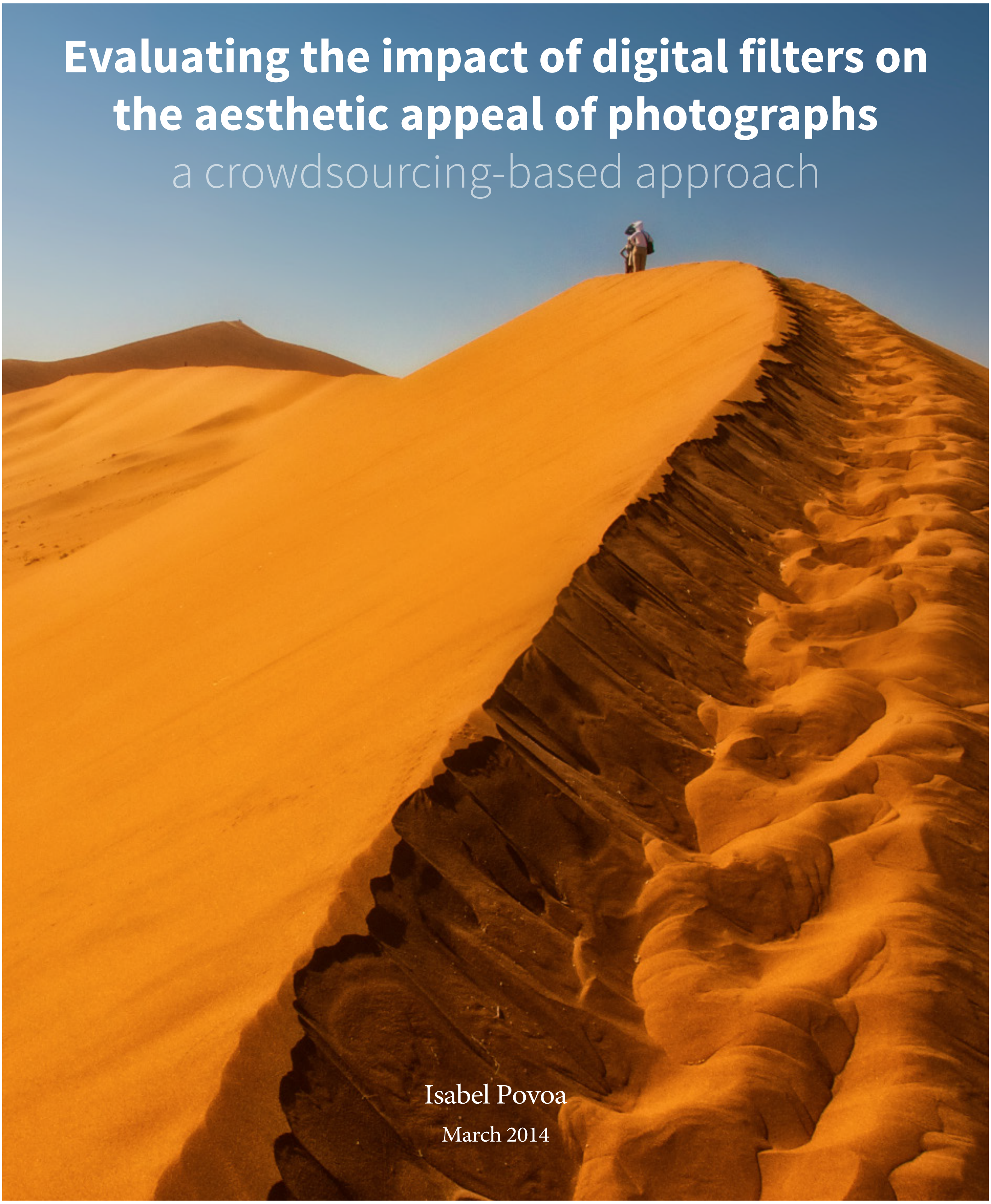
MSc Thesis



Evaluating the impact of digital filters on the aesthetic appeal of photographs

a crowdsourcing-based approach

Isabel Povoa

March 2014

Master of Science in Embedded Systems
Interactive Intelligent Systems Group
Department of Electrical Engineering, Mathematics and Computer Science

Evaluating the impact of digital filters on the aesthetic appeal of photographs: a crowdsourcing based approach

Isabel Povoa

Supervised by Dr. J. A. Redi

March 2014

Evaluating the impact of digital filters on the aesthetic appeal of photographs: a crowdsourcing-based approach

M.I. Campos Sarmento Povoa
4102622

Master Thesis
Embedded Systems
March 2014

Graduation Committee:

Prof. Dr. I. Heynderickx	Philips Research/Delft University of Technology
Dr. J. A. Redi	Delft University of Technology
Dr. Martha Larson	Delft University of Technology

Electrical Engineering, Mathematics and Computer Science Department
Interactive Intelligent Systems Group

Preface

In this final project at TU Delft, I got closer to crowdsourcing as a methodology to employ in the relatively new field of Computational Aesthetics. I have gathered quite a lot of insight during my research and that made more passionate about my work.

On the course of this work, I would like to show my gratitude to my parents, **Carlos Povoa and Teresa Sardo**. They are both loving parents and I know it must have been difficult for them to guide me through this last process. Although, they have always allowed me to do what I wanted without questioning my reasoning. I am really grateful for all their comfort, advice and for giving me the opportunity to study abroad.

Then, I would like to express my gratitude to **J. Redi** for her guidance, brainstorm, constructive feedback and for keeping me from writing a PhD thesis instead of a MSc one. In addition, for the opportunity to attend the ACM Multimedia 2013 conference in Barcelona. Likewise, the s**Delft University Funds** for funding my travel costs.
In addition, I am grateful to J. Redi and my parents for their patience and for the time spent reviewing my thesis. I would also like to express my gratitude to the **multimedia research group** for their assistance every time I had a doubt, and to **Ingrid Heynderickx** and **Martha Larson**, who have been present in different stages of this project, for their help and advice.

Importantly, I want to leave a special thanks to **Mark Dekker**, who has been always present to assist me with his creative mind, challenging me to do my best and for donated most of the image material used in this study. Also, for the time, effort and patience spent helping me on the design of the visual parts of this work and teaching me how to use Matlab. I am really thankful for having him in my life in this period and for the effort spent cheering me and advising me, more than anyone.

Last but not the least, I would like to extend my gratitude to all my friends: those I have made here in Delft for the wonderful moments together and for optimistic support throughout this adventure and those back in my hometown for their love and time when I most needed them. All of you have had some sort of influence in me during the process of this work.
I will use this opportunity to mention some names and if your name has been left out, please accept my apologies but, the time is constrained.

Friends made in Delft:

Erasmus group: Memet, Jose and Javi thank you for still keep in touch even though the different countries that we all are.
Suit up group: Uj, Trivik, Carlos and Paola, Cristina, Eva, Gonçalo, Santiago and Ago thank you for all the dinners parties and hangouts in Klooster, Doerak or Marcushof that we have shared.
The "Greeks": Diana, Bjorn and Tasos for their last long friendship and for the lekker dinners in Rotterdam. Muareshis poli! ☺
Arnoldstraat roommies: Roger, Mark, Paolo and Bassem thank you for all the nice evenings together.
Stukafest gals: Q, Roos and Mad thank you for your patience when I was stressing or multi-tasking, for the funny meetings with "done" and "hot dogs" and for the great teamwork that lead to a great evening!
Marilia, Felipe and their friends that I got to know through them: Mariana, Francisco, Jason, Jamie, Riccardo, Andrea and Deborah thank you for accepting me so easily in your group of friends and for the hard studying evenings in the library that your company made more fun.
KVZ roommies*:* Wyb, Melieke, Ilse, Michou, Nicky, Gert, Martijn and Lied thank you for having me at your place, for the shared dinners and *gezellig* tête-à-têtes.
Mafalda: Thank you for your company over coffee and for helping me print this thesis.

Friends from my hometown:

Childhood friends: Celina, Patricia, Ana and Rita thank you for being present even though difficult from far and for the effort made to meet when I visited.
Lendarios group from my BSc: Gonçalo, Pina, Beto, Kiko, Lazer and Fred thank you all for your effort to keep in touch and for our geeky chats.

Bento: thank you for all the late night chats and for your support through some difficult stages.

I would also like to express my gratitude to all my participants in both setups: laboratory and crowdsourcing. Without them this thesis would not have been possible.

Finally, thank you for reading my thesis, I have spent some time in it and I hope you enjoy the reading.

Isabel Povoa
March 2014

Abstract

In particular grouping by aesthetics and quality of the media has brought along new challenges for Computational Aesthetics research such as what makes an image beautiful, what means beautiful and how do you quantify beautiful. to meet those challenges, researchers have tried to come up with several algorithms based in different metrics to bridge the gap between the quantitative aspects of what is called beauty and what people call beauty.

In order to fill part of this gap we studied the effect of digital filters in photographic aesthetics so widely used in the social networks nowadays. Taking in consideration the popularity of digital filters among many social network users, it was a surprise to understand that most participants in the experiment preferred the images with no filter.

In any case measuring what is beautiful always requires collecting aesthetics scores from people. Doing that collection process in a laboratory environment is the most effective approach. The main reasons are the highly controlled environment that leads to good data quality. The downside is cost, time and restriction of participants to the people available nearby.

Therefore another issue addressed in the study was the use of crowdsourcing to minimize time and cost, as well as to expand the scope of participation, in the process of collecting image scores from users.

To test that possibility a 4 step process step was designed and implemented. First preference scores were collected in a lab environment over a previous selected dataset. Afterwards the crowdsourcing experiment was planned what included an optimization of the dataset (ground truth dataset). Subsequently three digital filters were then applied to the collection and an online experiment followed to once again collect preference scores. In phase, we developed the experiment in the context of Microworkers and as a Facebook app interface enriched with a playful visual interface. The last step included a process to filter the suspicious participants and check results consistency.

The results show that implementing an experiment to collect preferences of image quality in social media is a good methodology for Computational Aesthetics, if appropriate planning and management is adopted.

Index

1. Introduction

The practice and technology of photography has seen intense transformations over the last century. The most evident and noticeable of these changes have taken place in the last decades of the 20th century, with the shift from analogue to digital photography. Since then, technological advances had made it easier and simpler to produce, store, manipulate and share personal photos with anyone from inside and outside one's network. Think about smartphones: their introduction has led to an explosion of photos, which was amplified by social networking sites that enable users to share and upload them on the spot As a consequence, photos have become a big proportion of the web content[1] (Bhattacharya et al 2010), and social networking platforms such as Instagram[2], Facebook[3] and Flickr[4], are tightly dependent on users sharing as much data (including pictures) as possible, since every posted and re-posted piece of data represents market value for their advertisers. Especially Instagram rapidly gained popularity with over 150 million monthly active users, 55 million average photos per day and 1.2 billion daily likes on January 2014[5]. Instagram was first released in 2010 as a smartphone application that allow users to take photos and videos on the mobile, add some effects on the fly and share them. These effects are commonly called digital filters and consist in signal processing algorithms that alter photo properties such as saturation or contrast. In fact, digital filters have been commonly used by professional photographers to manipulate their images' proprieties through imaging software as Adobe Photoshop[6] and Aperture[7] for quite some time now, and with the goal of concealing imperfections or enhancing some specific image aspects. Instagram digital filters are peculiar in the sense, each of them provides a specific combination of effects to emulate the output of some old-fashioned analogue cameras, which recall in some way the nostalgic return to analogue photography witnessed in these days by e.g. the rise of the Lomographic movement (Dowling 2012).

In practice, Instagram marketing pitch promises beautiful photos fast[8], which the rising number of users seems to confirm. If Instagram is really able to generate beautiful images faster, then amateur photographers will have their life simplified and the difference between professional and amateur photography will itself get like an image under a blurred filter. As a matter of fact, the usage of digital filters that reproduce the photographic quality of an old camera has become so popular that most imaging software offer now a built-in set of Instagram-like filters. Hence, we can ask ourselves, is this a new revolution in the field of photography? Is it something that will last in terms of perception of beauty or is it only a fashion? And more specifically: do digital filters assure the beauty of any photograph?

Within this last question, several further issues open further: if filters are essential and its effects have a noticeable effect on the perception of beauty, how should we use them? And which filter do enhance the beauty of photograph the most?

A quick search reveals that Instagram fans use most frequently the original version[9], i.e. no filter. Plot twist! So what is the added value of digital filters, if any?

1.1. Image Aesthetic Appeal prediction

Taking pictures that are beautiful is often described as an art; in the sense of something we don't understand exactly how to make. As such, we can easily pick good-looking photos to show from our photo albums and tell photos taken by professionals from those of amateurs, which would be a difficult task for a computer. But with the bursting amount of digital pictures in our daily life, selecting the most beautiful photos is becoming increasingly difficult and time consuming. Therefore, automatically assessing the beauty of photos in a way that is consistent with human preferences would bring enormous advantages (Yiwen et al 2008)(Joshi et al 2011), for example, for web search engines to display relevant and high quality images or for an amateur photographer to improve his technique, or even further to automatically enhance the beauty of existing pictures.

[1] http://kpcb.com/insights/2013-internet-trends
[2] http://instagram.com
[3] http://facebook.com
[4] http://flickr.com
[5] http://instagram.com/press/
[6] http://photoshop.com/
[7] http://apple.com/pt/aperture/
[8] http://blog.instagram.com/post/8755384810/fast-beautiful-photo-sharing-now-with-foursquare
[9] http://web.stagram.com/hot/

In fact, Google[10] has recently integrated its social network Google + with an algorithm to predict the best pictures (called highlights) of an user-generated photo album[11] (Smith 2013)). Although of undoubted value, the performance of this algorithm is still questionable, in terms of its agreement with what humans would consider the most beautiful pictures: we tested it on an image set (Redi et al 2013)for which we had available subjective judgments of aesthetic appeal, and check if the highlighted images from Google+ would correspond to the images with higher aesthetic appeal scores. The highlighted pictures were mainly related with famous contents such as the Eiffel tower and the Big Ben (see Figure 1), whereas the pictures rated as beautiful by humans were more diverse in terms of content (also visible in Figure 1). Although most of the highlighted pictures got also a high score in the lab, the difference in results show that available applications to assess the beauty of photos have a wide margin for improvement.

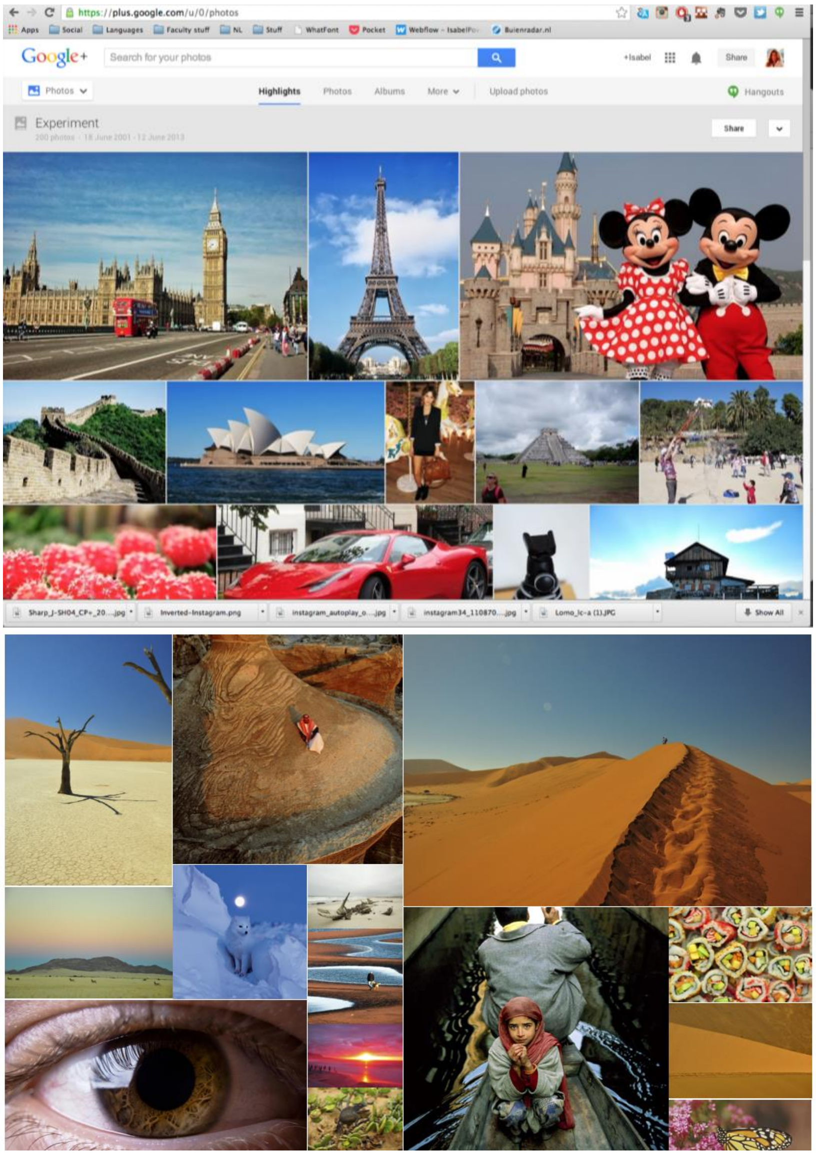


Figure 1: Output from the highlights feature designed by Google to retrieve the best images (top) versus the most beautiful images according to the subjective scores (bellow), both from the same dataset of (Redi et al 2013).

[10] http://google.com

[11] http://plus.google.com

1.1.1. Computational Aesthetics

Aesthetics comes from a Greek word that means "perception" and refers to a branch of philosophy dedicated to the study of the characteristics of what people call beauty (Merriam-Webster 2014). Aesthetics is therefore relevant to understand what people like and why. This is achieved by studying the process of an ***aesthetic experience***, i.e. the act of perceiving an aesthetic object (Joshi et al 2011)(Locher et al 2011), which involves complex and subjective interactions between various mechanisms such as the emotional state emotion and the personal taste of an observer.

Although difficult to quantify and analyse, the multimedia processing community together with vision and computer scientists has recently increasingly tried to address the problem of computational assessment of image aesthetic appeal. A typical approach entails the use of image features and human visual system models to predict aesthetic judgments (Mansilla et al 2011). To better understand these studies, it is useful to first look at the Image Quality Circle framework (Engeldrum 2000) , which has been used to optimize image quality of imaging systems (see Figure 2).

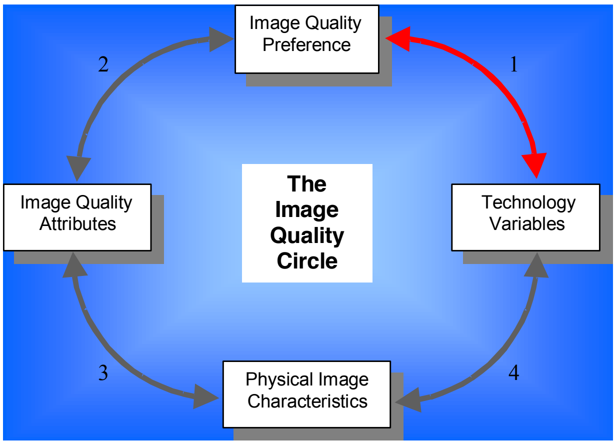


Figure 2: Image Quality Circle framework from (Isola et al 2011).

The functioning of the framework can help to investigate the problem of computational inference of aesthetics.

This framework offers an effective methodology to link Image Quality Preference (the overall image quality as judged by observers) with the technological variables of an imaging system**. Technological Variables** are the manipulated variables that we need to vary to produce a certain effect on image quality, such as pixel size which affects resolution, and choice of camera colour filter mosaic which affects colour reproduction. Therefore, given a change in technological variables, this framework helps us understand its relation to image quality preferences (link 1 from Figure 2). For that purpose, one needs to use the framework in a counter clockwise direction from Image Quality Preference to Technological Variables.

In our case (see Figure 3), Image Quality Preference is instead called **Aesthetic Appeal** and it is related to **Aesthetic Appeal Attributes** (or Image Attributes). These are characteristics of the image as *perceived* , and therefore better quantifiable through subjective measurements of the aesthetic appeal of the image. Each attribute is then related to **Features**, i.e., those objective quantities that are measurable from pixel information (or metadata) and can be computed by algorithms. Image features contribute to the overall appearance of the image, and typically influence several Aesthetic Appeal Attributes. Finally, the last relation is usually trivial and known by researchers. Overall, by measuring image features, mapping them into attributes, we can predict the effect that the changes in technological variables have on the perceived aesthetic appeal.

Nevertheless, how to relate aesthetic appeal with its attributes, and the attributes with their features is not yet known and the first step is to observe how the relationships that those two links represent function in humans and to do so one needs to observe them empirically through subjective studies.

In these studies, users are typically asked to scale different images according to one or more of their attributes, and/or aesthetic appeal. Based on these studies, the goal of Computational Aesthetics is to create algorithms or models that, given the observed relationships features-attributes-aesthetic appeal, can model them with acceptable accuracy. In other words, based on the processing of physical characteristics of the images, Computational Aesthetics tools can predict the aesthetic appeal of the image in a way similar to what a human would do, yet without the need of human intervention.

In the scope of our work, we want to understand the added value of digital filters in aesthetic appeal. In other words, we want to evaluate how a manipulation (like the ones offered by Instagram) of image features impacts images aesthetic appeal. As a result, the Image Quality Circle (IQC) framework can be applied in this problem taking the digital filters as technologies variables. If one knows how image features influence attributes and attributes aesthetic appeal, then we can use the filters to change the features of the image and consequently study how the aesthetic appeal changes. Thus, this thesis work will build up along the IQC framework links to understand this relationship.

