estimation methods reasonably differ from image saliency methods. While an image can be viewed for a long time, a video frame is only observed for a fraction of second. This difference, among others, turns dynamic saliency into an even more challenging concept which is gaining interest nowadays. Although there exist approaches for video saliency detection that specifically address the only use of dynamic features such as speed or direction [54], most of them extend the static image saliency framework by considering motion and including related dynamic features [14, 69]. Despite dynamic features can be strong predictors of eye movement behavior [30], other works such as [74] claim that it may not generalize to natural behavior when watching sequences. Particularly, movie-style edited video clips are found to be problematic stimuli because of their frequent editorial cuts. These cuts present an unusual and artificial situation for the visual

system which hampers normal scene perception. More recently, Nguyen et al. [57] have introduced a comprehensive comparative study of dynamic and static saliency providing two key findings: first, although different, video saliency is yet quite related with image saliency; second, camera motions, such as tilting, panning or zooming, affect dynamic saliency significantly.

2.3 Applications of visual saliency and attention

Visual saliency modeling has captivated researchers since the early 80's with the Feature Integration Theory [75]. The original application (and motivation) of saliency maps was focused on attention and the stimulus and factors affecting it. Hence, many biologically motivated computational models of visual selective attention have been proposed since then [5], for example, to examine the degree to which stimulus salience guides the allocation of attention [63], or alternatively to modify its natural behavior by trying to redirect the human attention to specifically imposed regions [71].

Although it can be considered a relatively old research topic, it still remains very active beyond traditional research involving eye-tracking and simulations regarding visual attention. For example, there are numerous examples of applications which make use of saliency maps in data processing. In this regard, communication systems dealing with visual data (i.e. video streams or images) can be improved by identifying which parts of the information should be prioritized in data treatment. The principal advantage of this would be optimizing the use of computational resources needed to deal with visual data, mainly by allocating the most for those parts of the image that require more detail in line with the perception of the viewer [62]. Similarly, simulating visual perception for a synthetic human character and a video surveillance application [13], improving a web usage mining methodology for finding the most important objects in a web page for helping the designers in the website creation [76], or improving the accuracy of gaze tracking systems in the context of interactive 3D applications [27], are just another examples of the saliency utility.

In addition, and as a result of this intense activity, a new trend has also emerged introducing saliency models as the ground for novel paradigms empowering traditional frameworks such as image retrieval [4, 78], object recognition [22] or activity recognition [18, 60, 77]. Particularly, it has been demonstrated that features derived from visual saliency models can be useful at other applications such as multi-class automatic video classification tasks. For example, salient regions have proven to be successful for deriving global descriptors from which to perform the classification process of a collection of 924 video clips showing 7 different kinds of sports [67]. Monument recognition models have been also implemented based on GBVS saliency maps [38]. In this case, the matching process done for new images of monuments taken from different angles or zooms can be improved, both in time and accuracy, by using local visual features, such as Scale Invariant Feature Transform (SIFT) or Speeded Up Robust Features (SURF), specifically extracted from salient regions according to the GBVS maps. Finally, feature coding based on saliency detection has also demonstrated its effectiveness for image classification in elevator videos as well [51] (i.e. overload or violence detection).

2.4 Aesthetics assessment

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 Savakis et al. fifteen years ago [68]. 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 with the conclusion that image appeal had to be addressed through metrics others than those used for measuring image quality. More recently, Datta et al. [15] proposed 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. Khan and Vogel [39] carried out a higher-level analysis to assess the aesthetic quality of photographs, Marchesotti et al. [55] extended the study in 2011 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 [56]. 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 [15] and extended to the temporal dimension, obtaining an accuracy of 73%. Yang et al. [82] 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 [3] proposed a model with features based on psycho-visual statistics. Furthermore, Fernández-Martínez et al. [20] proposed some new features at the video-level based on cinematographic and photographic notions and a model which automatically annotates the videos 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 [36], in which a wide range of multimodal features is proposed and [19] which offers a comparative study of the performance between visual and acoustic features.

2.5 Motivations and proposal

Visual saliency analysis of a video serves for identifying where the areas of interest are located. Then, the observed saliency information allows to infer relevant visual cues on how viewers are perceiving the video. Particularly, salience features, extracted from those areas prioritized by visual attention, may be related to their movement, shape, size and other characteristics that could potentially affect the perception of the viewer. Hence, we will test whether saliency descriptors can be good indicators of the video impact (i.e. positive or negative) upon viewers and, in turn, whether they can help to refine and improve the classification results previously obtained by other methods such as our baseline, purely based on aesthetics [20].

To the best of our knowledge, this is the very first time that a computational model targets visual saliency as a successful exploit for the task of classifying commercials based upon their impact on viewers. Most similar work was performed by [52] who proposed an approach to advertisement evaluation using salient regions based on foveated imaging. Nonetheless, the work was strictly focused on static images and its evaluation was primarily focused on validating the saliency estimation process rather than actually assessing the advertisement impact.

3 Visual saliency features

The feature extraction process for a video frame begins with a pre-processing step where the black bars at the borders of the video frames are first removed. Then the process continues by computing its corresponding saliency map.

For that purpose we have initially decided to use the GBVS[1] framework [26]. This saliency detection algorithm has been considered one of the top-performing ones in major benchmarks [6], although it has already been surpassed by current DNN-based state of the art algorithms. In this regard, it is important to clarify that the goal of the present work is not analyzing what detection algorithm yields best results but simply applying a good-on-average algorithm (such as GBVS) correctly tuned for a reasonably good (though not necessarily perfect, as we will discuss later) operating point, to demonstrate that our novel application of saliency is sound and adequate. Nonetheless, and to confirm that the suggested approach exhibits good generalization capabilities when using other methods, alternatively, we will also evaluate the use of a current state-of-the-art DNN-based visual saliency detection framework. Particularly, we will make use of the deep contrast convolutional neural network (CNN) proposed by Li and Yu [42].

Once the map is available, a number of different salience features can be extracted from it. In this regard, it is important to point out that computing a saliency map for every frame in a video can be a really demanding process. This could affect the choice of the visual saliency algorithm in an attempt to better balance the trade-off between quality and computational cost [28].

In this work, cost can be reasonably found to be significant given that we process all the constituent frames of a video. However, although this strategy could be surely optimized, for instance by reducing the amount of frames to work with by exploiting the high temporal redundancy in videos or by testing other different and more efficient saliency detection algorithms, lessening this cost was out of scope for our research which main goal was, instead, to measure visual saliency in videos and determine its effect on viewers.

3.1 Saliency blobs segmentation and extraction

In the original normalized saliency map obtained for each video frame there exists a set of high saliency portions in a low-saliency background. The feature extraction process targets these high saliency portions or saliency blobs, which are considered the informative parts. Hence, in order to locate them and separate them from their background we need to perform a segmentation of the saliency map. Particularly, the saliency values are simply thresholded yielding only two types of regions (i.e. salient or non-salient).

Although more sophisticated segmentation algorithms are reported in the literature, such as the iterative fitting method proposed by [12] or the mean-shift method proposed by [1], saliency map thresholding is surely the simplest and most extensively used way to get a binarized version of a saliency map with the segmentation of salient objects. This method typically implies evaluating different threshold values so that precision and recall curves (PR curves) can be used for quantitative evaluation. Particularly, the threshold is varied to reliably compare how well various saliency detection methods highlight salient regions in images. Depending on the level of contrast desired, such kind of evaluation usually requires

[1] An official Matlab implementation of the algorithm is freely available for education purposes at: http://www.klab.caltech.edu/~harel/share/gbvs.php.

large datasets with marked bounding boxes delimiting the salient regions [50], or with even finer resolution by accurately labelling pixels instead [12]. Typical average precision, recall, and F-Measure (a weighted harmonic mean of precision and recall), or other recently emerged performance measures such as Pos@Top (a robust and parameter-free alternative proposed by Liang et al. [45, 47, 48] and particularly suitable for evaluating the performance of image retrieval systems), among others, can be finally compared against the entire ground-truth database, thus helping us to find the optimal solution.

However, although of great importance, evaluating or benchmarking different saliency detection methods, including the adopted ones (GBVS and the deep contrast CNN), has already been extensively studied in many other publications and is beyond the scope of the current work. Hence, instead of relying on extensive threshold sampling, particularly relevant when comparing the overall performance of different algorithms, we have rather defined a fixed reliable threshold value under several practical assumptions.

First, the proposed approach simply aims at measuring and characterizing attention in scenes without the need for accurately recognizing shapes and figures. In this regard, our novel application of saliency differs, for example, from typical object detection and recognition tasks, which usually requires significantly higher accuracies. Hence, although it certainly relies on saliency map estimation and segmentation, the required performance for salient regions detection does not need to be perfect for the approach to reasonably succeed (as it will be confirmed in the experimental section).