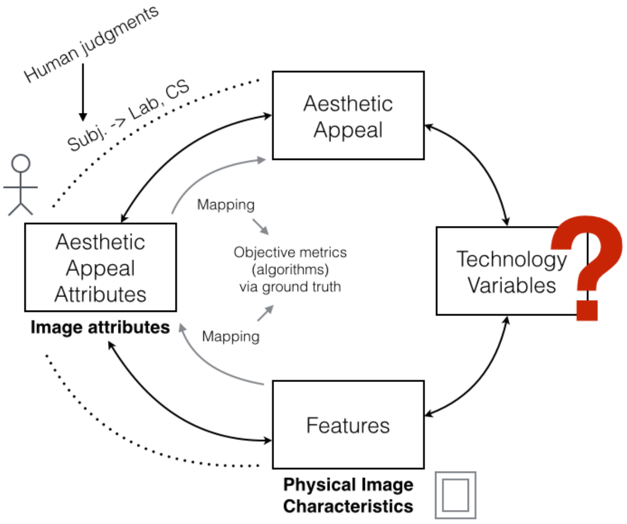


Figure 3: Adapted version of Image Quality Circle in terms of Computational Aesthetics

1.2. Scope of the thesis

Overall, research has been trying to understand and to model what makes an image beautiful to the human eye, but so far the topic of digital filters to enhance the perceived beauty of an image has not yet been targeted. To be able to do so, it is necessary to study the relationship digital filter-image feature-image attributes-aesthetic appeal. Since most of these links have not been studied in an empirical way nor computationally modelled, we will focus in this thesis on the observation of the above mentioned relationships by means of subjective studies. This opened up a number of research questions:

- Which image features are the most impacted by the application of digital filters?
- Which image attributes are mostly contributing to the generation of an aesthetic appeal judgment?
- Which image features do impact the abovementioned attributes and how?
- Which types of images should be included in a study investigating the effect of digital filters on aesthetic appeal?

To answer those questions, we first performed a thorough literature review, through which we identified potential key attributes and features to assess. Then a pilot study was designed and implemented in a laboratory environment to collect scorings on the elected attributes regarding on a large set of images. With the collected data from this experiment, two papers were published (see Appendix A and B).

Further, after analysing the scorings collected, the results were employed to create a representative dataset for the second, core experiment of this thesis: that investigating the impact of digital filters on aesthetic appeal. The image set for this second study was designed to include a smaller number of images, properly balanced in terms of features, attributes and aesthetic appeal levels. The latter was an important tactical detail for the success of the study: a representative and sufficiently diverse (in terms of aesthetic appeal, features and attributes) set of images was core to be able to draw conclusions as general as possible from an experiment investigating the added value of digital filters to aesthetic appeal. Once again, in designing this second experiment, a set of questions needed to be answered, starting with the number and type of filters to be included in the study. Having decided for testing three popular filters, we realized the number of images to be tested was too big. Besides, since cultural background influences aesthetics (Rhodes et al 2001), it is essential to reach a highly diversified population at a fraction of the cost that a lab experiment for the same number of participants would entail. As a consequence, traditional laboratory methods will not work, this led to more research questions:

- Can crowdsourcing help and how?
- Can crowdsourcing be reliable?
- Can I reproduce laboratory results in a crowdsourcing setting?

To sum up, the methodology adopted to understand the impact of digital filters on aesthetic appeal involves a two-step process of aesthetic evaluation, each one comprising an experiment: a pilot study in a laboratory environment, followed by a crowdsourcing-based experiment with an optimized image dataset, all of that to finally answer:

- Do filters improve the aesthetic appeal appreciation of an image?
- What relations can be established between the usage of the most popular filters and the visible attributes and computed features of a photograph?

1.3. Structure of the thesis

This thesis consists of 6 main chapters. After this introduction, we start with a literature review (Chapter 2) briefly covering the evolution of photography from analogue to digital, then focusing on quantification of aesthetic appeal from a methodological point of view first (reviewing benefits and drawbacks of lab- and web-based testing), and from computational point of view later. After the principles of Computational Aesthetics have been outlined, we shall then go on to present the preliminary experiment we performed to have more insight on the relationship image feature-attributes-aesthetic appeal (Chapter 3). The core research question of this thesis is addressed in the following chapter (Chapter 4), where an extended crowdsourcing-based experiment is reported that investigates the impact of digital filters on the aesthetic appeal of images. Based on the experimental results conclusions are drawn in Chapter 5 and recommendations for future extensions of this research is presented in Chapter 6.

2. Background

This chapter presents the background study on which this research project is based. This study uses notions inherent to the photographic, Computational Aesthetics and crowdsourcing domains. Hence, a brief perspective on each of them will be provided.

We begin with a short introduction to the evolution of photography, being its quality the subject of this thesis. In particular, we review its history, starting from the roots and describing its development into a social, mass activity (section 2.1). Then, in section 2.2, we dive into the concept of aesthetic appeal, focusing on existing empirical and automatic methods for the quantification of the aesthetic appeal appreciation of images (Computational Aesthetics). Finally, to conclude this chapter we target crowdsourcing as an increasingly popular methodology in user preference research, which may be beneficial for the investigations we plan to perform in this thesis.

2.1. From photography to Instagram

The practice and technology of photography have witnessed various changes over the last decades. One of these noticeable changes was the leap for amateur photographers in 1948 with the release of the first Land camera in the market by Polaroid, best known for its self-developing film (Bonanos 2012). Before the release of this rapidly popular instant camera, you would only have two options to process your images: work in your own darkroom or get your film to a lab, not always easy to reach. Then, when the polaroid hit the market, people saw the advantage of processing their own film easier and faster.

A second, epochal change took place in the last decade of the 20th century, when the shift from analogue to digital photography, in fact started already in 1975 (Prakel 2009), became reality. In 1991, Kodak introduced the Digital Camera System (DCS 100) (McGarvey 2004). Henceforward, the main camera manufacturers like Sony and Kodak continued introducing to the market more and more advanced models with more and more features. Cameras with communication capabilities such as with cellular phone transmission capability or wireless communication to the Internet began to emerge in the 1990s by several companies such as Sony and Olympus and by around 2000 the first mobile phones with integrated cameras appeared in the industry (Peres 2007). These and other early models suffered from low quality, low-resolution and shutter lag[12]. The subsequent period was characterized by an exponential development of technologies such as imaging software, consumer cameras and camera phones. This expansion led an enormous increase of photos production by individuals, thus becoming a way of communication that, simultaneously, motivated and helped the rise of social media platforms (Abraham et al 2009) like Flickr, and later Instagram, solely focused on Photographic material sharing.

Meanwhile, technology continued to evolve and applications changed the way people use technology to interact also with photos. After its first release in 2007 (Honan 2007), the iPhone has become the most popular "camera" in the Flickr community[13] (Truta 2013) and as for Instagram; it is now one of the most popular social photography applications available for smartphones since its release in 2010. It currently has a growing database of over 100 million users and was one of the largest Facebook's acquisitions to date[14,15].

One might say that analogue photography is as good as dead. The rapid reduction of film camera output has driven both Polaroid and Eastman Kodak into bankruptcy (Bonanos 2012). As a matter of fact, the legacy of analogue photography has not yet been forgotten. Lomography is a pretty recent type of photography invented by the Lomographic Society, founded in the early nineties in Austria. Lomography is considered an art movement and the community of Lomographic photographers promote creative and experimental film photography[16]. Additionally, since the year of 2010, instant film materials for vintage Polaroid cameras have again become available on the market, developed and manufactured by a group called The Impossible Project, at the former Polaroid production plant in Enschede, The Netherlands[17].

To summarize, photography today seems more alive than ever. It has changed throughout the years but never lost its importance since its first manifestation. Today it is used for various purposes such as a hobby, art, for commercial practice, scientific and forensic documentation or as an educational,

[12] http://economist.com/node/15865270

[13] http://flickr.com/cameras/

[14] http://instagram.com/press/

[15] http://blog.instagram.com/post/13928169232/were-the-2011-app-store-iphone-app-of-the-year

[16] http://lomography.com/about

[17] http://the-impossible-project.com/about/

institutional or judicial tool. It provides new opportunities for personal expression as well as new forms of citizen journalism. Undeniably, the camera phone has introduced the power of creating and sharing photos within seconds and information today is a collaborative, participatory, and interactive experience. As for Instagram, it seems to have created a new form of momentary or everyday aesthetics to the practice of photography combining the looks of old vintage photos with the new developments of digital imaging software. In Instagram, as end-users we can have the effects of an old analogue camera without the weight of carrying an analogue camera or the need of processing the film. Besides, all the process of sharing with friends is nowadays easier with the help of social networks. Nowadays photography is done with faster, lighter and simpler cameras and if the photographer makes a mistake, it is easy to review the image and take a new one or even to post-process it at home with help of an photo editor as Adobe Photoshop. In old photography, instead, we would have a heavy camera, slow and difficult to hand in a way that photos could end up being shaky or ruined by handling the film in a wrong way. But, an old photography brings nostalgia and it is becoming an hype among the youth. Instagram addresses then the problem of old cameras and brings the old nostalgic looks to your phone with which one can take a similar photo within seconds.

2.2. Understanding the appreciation of the aesthetic appeal of photographic images

2.2.1. Aesthetics everywhere

Unveiling the magic of an aesthetic experience has been the subject of study of numerous disciplines. For example, in philosophy, where several theorists have been discussing aesthetics in countless dissertations (Lind 1980)(Hoenig 2005), in psychology, where the aim of empirical aesthetics is to study factors that influence the aesthetic experience (Mansilla et al 2011) and within the field of Human-Computer Interaction (HCI), where aesthetics has also been extensively studied as an added value to product design (Mansilla et al 2011)(Hartmann et al 2008). Likewise, more and more domains of research have integrated the understanding and quantification of aesthetic appeal in their scope of study.

One such domain is Multimedia, being crucial for a number of applications, from multimedia information retrieval to computer graphics (Joshi et al 2011), to be able to assess in an automatic way the aesthetic appeal of images. Furthermore, recently it was found that image aesthetic appeal has also an impact on the overall visual experience in image observation, and affects the way technical quality (.e. on "the satisfaction of the user with respect to the purely perceptual impact of the media, independent on its content or context of usage") is judged (Redi 2013). The latter was an interesting result for a community that often associated image viewing experience with the technical specifications of imaging systems and network technologies (Engeldrum 2000) . That simplistic approach is not compatible anymore with today's visual media experiences, that can be provided in a wide diversity of contexts and through immersive technologies such as 3D displays and augmented reality. This leads the end-user to have higher expectations in relation to the experience provided by visual systems; therefore, technical specification of imaging systems are not sufficient to predict alone the quality of the overall experience (Le Callet et al 2012). Therefore, to be able to optimize such systems so that they fulfil users' expectations, it is important not only to consider the impact of system technical specification on the human visual system, but also to take into account other elements of visual experiences more related to affective and cognitive processes, such as aesthetic appeal is (Redi 2013). As a result, it is crucial for multimedia systems to be able to model user aesthetic appeal appreciation mechanisms.

2.2.2. Existing models of aesthetic appeal appreciation

A first step toward computationally modelling aesthetic appeal is to have an overall model of aesthetic appeal appreciation in users. Several psychological models have been proposed to explain the processes behind an aesthetic appeal appreciation, for example in (Mansilla et al 2011) and (Carbon et al 2011). In (Leder et al 2004), a five-stage model of aesthetic appeal appreciation in humans is described:

- **Perceptual analyses** deals with the images' inherent visual attributes (such as shape, colour, contrast, etc.), which are easily spotted and measured at this stage. These visual attributes of the image are the main focus in Computational Aesthetics research.
- **Implicit memory integration** refers to the stage in which implicit experiences and familiarity come in the picture. Implicit experiences are, in this context, previous exposure to specific

situations that influence a viewer's mood at the time of judgment. For instance, familiarity deals with repeated contact to a determined content.

- **Explicit classification** refers to the act of explicitly judging the semantic and type of content where the knowledge and the expertise of the viewer have an impact.
- **Cognitive mastering** entails the development of knowledge, usually learned through education or acquired with practice. This phase is a result from the previous in which a successful classification of semantics offers a self-rewarding intellectual experience.
- **Evaluation** is the result of the cognitive mastering stage. These two last stages reflect the process of expertise and are connected through a feedback-loop; the observer understanding is evaluated and if not successful, the observer will return to the previous stage.

The work presented in (Carbon et al 2011) proposes instead a dynamic two-step model to describe aesthetic experience. In the first step of the above mentioned process, the observer meets the images spontaneously, without any pre-conceptions; in the second step, the observer through repetitive exposure adapts to the images by integrating it into the subject knowledge (Carbon et al 2011).

This model results into a temporary state called taste, which will be constantly changing in function of adaptation and shared taste processes. Shared taste refers to general, interpersonal agreement on aesthetic traits independent of demographics. Adaptation refers to changes in preferences due to the media we consume (Carbon et al 2011), which are not only related to what is visible, but also to what is expected. This model can explain the often documented shared preferences (Faerber et al 2010)(Rhodes et al 2001) as well as personal predilections of some cultures, e.g. (Rhodes et al 2001) reports on facial averageness and symmetry attractiveness for non-Western cultures and shows that there was no preference for own-race (Chinese) averaged combinations over other-race (Caucasian) combinations.

Besides, the model has been used in product design to generate and trigger the development of trends. Different people will have contact with the same product repetitively by means of publicity and shared preferences will then emerge.

In truth, none of the models that have been presented to explain aesthetic appeal appreciation will be fundamental or absolute. But they represented very important attempts that can be corrected, thus supporting the development of improved models (Joshi et al 2011).

2.2.3. A computational approach to aesthetic appeal quantification

Computational Aesthetics is a branch of computer vision or multimedia signal processing that deals with the automatic prediction of aesthetic appeal. This is typically accomplished by extracting low level features of the images as well as from its metadata like from their EXIF files. These information can after be merged into a computational model, which can be based in machine learning techniques and given an image, can automatically quantify its aesthetic appeal, as it would be judged by humans.

Nevertheless, to be able to create robust models that can accurately predict aesthetic appeal, it is necessary to understand what makes an image beautiful for humans: how are they looking at images, what are they feeling towards then.

To comprehend what is beautiful for humans, in addition to studies found in psychological (Fedorovskaya et al 2013) and philosophical literature (Lind 1980), empirically studies (user studies) are typically the most reliable way to identify what matters for aesthetics (Carbon et al 2011)(Wagemans 2011). These studies are the key to understand how changes in the features impact the perceived attributes and aesthetic appeal because we can manipulate the features of the images or select images that vary in these features. But in doing so, there are multiple challenges that need to be overcome such as which methodology to use or the stimuli selection.

2.3. Understanding aesthetic appeal: methodologies for empirical studies of aesthetic appeal appreciation

Computational Aesthetics relies on the outcomes of subjective studies, for which robust methodologies are needed. The development of methodologies for subjective testing is important because a reliable empirical study will in turn provide useful data for the development of algorithms that predict aesthetic appeal. In turn, the output of these algorithms will also need to be compared with subjective testing data to check its accuracy. Thus, subjective studies helps us to understand which processes influence aesthetic appeal appreciation and which attributes have impact on it as well as to create ground truth for computational models.

Further, methodologies for subjective testing are developed based also on the environment in which the tests are performed. Whereas subjective studies have been performed within controlled lab environments, nowadays we are witnessing a paradigm shift also in this aspect because web-based testing is becoming more and more popular. That is because even though the traditional (laboratory-based) approach is effective, it is also expensive and time consuming.

Thus, in this section we are going to review how both environments support user studies as well as their advantages and disadvantages. Consistently, we will then analyse the process of stimuli selection, which is one of the most difficult parts for both crowdsourcing and laboratory studies. Failure to select a suitable sample can harm the quality of the data (Engeldrum 2000) .

2.3.1. The laboratory approach

Typically in subjective studies, we want to be able to consider different stimuli and sort them along a certain dimension under investigation, with a certain degree of confidence. These dimensions can vary widely from aesthetic appeal, image quality, or some other image attribute like colourfulness or sharpness. For example, when we ask participants to assign a value to an image on aesthetic appeal (the underlying dimension), we want to know where that image will be placed by each participant on an aesthetic appeal scale.

Psychometric scaling is then the process of sorting stimuli along a psychological continuum (dimension) and also of assigning them to a value on a scale. In our case, stimuli are images and the scale is aesthetic appeal and its attributes.

There are three classes of measurements used to scale the stimuli that were at first developed to be used in laboratory experiments:

- **Performance measurements** that are based on the quantification of the success of the participant in performing a task. In performance measurement studies, the position of the images on this scale of aesthetic appeal is measured through human performance at doing some task with an image, like reading the content in it (Komar et al 1997) . Thus, these entail objectively measurable outcomes of the experiment. For example in (Szechter et al 2007) human performance was evaluated at sorting images into similar groups.
- **Physiological measurements,** where reactions of the human body are measured like eye movements or the body postural adjustments when observe images or paintings (Wallraven et al 2009)(Mantel et al 2013)(Locher et al 2011). In turn these measures help us understand human visual quality preferences. In (Locher et al 2011) a conducted investigation on the viewers' postural body movements while observing paintings with compositional movement is reported. As expected, the body postural adjustments by the viewers were significantly more pronounced for the composition containing a higher degree of depicted motion. On the other hand, eye movement studies can help us better estimate image quality, e.g. (Engelke 2011) artefacts in the most attractive regions (extracted through recordings of eye fixations) were showed to be more annoying than those in the background.
- **Judgment measures** where participants are asked to assess aesthetic appeal or an attribute (Bhattacharya et al 2010)(Congcong et al 2009) or. These are the typical used methodologies to quantify image quality.

The focus of this work will only comprehend physiological measures and judgment measures.

In what concerns to physiological measures, these can give insight on involuntary reactions to images, that often can reveal more about the affective state of an observer (Locher et al 2011). Specially eye movements' studies can provide relevant data to identify the most important regions in an image which have been proven to be related to aesthetic appeal (Engelke 2011). Furthermore, more and more subjective studies based on eye movements studies have been increasingly emerging in the literature (Le Callet et al 2013). While observing a scene, due to the brain's constrained processing capacity, our eyes collect a limited amount of information. As a result, unconsciously we perform some sort of "scanning" in which our eye holds its attention on particular highly informative areas (fixations) and shifts to a different location when enough information has been collected from the current position (saccades). The fixation regions are then called salient and can be used as a measure of "attractiveness" between the eye and an image (Itti et al 2001).

In addition, recording eye movements is a valuable technique because it shows where the observers fixate when evaluating composition (Leder et al 2004) or other attributes (Redi et al 2011). Then, saliency refers to the most visually relevant parts of an image (Eickho et al 2012), that likely carry most of the image semantic information (ROI). Further, in (Alers et al 2010) it was showed that distortions in the regions-of-interest of an image are more likely to be seen and consequently more annoying for observers.

Overall, physiological measurements are important for the investigation of the processes behind aesthetic appeal appreciation because, when appreciating the aesthetic appeal of an image, there are underlying neural, perceptual, and cognitive processes involved that come into play (Locher et al 2011).

Judgement measures involve explicitly asking the participant to express a judgment on a stimulus in terms of a given criteria (dimension). The judgment can then be translated into its position on the scale (psychological continuum). This can be asked in many different ways, using different judgment methodologies such as the single stimulus method, direct scaling method or even the paired comparison method (Engeldrum 2000) . For example, for the paired comparison method, the participant is presented with two images and has to choose the image that best corresponds to what is asked. The participant will not assign a value but will anyway express a judgment and with the judgment we can retrieve a value. Of the mentioned judgment methodologies, the direct scaling method is also worth mentioning. In this method, participants are asked to pick a score (numerical judgment) for a given stimulus. The scores provided by the participants allows for a direct sorting of stimuli on the psychological continuum and therefore a straightforward data analysis. Even though there are methodologies that are more accurate, these are usually more complicated to analyse and, consequently, more timing consuming like the paired comparison method. Additionally, the single stimulus and direct scaling are standardized procedures by the ITU (ITU 2012) and quite convenient to use in subjective studies. The main variables included in these different methodologies are then what needs to be investigated that bring out a result of the dependent variable (aesthetic appeal), from a given value of one or more independent variables (features and attributes). At first sight, it seems quite simple and easy to ask observers to express their preference about a set of images. However, conducting a scaling study without a thorough plan is likely to generate useless and wrong results. To avoid this, thoughtful considerations are needed to assure data validity (Engeldrum 2000) . The planning then entails the next steps:

- Selection of the images
- Preparation of the images for observer judgement
- Selection of observers
- Preparation of the judgement task
- Presentation of the images to the observers for their judgment
- Collection or recording of responses
- Analysis of observers' data

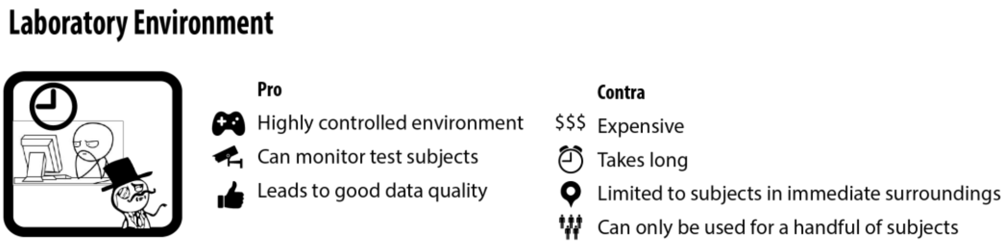
During the scaling study, viewing conditions and the way the images are presented are also important factors for data accuracy, precision and efficiency of the scaling process. For instance, the sequence of the images presentation can affect observers' judgements, so it is recommended to randomize the sequence (Engeldrum 2000) . Additionally, the viewing distance and ambient illumination needs to be taken in consideration during observers' judgements in order to achieve precision of the measurements (Engeldrum 2000) . Viewing distance can alter visibility of image characteristics, which in turn can cause alterations in the observers' judgments, which, in turn, may alter the outcomes of the scaling task. Plus, if the light source used is different amongst participants, then the colours of the images can be perceived differently in, which may also alter the outcomes of the scaling task. This is especially true for traditional visibility and image quality tests, where viewing distance and illumination influence the visibility of artefacts.

Likewise, the laboratory overall environment also needs to be controlled, specifically psychological and physical comfort, noise and surround. These can affect the way tasks are performed. It is up to the researchers to ensure that the observer is comfortable while performing the task and is clear about what to do. Moreover, the experimenter should monitor the observers closely but without making them feel that are being tested or watched (Engeldrum 2000) . Clearly, planning a scaling study in a laboratory where participants are highly monitored can be a challenge. With appropriate planning, laboratory data delivers highly reliable data thus supporting methodologies to set up an algorithm capable of predicting aesthetic quality.

Indeed, the laboratory setting has been used for long time and, consequently, researchers have managed to develop accurate methodologies to investigate the impact of a certain attribute on aesthetic appeal and to control the surrounding environment. As we mentioned, the environment can have a big effect on the outcomes of the scaling task. Then, a big advantage about the lab is the direct interaction between participant and experimenter, in which instructions can be re-iterated. This is important because it can be that the task is not so straightforward and thus, that allows for a higher level of control.

On the other hand, as already mentioned all of this effort takes time and is costly (Kittur et al 2008). Costs include for instance the facilities, the observers' rewards for their voluntary efforts or any special equipment like an eye tracker or a special lighting. Also, the entire process can require long time depending on the type and number of images, on the number of participants recruited, the participants' task and on what researchers need to set up and control. As a result, researchers have to compromise either the number of participants and/or the number of images to be tested due to time and monetary constraints (Kittur et al 2008). Then, the lack of statistical rigor associated with a small sample size is also another downside of these studies (Kittur et al 2008).

Nevertheless, it is important to remark that the fact that the environmental factors are highly controllable in a laboratory set-up to ensure robust results (along with the fact that participants can be instructed thoroughly and have the option to refer to the experimenter if their task is not clear), the mentioned drawbacks have led researchers to look for alternative practices. In Figure 4, one can see the advantages and disadvantages of laboratory experiments summarized.



Laboratory Environment

Pro	Contra
Highly controlled environment	$$$ Expensive
Can monitor test subjects	Takes long
Leads to good data quality	Limited to subjects in immediate surroundings
	Can only be used for a handful of subjects

Figure 4: The advantages and disadvantages of laboratory experiments.

2.3.2. The crowdsourcing approach

With the advent of internet, practitioners have turned their heads to new ways of collecting observers' responses from the Web (Kittur et al 2008) such as through online surveys or standalone platforms. However, these approaches rely on the capability of the researcher to recruit participants and thus, the pool of participants can be limited.

Crowdsourcing provides an alternative paradigm to the traditional way of collecting observers' responses (Kittur et al 2008)(Soleymani et al 2010). It refers to the process of outsourcing work to an unmanaged large crowd of anonymous workers in the form of an open request. Crowdsourcing tasks, sometimes referred to as micro-tasks, can be accomplished within a few minutes or even seconds and do not require a long-term contract (Hossfeld et al 2013). Tasks can vary largely, they can either be very simple as submitting a vote for a webpage or more repetitive as labelling consecutive images. When a worker has finished a task, a proof of that needs to be submitted so that the employer can check the work done and if satisfied, compensate the worker a monetary or non-monetary (e.g., reputation) reward. The reward amount is based on the time and effort required to finish the task, but because tasks are very easy and simple, workers are paid very little, on the order of a few cents.

Micro-tasks are grouped in campaigns. In order to create and manage a campaign, the employer needs to make use of a platform where the crowd is logged in. Some platforms allow the employers to confine the anonymous crowd, based either on their current country, previous performance or reliability on other campaigns (Hossfeld et al 2013). The Amazon Mechanical Turk[18] and the Microworkers[19] are two popular examples of crowdsourcing platforms, but Facebook and other social networks can also be used to recruit test participants (Hossfeld et al 2013). In practice, when used for subjective testing, crowdsourcing platforms act as an extra layer between researchers and participants, which handle the participant recruitment and payment, letting the experimenter focusing on the challenge of how to design and manage the experiment.

Mechanical Turk has its own API and requires employers to have a US Bank account, thus excluding non-US employers. Besides, the workers are mainly located in USA and India because the platform only pays the people on these two countries by cash. Workers from other countries are paid with Amazon.com gift certificates (Hossfeld et al 2013). On the other hand, Microworkers allows

[18] http://www.mturk.com/mturk/

[19] http://microworkers.com

international employers to pay cash to international workers, but it does not have its own API. As a result, employers need to redirect workers to their own web-based applications, that can be accessed via common web browsers, e.g. Firefox, Internet Explorer, or Google Chrome. The latter approach leads to more broad task designs and possibilities (Hossfeld et al 2013). Additionally, it does not offer features such as filters and qualification test mechanisms to select and build specialized worker groups as Mechanical Turk does (Hossfeld et al 2013). As a result, the risk of hiring unreliable workers, performing tasks without proper commitment, is higher.

Crowdsourcing can be an alternative method to collect ratings from images that otherwise would have been rated in a laboratory environment. It addresses the drawbacks of a laboratory-based experiment by outsourcing tasks via the Web to a global worker pool, with reduced costs, larger diversity of the test participants, and faster return from test campaigns (Kittur et al 2008)(Vliegendhart et al 2012). In fact, even though the three classes of measurements presented in the previous section were first developed for lab experiments, they are nowadays used in crowdsourcing set-ups as well. For example, in the work of (Kittur et al 2008) judgement measures were used where participants were requested to rate articles on a 7-point Likert scale according to a set of dimensions such as how well written and how well structured the article was.

On the other hand, reliability is a major issue. No matter the platform used, workers can submit dishonest results or even dishonestly complete tasks as fast as possible, without committing to it, for the sake of increasing their pay. Thus, crowdsourcing is best suited for tasks in which there is a bona fide answer. In that case, it is easy to spot deceitful answers submitted by cheating workers (Kittur et al 2008). However, when collecting ratings on aesthetics there is no right and wrong answer, making it difficult to identify the unreliable workers and creating a need for sophisticated statistical methods. Moreover, the diverse and anonymous nature of the crowd is both a strength and a weakness (Kittur et al 2008). Since workers are drawn from all over the globe, results have the potential to predict more accurately the population than a small sample of users from a limited geographic pool characteristic from laboratory experiments. Still, the insufficient information from the crowd such as age and expertise as well as the limited experimenter contact with the crowd raises the question of the possibility to extract useful data from this novel method (Kittur et al 2008).

A major challenge we have to tackle is designing the crowdsourcing task for research use (Vliegendhart et al 2012), since it is not as straightforward as implementing of existing subjective testing methodologies in a Web-based environment. One needs to address changeless like environment factors that might an effect on the outcome of the of the scaling task and the risk of malicious users. Thus, there is a need for re-adapting traditional laboratory techniques to crowdsourcing, taking into account the literature recommendations, such as the constraint on short task duration (while typical laboratory experiments can last hours) (Soleymani et al 2010).

Another issue with crowdsourcing studies is the lack of an experimenter to guide the workers through the tasks, but that can be easily overcome via a simple and easy interface. The training session present in laboratory setups here is replaced by qualification tests. Qualification tests are simple tests used to select and build specialized worker groups e.g. obtained by computation of simple text equations: “4 plus two=?”, or overall performance, or given attributes like country (Hossfeld et al 2013).

Once again, statistical methods need to be used to qualify or not the workers at this stage.

In short, the main disadvantage of crowdsourcing is the risk of unreliable data due to lack of an appropriate control and validation strategy to monitor user responses.

In Figure 5, one can see the advantages and disadvantages of crowdsourcing-based experiments summarized.

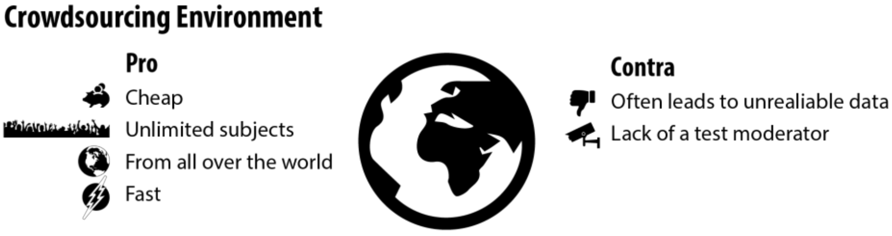


Figure 5: The advantages and disadvantages of crowdsourcing experiments.

A possible strategy to minimize the problems identified so far in using crowdsourcing to collect judgment measurements on aesthetic quality is to implement the test within in social networks as Facebook. This would allow combining tasks with amusement. Therefore if the user study is designed in a pleasing manner, social networks can provide a large number of test subjects for free (Hossfeld et al 2013). That possibility may further reduce costs and can be used to motivate honest participation thus enhancing data reliability. Furthermore, if one integrates a user study with a Facebook application that enables access to the users' demographics provided in their profiles that allows the experimenter to have some insights on the participants. Of course, authorization from the users is required.

However, designing a scaling study in an attractive way and integrating it with Facebook requires a significant amount of effort and time that is not always possible. Besides, participants from a social network might be biased in terms of expectations if they are familiar with the creator of the test (Hossfeld et al 2013).

2.3.3. Addressing Reliability issues in crowdsourcing

Reliability is a big challenge to address in crowdsourcing studies and thus researchers have focused on metrics and methods proposals to overcome it. User ratings might not be reliable due to (Hossfeld et al 2013):

- Technical errors – bugs introduced by the application or due to incompatibilities of the software/hardware
- Test instructions – too complex to understand
- Language problems – may occur for international participants
- Cheating participants – to maximize their payment over time

Many studies have proposed design considerations and best practices to encourage participants to answer genuinely and remain committed (Soleymani et al 2010)(Kittur et al 2008)(Vliegendhart et al 2012) through incentives and proper interface design (Eickho et al 2012). For example, (Soleymani et al 2010) offers a set of recommendations and best practices to design an experiment in a crowdsourcing environment such as:

- The experiment should be conducted in two steps: the first identifies reliable workers to recruit to the second step, where the actual tasks from the experiment are performed.
- When workers have to perform several tasks, it is suggested to recruit during the first step five times more workers than needed to complete the main tasks.
- Expect that 75% of the workers recruited will not be interested in a long time commitment. A date can be specified by which the workers need to accept the main tasks and, after that, the employer can remove them and recruit others if needed.

Then, the practitioner is also encouraged to implement possibility for the participants to give feedback on the study (e.g. comments, contact form, forums, etc.). Since participants share between themselves their experiences with certain employers, appreciation proper design will be shared among participants, reinforcing the efficiency of the study (Hossfeld et al 2013).

Some studies have argued that scaling studies should be implemented in a joyful manner. Gamification has shown promising results, keeping workers motivated and reducing the number of fake ratings. The work in (Eickho et al 2012) reports that fake ratings can be reduced by 11.2% and that 57% more participants return to a game than to a regular crowdsourcing task.

Other studies instead have focused on developing statistical methods capable of identify unreliable workers. In (Riegler et al 2013) a reliability measure is presented to detect untrustworthy annotations in images based on the comparison between ratings and on the self-reported familiarity of the worker with the topic fashion of the annotation. The algorithm performs well although it is only suitable for bona fide ratings. In contrast, in the context of aesthetic preference it is difficult to unveil malicious participants from user ratings and opinions since there is way of validating an opinion or a preference.

In contrast, in (Hossfeld et al 2011) an indicator for the reliability of the results of a scaling tests is reported. The SOS hypothesis quantifies inter-observer reliability based on the standard deviation of the opinion scores (SOS) and on the mean opinion scores (MOS) both on the same stimulus:

$$SOS(I)^2 = a * (-MOS(I)^2 + 6 * MOS(I) - 5) \qquad (1)$$

$$MOS(I) = \frac{1}{N}\sum_{i=1}^{N} AA_i(I) \qquad (2)$$

$$SOS(I) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(AA_i(I) - MOS(I))^2} \qquad (3)$$

Where MOS(I) and SOS(I) stands for Mean Opinion Score and the Standard deviation of Opinion Scores respectively for the image I and a is a parameter that we want to retrieve from this equation. The MOS and SOS descriptions can also be seen in (2) and (3), respectively. Whereas N stands for the total number of participants, AA_i (I) stands for the aesthetic appeal score from participant i on image I.

The SOS hypothesis reflects the level of scoring diversity. The idea behind this hypothesis is that since all participants experience the same test conditions then, for properly conducted quality tests, user ratings differ only to a certain extent. For this reason, participant scoring diversity within a subjective study can be accurately described by the SOS parameter a. A good quality measure is therefore formed by the MOS together with the SOS parameter a, represented in (1). Further, in (Hossfeld et al 2011) it is showed that different application categories (Web, Video, Image, etc.) map to different ranges of the SOS parameter a. Hence, this parameter can be used as a reliability indicator for a whole subjective test, for a specific type of application being tested and for comparison of data sets from different quality studies.

Therefore, the parameter a needs to be compared to standard values reported in the study (Hossfeld et al 2011), where it was computed for various applications from several subjective measurement studies. If the indicator a is very large, it tells us that there is a high disagreement among observers across all images, which may in turn indicate that data was not yet completely filtered for unreliable participants.

Besides the above mentioned approaches, the existent studies refer the need to break down long tasks into small ones to prevent fatigue of the workers, to use multiple methods to track that the workers are doing a good job, for example, with a simple questions and recording observation time (Hossfeld et al 2013)(Soleymani et al 2010).

In short, one should combine these procedures to support an easy filtering with statistical methods like the SOS hypothesis for a more thoroughly inspection for malicious users.

2.3.4. The dataset selection

Independent on whether the study is to be run in a lab or crowdsourcing environment, quite a bit of attention should be paid to which image material will be collected. Quite often Computational Aesthetics involves experiences with images that become fundamental to get valid results. When investigating aesthetic appeal, researchers assemble their own datasets from on-line communities of professional and amateur photography such as Flickr (Bhattacharya et al 2010) (Jiang et al 2010) or Photo.net (Datta et al 2006)(Datta et al 2007). One example of such approach is referred in (Bhattacharya et al 2010), that describes an application that was developed to improve the aesthetics of an image based on composition rules such as rule of thirds and golden ratio. For the purpose of that work, 632 photographs were downloaded from Flickr and other free image sharing portals. It is important to note that, only single subject and landscape/seascape images were used. This can be seen as a limitation of the application that was developed or perhaps as a design consideration since those photography classes are the most commonly used to demonstrate composition guidelines in photography (Freeman 2007).

Another approach to build a dataset was implemented in (Jiang et al 2010) to explore automatic aesthetic estimation in two different tasks: the evaluation of fine-granularity aesthetic scores and the prediction of coarse-granularity aesthetic categories. For the purpose of the above cited work, 450 images were collected from a number of different sources: Flickr, Kodak Picture of the Day, study observers, and a collection of recently captured consumer image sets. The dataset had images including several classes: with and without people as the main subject, with one of two levels of main subject size, with one out of six possible levels of type of perspective cue, and including indoor, outdoor and natural and man-made subject matter. Likewise, in (Cerosaletti et al 2009) 450-image dataset was selected from the same sources and used to investigate the effect on aesthetics of main subject size, presence or not of people, and the type of perspective. Going now to the research described in (Isola et al 2011) that tried to determine what makes an image memorable, the dataset used was a random sample of categories from the SUN Database (Xiao et al 2010).

The SUN dataset as well as other famous datasets such as the Caltech 256 (Griffin et al 2007) and the Lotus Hill (Everingham et al 2010) datasets provide a standard collection of images and respective annotations for semantic categorization. Such databases are typically composed of images obtained by web crawling and annotated by crowdsourcing and can be handy for future work as exemplified in (Isola et al 2011). Nonetheless, when opting for one of them, it is important to grasp their limitations. For example, many of the images in the Caltech 256 dataset only cover single participants, and only a small part of the Lotus Hill is made freely available to researchers.

The IQLab database (Redi et al 2011)(Redi et al 2011) and the LIVE database (Sheikh et al 2006)(Wang et al 2004)(Sheikh et al 2005) are two worthwhile mentioning datasets since part of the images we used in your research were taken from these. While the IQLab database (Redi et al 2011)(Redi et al 2011) was designed to explore the deviations of quality scoring saliency from free looking saliency in (Redi et al 2011) and the LIVE database (Sheikh et al 2006)(Wang et al 2004)(Sheikh et al 2005), a dataset built to calibrate test quality assessment algorithms.

In summary, a clear definition of the research question should precede any decision on the dataset to be used. The more specific the research question is, the more carefully the dataset needs to be chosen. For example, the research of (Isola et al 2011) modelled how good people are at remembering images and some of their visual details, despite the overflow of visual information. Therefore, the dataset was chosen to include a wide and random range of the scene categories from the SUN database (Xiao et al 2010).

2.4. Quantifying aesthetic appeal: aesthetic attributes and images features

The decision of what is beautiful and what is not is definitely based on individual preferences.

When observers judge aesthetic appeal, users do not “see” features, they “see” attributes. Then, features are quantitative and physically measured from an image. Features can be linked to attributes by the use of algorithms or models. As an example, the definition of CIELAB system of colour coordinates uses a visual algorithm for Chroma (Engeldrum 2000) . The feature can be then the spectral radiance property of a coloured image. Using the adapted Image Quality Circle in section 1.1.1, Figure 3, we can move clockwise from Features to Attributes and map the spectral property to Chroma perceptual attribute. First we need to calculate the value of the reference white from spectral properties of the image and then we can apply the CIE defining equations (Engeldrum 2000) to achieve Chroma, i.e. the perceived intensity of colours in an image.

Aesthetic preference encompasses several attributes. One should first understand how various attributes, such as brightness, contrast, colourfulness or sharpness contribute to the overall image aesthetic appeal. Then, one should determine how each attribute, on its turn, is related to the objective measures of the image, i.e. the features. As a result, a set of features can be composed in an algorithm to predict an aesthetic attribute. This algorithm can be a formula, model, or computer program.

In this paragraph, we review the two possible ways of quantifying aesthetic appeal appreciation: the attributes of aesthetic appeal and the image features for predicting aesthetic appeal. Note, that we will not provide an exhaustive description of low-level visual attributes and features, but rather discuss meaningful features usage patterns.

2.4.1. Attributes of aesthetic appeal

Attributes are observers’ (visual) perceptions that form the basis of the aesthetic preference. An attribute is then a characteristic of an image as we perceive it, and based on which our judgment on the aesthetic appeal of the image may change such as the intensity of the colours in an image.

In fact, the overall image aesthetic appeal may be hypothesized to be a weighted combination (Axelsson 2007) of all relevant attributes, where the weights express the relative importance of each attribute to the overall image aesthetic appeal. Hence, determining the relation between image aesthetic appeal and its attributes implies that one has to measure the importance coefficients for all attributes. However, these weighted coefficients may depend on a number of aspects, among which e.g. the image content, the ambient light, the viewing distance, etc. This makes the determination of the relation between image aesthetic appeal and its attributes a complex multi-dimensional research problem. Several scenarios have been used since the traditional verbal judgments and multidimensional scaling of aesthetic value with other related attributes, to measuring behavioural, psycho-physiological, and neurophysiological responses to images in controlled and free viewing conditions (Joshi et al 2011).

At the present stage, no complete taxonomy exists of aesthetic appeal attributes. Nevertheless, several studies have tried to understand what more prominently affects aesthetic appeal judgments.

For example, in (Redi et al 2012) it was examined whether image integrity impacts the viewing behaviour, when scoring aesthetic appeal. They compared the aesthetic judgments obtained from the two groups of observers: one evaluated the aesthetic appeal of high integrity images, and the other the same images, but with lower integrity. Although asked not to take quality into account, the second group scored the low integrity images with lower aesthetic appeal than the first group. Namely, lower for highly distorted images and higher for slightly distorted images.

In (Bech et al 1996) based on interviews with a large number of people, the most important identified attributes for image quality were brightness, contrast, colour rendering and sharpness. An image should be bright, should have a high contrast, should be sharp and should have nice colours. In addition, it should not have any artefact, such as spatial distortions or shape deformations (Bech et al 1996). Besides brightness, all others attributes are complex; according to the opinion of interviewees the remaining attributes consist of a combination of various sub-attributes. In the case of colour, some of its sub-attributes were the saturation of the colours, the naturalness of the colours, and the homogeneity of the colours over larger areas. Certain attributes are also mutually related, as in the case of perceived contrast and sharpness. In theory, contrast is determined by how well a contour of an object is observed. This does not only influence the way we perceive contrast, but also the perceived sharpness. In a similar way, in (Bech et al 1996) contrast was clustered with *ratio between light and dark parts*, which shows that brightness and contrast are mutually related as well.

In the study (Peters et al 2007), six attributes that constitute aesthetic appeal appreciation are derived from the modularity of the human visual system (HVS): colour, form, spatial organization, motion, depth, and the human body. Here, attributes are explored in sub-attributes in a reflective study on how to enhance aesthetic appeal justified by the way how sensory information is processed by the HVS, as well as by traditional practices of artists.

Colour is referred to as an element that has a major influence on observers' judgments of aesthetic appeal (Axelsson 2007). Complementary colours and the use of dynamic range play with the perceived contrast, i.e. with the difference in colour or luminance that makes an element salient. Complementary colours are widely used in design and when placed next to each other, create a stronger contrast and reinforce each other, attracting our attention. Dynamic range describes the contrast ratio between the maximum and minimum measurable light intensities and it is commonly used in photography. More specifically, dynamic range is the amount of light that a camera can capture, which is usually lower than the human eye. In order to exploit the dynamic range with a camera, one needs to take multiple photos of a scene, at different shutter speeds, so that we can combine all the full range of light and create a beautiful image (Peters et al 2007).