Additionally, and aside from ranking well, the salient regions detection algorithm needs determining an adequate operating point which strongly depends upon the specific application and which often happens to be ignored by previously mentioned benchmarking tests. The operating point is basically determined by the adopted segmentation threshold, which value yields a particular combination of precision and recall rates. This threshold offers a way to control the existing trade-off between precision and recall. Particularly, a higher threshold will typically mean better precision and worse recall than a low threshold.

As an example of the importance of selecting an adequate threshold, work by [1] remarks that, in spite of a very poor recall, algorithms yielding high precision results may be better suited for some particular applications (e.g. gaze-tracking experiments), than others (e.g. perhaps not for salient object segmentation). In this regard, and also as discussed by [50], recall rate is not as important as precision for attention detection (recall favours attention regions to be as large as possible, for example, a 100% recall rate can be achieved by simply selecting the whole image). The real challenge for attention detection is commonly referred to as locating the position of a salient object as accurately as possible, i.e. with high precision. Therefore, and as suggested by many works [7], the evaluation measures (e.g. F-measure) are often accordingly set to raise more importance to precision.

Finally, it is also important to highlight the fact that, in view of typical PR curves obtained for top-performance salient regions detection algorithms [12], threshold values in the range between 0.7 and 0.8 mostly provides a reasonably good operating point, successfully meeting the presented performance criteria of maximizing precision on the expense of recall, but, at the same time, preserving a reasonable recall level.

Hence, based upon the above-mentioned considerations, the threshold has been determined to 0.75 (maps are normalized between 0 and 1, hence the adoption of such threshold implies that only particularly salient regions are considered). The adoption of such threshold has also been validated by means of a qualitative analysis of the saliency maps obtained for several representative videos of our dataset (which is not labelled with

saliency information). As a result of this analysis, we have found the observed saliency detection results to be satisfactory and the adopted threshold to be not too sensitive to small variations (it can be varied by 10% of its value without significantly affecting the segmentation results).

Once the segmentation has been performed, in the next step the blobs are extracted by applying typical connected component labeling to the resulting binarized map [25] (some typical map examples of the two video classes, "positive" and "negative", are presented in Fig. 5). Features are then to be calculated from the identified blobs in every video frame by measuring different properties about their size, shape, movement, and others.

It is important to point out that segmentation typically produces several fragmented salient regions. In this regard, two different types of descriptors will be extracted from every frame. First type will be referred to the whole set of blobs by defining a global and unique measurement of overall saliency computed over all the blobs (e.g. overall salient perimeter computed as the sum of the perimeters of all the blobs). For the second type, the blobs will be sorted upon their corresponding area size so that the descriptors will only be computed over the biggest blob of the frame (i.e. the biggest salient area). Figure 2 shows the result of the blob segmentation and extraction process for two different frame samples. The whole set of features proposed in this project is described below.

3.2 Frame-level saliency features

In this subsection we will describe the set of visual saliency based descriptors that we will use for constructing our impact prediction model. These video descriptors are to be calculated from the identified blobs in every video frame by measuring different properties. All the proposed descriptors are inspired and supported either by cognitive psychology [70] or well-known photographic composition rules [53]. Particularly, descriptors have been organized into seven distinct families according to their nature, namely population, shape, geometry, orientation, location, movement and photographic composition related features.

3.2.1 Population features

This subset of features is basically related to the amount of saliency blobs extracted from the saliency maps. By measuring the number of blobs or attention spots we can categorize the visual attention either as *focused* or as *divided* [70]. Concentration on one spot to the exclusion of any other is known as focused attention. On the contrary, in cases of divided attention, more than one source is attended. Since our attentional resources are limited,

(a) Single blob sample. (b) Multiple blobs sample.

Fig. 2 The image to the left concentrates attention on one single blob, while the image to the right does on two

divided attention typically results in a drain on our overall attentional capacity, thus causing an information loss due to the increased mental effort. Hence, we expect such mental or attentional effort to be indicative of the viewers' perception. Presumably, the more attention spots, the more the effort required, thus possibly exceeding the available capacity of the viewers and, therefore, producing a failed (or worse) perception.

Specifically, we have defined the following features:

- **num-blobs**: the number of saliency blobs extracted from each video frame.
- **no-blob-frames-percent**: percentage of frames with no blobs at all.
- **focused-attention-frames-percent**: percentage of frames whose biggest salient area (normalized by the frame area) is lower than a defined threshold (i.e. 3%).

Please note that percentage features are already defined at the video level. Hence, no pooling technique will apply in both cases, as it will be presented in Section 3.3.

3.2.2 Size features

Among the various factors influencing the saliency of a region, works on visual saliency often rely on two: its size and location. Intuitively, and consistent with human visual perception, larger image regions closer to the image center are more salient. Hence, different methods [86] have been proposed by defining the saliency of a region as the product of its size and centerness. In this regard, this work also tries to address these two issues by introducing and computing additional descriptors linked to such concepts.

Particularly, the following size measurements will be part of our model:

- **blob-area**: a scalar that specifies the number of pixels that define the entire blob. This feature has been normalized by the frame area (i.e. $W \times H$, W being the frame width, and H being the frame height, both measured in pixels).
- **blob-perimeter**: a scalar that specifies the number of pixels which defines the boundary that encloses the blob. Since the blobs are continuous regions, a boundary can be defined for each blob with no exceptions. This feature has been normalized by the frame perimeter (i.e. $2W + 2H$).
- **blob-ellipse-area**: a scalar that specifies the number of pixels that define the ellipse that has the same normalized second central moments as the blob region. This feature has been normalized by the frame area.
- **blob-major-axis-length**: a scalar that specifies the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the blob region. This feature has been normalized by the longest line segment that could be described within the frame (i.e. the diagonal of the frame, $\sqrt{W^2 + H^2}$).
- **blob-minor-axis-length**: same as previous one but for the minor axis.
- **blob-extent**: a scalar that specifies the ratio of pixels in the blob region to pixels in the corresponding bounding box (i.e. the smallest rectangle containing the whole blob). It is computed as the blob area divided by the area of the bounding box.

3.2.3 Location features

Previously, we have highlighted the importance of size and location for a region to be salient. Some size related descriptors have already been introduced. In this case, we also propose to improve our saliency based model by adding an extra set of location cues. Recent works have shown that subjects tend to fixate the screen center when watching natural scenes [73]

or natural edited videos [16]. These are the different features we propose for modeling such behavior:

- **blob-centroids**: returns the 2-D coordinates that specify the center of mass of the blob. These coordinates have been normalized by the frame width (i.e. W) and the frame height (i.e. H), respectively.
- **blob-extrema**: this set of features returns an 8-by-2 matrix that specifies the extrema points in the blob region, detailed in Fig. 3. Each row of the matrix contains the x- and y-coordinates of one of the points. Same normalization applies for these coordinates as well.

3.2.4 Geometrical features

Aside from size and location, there are many other factors that contribute to saliency. As an example, different feature maps can be defined either at the pixel (e.g. color, intensity, and orientation [33]), object, or semantic level (e.g. a face tends to attract attention more than other objects [11]). Among these, object-level information has a significant importance in the prediction of visual attention.

Usually, object categories are added into the saliency models to improve the prediction of attentional selection [37]. However, despite improving the performance, having an object detector for each individual possible object does not seem plausible. As an interesting alternative, recent works have proposed an attribute-based framework where each attribute captures inherent object information that is important to saliency [81]. Inspired by these approaches we will encode some additional visual cues by handling saliency blobs as objects from which to calculate a set of similar local shape attributes. These proposed attributes have already shown to be strongly correlated with attention selection. For example, solidity is an important object-level attribute that describes the shape of the objects, and objects with low solidity values may indicate occluded objects. These are the features we are extracting from blob contour segmentation:

- **blob-complexity**: the complexity of a particular blob is denoted as the ratio between its perimeter and area. With the area of the blob fixed, the complexity is higher if the contour is longer. A circle has minimum complexity.
- **blob-solidity**: formally, the blob solidity is a scalar specifying the proportion of the pixels in its convex hull (i.e. the smallest convex polygon that can contain the blob) that are also in the blob. The solidity attribute is intuitively similar to the typical convexity measure (referred to perimeters instead), but it also measures holes in blobs. Hence, if a blob is convex and without holes in it, it has a solidity value of 1.