Another important attribute reported in the literature is composition (Axelsson 2007). Composition denotes the placement or arrangement of visual elements or ingredients in a work of art and can be referred as spatial organization in (Axelsson 2007) or dynamics in (Peters et al 2007). Image composition consists of guidelines that have been widely used by many artists to create aesthetically pleasing images by easing the viewing behaviour of users when observing the image. The most popular guidelines have to do with the simplicity of the scene, the balance among visual elements and geometry (Obrador et al 2010)(Peters et al 2007), such as:

- The golden mean - the particular ratio of an asymmetric line division that can be seen in white in Figure 6 and it widely used in photography and arts (Peters et al 2007).

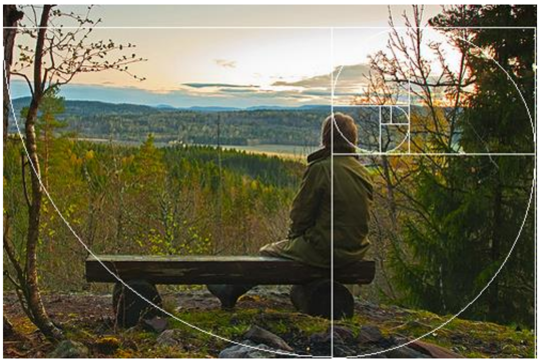


Figure 6: Example of the gold mean rule applied in photography.

- The rule of thirds - proposes that an image should be imagined as divided into thirds both horizontally and vertically, like a tic tac toe grid, and our subject should be aligned with the points where those lines cross (see Figure 7) (Obrador et al 2010).

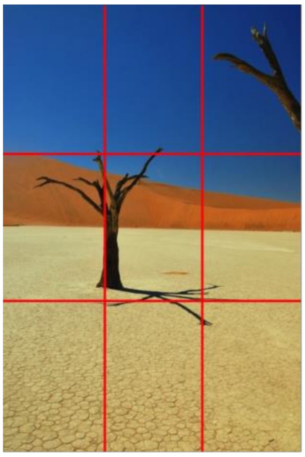


Figure 7: Example of the rule of thirds usage.

- Simplicity of the scene – the simpler an image can be the more pleasing it has been shown to be (Obrador et al 2010)(Peters et al 2007). As example, the Figure 8 presents an example of simplicity and beauty.

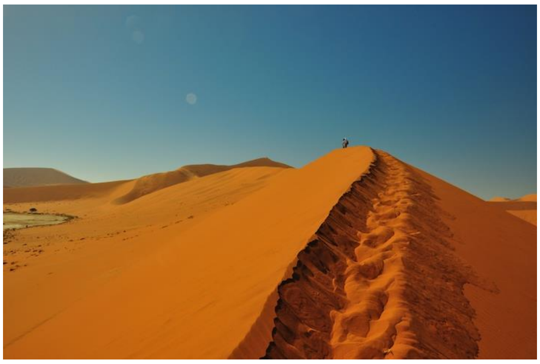


Figure 8: Example of simplicity of a scene.

In subjective aesthetic appeal judgment literature, one can find multiple references to familiarity with the content depicted in images (Lind 1980)(Carbon et al 2011)(Wagemans 2011)(Axelsson 2007)(Redi 2013)(Shapiro et al 2001) and, recently, recognisability and semantics of the content. Familiarity is described in (Shapiro et al 2001), as a sense of insight based on previous experience with an object or person. In general, familiar images are viewed as more attractive and processed faster and elicit less attention than novel images. When presented with two images at the same time, a familiar and a novel one, the familiar was rated more pleasant than the novel one (Mansilla et al 2011). Then, familiarity can also be related to experience (Mansilla et al 2011). In (Szechter et al 2007) the quality of children's aesthetic understanding of photographs was assessed, as well as the social interactions in the aesthetic domain between parents and children. Children's performance showed to be related to age, but neither to parents' art experience nor to the qualities of the photographs described by the parent. Also, individual artistic experience was correlated with aesthetic preference in agreement with the literature (Szechter et al 2007). On what concerns to content semantics, recognisability has also been shown to have an influence on aesthetic appeal appreciation in (Mansilla et al 2011)(Lassalle et al 2012) . Further, people's dislike for abstract paintings has been reported in (Congcong et al 2009)(Komar et al 1997) .

We should also consider saliency as an attribute of the image because it gives the perceived visual property of what is visual important and what is not. In fact, saliency was found to be related to aesthetics (Redi 2013). By comparing saliency data with aesthetic judgments in (Redi et al 2012) it was shown that the viewing behaviour when scoring aesthetic appeal differs from the free looking viewing condition. Further, we should bear in mind that it is not measured at one dimensional scale like the before mention attributes but it needs to be measured as a psychological measure.

2.4.2. Image features for the prediction of aesthetic appeal

Features are objective measures that can be computed/measured from low level properties of an image such as pixels and textures.

The semantic gap between these low-level computable visual features and high-level human-oriented attributes, visible in the adapted Image Quality Circle in section 1.1.1., Figure 3, presents a key challenge for researches.

The highly subjective nature of attributes related with aesthetics, makes it difficult to relate low-level computable visual features with these (Joshi et al 2011). Nevertheless, various research attempts have been made to model aesthetic judgment. For most features, we know how they affect a particular attribute. The opposite is not so certain. Take as an example the work in (Isola et al 2011) on predicting image memorability.

This works shows that image memorability depends on object and scene semantics. This, however, does not imply that we can predict memorability with only those and therefore for future research it is encouraged to investigate other attributes such as saliency and image quality.

The work in (Hasler et al 2003) tries to quantify colourfulness first using the CIELab colour space and later sRGB for a computationally more efficient approach. But first a category scaling experiment is setup where observers' judgments are collect on colourfulness in a 7-point category scale for ground truth purposes. Moreover, it is consider that a greyscale image has no colourfulness. To compute a colourfulness metric, the distribution of the image pixels for each colour space is studied. Additionally, it is assumed that image colourfulness can be represented by a linear combination of a quantities such as the trigonometric length of the standard deviation (5) in the variation across the red and green axis (7). The parameters for a linear combination are then found by exploiting the correlation between the experimental data and the metric. By combining different subset of quantities, the authors chose the best correlation. As a result, assuming that the image is in the sRGB colour space, the following metric is defined:

$$\widehat{M}^{(3)} = \sigma_{rgyb} + 0.3 * \mu_{rgyb} \qquad (4)$$

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} \qquad (5)$$

$$\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \qquad (6)$$

$$rg = R - G \qquad (7)$$

$$yb = \frac{1}{2}(R + G) \qquad (8)$$

Where each of the σ and μ represent respectively the standard deviation and the mean value of the pixel cloud along direction $\boldsymbol{rg}$ or $\boldsymbol{yb}$. The correlation between this metric $\widehat{M}^{(3)}$ and experimental data achieved 95.3%.

Clearly, low-level features apprehend useful information and using pixels characteristics and spatial relationships in images among them can be very helpful. Take the example of perceived contrast. It is not clear yet how to convey it through a combination of features. Perceived contrast is not only related to the ratio between white and black, but also on local luminance differences between grey scale values. However, in (Matkovic et al 2005) overall contrast is defined by computing contrast ratio at various resolution levels. This proposed metric was designed towards a more complete solution based on a previous concept and conducted with a subjective experiment. Referred as Global Contrast Factor (GCF) it can be estimated as a weighted average of local contrast factors at various resolutions:

$$GCF = \sum_{i=1}^{N} w_i * C_i \tag{9}$$

$$w_i = \left(-0.406385 * \frac{i}{9} + 0.334573\right) * \frac{i}{9} + 0.0877526 \tag{10}$$

$$C_i = \frac{1}{w * h} * \sum_{i=1}^{w*h} l_{c_i} \tag{11}$$

$$l_{c_i} = \frac{|L_i - L_{i-1}| + |L_i - L_{i+1}| + |L_i - L_{i-w}| + |L_i - L_{i-w}|}{4} \tag{12}$$

$$L = 100 * \sqrt{l} \tag{13}$$

$$l = \left(\frac{k}{255}\right) * \gamma \tag{14}$$

Where k, with $k \epsilon \{0 \dots 255\}$ denotes the original pixel value, $\gamma = 2.2$ stands for the gamma for standard displays. First, the pixels' original values are scaled to [0,1] range and corrected with the gamma correlation, leaving us linear luminance l. The perceptual luminance L is then computed (13), followed by local contrast l_{c_i}. Assuming the image to be w pixels wide and h pixels high and to be organized as a one-dimensional array of row-wise sorted pixels, local contrast $l_{c_i}$is for each pixel, the average difference of L between the pixel and four neighbour pixels as in (12) for pixel i. For edge-pixels only the available neighbour pixels are taken into account. After, the average local contrast for current resolution C_i is computed as the average local contrast l_{c_i} over the whole image. The C_i needs to be computed for various resolutions. In (Matkovic et al 2005) 9 resolutions were chosen, where each smaller resolution combines 4 pixels from the resolution before (except the original) into one super pixel. The drawback of this proposed feature is that it only focuses in grayscale images. Colour contrast is left for future work since it is a more complex problem and it is not only inferred by the grayscale contrast.

2.4.3. From image features to attributes and aesthetic appeal: existing models

When reasoning about aesthetic judgments, observers usually support a high rating with "interesting composition" (Joshi et al 2011). Therefore, researchers have engaged in addressing more challenging problems such as relating aesthetic appeal judgments in images with low-level image composition (Bhattacharya et al 2010)(Datta et al 2007)(Datta et al 2006). The study in (Bhattacharya et al 2010) has focused on enhancement of image aesthetics through suggestions of better composition based on two famous composition guidelines learned rules, golden mean and rule of thirds. With a different goal, the study in (Datta et al 2006) has also used composition guidelines, specifically the rule of thirds, to analyse its relevance to aesthetics. The rule of thirds was measured through the average hue, saturation, and intensities within the inner third region of a photograph. This work assessed the quality of score prediction using images visual content as a two-class classification problem for high and low scores, respectively. This can be considered a reasonable choice because intuitively it is difficult to distinguish between some variance within user ratings. The regression problem was therefore changed to one of classification, by threshold on the average scores to create high-versus low-quality image classes. It has then found that if the score-gap between the two classes is are then more easily separable, the classification performance improves (Datta et al 2006).

An alternative might be keep a sample of aesthetic pleasing photos or eliminate low-aesthetic ones, as proposed in (Datta et al 2007) from a wide dataset. In that case, it is essential to guarantee that the selected images are of high aesthetic appeal even though many of those not selected may be of high aesthetic appeal as well. The work in (Datta et al 2007) has proven this approach to be more successful than the general two-class classification problem. Further, the limited success of the first approach acknowledges the fact that score prediction is a difficult task (Datta et al 2006).

As far as predicting the aesthetic quality of paintings is concerned, the work in (Congcong et al 2009), with a similar approach as in (Datta et al 2006), tries to infer aesthetics of paintings based on judgments from non-expert observers with a two-class classification problem as well. Here, feature selection is based on the belief that people use a top-down approach to appreciate art. A more general impression is first gathered, followed by an examination of the details. A combination of global and local features has then been analysed. Supported by art and psychology, the set of features chosen include blur distribution, seen as an important artistic effect, and detection of edges, used by artists for emphasis (Congcong et al 2009). Further, the work in (Congcong et al 2009) and in (Datta et al 2006) make use of machine learning methods to learn two-class (high and low aesthetic appeal) classification patterns among observers' judgments. An important future research direction should anyway be to incorporate cultural, social, and personal differences into the aesthetic appeal appreciation models. Whereas there are some rationalization entailed for feature design with respect to the aesthetic attributes inference problem, designing features capable of capture emotions is a bigger enigma (Joshi et al 2011).

3. Investigating the relationship between physical image characteristics and aesthetic preferences

As a first step towards investigating the effect of digital filters on image aesthetics, it was necessary to create a representative image dataset that would include sufficient variability in image features as well as perceived attributes and aesthetic appeal. To do so, it was necessary to first clarify the relationship existing between these three quantities. As reported in Chapter 2, this relationship is described in an incomplete way, in literature. Thus, as a first step, we designed a subjective experiment devoted to get more insight into the link between physical image characteristics, image attributes, and aesthetic preferences. Since context factors (Le Callet et al 2012) like the illumination of the screen, luminance in the room, viewing distance and viewing angle of the display may have impacted the results, we decided for this first step to perform the experiment in a controlled lab environment. Also, since we were interested in investigating the relationship between visual attention deployment, image characteristics and aesthetic preferences, our experiment involved the use of eye-tracking equipment, thus performing it in a lab setting was the most obvious choice.

Below we detail first which image attributes and physical characteristics we were interested in investigating; we describe then the setup of our experiment, the procedure adopted to process the data, and finally we report the outcomes of the analysis of the data collected.

3.1. Image attributes

As detailed in the section 2.4.1, attributes are those *perceived* characteristics of an image that influence the appreciation of aesthetic appeal. We sought to collect subjective ratings on the dataset in terms of the attributes colour likeability, familiarity, recognisability and also saliency, as well as to infer their relationship with aesthetic appeal.

- Colour likeability reflects the user preferences in terms of colour rendering. It is interesting to investigate because the impact of colour on aesthetic appeal appreciation has been reported often in subjective studies literature, e.g. in (Peters et al 2007)(Joshi et al 2011). These studies can almost be grouped into two strategies: studies where participants are asked to justify their preference choices and studies where the images is chosen based on colour (Joshi et al 2011). Few are the studies that have inquired participants directly on their colour preferences in the field. Thus, we decided to add this attribute in our experiment to acknowledge the reported effect using another approach..
- Familiarity reflects the extent to which the content of an image is familiar to the observer, i.e. how often has the participant seen the content of the image. As for colour, familiarity has also been shown to be correlated with aesthetic appeal appreciation (Lind 1980)(Datta et al 2006), but few have been the studies that assessed participants’ familiarity with the images with the purpose of studying it (Congcong et al 2009).
- Recognisability is defined as the ease of recognition of the image content by the observer, i.e. how clear the subject of the image is to the observer. For example, the content of images presenting some kind of distortion, such as blur, might be difficult to recognize. In fact, in (Mansilla et al 2011)(Congcong et al 2009), content recognisability was found to have an influence on aesthetic appeal. It should be noted that what is familiar has to be recognizable, whereas what is recognizable may or not be familiar. In other words, when designing this experiment it was expected that participants could experience some difficulties when asked to distinguish recognisability from familiarity. Nevertheless, having to rate both attributes separately would force participants to conceptualize them, thus helping understanding what familiarity is and how does it relate to recognisability .
- Saliency - This attribute describes which areas or parts of an image are most likely to catch a viewer’s attention. Visual attention has shown to have a high added value in quality assessment (Engelke 2011), e.g. in (Alers et al 2010), it was found that artefacts located in the regions-of-interest (ROI) of an image are more likely to be noticed and therefore more annoying for observers. The ROI are the most important parts of an image which mainly attracts observers’ attention (Wang et al 2010) . Therefore, computing the distortion visibility

measurements with saliency information only in the ROI might lead to computational performance improvements and cost savings (Engelke 2011). Similar principles could be applied in Computational Aesthetics. Furthermore, (Lai-Kuan et al 2009)(Bhattacharya et al 2010), one can relate compliance with compositional rules to the deployment of visual attention (or models of it).

3.2. Physical image characteristics (Features)

Digital Filters change physical quantities in an image such as colourfulness and contrast. Understanding the relationship between these quantities and attributes, and later aesthetic appeal, would be therefore of added value for setting up our main study. Thus, we used features from (Hasler et al 2003) and (Matkovic et al 2005) to quantify colourfulness and contrast attributes through the equations (4) and (9) described in section 2.4.2., respectively.
The contrast feature was also in this case computed at 9 different resolutions, and the coefficients C_i from equation (9) in section 2.4.2., where $i \epsilon \{1 \ldots 9\}$ were combined in a single contrast score by using the weighting factors indicated in (Matkovic 2004).

A second interesting aspect to look at was the link between image characteristics, composition, saliency and eventual aesthetic preferences. To quantify this we defined two features that could shed light on this relationship.
In a previous work (Redi et al 2013), we investigated the relationship between aesthetics, composition and visual attention deployment focusing on subject simplicity and on the famous rule of thirds commonly used in photography and described in section 2.4.1 (see Figure 7 in section 2.4.1 and Figure 9B below). It was found that the distance of the centre of attention from the vertical lines of thirds is predictive for aesthetics. Subject simplicity was also analysed in terms of low image clutter. The amount of clutter is an important feature that impacts main subject aesthetic appeal appreciation in consumer images (Cerosaletti et al 2011), which was confirmed (Redi et al 2013), to diverge attention and negatively related with aesthetic appeal.
Besides the rule of thirds depicted in Figure 9B), there are more compositional rules worth of investigating. The concept behind Figure 9A) is again simplicity but in terms of number of identifiable regions in an image. The lower the number of coherent regions in which the image can be divided, the higher the aesthetic appeal. Visual coherence denotes a sense of unity linking the different regions of an image (Berdan 2004). However, exaggerated use of visual coherence can also lead to boredom.

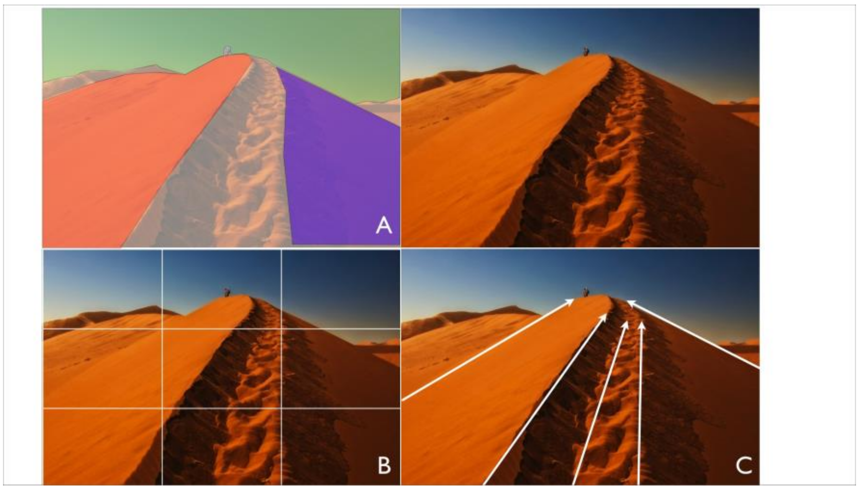


Figure 9: Some examples of interesting compositional rules.

The following steps were taken to compute image coherence in an image:

1. To distinguish the different regions that each image entails, a Gaussian filter was applied followed by a circular averaging filter, creating a double blur effect. While the first filter reduced the image noise and detail, the second filter smoothed the edges of the regions blurred by the first.

2.To then extract each region, we applied an algorithm that simulates the Adobe Photoshop tool magic wand[20]. This function allows the selection of connected pixels whose colours are within a defined tolerance of reference pixels and it was applied in several spread pixels of the image to spot all possible connected regions. Note that this function outputs a binary map for each extracted region.
3. All the extracted regions were afterwards checked for overlaps between each other, to bridge possible connected regions that were originally belonging to the same region but split due to different colour levels.
4. Finally, our simplicity factor was defined as the total number of regions identified in an image:

$$S = N_R \tag{15}$$

Where N_R stands for the number of regions outputted from the previous step.

Lastly, we designed a feature to quantify the compliance of the image to the composition rule called converging lines (visible in Figure 9C). Converging lines are used by photographers to guide the viewer's eyes through the image[21] towards the main image subject. Lines are often arranged to converge to the image main element, so that the observers' eyes can follow them to easily find the most important area of the image.
Again, we will list briefly the procedure we used to detect the converging lines and check whether they intersect with the largest blob in the ROI per image:

1. Again, a Gaussian filter followed by a circular averaging filter was applied to merge possible connected regions and smooth their edges.
2. Second, the Canny method was employed to find the lines in the image (edges) by looking for local maxima of the gradient in an image. This method was judged to be the most suitable for the task as it uses two thresholds, one to identify strong edges and other to identify weak ones. The weak edges are only incorporated in the output if they are connected to strong edges; therefore, it is more robust to noise. This is a desirable property for our application since allow us to follow a vague section of a given line and to discard a few noisy pixels that do not constitute a line but have produced large gradients.
3. Next, we ran the edge output through the Hough transform (Shapiro et al 2001) to extract the prominent lines in the image. The Hough transform is a feature extraction technique used in image analysis and designed to detect lines in an image, using the parametric representation of a line:

   $$r = x * \cos\theta + \mathrm{y} * \sin\theta \tag{16}$$

   Where r represents the distance between the origin and the line along a vector perpendicular to the line and θ the angle between the x-axis and this vector.
   The final result of the transform is a matrix, whose rows and columns correspond to these r and θ values respectively. The matrix peak values represent potential straight lines in the input image.
4. Finally, we determined how many lines intersected the centroid of the largest blob in ROI in the saliency map and we formalize our converging lines metric L as follows:

   $$L = \frac{Number\ of\ lines\ crossing\ the\ ROI}{Total\ number\ of\ lines\ extracted} \tag{17}$$

[20] http://mathworks.com/matlabcentral/fileexchange/4698-simulating-photoshops-magic-wand-tool
[21] http://digitalcameraworld.com/2012/04/12/10-rules-of-photo-composition-and-why-they-work/

3.3. Experimental setup

We decided to run an exploratory experiment to scale a large range of images according to the different attributes described above and aesthetic appeal. For that purpose, we decided to investigate attributes and aesthetic appeal through a within subject design pilot study. We showed a set of 200 images to all of our 14 experiment participants and asked them to rate the images on the mentioned dimensions through a numerical scale.