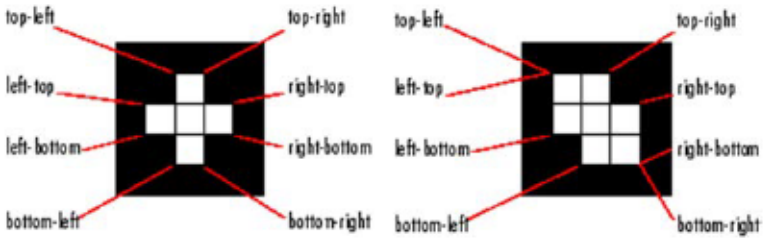


Fig. 3 Extrema feature coordinates. This figure illustrates the extrema of two different regions. In the region on the left, each extrema point is distinct. In the region on the right, certain extrema points (e.g., top-left and left-top) are identical. Source: http://es.mathworks.com/help/images/ref/regionprops.html

- **blob-eccentricity**: the blob eccentricity is represented by the eccentricity value of an ellipse that has the same second-moments as the blob region. The eccentricity value is computed as the ratio of the distance between the foci of the ellipse and its major axis length. An ellipse whose eccentricity is 0 is a circle, while an ellipse whose eccentricity is 1 is a line segment.
- **blob-circularity**: this feature measures how similar to a circle is the blob. It is computed as the ratio between the area of the blob and the area of a circle with the same perimeter.

3.2.5 Orientation features

Orientation is known to be one of the relevant properties of a scene that determines where to look [61]. Furthermore, it is one of the three more commonly used biologically plausible pixel-level attributes (i.e. together with color and intensity [33]) in saliency modeling. Either intentionally or not, the orientation of the blobs may initiate a reflexive shift of attention to a peripheral location [21]. Therefore, we have decided to explicitly model the following properties of the blobs that could direct our attention towards a particular direction:

- **blob-orientation**: a scalar that specifies the angle between the x-axis of the video frame and the major axis of the ellipse that has the same second-moments as the blob. The value is in radians, ranging from 0 to π (i.e. 0 to 180 degrees).
- **blob-orientation-bin**: orientations are quantized into four bins thus creating an orientation-based histogram. The histogram channels (i.e. $N = 4$) are evenly spread over 0 to 180 degrees (thus resulting into horizontal, vertical and the two diagonal orientations). The value of the feature is assigned the bin index k that corresponds to the closest orientation center given by $\theta_k = k \times \frac{\pi}{4}$, where $k = 0, 1, \ldots, N-1$.

3.2.6 Motion features

The HVS is highly sensitive to the relative motion. For instance, moving objects in a very cluttered scene are still able to attract our gaze very effectively, as shown by [10], where motion contrast accounts for most of the fixations.

By applying a spatio-temporal saliency model in our impact assessment framework we are assuming the temporal dimension of videos, their salient areas and their evolution, to be relevant in terms of their impact upon viewers. These are the features we propose for our impact model to also explicitly account for dynamic behavior:

- **speed**: our model incorporates the motion information by analyzing the magnitudes of the biggest blob motion in horizontal and vertical directions. As this feature is calculated by comparing the coordinates of the biggest blob in two consecutive frames, biggest blobs correspondence between frames is assumed to be maintained through spatio-temporal continuity.
- **acceleration**: horizontal and vertical acceleration values are also calculated for the biggest blob by comparing the corresponding speed values in two consecutive frames. Same assumption applies with regards to biggest blob continuity between frames.

3.2.7 Rule of thirds features

The rule of thirds (ROT) is one of the most important composition rules used by photographers to create aesthetically appealing photos. ROT states that placing important objects along the imaginary thirds lines or around their intersections often produces highly aesthetic

photos. Detecting ROT from a photo or an image may require complex semantic content understanding to locate important objects. ROT features were already successfully exploited as part of an impact assessment model based on aesthetics in [20]. Alternatively, this work relies on the observed visual saliency to introduce a novel interpretation of such aesthetically inspired features that specifically targets the composition of the important content (i.e. the detected salient regions in the video sequence).

Saliency analysis has already served the purpose of detecting ROT in photography [53]. Similarly, our method will approximate the important objects in the videos as the segmented and extracted blobs, and their location as their corresponding centroids. Therefore, our impact model will be improved by introducing a method to automatically determine whether the blobs follow the ROT principles and accordingly design a range of related features, under the assumption that the better the composition of salient regions, the better the impact upon viewers.

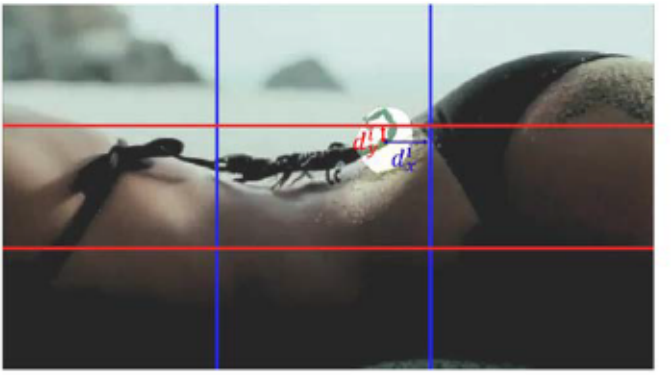
- **ROT-distance**: distances between the blob centroid coordinates and their nearest imaginary third lines are calculated as a measure of the degree of utilization of ROT (i.e. the shorter the distance, the higher the degree). These distances, the horizontal distance for the x-axis and the vertical distance for the y-axis, are normalized by the frame width (i.e. W) and the frame height (i.e. H), respectively. Fig. 4a illustrates this feature calculation.
- **ROT-score**: horizontal and vertical ROT scores are also calculated. Particularly, we split every frame into a 12×12 grid mesh such that it aligns well with the third lines and blob centroids can be assigned a score depending on their distance to third lines. Since every horizontal or vertical third is divided into 4 segments, distances are quantized into 4 discrete categories by assigning a 1-to-4 score (i.e. 4 when minimum distance to the nearest third line, 1 when maximum). Figure 4b illustrates our centroid location map and their corresponding horizontal and vertical ROT scores.

3.3 Feature pooling techniques for video-level features generation

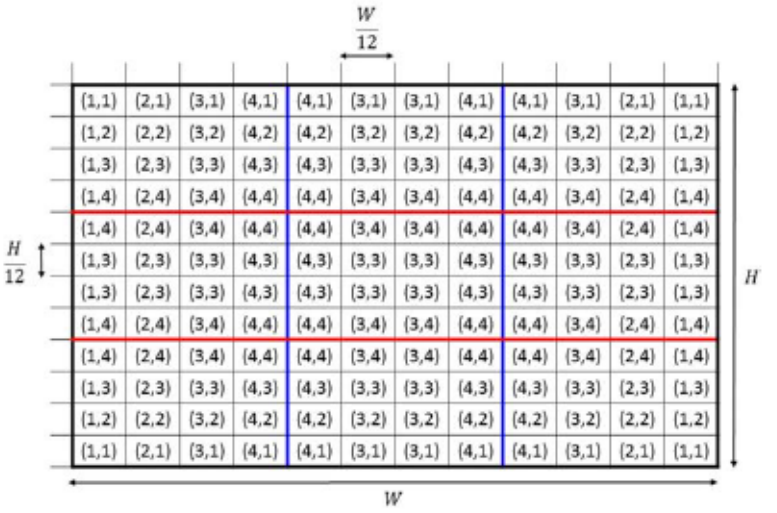
Given an input video, the 7 different types of frame-level visual features previously presented are computed on every video frame. Then, during a post-processing stage, these frame-level features are extended to the temporal dimension by aggregating and combining them to yield features at the video level. The procedures that we have used to pool the features at such level are very simple:

- Average (labeled as AVG), computed as the mean of the features across all the video frames;
- Standard deviation (labeled as STD), again computed across all the video frames;
- Median (labeled as MED), computed as the median of the features across all the video frames;
- The mode value (labeled as MOD), computed for the discretized attributes as the most frequently occurring value among the observed entries along the video sequence. In this regard it is important to mention that only some features have been pooled like this (i.e. horizontal and vertical ROT scores, the number of blobs, and the orientation bins).

The combination of the two different types of descriptors introduced in Section 3.1, namely: overall and biggest blob descriptors, together with the seven different families proposed, plus the four different pooling strategies yields a final set of 452 video-level features in total.



(a) ROT distances for blob $i \equiv (d_x^i, d_y^i)$



(b) ROT scores for blob $i \equiv (score_x^i, score_y^i)$

Fig. 4 The upper image is a sample frame with the distances between the blob centroid and the nearest horizontal and vertical third lines. The bottom image shows the quantization function or map used to assign horizontal and vertical scores to the blobs given their centroid location

4 Results and discussion

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 approach.

4.1 Experimental data

For experimental purposes we have adopted the same annotated set of videos presented in [20]. This dataset consists of 138 car commercials collected from YouTube and annotated as either *positive* or *negative* to closely relate to how viewers actually perceive each

(a) Positive samples.

(b) Negative samples.

Fig. 5 Some representative frames of samples of both classes showing their corresponding segmented salient regions overlaid

video.[2] Figure 5 shows some representative frames, together with their segmented salient regions overlaid, corresponding to samples of both classes.

4.2 Experimental setup

Our primary goal is demonstrating that saliency based features may be indicative of the impact of the video on viewers. For that purpose, we have evaluated the classification

[2]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.

performance (i.e. binary classification with the two classes previously introduced) of different impact models built from different numbers of saliency based features (i.e. different dimensionality). In this regard, a feature selection technique has been used to select the subset of the most relevant features for use in the model construction.

Besides, we have adopted the aesthetics prediction model based on low-level visual descriptors presented in [20] as our baseline model. Hence, our novel model based on visual saliency has been compared to this baseline. Both approaches modeling the impact of videos upon viewers, the baseline based on aesthetics and the current one based on visual attention, have been also combined together resulting in a third model or strategy that is also analyzed.

In order to confirm the significance of the proposed features and methods we have also decided to reevaluate our proposal on the basis of a completely different visual saliency detection framework: a state-of-the-art deep learning based method.

Moreover, the seven families of descriptors defined in Section 3.2 have been also compared in terms of classification accuracy. In order to make fair comparisons between the different families, we have performed identical classification experiments on each of the seven feature subsets corresponding to the different families.

Finally, a similar study has also been performed targeting the four distinct pooling techniques used in the video-level feature computation.

4.2.1 Baseline model

This model is based on a family of descriptors according to different visual aspects which proved to be suitable for automatically predicting aesthetics in our video dataset, namely video colorfulness (based on CIE Lab color histograms), descriptors related to the rule of thirds (portions above and below the horizontal imaginary third lines are compared as a measure of the degree of utilization of ROT), and typical intensity and entropy-based descriptors. Some temporal segmentation descriptors are also extracted (by detecting the cuts location in the videos). This baseline model relies on 21 different features in total, which we will simply refer to as *aesthetics features* in this paper.

4.2.2 Feature selection

To evaluate the performance of a model for a particular dimensionality we have carried out a feature selection analysis so that we can evaluate the worth of every feature, rank them, and select the subset of the best ones fitting the specified dimensionality (and providing best 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 [24]. This tool provides a set of feature selection algorithms, from which we have picked *SVMAttributeEval with Ranker* [23]. This algorithm evaluates the worth of every feature by using an SVM classifier and ranks them by the square of the weight assigned by the SVM. Once features have been ranked we can simply indicate the number of them to select.

4.2.3 Classification

For classification, as a necessary reference for a valid comparison with the common baseline approach purely based on aesthetics, we have used a Logistic Regression (LR) model with ridge estimator, based on the well-known method of le Cessie and van Houwelingen [41]. In addition, reference results obtained with LR models have been also compared with

other popular machine learning models, such as Support Vector Machines (SVM), where we have used a support vector classifier with polynomial kernels based on John C. Platt's sequential minimal optimization (SMO) algorithm [64], or standard probabilistic Naïve Bayes (NB) classifier. All the classifiers have been tested using the implementation of the WEKA machine learning software.

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 averaging 10 random repetitions of a 10-fold cross-validation (10 × 10-fold CV) on every data set.

4.3 Comparison between strategies: aesthetic versus attentional information

As previously introduced, our experiments covered the comparison between three different models or strategies in terms of the descriptors used to build them, namely: based on aesthetics features (the baseline), based on visual attention features (in this case derived from GBVS maps; we will specifically address the comparison between GBVS and CNN results in the subsequent section), and their combination.

Figure 6 presents the classification accuracy resulting for each strategy using different numbers of descriptors. Specifically, we have configured the feature (descriptor) selection algorithm to provide reduced subsets of top features ranking from 1 up to 50 descriptors (worse results were systematically observed beyond this point). Please, note that no further subsets can be selected for more than 21 descriptors in the case of the aesthetics model. Also note that, in the case of the combined strategy, feature selection is applied to the augmented

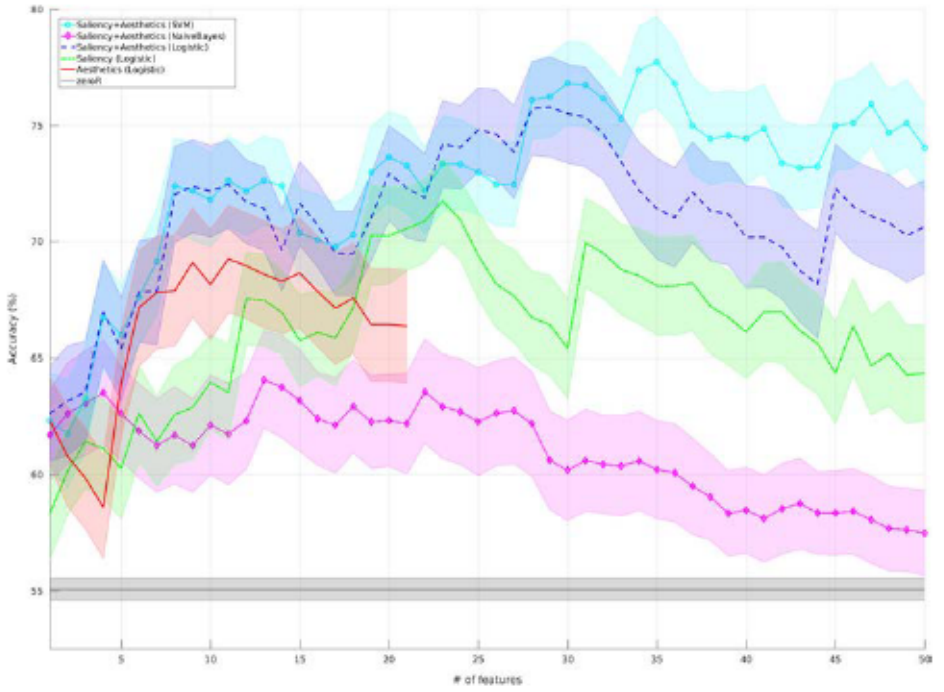


Fig. 6 Classification accuracy for different dimensionalities including 95% confidence intervals

set of features resulting from the aggregation of all the aesthetics and saliency descriptors together (an early fusion scheme is adopted so that the descriptors are combined in the feature space before the feature selection and classification).

Besides the accuracy, approximate 95% confidence intervals are also included for reliable accuracy assessment. Similarly, we have also considered the ZeroR classifier (i.e. a classifier that predicts the majority class), which is commonly used for determining baseline accuracies (i.e. 55.05% in our experiments) and serves well as another reference performance.

As we can observe, evaluation results have shown that the proposed saliency based features are indicative of the impact of the videos on viewers (the saliency based approach has clearly outperformed zeroR). Particularly, best classification accuracy achieved is 71.74% when using the top 23 saliency descriptors.

If compared to the baseline aesthetics model, the latter has a top performance of 69.24% achieved when using the top 11 aesthetics descriptors (also clearly outperforming zeroR). Observed difference between both top results is found to be non relevant (confidence intervals are overlapped) although the aesthetics based model seems to surpass the saliency one in configurations with fewer descriptors (i.e. for subsets ranging from 6 to 9 and also for 11 top descriptors).

Once it has been demonstrated that both of them, aesthetics and visual attention measurements, can be separately used to successfully model the viewers' perception (i.e. distinguishing between 2 impact or satisfaction levels: videos perceived as good or bad), the further combination of both should be performed as part of the construction of a more complete and effective inference model. In this regard, effectiveness has shown to be better as it can be confirmed straightforward from the best performance achieved by the joint use of both types, which yields an accuracy of 75.79% for the LR based approach when using the top 29 descriptors (i.e. a mix of aesthetics and saliency features).

The SVM classifier further improved this result to 77.72% using the top 35 descriptors, roughly the sum of both previously reported top saliency and aesthetics descriptors subsets, thus making the best of their combination. SVM are particularly effective in high dimensional spaces (as it can be observed in Fig. 6 SVM performs particularly better than LR for bigger subsets). In addition, LR converges to any decision boundary that can divide the training samples into positive and negative classes, whereas SVM objective causes the decision boundary to lie (geometrically) mid-way between the support vectors which usually means better generalization.

With regards to the other evaluated classifier, NB has consistently shown worse performance, thus suggesting that the independence assumption between features is general not true when modeling the video impact.

Finally, when compared to the top performance achieved by the sole use of saliency or aesthetics features, the combined result is found to be significantly better (confidence intervals do not overlap), thus confirming the synergy and complementarity of the two considered approaches.