3.3.1. The dataset

The experiment was targeted at understanding aesthetic preferences in regular consumer photographic images. We needed therefore to build a dataset which would constitute a representative sample of such type of images. To achieve this, it was mandatory for the dataset to reflect a wide range of subject categories, typical of both professional and amateur photography. We collected therefore images belonging to 16 of the most popular categories from the website 500px[22], namely:

- Abstract: The choice of including abstract images in our dataset was to encompass a low level of recognisability . Further, in (Congcong et al 2009), abstract paintings were found to be less likely appreciated by people with respect to immediate works of art.
- Animals: Animals are often present in our daily life and photos. Besides, this category is present in the literature (Bhattacharya et al 2010) and in object-annotated datasets used by researchers in computer vision, human perception, cognition and neuroscience, machine learning and data mining, computer graphics and robotics (Everingham et al 2010).
- Celebrities: Images of famous places and people were also included to encompass a high level of familiarity amongst the observers.
- City and Architecture: Architecture and cities attract many people for different reasons: for living, for tourism, for studying. This diversity of perspectives on the same on the theme will cause, according to the specific participant's experience, either a high level or low level of familiarity.
- Fashion: We decided to have images depicting fashion because of their dominant presence in social networks (Hoffman 2013), contributing for a representative dataset of the images one can find in the Web.
- Food: Likewise, food is also a dominant content in the Web (Thomas 2013) and social networks, which makes this category to be considered important.
- Landscapes: This is another category considered in Computational Aesthetics literature and in the Lotus Hill database (Everingham et al 2010) for computer vision research. Also, observers usually tend to prefer landscapes images (Wagemans 2011)(Joshi et al 2011). On the other hand, the definition of landscape photography is broad and may include urban settings, industrial areas and nature photography.
  Because of that, we have decided to use both landscapes and nature categories.
- Macro: Macro is a common photography category and entails extreme close-ups of usually very small participants. That means that the distance range from a camera (Depth of Field, DOF) is very small (Joshi et al 2011) and, in (Datta et al 2006), a low DOF was found to be aesthetic pleasing in appropriate context. Adding this images to our collection will tell us if macro photography is an "appropriate context".
- Nature: Similar to landscapes, images from this category is mainly used in Computational Aesthetics studies (Joshi et al 2011) and in a computer vision databases like the one in (Xiao et al 2010).
- People: Besides, being another important category used in research and object-annotated databases, it is interesting to study how people assess other people. At the same time, we add another level of low familiarity.

[22] http://500px.com

- Sport: This category has not yet been explored by aesthetics research although we decided to include because of the high amount of people following football, tennis, and so on..
- Still Life: Also, used in literature (Ishai et al 2007), this type of images depicts inanimate subject matters.
- Street: This category originates from street photography and features the human condition within public places, with or without people, a street or even an urban environment. It is widely explored by professional photographers.
- Transportation: This category contains images with vehicles. The reason why we decided to include these is because frequently we are bombed with images trying to sell us a car. Besides fashion and food, cars can also be found in our social networks, but less.
- Travel: As transportation, travel photographs were included as well because they are also present in our social networks, in general advertisement and in our travels to exotic places. Therefore, we decided to include it as well.
- Urban Exploration: Refers to the exploration of man-made structures, usually abandoned ruins or not usually seen components of the man-made environment. This category was then added due to its popularity amongst photographers.

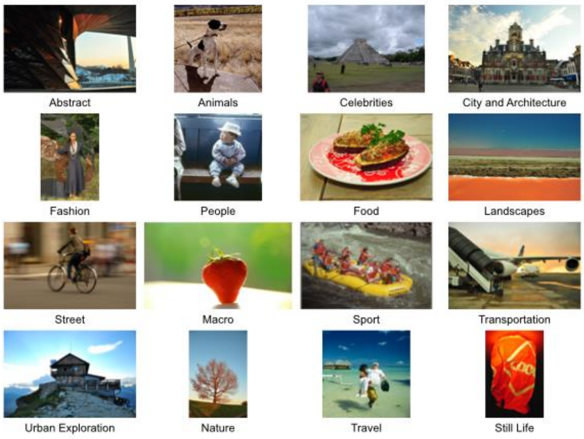


Figure 10: Sample of representative images for each image category.

Overall, these categories were selected so that they would (1) resemble semantic categories used in computer vision literature like the LHI (Lotus Hill) Image Dataset (Everingham et al 2010), such as Landscapes and Sport (2) frequently recur in social networks, (e.g. Food (Thomas 2013) and Fashion (Hoffman 2013)) and (3) encompass different levels of familiarity, as it will be the case of Abstract at one end and Celebrities at the opposite end. We adopted the first criteria to be able to compare our results with the documented ones. At first, we sought at only labelling the images from the MIT SUN dataset (Xiao et al 2010) on scene, places and objects within but soon we realised the difficult entailed on the task. Not only would be difficult to compare scenes with places but also difficult to have a significant number of types of scenes, places or objects, i.e, having a set of images of only auditoriums and still be able to have a balance range in colour, recognisability and familiarity. Thus, the dataset was selected based on the above-mentioned 16 categories from the website 500px for both expert and amateur photography. The sample size for this experiment has been limited to about 200 image. From those 200:

- 118 from an amateur photographer collection[23]
- 56 from the IQLab database (Redi et al 2011) and the LIVE database (Sheikh et al 2006)(Wang et al 2004)(Sheikh et al 2005)
- 26 from the Google image engine

In order to meet our goal to collect subjective scores on the mentioned attributes, one should consider a balanced range of each attribute in the image collection. Therefore, we took in consideration content that usually is associated with low familiarity or recognisability in the literature (Congcong et al 2009) and a big set of raw images, straight from the amateur photograther's camera without any post

[23] http://gplus.to/markdekker

processing, containing all the information that the sensor captured (usually less colourful and with less contrast).
Finally, after labeling the images, we asked two participants to re-label them, to achieve a better categorization on the image dataset.

3.3.2. Apparatus

The experiment took place in a room with constant illumination at approximately 70 lux, each participant accessed the images on a 23” LED backlight monitor having a resolution of 1360x768 pixels. During the quality scoring, the eye-movements of the participants were recorded to later measure saliency. For that purpose, a *SensoMotoric Instruments* GmbH Eye Tracker with a sampling rate of 50/60 was used, with a pupil tracking resolution of 0.1◦, a gaze position accuracy of 0.5 to 1, and an operating distance between the subject and the camera of 0.4 to 0.8 meters. Also, participant’s face movements had to be constrained by a chinrest place at a distance of 0.7 meters from the display.

3.3.3. Experimental methodology

Participants were asked to rate colour likeability, familiarity, recognisability and aesthetic appeal using the Single Stimulus (SS) method (from ITU-R BT.500 recommendation (ITU 2012)) with a discrete numerical scaling. The experiment was run one subject at a time, and each subject was requested to assess 200 images thus following a within-participants design.

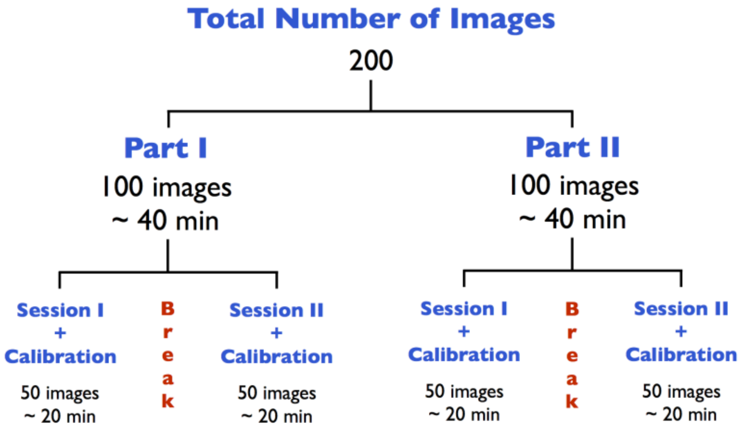
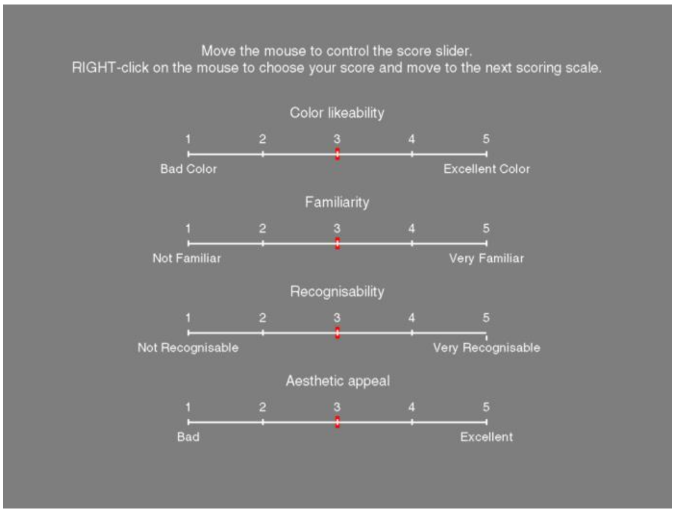


Figure 11: Overview of the structure of the experiment.

To minimize memory and fatigue effects, the experiment was split in 2 parts of approximately 40 minutes with 2 sessions each. Each session started with a calibration of the eye-tracker on the participants gaze based on a 13-points grid. After, the participants went through 4 short training sessions (one per attribute, with 3 images each) to make them acquainted with the scoring task interface and the meaning of the attributes they were asked to score. Each attribute was scored through a slider with a discrete 5-point scale from 1 (lower bound) to 5 (upper bound), with 3 being the mid value. Once the training was concluded, the actual experiment started. In a following similar set ups in the literature (Redi et al 2011), before presenting each image, a white cross was shown at the centre of the screen for 1 second. Images were shown in a randomized order, different for each participant and the observation time was not constrained, but recorded. After examination of each image, participants could access the scoring scales in a separate scoring screen (see Figure 12). Participants could enter their judgment by simply positioning the slider on the appropriate score number through the computer mouse. After scoring all attributes, another image would be displayed, and the process would repeat until the training or the experiment completed for each participant. Moreover, the exact step-by-step experimental procedure for each participant, which can be seen online[24] .

[24] http://prezi.com/-jz3c0we3vcc/?utm_campaign=share&utm_medium=copy&rc=ex0share



Move the mouse to control the score slider.
RIGHT-click on the mouse to choose your score and move to the next scoring scale.

Figure 12: The scoring screen used.

3.3.4. Participants

In total 14 participants were recruited through the poster (see Appendix C) from the student population at the TU Delft. From 14 participants: 6 expert observers and 8 naive observers. In this context, expert observers mean observers with some knowledge in photography post-processing, which are more sensitive about photo content than a normal observer (naive).

3.4. Data analysis

Once the data collection process from the lab experiment finished, we got the following types of data per image and participant:

- Four scores, for familiarity, recognisability , colour-likeability and aesthetic appeal, respectively
- Observation time (excluded the scoring time)
- A set of fixation points and saccades

It was then time to analyse the data to set the foundation for the main experiment. Each collection of single aesthetic appeal, colour-likeability, recognisability and familiarity scores, each ranging from 1 to 5, was processed according to (ITU 2012). First, they were scanned for outliers. One outlier participant was detected on the recognisability scores, and all his scores were then excluded from the following analysis. Scores were then normalized per participant; the z-scores ranges of each attribute collection of z-scores can be seen in the Table 1.

	Aesthetic Appeal Z-scores	**Colour likeability Z-scores**	**Recognisability Z-scores**	**Familiarity Z-scores**
Minimum	-3.01	-2.73	-5.11	-4.99
Maximum	3.47	3.01	1.58	2.51

Table 1: the z-scores ranges of each attribute collection of z-scores.

To analyse scoring inter-observer consistency, we computed, per each attribute separately, the standard deviation across the z-scores given to the same image by the different participants, and then averaged it across all image. This resulted into the values reported in Table 2. Inter-participant consistency states the level of understanding of the task and of the underlying image construct to be rated (either recognisability, familiarity, colour likeability or aesthetics). Clearly, participants were very consistent when rating each of the four attributes.

	Aesthetic Appeal Z-scores	**Colour likeability Z-scores**	**Recognisability Z-scores**	**Familiarity Z-scores**
Inter-observer	0.82	0.82	0.66	0.77

consistency				

Table 2: Inter-observer consistency for each attribute.

The value of 0.82 on aesthetic appeal z-scores (AAZ) corresponds to 12% of the aesthetic appeal range covered by the AAZ, in line with previous results in the field (Redi 2013).

Our next step was to transform the normalized scores into normalized mean opinion scores (MAOZ) per image and per attribute, according to the following formula:

$$MAOZ(s,I) = \frac{1}{N}\sum_{i=1}^{N} A_0(s,i,I) \qquad (18)$$

Where N denotes the number of participants, $A_0(a,i,I)$ stands for a single score given by the participant i with $i \epsilon \{1 \dots 14\}$ per image I with $I \epsilon \{1 \dots 200\}$ on scale s with $s \epsilon \{f,r,c,a\}$, one of the four attributes familiarity, recognisability , colour likeability and aesthetic appeal.

3.4.1. Eye tracking data analysis

We processed eye-tracking recordings in order to collect information on both eye movements and attention deployment. With respect to the latter, we processed fixation data according to (Redi et al 2011) to obtain, per each image, visual importance information in the form of saliency maps. Saliency maps (Koch et al 1987) represent the probability, pixel per pixel, that a location in the image is attended by the (average) observer. As such, they outline the areas in the image, which attract most attention. We believe this information can be helpful in our analysis for two main reasons. First, they may provide a powerful tool to estimate simplicity, in terms of how visually crowded (how many areas of the image attract attention, as a measure of clutter, or low simplicity) is the image. Second, it is commonly assumed that highly salient areas correspond to the most important elements in the image. Photographers intentionally compose images so that visual attention is driven to these elements; an analysis of salience could therefore reveal the compliance of an image to these rules, to be later matched to an actual benefit in terms of aesthetic appeal.

The following steps were performed to create saliency maps from raw eye-tracking data:

1. All fixations lasting less than 100 ms were discarded from the recordings
2. For each image I of size W_I x H_I, locations fixated by every observer were identified and added to a fixation map $FM^{(I)}(x,y)$, eventually gathering all fixation points from all observers
3. $FM^{(I)}(x,y)$, was then smoothed by applying a grey scale patch with Gaussian intensity distribution whose variance (σ) was approximating the size of the fovea (~2° of visual angle). The resulting saliency map element $SM^{(I)}(k,l)$, at location (k,l) was therefore computed as:

$$SM^{(I)}(k,l) = \sum_{f=1}^{N_f} \exp\left[-\frac{(x_f-k)^2+(y_f-l)^2}{\sigma^2}\right] \qquad (19)$$

   with (x_j, y_j) being the pixel coordinates of the fth fixation ($f=1 \dots N_f$) in $FM^{(I)}(x,y)$, and $k\in[1, W_I]$, $l\in[1, H_I]$.

We also produced binary versions of the saliency maps SM, in order to isolate the Region(s) of Interest (ROI) of the image. To compute our Binary Maps (BM) we performed the following extra steps:

4. A saliency threshold th^S was determined, common for all maps, as one third of the maximum saliency value across all maps. A threshold for saliency was preferred over a threshold for the size of the ROI area (as used in other works, e.g., (Alers et al 2010)), in order to isolate areas that were equally salient across all images. Of course, the value of the threshold was established in a somewhat arbitrary way and changes in the threshold may affect the results reported in the following section. We demand to future studies further investigations on these aspects.
5. For each image I, its binary map $BM^{(I)}$ was determined as:

$$BM^{(l)}(x,y)=\begin{cases}1 & if \quad SM^{(l)} > th^{S}\\ 0 & otherwise\end{cases} \qquad (20)$$

3.5. Results

The first step to analyse the relationship between the defined features, attributes and aesthetic appeal was to check the mean opinion scores distributions for normality with the Shapiro-Wilk test. This test tells us the probability of which the collected scores were obtained from a normal distributed population. This test revealed that, except for familiarity and colour likeability scores, the distributions to be not normal.

Second, we checked for inter-participant reliability.

Reliability in a psychometric test can be measured with the Cronbach's alpha coefficient (Robson 2002) on the z-scores. Besides inter-participant reliability, we used the Cronbach's alpha also to check if all the samples were distinct measurements or not. A 0.7 or higher value is usually acceptable for consistency.

A lower level of alpha was found between all the attribute scores, whereas when computing it between participants, separately per each attribute, the value was found to be consistently above 0.7 (see Table 3). These values confirm that the four attributes we considered portray independent underlying constructs and that were sufficiently consistent when rating each of them.

	Aesthetic appeal	**Colour likeability**	**Recognisability**	**Familiarity**	**All**
Reliability	0.849	0.844	0.926	0.890	0.594

Table 3: Cronbach's alpha for each attribute separately and joint.

Finally, we tested whether the level of expertise of the participants (Photographers vs naïve) did have any effect on the judgment strategy. A Mann-Whitney U-test revealed that there was no statistically significant difference between the expertise levels amongst observers when scoring aesthetic appeal (U = 797002, p = .872), colour likeability (U = 794802, p = .780), recognisability (U = 793146, p = .713) and familiarity (U = 781138, p = .311).

3.5.1. Relationship between attributes and aesthetic appeal

To better understand the relationship between the attributes we measured and their impact on aesthetic appeal, we computed the Pearson correlation coefficients between the MAOZ of different attributes on the aesthetic appeal MAOZ. These can be seen in Table 4, where statistically significant correlations ($p < 0.05$) are represented in green cells and not significant in red cells.

During the experiment, we recorded the participants' time spent observing per image. Accordingly, we were able to explore the association between the average time spent observing an image and each attribute scoring behaviour. For instance, familiar images has shown to take less time to be observed (Mansilla et al 2011). Although, we have not found any statistically significant association between average observation time and the four attributes mean opinion scores, so correlation of attributes with observation times is not reported in the tables and graphs below for ease of reading.

	Colour likeability	**Familiarity**	**Recognisability**
Aesthetic appeal	0.856	0.093	0.189
Colour likeability		0.061	0.044
Familiarity			0.704

Table 4: Pearson correlation coefficients between the different attributes.

Quite interestingly, colour likeability was found to be highly predictive for aesthetic appeal, whereas recognisability and familiarity were found to be weakly or not significantly correlated with it. On the other hand, familiarity and recognisability were found to be related ($r > 0.7$).

To get more insight in these results, we analysed partial correlations for all the statistically significant associations. This analysis allows to infer the correlation between two variables, when the linear effect of another independent variable has been removed from that relation (Norusis 1990) . Thus, we computed the partial correlations for each of the four the statistically significant relations, controlling all the possible combinations of the remaining attributes. Figure 13 presents the correlations between colour likeability and aesthetic appeal and between recognisability and familiarity, respectively. These relationships were not affected by the control of other attributes. As a consequence, it is clear that the correlation between colour likeability and aesthetic appeal is not due to their relation with recognisability or familiarity and the same applies for the correlation between familiarity and recognisability in terms of aesthetic appeal or colour likeability. Figure 13 also shows the partial correlations between recognisability and aesthetic appeal. It is clear that the original weak positive correlation gets stronger when colour likeability is held constant. That means that colour likeability is a suppressive variable, meaning that people are more attracted to recognizable images when they like their colour.

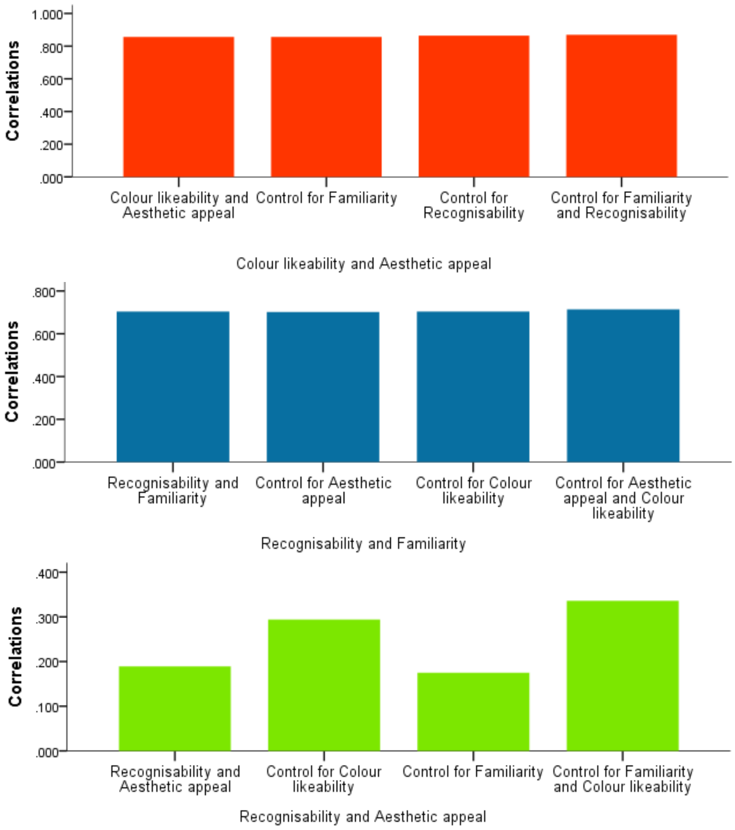


Figure 13: Presents the correlations between colour likeability and aesthetic appeal, between recognisability and familiarity and between recognisability and aesthetic appeal, respectively.

3.5.2. Relationship between image features, attributes and aesthetic appeal

The second important goal of this study was to understand better how physical properties of the image would relate to attributes and then aesthetic appeal. Thus, we computed the colourfulness (Hasler et al 2003) and contrast (Matkovic et al 2005) features for each image in the dataset and observed whether they had a linear relationship with the corresponding attributes and aesthetic appeal MAOZ. Table 5 shows the non-significant correlations ($p > 0.05$) in red cells, significant correlations ($p < 0.05$) in green cells and moderate significant correlations ($r \gtrsim 0.3$) in bold.

	Contrast	**Colour likeability MOAZ**	**Familiarity MOAZ**	**Recognisability MOAZ**	**Aesthetic appeal MOAZ**	**Average observation time**
Colourfulness	-0.157	**0.348**	0.132	-0.102	0.106	-0.184
Contrast	-	**-0.304**	0.127	0.175	**-0.292**	**0.284**

Table 5: Pearson correlations between features and attributes, with non-significant correlations (p > 0.05) in red cells, significant correlations (p < 0.05) in green cells and moderate significant correlations in bold.

We expected that contrast would be similarly correlated with aesthetic and colour likeability scores, because these two are strongly correlated in turn. However, for colourfulness there is a weak non-significant correlation with aesthetics, the correlation is significant moderate with colour likeability.

Possible reasons might have to do with the way the feature is computed or because the sample size was too small to infer a similar correlation.

To identify possible intervening attributes and detect hidden relationships in the four moderate correlations, we again computed partial correlations, as previously done in section 3.5.1. The statistically significant partial correlations were plotted in Figure 14.

From examination of the partial coefficients of each Figure, it is worth mentioning that:

- The correlation between contrast and aesthetic appeal seems to get stronger or unaltered in most cases, except when controlling for all variables. In particular, it seems to get stronger for recognisability. It might be then that recognisability is a suppressive variable and thus, people prefer high contrast images when the subject in the image is clear to them (see section 3.1. for the meaning of recognisability).
- The initial negative moderate correlations between contrast and colour likeability gets visibly weaker when controlling for aesthetic appeal, meaning that this correlation is spurious. This means that effects of aesthetic appeal on the two other variables leads one to conclude, in error, that the two other variables are causally linked.
- The small variation of the different correlation coefficients between contrast and observation time is too small to understand if there is a spurious correlation or a hidden correlation.
- Further, the initial weak correlation between colourfulness and colour likeability gets stronger, when controlling for aesthetic appeal. Thus, it might be that aesthetic appeal is a suppressive variable, which means that people like more the colour of more colourful images when they are attracted to them.

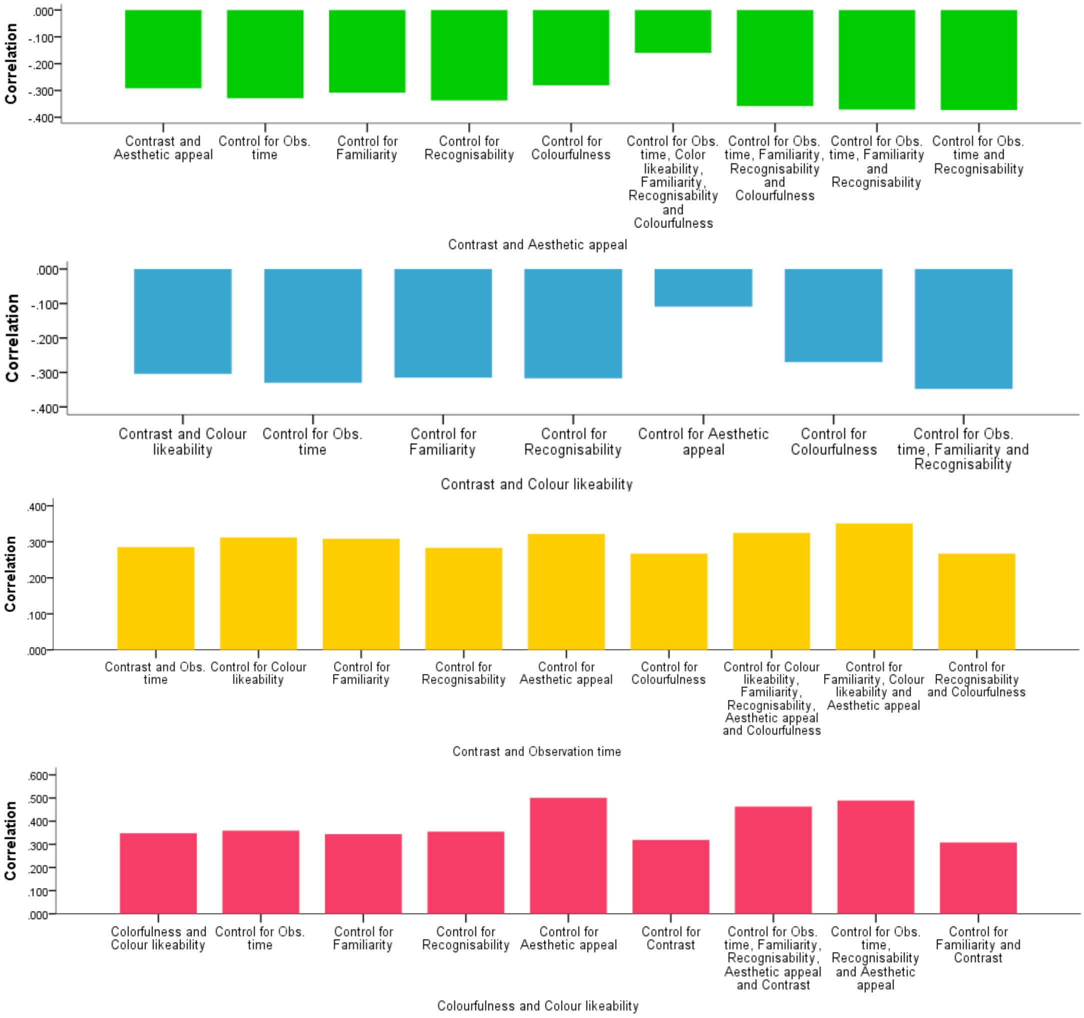


Figure 14: The statistically significant partial correlations plotted.

To better appreciate the above significant relationships, we discretized the feature values into a limited number of colourfulness and contrast categories. For colourfulness, we mapped our results into a 6-point scale based on the 7-point category scale reported in (Hasler et al 2003), from 1 (not colourful) to 7 (extremely colourful). We excluded the first category that would return 0 for images not colourful, i.e. in grayscale, due to absence of this category in our dataset. For contrast, it was no categorization scheme was proposed in (Matkovic et al 2005) and thus we mapped the different levels of contrast of our results into 5 groups depending on their (increasing) contrast value.

We then plotted the single z-scores (i.e. normalized individual scores given by each participant to each image) into boxplots in Figure 15 to compare how the different levels of each feature were distributed in relation to the presented moderately correlated attributes.

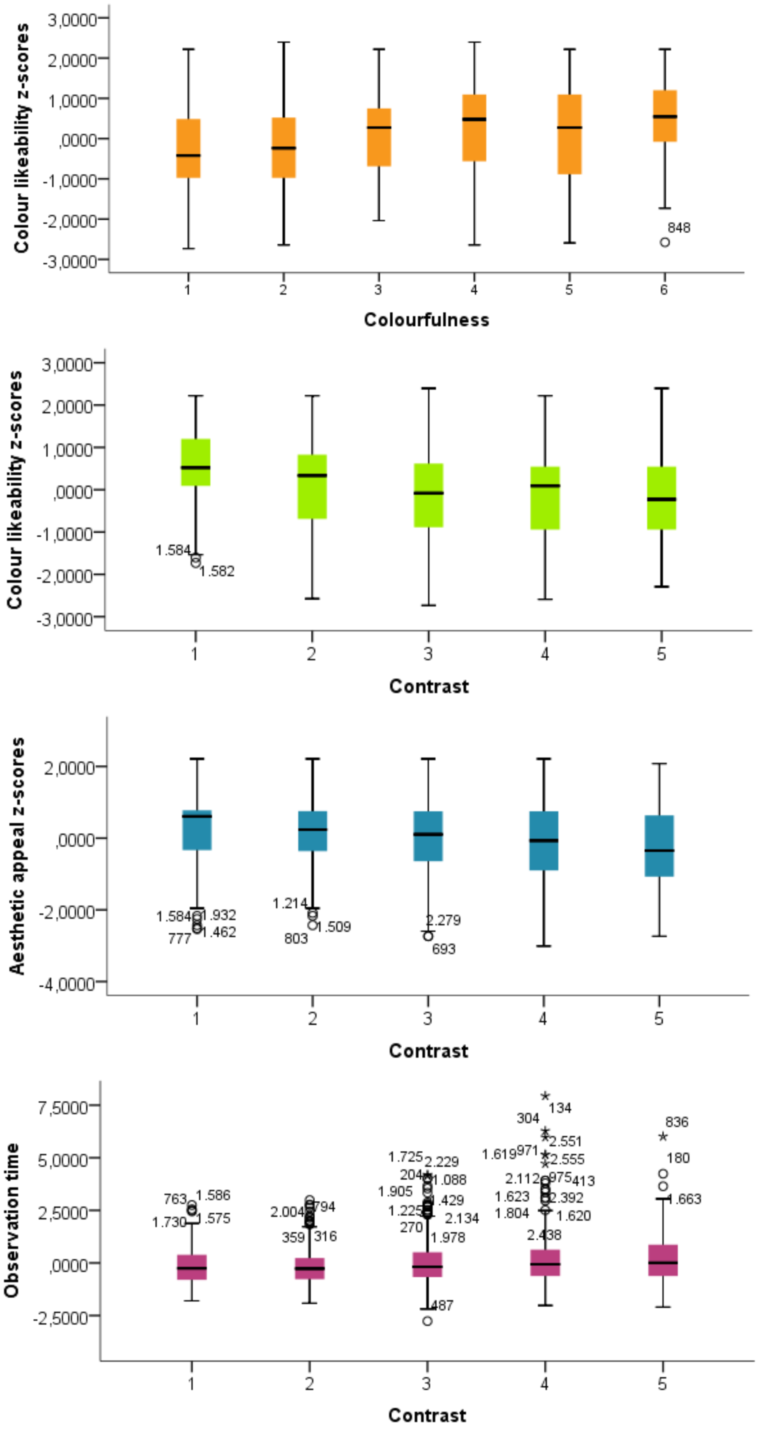


Figure 15: Boxplots of moderate correlations found in the previous analysis.

As shown earlier in this section, colour likeability scores were normally distributed. Therefore, to understand whether colour likeability scores differed based either in contrast or in colourfulness levels, we performed two one-way ANOVAs which determined that there was a statistically significant difference in between levels of contrast (F(4,195) = 5.178, p = 0.001) and colourfulness (F(5,194) = 7.281, p = 0.000). A Turkey post-hoc tests revealed that:

- Colour likeability was statistically significantly lower for images with a contrast level of 3 (p = .002), 4 (p = .001) and 5 (p = .004), with respect to images with a contrast level of 1
- Colour likeability was statistically significantly higher for colourfulness levels 4 (p = .002) and 6 (p = .004), compared to images with colourfulness level 1; also, images with colourfulness level 3 (p = .029), 4 (p = .000) and 6 (p = .003), had significantly higher colour likeability than images with colourfulness level to 2.

To sum up, colour likeability seems to diminish for images with higher contrast, and grow instead with colourfulness. This shows that people enjoy more the colours on more colourful images and in low contrast images. Indeed, psychological studies have shown that aesthetic response to a picture may depend colourfulness (Peters et al 2007).

We executed a Kruskal-Walis H test to investigate whether either observation time or aesthetic appeal scores (both non-normally distributed) differed based on the different contrast levels. This test is the nonparametric test equivalent to the one-way ANOVA and permits us to understand whether he scoring differed based on the categories selected. Note that the Kruskal-Wallis test is an omnibus test and thus it cannot tell us which specific categories were significantly different from each other; it only tells us that at least two categories were different. The test reported a statistically significant difference between the different contrast levels for observation time (chi = 12.687, df = 4, p = 0.013) and for aesthetic appeal scores (chi = 14.623, df = 4, p = 0.006). Determining which of these categories differ from each other is important and can be done using a Mann-Whitney U test.

To examine which specific levels were significantly different from each other, we ran a Mann-Whitney U test for each level of contrast. We found that:

- Low contrast level 1 an level 2 ranked significantly higher in aesthetics (U = 1039.0, p = 0.024) and (U = 2529.0, p = 0.027) respectively
- High level of contrast (level 5) ranked significantly lower in aesthetics (U = 1203.0, p = 0.031)
- Level 2 of contrast increased significantly the observation time (U = 2346.0, p = 0.006)

From this, we can conclude that, in consistency with colour likeability, aesthetic appeal seems to rank lower for images with higher contrast and higher for images with low contrast. Hence, it seems that people prefer low contrast images, which goes against with what found in the literature (Joshi et al 2011).

Further, observation time appears to be higher for low contrast images which might have to do with the difficulty to distinguish objects when contrast is lower. Thus, people spend more time observing an image to understand what is being presented (see Figure 16).

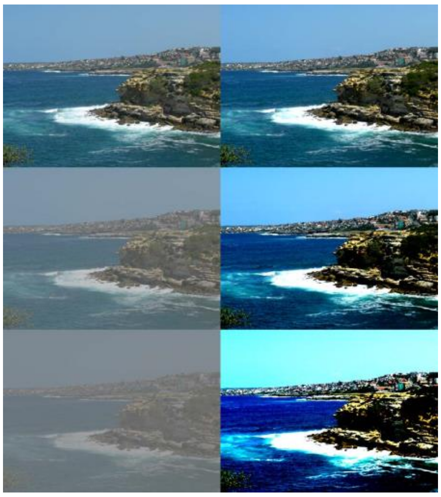


Figure 16: Different levels of contrast where the original image is on top left, then with less contrast to the left and with more to the right[25].

Finally, we quantified the relationship between the saliency features designed in section 3.2. and aesthetic appeal again through Pearson correlation coefficients. These correlations were found not to be statistically significant as one can see in Table 6 (in red cells), therefore, we cannot infer anything from these saliency features.

	Simplicity factor	**Converging lines**
Aesthetic appeal MAOZ	-0.070	0.097

Table 6: Pearson correlation coefficients between the saliency features and aesthetic appeal.

3.5.3. Relationship between image content, attributes and aesthetic appeal

Finally, we were interested in understanding whether the semantic connotation of the image content would play a role in the aesthetic appeal as well as in the attribute judgment. For a comparison between the different sets of scores that come from different image categories, we used Kruskal-Wallis H test., as our scores had shown to be non-normal. The Kruskal-Wallis H test revealed that there was a statistically significant difference between the 16 categories when scoring aesthetic appeal (chi = 33.814, df = 15, p = 0.004), colour likeability (chi = 31.822, df = 15, p = 0.007), recognisability (chi = 71.802, df = 15, p = 0.000) and familiarity (chi = 89.571, df = 15, p = 0.000).

To compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed, we performed the Mann-Whitney U test between the scores given to a single category and the rest of the images (i.e., we checked whether the median of the scores given to one category, e.g., aesthetic appeal scores for macro, was significantly different from the median of the aesthetic appeal scores assigned to all the images that were not macros). As for the categories, since we had 16 categories, we ran 16 Mann-Whitney U tests to understand whether each attribute scores differed significantly according to the category being scored. In this case, we tried to address the following questions, for each category:

- Were the images from this category scored significantly higher in aesthetic appeal than the rest of the images?
- Were the images from this category scored significantly higher in either of the considered attributes with respect to the rest of the images?

For ease of result summarization, we are going to report the outcomes of our statistical tests in Table 7 and with the following notation:

- If the answer to the above mentioned questions is **yes** (significant increase in aesthetics/attribute value for the images in the currently investigated category), we represent the outcome as:
    - ++ and in colour green, for statistically significant difference with $p < 0.01$
    - + and in colour yellow, for statistically significant difference with $p \in [0.01, 0.05]$
- If the answer to the above mentioned questions is **no**, then:
    - If there is a significant decrease in aesthetic appeal/attribute value for the images in the currently investigated category, we represent it as:
        - -- and in colour red, for statistically significant difference with $p < 0.01$
        - - and in colour orange, for statistically significant difference with $p \in [0.01, 0.05]$
    - If no statistically significant difference was found between the aesthetic appeal/attribute for this category, then we use no colour or symbol coding

[25] http://en.wikipedia.org/wiki/Contrast_%28vision%29

	Colour likeability	Familiarity	Recognisability	Aesthetic appeal
Macro	++	++		++
Landscapes	++		++	++
Travel	++		++	++
Animals	++	++	++	++
Nature	++		++	++
Street	--		-	--
Sport	--	++	+	--
Still life		+		--
Fashion	--		++	--
People	--	--		--
Celebrities	--	++	++	--
Food	++	++	-	-
Urban exploration	--	--	--	
City and architecture		++		
Transportation		++	++	
Abstract	-	--	--	

Table 7: Overview of the different results of the Mann-Whitney U analysis for each image category in terms of each attribute. These categories were sorted according to their impact on aesthetic appeal.

Aesthetic appeal was consistently higher for categories that seem possible to group according to two criteria:

- Images depicting something that could evoke to some extent a sense of evasion in the participants mind, as in the case of landscapes and travel.
- Images containing natural elements (landscapes, nature, macros, animal)

At the other end of the table, we can find categories that seem possible to group according also to other two criteria:

- The first refers to something that is present in everyday' s life (i.e., the opposite of the evasion feeling). That will be the case of street, sport, people and food.
- The second is related to something humans are doing or exhibiting (e.g. beauty and wealth) that is far away from the participant's reach. That will be the case of fashion and celebrities. This may relate to some extent to the research reported in (The Economist 2013) . After surveying hundreds of Facebook 20-year old participants, it was concluded that the most common emotion triggered by using Facebook is envy and social upward comparison. Thus, a replica of the psychological mechanisms underlying human reactions towards others leisure, level of social interaction and (apparent) happiness (one of the main factors that trigger envy according to the mentioned research) might influence the scores of images that convey similar messages to the participants.

In general, the presence of humans in the images seems to drive to lower aesthetic scores, which is in contrast with what found in literature (Axelsson 2007). On the other hand, abstract images show to score low in aesthetic appeal, which is in accordance with what found in the literature so far (Congcong et al 2009).

Colour likeability was consistently scored higher for categories where the images are usually conceived with a strong and immediate appeal to the senses. That will be the case of landscapes, nature, travel, animals or even food and macros. In the lower part of the table are categories that, with one exception (urban exploration), scored poorly in aesthetic appeal, which could be expected, as the two quantities are highly correlated (see section 3.5.1). It is also interesting to note that, again, most of these categories had some relation to human presence. That will be the case of sport, celebrities, fashion, people and to some extent also street or urban exploration.

Familiarity was consistently higher for images in categories that seem possible to group according to two criteria: either something that is often present in media to or something with small or well-defined information content. Sport, celebrities, food, city and architecture, transportation as well as animals fall in the first condition, while still life and macro fall in the second condition. At the other hand of the scale are categories that by definition refer to what is not known or requires thorough inspection. Those categories include abstract, urban exploration and people (that the participants have not met before).

It is important to point out, at this stage that a high recognisability score implies that a participant identifies easily an image from knowledge of appearance or characteristics, whereas a high score in familiarity reveals that an image is close to a participant's daily experience.

Further, there is a difference between the two variables and some categories showed dissimilar scores for each of them. For example, the images presented in Figure 17 were scored differently for familiarity and recognisability, e.g. the one at the left side has a familiarity MAOZ of -.5727 and a recognisability MAOZ of 0.4284.

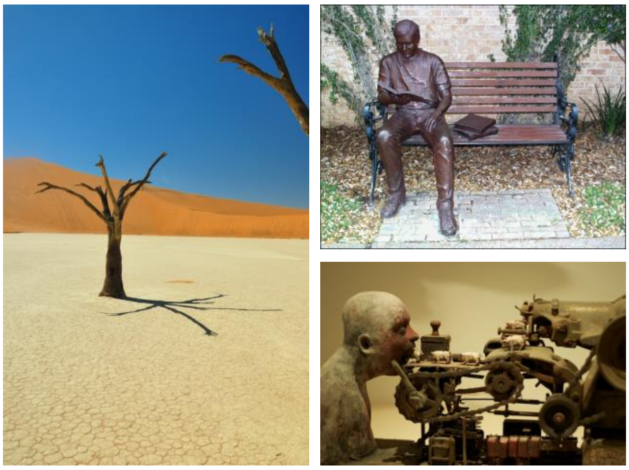


Figure 17: Examples of images scored differently in familiarity and recognisability.

Recognisability was consistently higher for categories that generate a lot of traffic in the social networks. That is the case of fashion, nature, celebrities, animals, transportation and travel, according to photo labels' statistics in Instagram[26]. Conversely, categories like urban exploration, abstract, street and food obtained low recognisability scores. Abstract images have no true content or no clear subject involved and so it is clear that scored low on recognisability . On what concerns to food, a possible explanation for the low scores might be due to the fact that dishes and spices may have a confusing appearance, and could be recognizable given specific cultural backgrounds and previous cooking/food experiences. As an example, the top images in Figure 18 represent two of the images categorized as food. Whereas the right one shows a glass with a jelly dessert (recognisability MAOZ of -1.5675), the left one shows a shot glass with cherry (recognisability MAOZ of -1.3165).

Street and urban exploration images' examples can also be seen at the bottom of Figure 18, left and right respectively. Likewise for the food category, participants might have wondered what exactly was been presented. While the first picture recreated a street in motion (recognisability MAOZ of -1.1684), the second shows a LED light seen from close (recognisability MAOZ of -2.4262).

[26] http://hashtagig.com/top-hashtags-on-instagram.php

Figure 18: Street and urban exploration images' examples.

3.6. Conclusions

In this chapter we developed a methodology to support the creation of a representative image dataset that would include sufficient variability in image features as well as perceived attributes and aesthetic appeal. To do so, it was necessary to first clarify the relationship existing between image physical properties, perceived attributes and aesthetic appeal. For that purpose, we collected preference ratings from 14 participants in terms of colour likeability, familiarity, recognisability and aesthetic appeal on a set of 200 images sorted into 16 categories commonly used in photography and research. Further, we used an eye tracker to record the participants' eye movements and thus, measure salient regions in an image.

This data was then prepared by check for outliers following the recommendations of ITU (ITU 2012), after which the data was submitted to normalization followed by internal consistency and normality tests. We were interested in analysing the relationship between these two and the aesthetic appeal and image attributes. The eye tracker data were also converted according to (Redi et al 2011) into saliency maps. These were used to analyse the link between compositional rules (simplicity of regions and converging lines), saliency and eventual aesthetic preferences.

We found two strong positive correlations between colour likeability and aesthetic appeal and between recognisability and familiarity, and a weak positive correlation between aesthetic appeal and recognisability. The remaining correlations have not shown to be statistically significant.