4.4 Comparison between saliency estimation methods: GBVS versus DNN-based

Although the experimental results presented in the last section have demonstrated that visual saliency can effectively help to improve the accuracy of predicting the impact of a video upon its viewers, we have also tried to find out whether these results could be further improved by adopting top-performance saliency detection methods such as current state-of-the-art DNN based approaches.

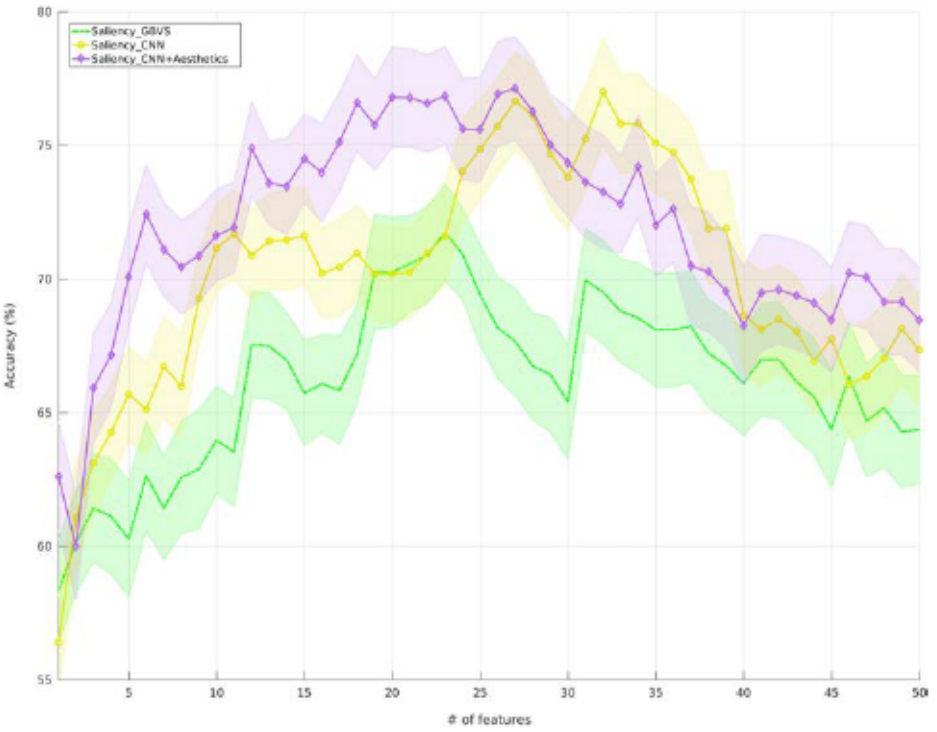


Fig. 7 Classification accuracy for different saliency detection methods including 95% confidence intervals

In this regard, we have adopted the aforementioned CNN proposed by Li and Yu [42] as our deep learning solution for visual saliency detection. Specifically, instead of training a CNN from scratch on our small video dataset (which could lead to a model over adjusted to the peculiarities of the car commercials), we have rather preferred to directly use their CNN as a generic large-scale deep learning model, trained on a large-scale dataset and publicly available for download.[3] This CNN has been run on our video frames to produce their corresponding normalized saliency maps. These maps have been then post-processed by applying exactly the same blob segmentation and extraction procedure as with the GBVS maps. Finally, the same the set of visual saliency based descriptors have been calculated.

Figure 7 presents a series plot of results covering the comparison between three different models in terms of the saliency detection method used, namely: based on salience descriptors computed from GBVS maps, based on salience descriptors computed from state-of-the-art CNN maps, and finally a third alternative resulting from the combination of CNN-based salience descriptors together with aesthetics descriptors (our proposal). Specifically, we have adopted the same experimental setup as in the previous section: first, feature selection to retrieve the top-rank feature subsets from 1 up to 50 descriptors, and then classification using our reference LR classifier and 10×10-fold cross-validation.

As we can observe, thanks to the improved spatial coherence and contour localization, the experimental results have demonstrated that the salience descriptors derived from the

[3] https://drive.google.com/file/d/0BxNhBO0S5JCRbUt0NHBtQWtZb2c/view.

CNN model (i.e. Saliency_CNN in the Figure) are more powerful than those derived from the baseline GBVS model (i.e. Saliency_GBVS). Particularly, the deep CNN model has achieved a top performance of 77% when using the top 32 salience descriptors, clearly outperforming the best result obtained with the GBVS method, only 71.74% when using the top 23 descriptors, as we already presented in the previous section.

Finally, once again, we have also evaluated the combination of these CNN-based salience descriptors together with the aesthetics descriptors that were previously introduced as our baseline. In this case, the best accuracy found has been 77.12% when using the top 27 subset, a small improvement over the sole use of CNN-based salience descriptors but again outperforming the best resulting accuracy achieved so far with the GBVS alternative (the joint use of both aesthetics and GBVS-based saliency features yielded a top performance of 75.79% when using the top 29 subset, as it was presented in Section 4.3). Our results also show that the impact of combining both is particularly evident when making use of more simple and compact models (i.e. less parameters or descriptors), which have been able to attain similar levels of accuracy as seen with the top-performing one.

These results also confirm that our approach has consistently demonstrated for both detection methods, GBVS and DNN based, that visual saliency can be used as a successful exploit for the task of classifying videos based upon their impact on viewers. As could be expected, the latter has yielded better and particularly encouraging results since, as we have mentioned earlier, we are simply reusing a large-scale generic CNN model without any change nor adaptation to our specific problem. This suggests the interesting possibility to explore using this pre-trained model to bootstrap a better adapted saliency model out of our very little data. Furthermore, no optimization has been performed regarding the operating point of the CNN model (i.e. we have used exactly the same segmentation threshold as with the GBVS approach). Hence, a better setup could be specifically tuned up for it, thus supporting further improvements.

4.5 Comparison between families of saliency features

Once we have confirmed the validity of the suggested approach, it is also interesting to gain further insight into which features are the most valuable. Therefore, we have also performed an additional analysis targeting the comparison between the different families of saliency based descriptors proposed.

In order to simplify the analysis we have adopted a similar set-up as before: first, feature selection to retrieve the top descriptors for each family, and then classification with 10×10-fold cross-validation. The only difference is that, instead of presenting another series plot of results for different families, numbers of features, and classifiers, we have decided to simplify the analysis by only considering our baseline GBVS based approach, our reference LR classifier, and making use of a different type of graph to better illustrate the strength (or weakness) of each individual family: a start chart (also known as spider chart [72]).

Figure 8 presents such chart where each spoke represents the top performance achieved by one of the defined families (concentric grid lines have been included to help visually comparing the different spokes). Families wind counterclockwise around the chart in accordance with the introduction order. The number of features required by families to achieve such results varies in a range from 5 to 15. ZeroR results have also been included (i.e. inner heptagon) together with 95% confidence intervals for true accuracy assessment.

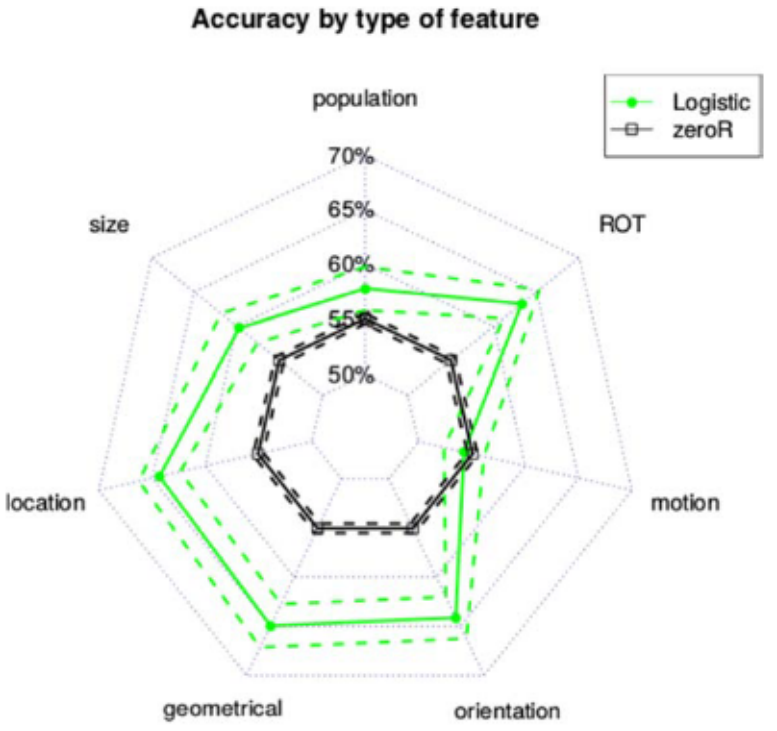


Fig. 8 Top classification accuracy distribution for each specific family of features including 95% confidence intervals

If we observe the figure we can first mark geometrical, orientation and location families as the best ones, closely followed by ROT. Statistical evidence does not suffice for performance to be considered different among these top families, which can be interpreted as all being similarly useful to model the impact on viewers.