We then found that the initial weak positive relation between aesthetic appeal and recognisability was due to colour likeability. This imply that people are more attracted to recognizable images when they like their colour.

On what concerns to the relationship between aesthetic appeal and its attributes and image features, three moderate correlations were found between contrast and aesthetic appeal, contrast and observation time and colourfulness and colour likeability. Further, from those we have found that:

- Aesthetic appeal might be higher for high contrast images when the subject is clear for the observer.
- Observers appreciate better the colour of high colourful images when the image is aesthetically appealing.

A further analysis showed that generally colour likeability decreases after a contrast increase and increases when colourfulness increase, which is in agreement with psychological studies. Also, it seems that low contrast images rank higher in aesthetics, which goes against what has been showed in the literature. Moreover, low contrast images might take longer to observe because its content is harder to distinguish.

Regarding the relationship between compositional rules implemented and aesthetic appeal, this showed to be not significant. One possible reason might be because during the selection of images, we did not address these compositional rules, i.e. some rules are applicable to a certain type of images and we did not address that factor in our study.

As a final point, our results showed that indeed image content plays a role in the aesthetic appeal as well as in the recognisability , familiarity and colour likeability. In agreement with the literature, we found abstract images to be ranked lower in aesthetics (Congcong et al 2009). Surprisingly, images with people showed to be ranked lower in aesthetics in contrast with the study in (Axelsson 2007).

4. Crowdsourcing-based evaluation of image aesthetics by gamification using a Facebook application

This chapter reports the core experiment of this research. The goal of this thesis was to better understand the impact of digital filters on aesthetic appeal of images, and to do so by using crowd-testing techniques (Hossfeld et al 2013). Based on the results reported in Chapter 3, we first identified, a representative subset of images to which apply a representative set of Instagram-like filters. The resulting testbed was eventually made of 360 images diverse in aesthetic appeal, colourfulness, contrast and composition.

Because of the large number of images we needed to test, as well as our intention to reach out to a user population as diverse as possible (Redi et al 2013), we aimed at using crowdsourcing as a methodology to collect aesthetic appeal of filtered images. Although, crowdsourcing does not ensure the same degree of experimental control as in laboratory environment and thus, more precautions are necessary to guarantee the reliability of the collected data.

Our approach focuses on the willingness of the participants to participate and at improving the quality of the work conducted by the participants with integration of a joyful interface in a familiar platform for many users that is Facebook. Facebook is a popular social network that connects people with friends and others like work colleagues and family. Besides posting, commenting and sharing events, Facebook also offers the possibility for game developers to connect their games to the platform. In this way, games are provided with a second layer of security and users can then share their score or achievements in the network. Facebook has been proposed in the past as a good platform to develop gamification of micro-tasks, through it is expected to engage the workers more, and increase their willingness to conduct the task seriously (Hossfeld et al 2013).

In this experiment, we developed a Facebook-based aesthetic appeal platform, which we then coupled with Microworkers (see section 2.3.2. for more details on crowdsourcing platforms such as Microworkers).

Below, we report the steps we took to perform our investigation on the added value of digital filters. The selection process for both images and filters is described in section 4.1. We describe the experimental setup and the implementation of the Facebook-based crowd-testing application through which we performed our subjective study in section 4.2. and section 4.3. Then, we analyse aesthetic appeal preferences from over 600 users, of which roughly half were paid through a crowdsourcing scheme, whereas the others were volunteers. Section 4.4. reports an analysis of the differences in reliability between paid and volunteer participants and on the relationship between image physical properties, attributes, aesthetic appeal and digital filter application.

4.1. Image material preparation

4.1.1. Selection of digital filters

Our first idea was to study actual Instagram filters because of its recent popularity[27]. However, Instagram is a smartphone-based app and the algorithms to apply the filters have property rights[28]. One solution to this problem would have been creating a spare Instagram account, upload all the images to our phone and then apply these filters one by one to each image and upload them back from a phone to our computer. However, this would not give us an understanding of what these filters were actually doing. An alternative approach was to look for Instagram-like filters. Due to Instagram popularity, more and more apps started to appear for phone and desktop such as pixlr-o-matic[29] or aviary[30], yet without open-source code for their filters. We finally found Daniel Box's (Box 2011) Photoshop-based implementation of Instagram-like filters. Filters are implemented in a set of actions, that is, basic manipulations that allow automation of repetitive changes in image like in contrast or brightness, commonly used by photographers. In Figure 19, one can see that the output of these Photoshop actions is visually close to Instagram filters output for filters Early bird and X Pro II respectively.

[27] http://instagram.com/press/

[28] http://quora.com/Instagram/How-does-Instagram-develop-their-filters

[29] http://pixlr.com/o-matic/

[30] http://aviary.com/

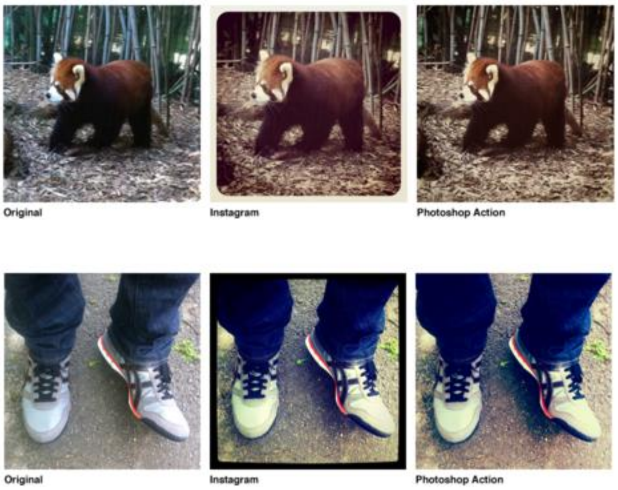
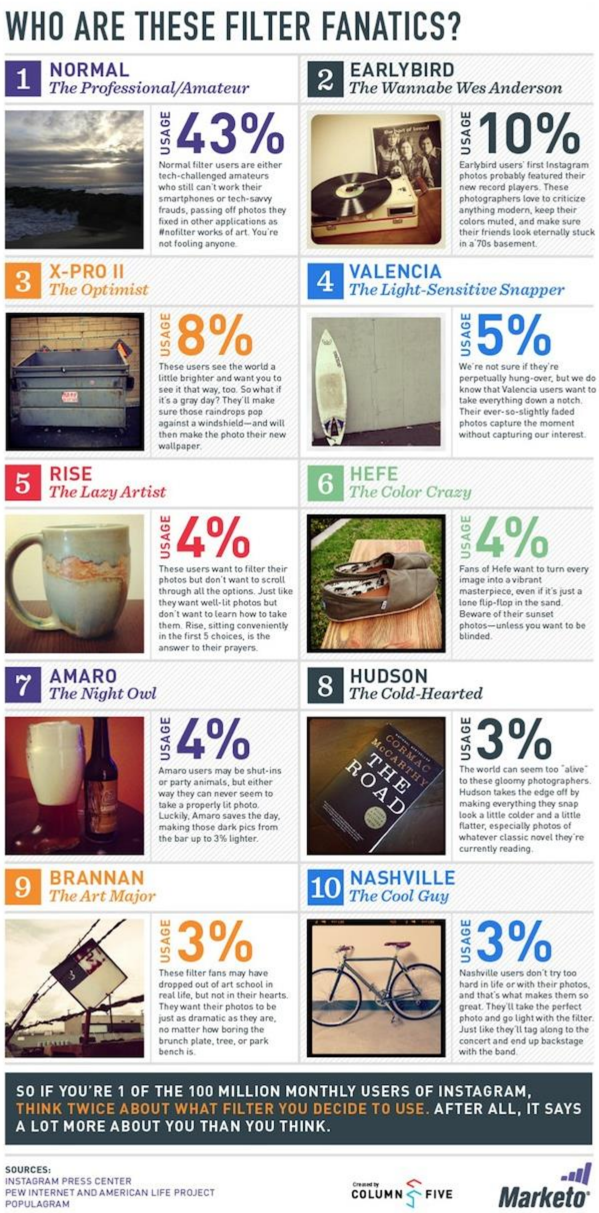


Figure 19: Comparison between the real Instagram filters Early bird on top and X Pro II bird on bottom with their respective dbox version.

The next step was to determine which filters to study. One should note that the higher the number of filters used, the larger the dataset. If we had kept our dataset with 200 images and chosen 4 filters, we would have ended up with 200 * 4 = 800 images, which is a rather conspicuous number, even for crowdsourcing-based tests. We therefore resolved to use the 3 filters: Early bird, X Pro II and Hefe. This choice was based on the limited range of implemented filters by Daniel Box (9 stable filters implementations at update #4) and due to their noticeable and different popularity in Figure 20. Further, Figure 21 shows examples of the same image processed with all three filters. In general, all of these filters alter brightness, contrast and fill the image with a different colour that contrasts with a given image. Hefe and Early bird alter hue and saturation as well. While Hefe also softens the edges of the image by making the edges gradually fade out until it becomes transparent, in contrast Early bird refines the edges of an image to make them finer. X Pro II and Early bird adjust intensity levels of image shadows, midtones, and highlights. In terms of complexity, Early bird post-processes more an image than Hefe and X Pro II.



WHO ARE THESE FILTER FANATICS?

1 NORMAL
The Professional/Amateur

USAGE 43%

Normal filter users are either tech-challenged amateurs who still can't work their smartphones or tech-savvy frauds, passing off photos they fixed in other applications as #nofilter works of art. You're not fooling anyone.

2 EARLYBIRD
The Wannabe Wes Anderson

USAGE 10%

Earlybird users' first Instagram photos probably featured their new record players. These photographers love to criticize anything modern, keep their colors muted, and make sure their friends look eternally stuck in a '70s basement.

3 X-PRO II
The Optimist

USAGE 8%

These users see the world a little brighter and want you to see it that way, too. So what if it's a gray day? They'll make sure those raindrops pop against a windshield—and will then make the photo their new wallpaper.

4 VALENCIA
The Light-Sensitive Snapper

USAGE 5%

We're not sure if they're perpetually hung-over, but we do know that Valencia users want to take everything down a notch. Their ever-so-slightly faded photos capture the moment without capturing our interest.

5 RISE
The Lazy Artist

USAGE 4%

These users want to filter their photos but don't want to scroll through all the options. Just like they want well-lit photos but don't want to learn how to take them. Rise, sitting conveniently in the first 5 choices, is the answer to their prayers.

6 HEFE
The Color Crazy

USAGE 4%

Fans of Hefe want to turn every image into a vibrant masterpiece, even if it's just a lone flip-flop in the sand. Beware of their sunset photos—unless you want to be blinded.

7 AMARO
The Night Owl

USAGE 4%

Amaro users may be shut-ins or party animals, but either way they can never seem to take a properly lit photo. Luckily, Amaro saves the day, making those dark pics from the bar up to 3% lighter.

8 HUDSON
The Cold-Hearted

USAGE 3%

The world can seem too "alive" to these gloomy photographers. Hudson takes the edge off by making everything they snap look a little colder and a little flatter, especially photos of whatever classic novel they're currently reading.

9 BRANNAN
The Art Major

USAGE 3%

These filter fans may have dropped out of art school in real life, but not in their hearts. They want their photos to be just as dramatic as they are, no matter how boring the brunch plate, tree, or park bench is.

10 NASHVILLE
The Cool Guy

USAGE 3%

Nashville users don't try too hard in life or with their photos, and that's what makes them so great. They'll take the perfect photo and go light with the filter. Just like they'll tag along to the concert and end up backstage with the band.

SO IF YOU'RE 1 OF THE 100 MILLION MONTHLY USERS OF INSTAGRAM, THINK TWICE ABOUT WHAT FILTER YOU DECIDE TO USE. AFTER ALL, IT SAYS A LOT MORE ABOUT YOU THAN YOU THINK.

SOURCES:
INSTAGRAM PRESS CENTER
PEW INTERNET AND AMERICAN LIFE PROJECT
POPULAGRAM

Created by COLUMN FIVE

Marketo

Figure 20: Popular filters.

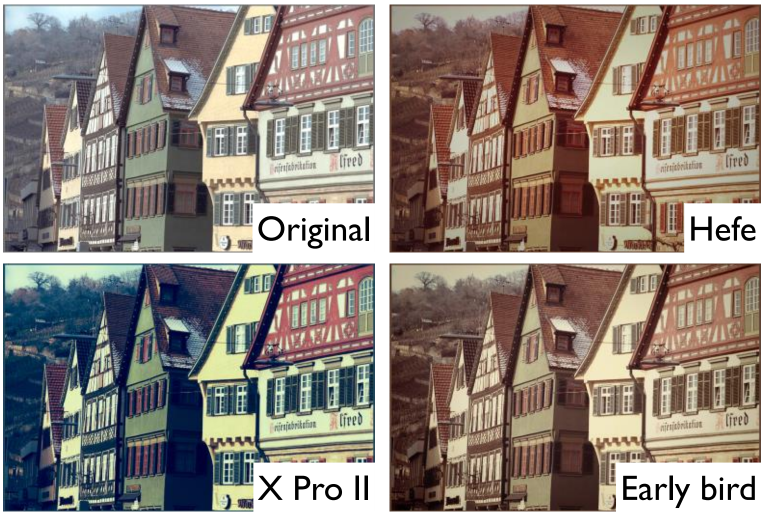


Figure 21: Examples of the same image processed with all three filters: Hefe, X Pro II and Early bird.

4.1.2. The dataset selection

So far we had collected user preference scores on recognisability , familiarity, aesthetic appeal and colour likeability, and then quantified their relationship with image features such as colourfulness, contrast and saliency distribution features. Based on this relationship, we were now ready to calibrate our dataset to be used in the second experiment. The main reason behind the need to calibrate the original dataset was its size. The original dataset had 200 images in total. Since we decided to test the effect of the three filters mentioned above, this would have implied considering in the next experiment four different versions of the original images (the original, plus the three filters), ending up with a of 200 * 4 = 800 images. In scaling studies, a high number of images (images) implies a longer duration of the test, which can lead the observer to lose interest and hurry through the scoring task, delivering unreliable judgments (Engeldrum 2000) . An alternative solution is to conduct several scaling studies, each including a subset of the images, with the subsets overlapping to some extent for realignment purposes (Redi et al 2013). This implies running multiple sessions; when dealing with 800 images, the number of sessions may be consistent, making it difficult to recruit participants willing to return for every session. The use of crowdsourcing would make the recruitment easier, but would require the establishment of a large number of campaigns, since the duration of a micro-task should not exceed 10 minutes (Hossfeld et al 2013). As a result, reducing the images in the dataset to a smaller yet representative subset becomes an important step towards our goal.

As far as the size is concerned, we aimed at cutting the original number to half, thus around 100 images. We also wanted to gather a number divisible by 15, in order to implement an experimental setup similar to that employed in our previous work in (Redi et al 2013). In (Redi et al 2013), from the 200 images, we used 5 images as anchor images and the remaining were grouped in groups of 15 to keep each rating task short. Thus, each task would comprehend 20 images (5 anchors, same for all tasks, of different aesthetic appeal levels, and 15 changing) to be rated on a 5 point scale in aesthetic appeal. These tasks could be completed in less than 10 minutes, in agreement with crowdsourcing best practices (Hossfeld et al 2013) (for more details, see appendix B). Given the positive results of this setup, we decided to maintain it for this experiment. As a result, we aimed at collecting 90 images for this new dataset.

To make this subset representative for the variability in attributes and features observed so far, we selected images based on their subjective and objective properties as analysed in Chapter 3.

To begin with, a desirable property for our subset was to uniformly span the aesthetic appeal range observed so far. To ensure this, the aesthetic appeal MAOZ from the laboratory experiment were first discretized into five levels, by taking as thresholds for these levels the integer values of the five-point scale on which the aesthetic appeal was scored. Thus, Images with a MAOZ within (Bhattacharya et al 2010)(Wallraven et al 2009) were assigned to the aesthetic appeal level 1, images with a MAOZ within (Wallraven et al 2009)(Peters et al 2007) were assigned to the level 2, and so on . Since the lowest bin of aesthetic appeal had seven images, we took seven images of each of the aesthetic appeal classes and thus, 7 * 5 = 35 images representing the full range of aesthetic appeal. Five of these images represented the 0th, 25th, 50th, 75th, and 100th percentiles of the distribution of all aesthetic appeal MAOZ (Figure 23).

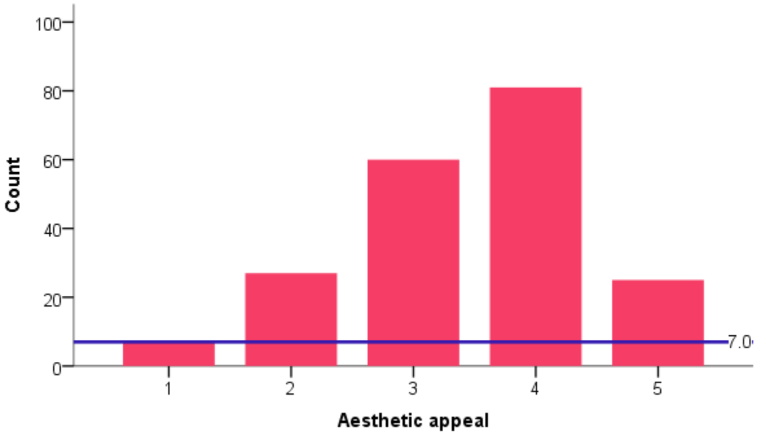


Figure 22: Selecting a representative range of aesthetic appeal for the baseline dataset.

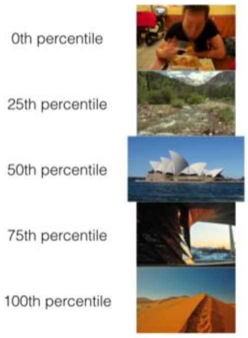


Figure 23: Five of these images represented the 0th, 25th, 50th, 75th, and 100th percentiles of the distribution of all aesthetic appeal cores from the previous experiment.

We then considered the saliency distribution in the image. In our previous work in (Redi et al 2013), we had split each of our images in nine equally sized regions following the rule of thirds (Figure 24), and we showed how the location of the image ROI with respect to those regions was relevant for the aesthetic appeal of the image. To ensure diversity with respect to saliency distribution in our dataset, we divided the images in classes according to the region of the 9 in which the centroid of the main ROI was located (Figure 24).

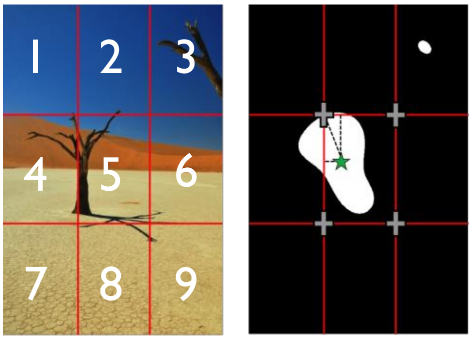


Figure 24: Example of the procedure from (Redi et al 2013) where the rule of thirds was associated with saliency data.

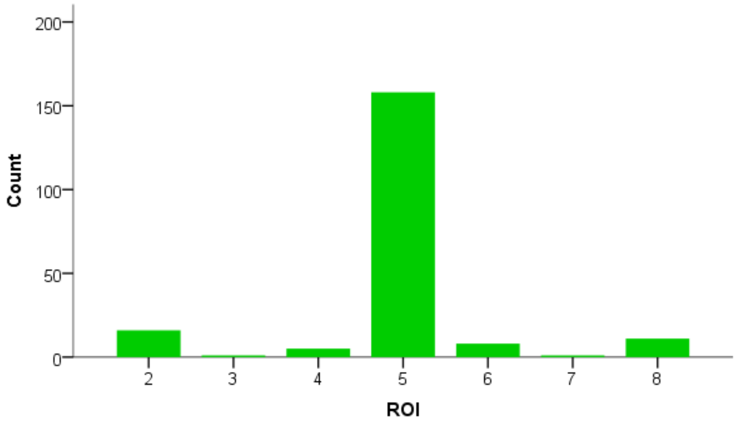


Figure 25: Number of images in each of the 9 regions.

In Figure 25, we have plot these groups and how many images each group has. One should notice that there were no images with the centroid of the largest blob in the ROI situated in the upper left region (group 1) and lower right region (group 9). Also, most of the images have the largest blob in the ROI located in the image centre (region 5). Targeting at a diverse and balanced new dataset, we decided to include in our dataset all the 42 images whose largest blob was not located in the central region of the image. From these, 5 images were already included in the first collection of 35 images. As a result, our dataset included (42 - 5) + 35 = 72 images, so we needed more images to reach the goal of 90 images. To collect these last 18 images, we looked at the colourfulness and contrast properties of the images. In fact, digital filters alter contrast and colour levels. Hence, it might be interesting to select the remaining images based on these features so that we could measure later how filters impact on features values first and on image appeal consequently.