Next level down would include size and population families, still above the zeroR level, thus demonstrating to be also reasonably good indicators.

Finally, worst performance has been obtained when using only motion features, which have failed to act as helpfully as we could have expected, since the HVS is known to be highly sensitive to motion. According to displayed results, the performance of motion features is even worse than chance. Hence, this reasonably indicates the need for further research on exploring how to consider motion and include related dynamic features. Nonetheless, it may be also concluded that most of the different types of features tested have attained notable success, complementing each other reasonably well as can be derived from the significantly better result obtained when combined together (i.e. 71.74%, as presented in the previous section).

4.6 Comparison between pooling strategies

Finally, a similar analysis to that presented per family in the previous section has also been carried out but this time grouping the features in accordance with the pooling technique

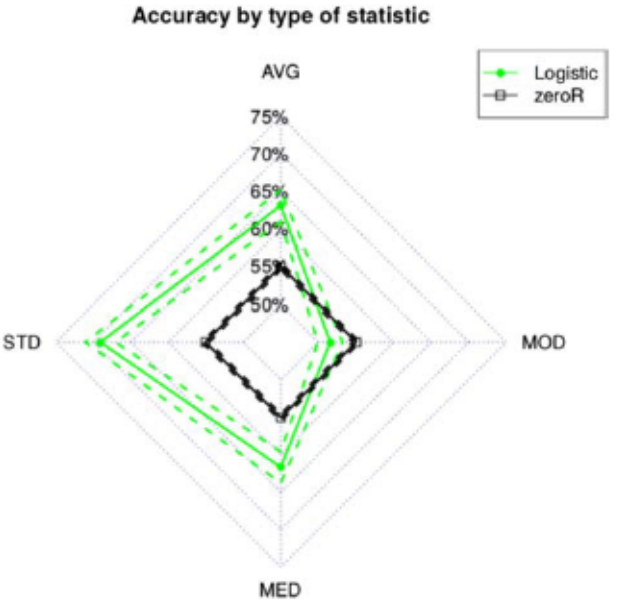


Fig. 9 Top classification accuracy distribution for each specific polling technique including 95% confidence intervals

used. Same experimental set-up has been adopted to find out which pooling technique achieves best result individually.

Figure 9 shows the corresponding start chart where STD clearly emerges as the top-performing polling technique. On the contrary, MOD turns to be the worst choice since its performance happens to fall below our zeroR reference. In between, we find AVG and MED, which do not differ significantly but both clearly succeed in outperforming the baseline, thus demonstrating their validity.

AVG and MED are different measures of central tendency: both try to capture the dominant value for each feature throughout the whole set of video frames and smooth out variation. Despite the well-known difference between them (i.e. MED is considered to be more appropriate when dealing with skewed data), both yield similar results thus suggesting little or no skewness in our data distributions so that resulting means and medians are closely similar.

MOD is another method occasionally used to find a typical value from a set of data. However, in this case performance is clearly conditioned by the reduced size of this feature subset: only ROT scores (horizontal and vertical), number of blobs, and orientation bins are utilized. Hence, to ensure a fair comparison between this polling alternative and the rest, mode values should be estimated for the same features as for the rest of the techniques.

STD conveys information about the feature variability, which seems to be even more informative than central tendencies when modeling the viewer perception of a video.

5 Conclusions and future work

In this paper we have presented a hybrid computational method for predicting the impact of 138 car advertisements videos on viewers by using both conveyed aesthetic and

attentional information. Widely used, but limited, low-level aesthetics descriptors are enhanced and complemented by novel visual saliency map based descriptors which reflect high-level semantics objectively and help improving the video impact categorization results.

Suggested set of visual saliency descriptors are inspired by the mechanisms that underlie the attentional abilities of the human visual system, and have demonstrated to be indicative of the measured impact upon viewers. Different classification experiments have been performed for the suggested features to be tested and validated.

First, our model based on features derived from GBVS maps has achieved an accuracy of 71.74% in distinguishing between positively and negatively perceived videos. This result is very similar to the performance produced by a set of low level features specifically defined and implemented for aesthetics in [19]. On the other hand, the combination of both models together, saliency and aesthetics, has yielded a top classification accuracy of 75.79%, which confirms that a more complete and effective visual model of how videos are perceived can be constructed from features modeling not only their aesthetic value, but also the mechanisms of human attention (i.e. visual saliency). This result has been confirmed and further improved to 77.12% when relying on a current state-of-the-art DNN based approach for visual saliency detection, which in turn has suggested the need for deeper analysis of similar approaches to be explored in the future. Particularly, such analysis should focus not only on performance and classification accuracy but also on efficiency as another major aspect for the feasibility and practical implementation of the suggested approach.

From a different perspective, saliency determines which part of the visual scene has to be processed and which ones will be discarded. Hence, saliency maps could be exploited to constraint the extraction of the aesthetic descriptors to only those particularly salient parts, which could be considered as a refined (or weighted by saliency) version of our initial aesthetics model.

Additionally, it would be also worth exploring the combination with new features. In this regard, it would be particularly interesting to work towards adopting a general attentional approach by extending visual to aural saliency and incorporating auditory scene analysis. Finally, future research should be extended to different video domains mainly to test whether the obtained results could be generalized and scaled to different scenarios.

Second, most of the families of salience descriptors suggested have been found to be suitable for automatically predicting the impact upon viewers with reasonable success. Particularly, geometrical or local shape attributes, well-known attention selection indicators, have also shown good performance at distinguishing between good and bad commercials. Similarly, location or orientation cues, extensively used in saliency modeling, now demonstrate their importance in modeling viewers' perception. With regards to composition rules, again ROT measures succeed as in previous studies, but now under the novel perspective of attentional selection. On the contrary, it is also remarkable that the subset of motion features proposed has failed. Related dynamic features have been included by temporally extending the static image saliency framework to consider motion. However, the poor results obtained in this case suggest the approach to be inadequate. In future work, we plan to consider alternative ways of accounting for the dynamic behavior of saliency, such as using a explicit video saliency estimation method or identifying camera motions (e.g. pan or zoom) which are known to affect saliency.

Third, standard pooling strategies have been successfully applied to collapse frame-level values into video-level descriptors. In this regard, the different degrees of variability captured by STD measures seem to be particularly well correlated with the perception level elicited by a video. (i.e. they have been the most successful predictors).

Obtained results enable further research following the suggested approach to improve, for instance, the performance of classification and recommendation systems based on salience and aesthetics characteristics.

From an applicability point of view, proposed solutions could pave the way for a new generation of recommendation systems that could change the way consumers interact with multimedia search engines by allowing them to actively use enhanced search features based on attentional and aesthetics features thus enabling retrieved content to be more related to the affective response and more personalized.

Similarly, automatic multimedia indexation and retrieval systems may elicit new taxonomies guided by the suggested visual descriptors or enable the retrieval of videos according to some specific characteristics (e.g. retrieve only particularly "good" videos). 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 or those that attract more attention).

Finally, anticipating the subjective value perceived by the viewers of any audiovisual content and inferring the level of attention and excitement potentially generated by this content to these viewers could be a huge competitive advantage. For example, video test screening processes, typically used to gauge audience reaction, have associated costs and efforts that could be drastically reduced by making use of the developed technology. Particularly, it could be exploited to automatically predict the expected success of the video instead of relying on the assessments provided by a recruited audience.

Acknowledgements This work has been partially supported by the National Grants RTC-2016-5305-7 and TEC2014-53390-P of the Spanish Ministry of Economy and Competitiveness. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE Conference on computer vision and pattern recognition, 2009. CVPR 2009, pp 1597–1604. https://doi.org/10.1109/CVPR.2009.5206596
2. Adomavicius G, Tuzhilin A (2005) 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
3. Bhattacharya S, Nojavanasghari B, Liu D, Chen T, Chang SF, Shah M (2013) Towards a comprehensive computational model for aesthetic assessment of videos. In: ACM Multimedia, Grand Challenge
4. Boato G, Dang-Nguyen DT, Muratov O, Alajlan N, De Natale FGB (2016) Exploiting visual saliency for increasing diversity of image retrieval results. Multimed Tools Appl 75(10):5581–5602. https://doi.org/10.1007/s11042-015-2526-4
5. Borji A, Itti L (2013) State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intell 35(1):185–207
6. Borji A, Sihite D, Itti L (2012) Salient object detection: a benchmark. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Computer vision – ECCV 2012 lecture notes in computer science. Springer, Berlin, pp 414–429
7. Borji A, Cheng MM, Jiang H, Li J (2014) Salient object detection: A survey. arXiv preprint arXiv:14115878
8. Brezeale D, Cook DJ (2008) Automatic video classification: a survey of the literature. IEEE Trans Syst Man Cybern Part C 3:416–430
9. Bruce N, Tsotsos J (2006) Saliency based on information maximization. In: Weiss Y, Schölkopf B, Platt J (eds) Advances in neural information processing systems, vol 18. MIT Press, pp 155–162
10. Carmi R, Itti L (2006) Visual causes versus correlates of attentional selection in dynamic scenes. Vis Res 46(26):4333–4345

11. Cerf M, Frady EP, Koch C (2009) Faces and text attract gaze independent of the task: experimental data and computer model. J Vis 9(12):10
12. Cheng MM, Mitra NJ, Huang X, Torr PHS, Hu SM (2015) Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell 37(3):569–582. https://doi.org/10.1109/TPAMI.2014.2345401
13. Courty N, Marchand E (2003) Visual perception based on salient features. In: 2003 IEEE/RSJ International conference on intelligent robots and systems, 2003. (IROS 2003). Proceedings, vol 1, pp 1024–1029
14. Culibrk D, Mirkovic M, Zlokolica V, Pokric M, Crnojevic V, Kukolj D (2011) Salient motion features for video quality assessment. IEEE Trans Image Process 20(4):948–958
15. Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. In: Proceedings of the 9th European conference on computer vision - volume part III ECCV'06. Springer-Verlag, Berlin, pp 288–301
16. Dorr M, Martinetz T, Gegenfurtner KR, Barth E (2010) Variability of eye movements when viewing dynamic natural scenes. J Vis 10(10):28
17. Egeth HE, Yantis S (1997) Visual attention: control, representation, and time course. Annu Rev Psychol 48(1):269–297
18. Fathi A, Li Y, Rehg J (2012) Learning to recognize daily actions using gaze. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Computer Vision – ECCV 2012, lecture notes in computer science, vol 7572. Springer, Berlin, pp 314–327
19. Fernández-Martínez F, García AH, Gallardo-Antolín A, de María FD (2014) Combining audio-visual features for viewers' perception classification of youtube car commercials. In: Proceedings of the second international workshop on speech, language, and audio in multimedia, SLAM'14
20. Fernández-Martínez F, García AH, de María FD (2015) Succeeding metadata based annotation scheme and visual tips for the automatic assessment of video aesthetic quality in car commercials. Expert Syst Appl 42(1):293–305
21. Friesen C, Kingstone A (1998) The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonom Bull Rev 5(3):490–495
22. González-Díaz I, Benois-Pineau J, Buso V, Boujut H (2014) Fusion of multiple visual cues for object recognition in videos. In: Ionescu B, Benois-Pineau J, Piatrik T, Quénot G (eds) Fusion in computer vision, advances in computer vision and pattern recognition. Springer International Publishing, Berlin, pp 79–107
23. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422
24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. SIGKDD Explor Newsl 11(1):10–18
25. Haralick RM, Shapiro LG (1992) Computer and robot vision, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston
26. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. In: Advances in neural information processing systems 19, proceedings of the twentieth annual conference on neural information processing systems. Vancouver, pp 545–552
27. Hillaire S, Breton G, Ouarti N, Cozot R, Lécuyer A (2010) Using a visual attention model to improve gaze tracking systems in interactive 3d applications. Comput Graph Forum 29(6):1830–1841
28. Hou X, Harel J, Koch C (2012) Image signature: Highlighting sparse salient regions. IEEE Trans Pattern Anal Mach Intell 34(1):194–201
29. Itti L (2004) Automatic foveation for video compression using a neurobiological model of visual attention. Trans Img Proc 13(10):1304–1318
30. Itti L (2005) Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Vis Cogn 12:1093–1123
31. Itti L (2007) Visual salience. Scholarpedia 2(9):3327
32. Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194–203
33. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
34. James W (1890) The principles of psychology. No. v 1 in American science series: Advanced course. H. Holt
35. Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: a discriminative regional feature integration approach. In: 2013 IEEE Conference on computer vision and pattern recognition, pp 2083–2090. https://doi.org/10.1109/CVPR.2013.271
36. Jiang Y, Wang Y, Feng R, Xue X, Zheng Y, Yang H (2013) Understanding and predicting interestingness of videos. In: Proceedings of the 27th AAAI conference on artificial intelligence (AAAI)

37. Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: 2009 IEEE 12th international conference on computer vision, pp 2106–2113
38. Kalliatakis G, Triantafyllidis G (2013) Image based monument recognition using graph based visual saliency. ELCVIA 12(2):88–97
39. Khan SS, Vogel D (2012) Evaluating visual aesthetics in photographic portraiture. In: Proceedings of the eighth annual symposium on computational aesthetics in graphics, visualization, and imaging, eurographics association. Aire-la-Ville, CAe '12, pp 55–62
40. Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. In: Vaina L (ed) Matters of intelligence, synthese library, vol 188. Springer, Netherlands, pp 115–141
41. le Cessie S, van Houwelingen J (1992) Ridge estimators in logistic regression. Appl Stat 41(1):191–201
42. Li G, Yu Y (2016) Deep contrast learning for salient object detection. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 478–487
43. Li Z, Shi Z, Zhao W, Li Z, Tang Z (2013) Learning semantic concepts from image database with hybrid generative/discriminative approach. Eng Appli Artif Intell 26(9):2143–2152. https://doi.org/10.1016/j.engappai.2013.07.004. http://www.sciencedirect.com/science/article/pii/S0952197613001322
44. Li C, Yuan Y, Cai W, Xia Y, Feng DD (2015) Robust saliency detection via regularized random walks ranking. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR), pp 2710–2717. https://doi.org/10.1109/CVPR.2015.7298887
45. Li Q, Zhou X, Gu A, Li Z, Liang RZ (2016) Nuclear norm regularized convolutional max pos@top machine. Neural Comput Appl 1–10. https://doi.org/10.1007/s00521-016-2680-2
46. Li X, Zhao L, Wei L, Yang MH, Wu F, Zhuang Y, Ling H, Wang J (2016) Deepsaliency: multi-task deep neural network model for salient object detection. IEEE Trans Image Process 25(8):3919–3930. https://doi.org/10.1109/TIP.2016.2579306
47. Liang RZ, Shi L, Wang H, Meng J, Wang JJY, Sun Q, Gu Y (2016) Optimizing top precision performance measure of content-based image retrieval by learning similarity function. In: 2016 23st International conference on pattern recognition (ICPR)
48. Liang RZ, Xie W, Li W, Wang H, Wang JJY, Taylor L (2016) A novel transfer learning method based on common space mapping and weighted domain matching. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI), pp 299–303. https://doi.org/10.1109/ICTAI.2016.0053
49. Liu Y, Zhang D, Lu G, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. Pattern Recogn 40(1):262–282. https://doi.org/10.1016/j.patcog.2006.04.045. http://www.sciencedirect.com/science/article/pii/S0031320306002184
50. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum HY (2011) Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intell 33(2):353–367. https://doi.org/10.1109/TPAMI.2010.70
51. Lv X, Zou D, Zhang L, Jia S (2014) Feature coding for image classification based on saliency detection and fuzzy reasoning and its application in elevator videos. WSEAS Trans Comput 13(1):266–276
52. Ma Z, Qing L, Miao J, Chen X (2009) Advertisement evaluation using visual saliency based on foveated image. In: Proceedings of the 2009 IEEE international conference on multimedia and expo ICME'09. IEEE Press, Piscataway, pp 914–917. http://dl.acm.org/citation.cfm?id=1698924.1699148
53. Mai L, Le H, Niu Y, Liu F (2011) Rule of thirds detection from photograph. In: Proceedings of the 2011 IEEE international symposium on multimedia ISM '11. IEEE Computer Society, Washington, DC, pp 91–96
54. Mancas M, Riche N, Leroy J, Gosselin B (2011) Abnormal motion selection in crowds using bottom-up saliency. In: 2011 18th IEEE International conference on image processing (ICIP), pp 229–232
55. Marchesotti L, Perronnin F, Larlus D, Csurka G (2011) Assessing the aesthetic quality of photographs using generic image descriptors. In: ICCV, pp 1784–1791
56. Moorthy AK, Obrador P, Oliver N (2010) 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. Springer-Verlag, Berlin, pp 1–14
57. Nguyen TV, Xu M, Gao G, Kankanhalli M, Tian Q, Yan S (2013) Static saliency vs. dynamic saliency: a comparative study. In: Proceedings of the 21st ACM international conference on multimedia MM '13. ACM, New York, pp 987–996
58. Niebur E, Koch C (1996) Control of selective visual attention: modeling the "where" pathway. In: Touretzky DS, Mozer MC, Hasselmo ME (eds) Advances in neural information processing systems. MIT Press, Cambridge, pp 802–808
59. Nothdurft HC (2005) Salience of feature contrast. In: Itti L, Rees G, Tsotsos JK (eds) Neurobiology of attention. Academic Press, pp 233–239

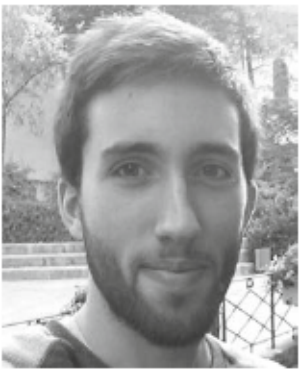
60. Ogaki K, Kitani K, Sugano Y, Sato Y (2012) Coupling eye-motion and ego-motion features for first-person activity recognition. In: 2012 IEEE Computer society conference on computer vision and pattern recognition workshops (CVPRW), pp 1–7
61. Palmer S (1999) Vision science: photons to phenomenology. A Bradford book, Bradford Bokk
62. Parkhurst D, Niebur E (2002) Variable resolution displays: a theoretical, practical and behavioral evaluation. Hum Factors 44(4):611–29
63. Parkhurst D, Law K, Niebur E (2002) Modeling the role of salience in the allocation of overt visual attention. Vis Res 42(1):107–23
64. Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel methods - support vector learning. MIT Press
65. Posner MI (1980) Orienting of attention. Q J Exp Psychol 32(1):3–25
66. Qin Y, Lu H, Xu Y, Wang H (2015) Saliency detection via cellular automata. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR), pp 110–119 https://doi.org/10.1109/CVPR.2015.7298606
67. Rapantzikos K, Tsapatsoulis N, Avrithis Y, Kollias S (2009) Spatiotemporal saliency for video classification. Signal Process Image Commun 24(7):557–571
68. Savakis AE, Etz SP, Loui ACP (2000) Evaluation of image appeal in consumer photography 3959:111–120
69. Seo HJ, Milanfar P (2009) Static and space-time visual saliency detection by self-resemblance. J Vis 9(12):15
70. Smith E, Kosslyn S (2013) Cognitive psychology: pearson new international edition: mind and brain. Pearson Education Limited
71. Su SL, Durand F, Agrawala M (2004) An inverted saliency model for display enhancement. In: 2004 MIT student oxygen workshop
72. Tague N (2005) The quality toolbox, 2nd edn. ASQ Quality Press
73. Tatler BW (2007) The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. J Vis 7(14):4
74. Tatler BW, Hayhoe MM, Land MF, Ballard DH (2011) Eye guidance in natural vision: reinterpreting salience. J Vis 11:5
75. Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12(1):97–136
76. Velásquez JD (2013) Combining eye-tracking technologies with web usage mining for identifying website keyobjects. Eng Appl Artif Intell 26(5–6):1469–1478. https://doi.org/10.1016/j.engappai.2013.01.003. http://www.sciencedirect.com/science/article/pii/S0952197613000134
77. Vig E, Dorr M, Cox D (2012) Space-variant descriptor sampling for action recognition based on saliency and eye movements. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Computer Vision – ECCV 2012, lecture notes in computer science, vol 7578. Springer, Berlin, pp 84–97
78. Wan S, Jin P, Yue L (2009) An approach for image retrieval based on visual saliency. In: 2009 International conference on image analysis and signal processing, pp 172–175. https://doi.org/10.1109/IASP.2009.5054642
79. Wang L, Lu H, Ruan X, Yang MH (2015) Deep networks for saliency detection via local estimation and global search. In: 2015 IEEE Conference on computer vision and pattern recognition (CVPR), pp 3183–3192. https://doi.org/10.1109/CVPR.2015.7298938
80. Wang J, Jiang H, Yuan Z, Cheng MM, Hu X, Zheng N (2016) Salient object detection: a discriminative regional feature integration approach. Int J Comput Vis 1–18. https://doi.org/10.1007/s11263-016-0977-3
81. Xu J, Jiang M, Wang S, Kankanhalli MS, Zhao Q (2014) Predicting human gaze beyond pixels. J Vis 14(1):28
82. Yang CY, Yeh HH, Chen CS (2011) Video aesthetic quality assessment by combining semantically independent and dependent features. In: ICASSP. IEEE, pp 1165–1168
83. Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. In: 2013 IEEE Conference on computer vision and pattern recognition, pp 3166–3173. https://doi.org/10.1109/CVPR.2013.407
84. Yarbus AL (1967) Eye movements and vision. Plenum, New York
85. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) Sun: a Bayesian framework for saliency using natural statistics. J Vis 8(7):32
86. Zhao L, Liang S, Wei Y, Jia J (2015) Size and location matter: a new baseline for salient object detection. In: Cremers D, Reid I, Saito H, Yang MH (eds) Computer Vision – ACCV 2014, lecture notes in computer science, vol 9005. Springer International Publishing, pp 578–592

F. Fernández-Martínez received the Telecommunication Engineering degree and the Ph.D. degree from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2002 and 2008, respectively. He has made several research stays as a visiting scientist and professor including: the University Lille 3 (Lille, France), IDIAP Research Institute (Martigny, Switzerland), Ulm University (Ulm, Germany), and more recently at the Department of Signal Theory and Communications of the Universidad Carlos III of Madrid, where he was also a member of the Multimedia Processing Group (GPM). Since September 2015, he has been an Associate Professor in the Department of Electronic Engineering of the UPM and also a member of the Speech Technology Group, currently attached to the Information Processing and Telecommunications Center of the same university. His main research interests include: HCI Systems, speech technology, affective computing, social signal processing, multimedia information retrieval, image processing, scene understanding and aesthetics assessment. In terms of research, his experience highlights an important record of research projects and contracts in the fields mentioned. As a result of this involvement, he has authored or co-authored more than 50 articles in both international journal and conferences. He also holds one software patent.

A. Hernández-García is a PhD student at the Neurobiopsychology research group at the Institute of Cognitive Science of the University of Osnabr?ck, Germany. He obtained his B.Sc. in Audiovisual Systems Engineering in 2014, as well as his M.Sc. in Multimedia and Communications in 2015 from the University Carlos III de Madrid, Spain. At this university, he was a research assistant at the Signal Theory and Communications Department from 2013 until February 2016, when he moved to Berlin, Germany, to start his PhD in visual emotion recognition from images, with a Marie Sklodowska-Curie ITN grant. His research interests include machine learning and deep neural networks in particular, affective computing, computer vision and visual neuroscience.

M. A. Fernández-Torres received the Audiovisual Systems Engineering degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2013, and the Master degree in Multimedia and Communications from Universidad Carlos III de Madrid, Spain, 2014. He is currently pursuing his Ph.D. degree at the Signal Theory and Communications Department in Universidad Carlos III de Madrid, Madrid, Spain. His current research interests include visual attention modeling, image and video analysis, medical image classification, and computer vision.

I. González-Díaz received the Telecommunications Engineering degree from Universidad de Valladolid, Valladolid, Spain, in 1999, the M.Sc. and Ph.D. degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2007 and 2011, respectively. After holding a postdoc position in the Laboratoire Bordelais de Recherche en Informatique at the University Bordeaux, he currently works as a Visiting Lecturer at the Signal Theory and Communications Department in Universidad Carlos III de Madrid. His primary research interests include object recognition, category-based image segmentation, scene understanding and content-based image and video retrieval systems. In these fields, he is co-author of several papers in prestigious international journals, two chapters in international books and a few papers in revised international conferences.



Á. García-Faura completed his Bachelor of Engineering in Telecommunication Technologies and Services Engineering in 2016 at Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid. He started collaborating with the Speech Technology Group, from the Electronics Engineering Department (UPM) in 2015, where he researched on computational models for video aesthetics. Álvaro is currently pursuing a Master's degree in Telecommunication Engineering at ETSIT, UPM. He is also working as an associate researcher for the Speech Technology Group, where his main tasks are related with image analysis, scene understanding and aesthetics modeling.

F. Díaz de Maria received the Telecommunication Engineering degree and the Ph.D. degree from the Universidad Politécnica de Madrid, Madrid, Spain, in 1991 and 1996, respectively. Since October 1996, he has been an Associate Professor in the Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Madrid, Spain. His primary research interests include video coding, image and video analysis, and computer vision. He has led numerous projects and contracts in the fields mentioned. He is co-author of numerous papers in peer-reviewed international journals, several book chapters and a number of papers in national and international conferences.