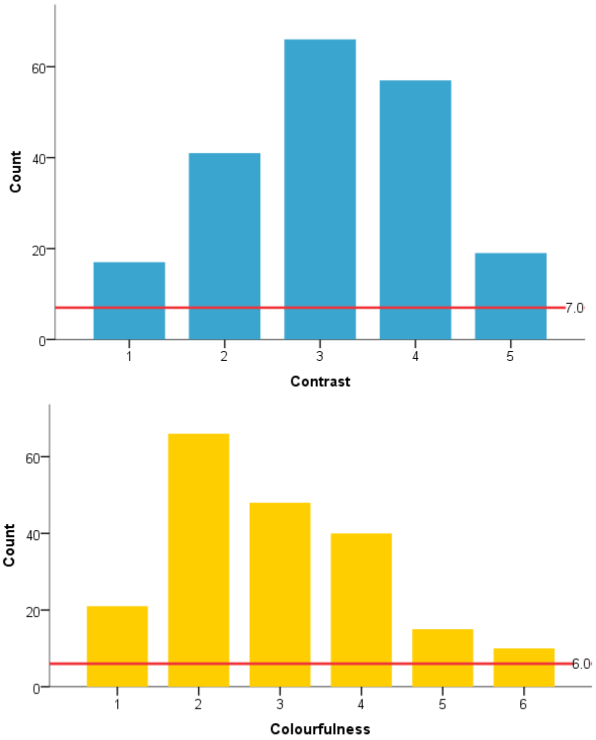


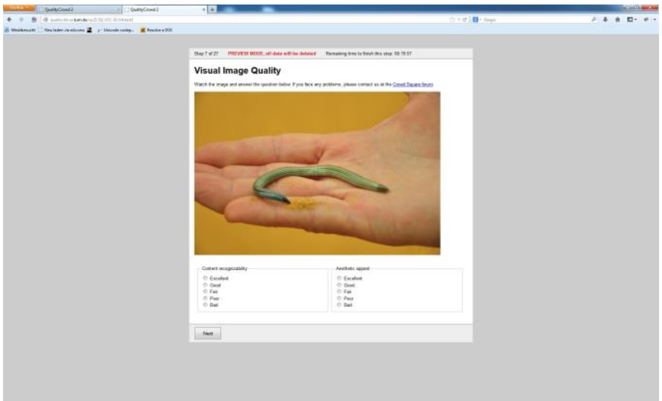
Figure 26: Selecting a representative range of images with colourfulness and contrast for the baseline dataset.

We thus considered the (discretized) contrast and colourfulness values already computed in section 3.5.2. First we looked at contrast values. As for aesthetic appeal, we decided to use 7 * 5 = 35 images representing 5 different levels of contrast (Figure 26). From these 35, 26 images were already included in the collection of 72, so we ended with a total of 72 + (35 - 26) = 81 images.
Finally, to assemble a dataset of 90 images, we collected 6 * 6 = 36 images representing 6 different ranges of colourfulness (Figure 26). Likewise the latter step, 27 images were already included in the previous 81 images, so as planned we met the 81 + (36 - 27) = 90 images.

To sum up, the dataset was reduced from 200 to 90 images balanced in terms of aesthetic appeal, location of the regions-of-interest, contrast and colourfulness. The three filters were then applied to the 90 images, creating a total of 360 images: the 90 originals plus the three versions with each filter applied (see section 4.1.1. Figure 21 for examples of the four versions of an image).

4.2. Experimental methodology

At this point in time, after building our representative dataset, we were ready to perform the main experiment of this work. And as a consequence, to finally answer what is the effect of digital filters on images aesthetics. With this in mind, we sought to collect aesthetic appeal scores via crowdsourcing. Our previous work in (Redi et al 2013) helped us to get acquainted with this novel approach and to study if crowdsourcing can be a reliable tool to provide subjective evaluations. In that study, we reproduced the lab experiment described in Chapter 3 in a crowdsourcing environment using simple HTML with a grey background to present each image and the attribute questions as showed in Figure 27.



Step 7 of 27 PREVIEW MODE, all data will be deleted Remaining time to finish this step: 00:19:57

Visual Image Quality

Watch the image and answer the question below. If you face any problems, please contact us at the Crowd Square forum

Content recognizability
○ Excellent
○ Good
○ Fair
○ Poor
○ Bad

Aesthetic appeal
○ Excellent
○ Good
○ Fair
○ Poor
○ Bad

Figure 27: Example of the experimental set-up in (Redi et al 2013).

Due to the nature of crowdsourcing, which requires short and clear tasks, some adjustments in the experimental procedure had to be considered to reduce the risk of unreliable ratings. Otherwise, possible boredom effects could be introduced on a long evaluation. Some of these adjustments consisted in, for example:

- Creating a higher number of sessions (campaigns), involving a smaller number of images and thus of shorter duration (20 images, rated in 5 minutes instead of 100 images originally, rated in approximately 30 minutes)
- Reducing the number of evaluated attributes from four to two, namely only aesthetic appeal and recognisability

Our results showed recognisability scores were obtained in lab- and crowdsourcing-based tests were consistent, but this was not the case for aesthetic appeal. We identified the reason for this in the fact that the two scoring scales were presented one close to the other (recognisability first and aesthetic appeal later), which may have lead participants to express the same score for both scales. In fact, we found recognisability and aesthetic appeal to be highly correlated in the crowdsourcing-based experiment, which was not the result we had for the lab (see section 3.5.1. Table 4). Thus, we concluded that crowdsourcing can be a useful tool to collect subjective preferences but careful considerations in the design must be made to accommodate the entailed risks.

Consequently, in the present work, we took into account the main three recommendations to design our test platform already mentioned (see section 2.3.2.): a two-step process with a feedback platform and a proper design interface.

Facebook is a social network and a crowd provider, which offers a good way to easily reach a large number of test participants for free. Then, if one integrates a user study as an application in Facebook, one can consequently access participants' demographics provided in their profiles. Additionally, Facebook offers a familiar feedback system for its users with comments and likes. Although, the test has to be designed in a joyful manner to attract participants and keep them motivated, which if done well, has the possibility to go viral.

We therefore designed therefore a gamified web application, which we named Phototo, only accessible to the users of the social network Facebook. In addition, by means of the use of clear instructions and a simple and easy-to-use interface, we tried to overcome possible problems due to the lack of an experimenter to guide participants (Vliegendhart et al 2012)(Hossfeld et al 2013). Furthermore, due to a received award on a submitted idea about the concept of this experiment in the CrowdMM 2013 workshop competition[31], we decided to couple the web application with the crowdsourcing platform Microworkers.

Thus, we had two pools of different participants:

1. Participants that had access to the application via Microworkers platform and were therefore paid for their ratings (we will refer to these participants as Microworkers, from now on)
2. Participants that either were from our personal network or saw the application online and decided to access it, which were not paid for their ratings (we will refer to these participants as volunteers)

This setup would have allowed us to compare the reliability of paid (Microworkers) versus volunteer (unpaid Facebook users) participants.

[31] http://crowdmm.org/

4.2.1. Experimental setup

In this experiment, we focused on the effect of the three filters selected on aesthetic appeal. Both preference of filter and aesthetic appeal were again evaluated within-subjects.

For that we decided for a 4-wise comparison design to reinforce direct visual comparison. In this way, participants are able to compare the four different versions of each image (original and with filters Early bird, Hefe and X Pro II - see Figure 28) to judge their aesthetic appeal.
Further, all the images from the dataset had different resolutions, so a general rule had to be found to retrieve all of them in the same way in the same space without changing their dimensions.

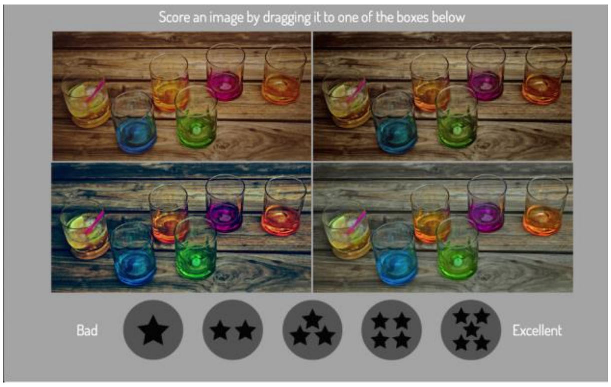


Figure 28: Training session.

Overall, in this experiment we tried to follow a similar design setup as the ones in (Redi et al 2013) and in Chapter 3. From it, we adopted:

- The same grey background found to have a neutral effect to display images
- The 5-point scale to rate the images
- The aesthetic appeal training

In our previous work in (Redi et al 2013), 0.3 USD were paid for each 20 images rated, and we had a total of 13 campaigns. In this experiment instead we split our 360-image dataset in 18 campaigns, with 20 images each. In this setup we did not use anchor images, although we did include the 5 images corresponding to the 0th, 25th, 50th, 75th, and 100th percentiles of the distribution of all aesthetic appeal scores as determined in the lab-based evaluation of the 200 images (Redi et al 2013). Thus, the 20 images per campaign were unique to each campaign. With this division, each worker/participant would only rate a subset of the original image set but would also be able to participate in multiple campaigns.

4.2.2. Gamification aspects

On the other hand, the appearance of the scoring scale and the scoring mechanism were different from the previous experiment (discrete sliders). With the purpose of gamifying the application, here we opted for a playful dragging the image to one of five containers with a number of stars based on the number of the score, e.g. five score would be represented by a container with five stars (see section 4.2.1. Figure 28). Furthermore, the gamified elements of this experiment were:

- The scoring task:
    - Participants can score an image by clicking on it (the image will resize to a smaller size) and dragging it to the container with the number of stars that is more suitable. Also, it was allowed to rate all the images with the same score.
    - If a participant wants to change a image rating, it is as easy as clicking in the image in the container and the image will jump back to the initial place
- The content game: After scoring 4*5 = 20 images, the participant will be prompted with a question on the content of last image for reliability purposes. To answer the participant needs to select one of four buttons with each a different answer
    - When this screen is presented to the participant, a timer starts a count down from 6 seconds. Inside of the timer, the participant can see the respective score that can increment if the question is correctly answered or decrement otherwise. The timer

    element was added as an adrenaline factor on our experiment to incentivise participants to earn more points. Additionally, the competitiveness element is an important factor to keep participants engaged, therefore, participants, after rating the images, can check the overall score of an image as well as how the other participants are doing in terms of score and images left to score.

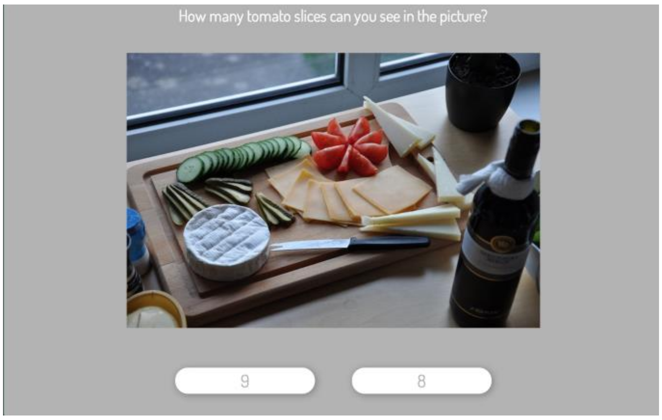
    - Once the participant clicks on one of the four answers, if the correct, the button will be filled with green and expand. Otherwise, will be filled in red while the correct answer button will be filled with green and expand at the same time.

- Once the participant has scored some images, he can compare the images' statistics of the images he scored as well as his friends' scores.

Additionally, all the visual elements of this experiment were also colourful and playfully designed to assure the pleasantness of the user interface and to have participants more engaged. These decisions on interface design took quite some time and effort. A mockup was first designed and showed to different people to understand how easy to use and joyful would this experiment be. Moreover, we got in contact with a Facebook marketing agency in Amsterdam, called "a friend of mine"[32] which helps companies on development of Facebook advertising, business pages and applications. They shared some of their expertise on how to remain users captivated, e.g. users tend to prefer dragging that clicking actions.

4.2.3. Experimental protocol

Because of the crowdsourcing-based experimental environment, we had to introduce (as already done for (Redi et al 2013)) a number of control steps to assess later the reliability of the ratings provided by each participant. So, as a first step, once a participant would log into the application and start the experiment he/she would need to answer a verification question (see Figure 29).

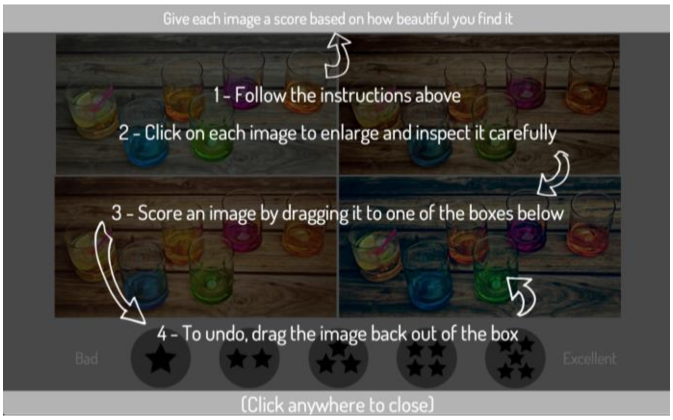


How many tomato slices can you see in the picture?

Figure 29: Verification question used.

Then, as already done for the lab experiment in Chapter 3 and the crowdsourcing study (Redi et al 2013), participants would go through a training phase, to get them acquainted with their task. To help explaining the scoring procedure, instructions were first displayed as in Figure 30. A summary of the instructions would also always be shown on the top part of the screen during every scoring trial.

[32] http://afriendofmine.nl/



Give each image a score based on how beautiful you find it

1 - Follow the instructions above

2 - Click on each image to enlarge and inspect it carefully

3 - Score an image by dragging it to one of the boxes below

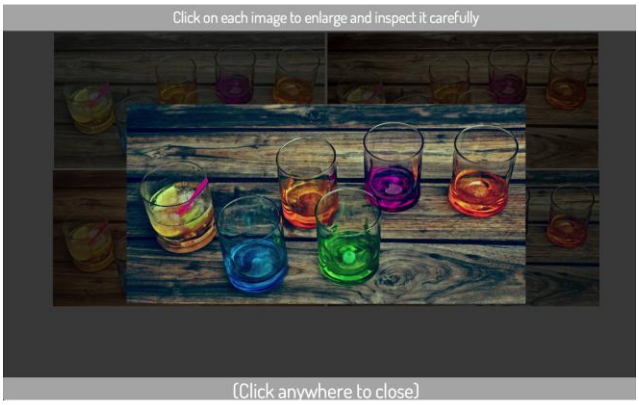
4 - To undo, drag the image back out of the box

(Click anywhere to close)

Figure 30: Instructions' screen.

Both during the training and the experiment, the scoring would go according to the following procedure. First, each participant was presented with 4 images, depicting the original and the three filtered versions of the same content, arranged as in section 4.2.1. Figure 28. Each time, the occupied position of each version (the original and the three filtered versions) was randomized, i.e. the Early bird version could likely appear on the top left, top right, bottom left or bottom right of the screen.

Due to the limited size of the scoring screen, when displayed in the side-by-side configuration of Figure 28, images would typically be re-scaled to fit the screen. In order to obtain reliable judgments, though, we wanted participants to evaluate the images at their full resolution. As a consequence, before being able to score them, participants had to first click on each of them to appreciate them at their original size (Figure 31). Then, once the four images had been observed, the scoring bar appeared.

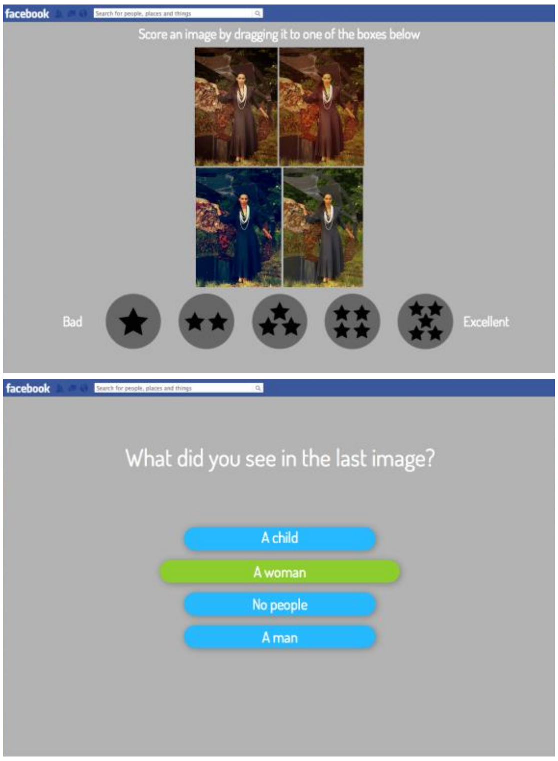


Click on each image to enlarge and inspect it carefully

(Click anywhere to close)

Figure 31: Example of how the image is presented in its original size.

This consisted of 5 circles with a number stars corresponding to the aesthetic appeal rating, ranging from 1 star with label Bad to 5 stars with label Excellent. The participant could then drag each image to one of the circles corresponding to the selected score. It was possible to assign the same rating to all 4 images. For the training, participants had to score in total 2*4 = 8 images, the same used as in the aesthetic appeal training of the laboratory experiment. After completing the training, the experiment started, including the scoring 5*4 = 20 images. After the last set of 4 images was rated, the participant was redirected to a screen with a content question about the last seen image (Figure 32). The participant had four possible answers and at the bottom right of the screen a timer constraining the time to 6seconds that the participant had to answer (this to add a bit of adrenaline to the game). The content questions were rather trivial and could be answered by simply pay attention to the images displayed. A wrong answer would thus indicate lack of attention of the participant to the task and could be used at a later stage to filter out unreliable participants. From the game perspective, if the participant answered the question correctly, he/she would be assigned 10 points; no points were awarded instead in case the answer was wrong or not given within the allowed time.



Score an image by dragging it to one of the boxes below

What did you see in the last image?

facebook
Search for people, places and things
Bad
Excellent
facebook
Search for people, places and things
A child
A woman
No people
A man

Figure 32: Example of one of the content questions with the image to which the question refers.

At the end of the experimental session, the participant would be given the choice to participate into another one (i.e., score another set of 4x5 images), and could go ahead with the scoring task until completing all 90 images in the dataset. This option was given only to the Facebook volunteer participants; it is important to mention that for a user from Microworkers platform, the option to score more images was not available, instead a payment code was displayed that could be submitted in the platform.

4.3. Implementation of the Facebook-based crowdsourcing platform

4.3.1. Design constraints

In order to control the influence of screen size on the results, we restricted this application to be only accessible through a computer browser. Although more contextual factors such as illumination of the screen, luminance in the room, viewing distance and viewing angle of the display may have had an influence on the results, we decided not to control for them due to time constraints. Controlling for these factors is still a main challenge for visual tests run through crowdsourcing; nevertheless, the assumption is typically that environmental and viewing conditions effect will average out given a sufficient number of participants.

Due to time constrains, the application was only compatible with the Chrome browser. Chrome is in fact the most widely used browser worldwide[33], as shown in Figure 33.

[33] http://statcounter.com

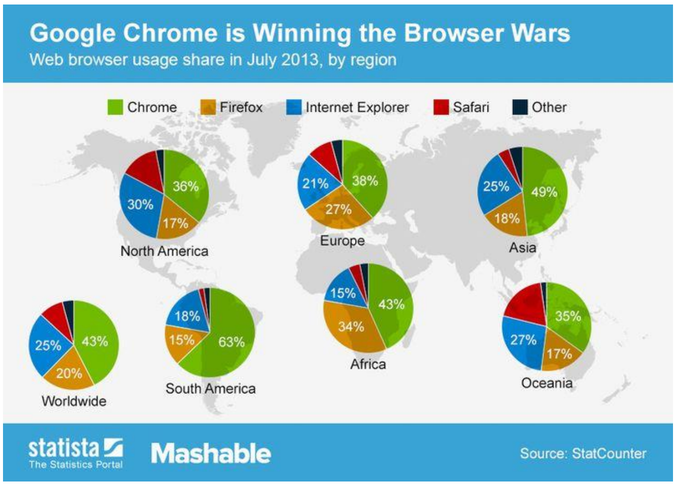


Figure 33: Browser usage statistics.

4.3.2. Implementation details

Phototo's architecture is split in three parts: front end, back end and database, visible in Figure 34. As many other systems, in the back end (PHP), the data collected is stored or retrieved by the server in or from the database (MySQL) respectively, while in the front end (JavaScript), the data is presented to the user.

The web-app was then hosted in a Faculty server[34].

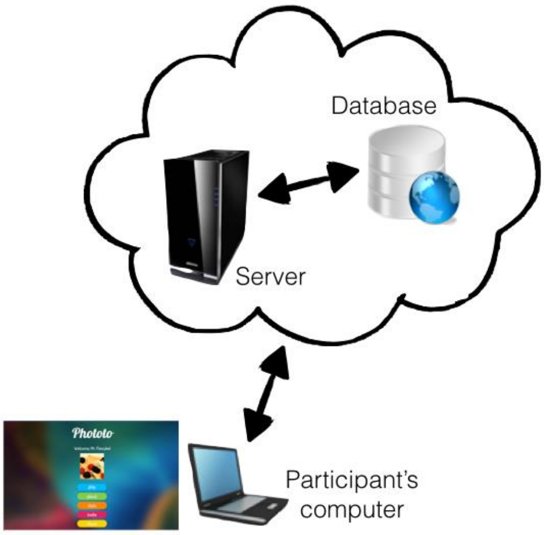


Figure 34: Architecture of the application designed.

Phototo's front end was developed by paying a lot of attention to the user friendliness of the interface. In order to test the usability of the tool, we first came up with a paper Mockup through which we could check how the participants interact with the interface. In that way, it would be fast and easy to fix the interface, if the user shows signs of confusion. To evaluate this mockup we asked the participant to pretend the application was real and interact with it through finger pressing and dragging. We took then the role of application and thus, we responded to the actions the participants performed in the paper mockup, e.g. if the participant pressed play we would show the next screen with the verification question. Our role was to observe the participant and not to interfere. An alternative approach used was to ask a user to perform a certain task in the application and observe how easy and fast was for the participant to find how to do it. At the end, we asked if they were happy with the interaction and to give us feedback. Overall the participants were happy with the tool, which gave us a green light to start implementing it.

[34] https://ii.tudelft.nl/

The database structure can be found in Figure 35 and has seven tables:

- **Table image** represent all the images from our dataset, thus it has 360 rows. Each row is an image and it has a filename, a photography category where it belongs to and the number of people who have so far scored it. Each image can also have either one of three filters applied or none, and either one question associated to its content or none. In the case the image has no filter applied or no content question associated, the corresponding table entry will be null. Eventually, there were only 18 images with questions (one per campaign, the one displayed last).
- **Table answer** holds the four possible answers that associated, hence it has 18*4 = 72 rows or answers. Each row includes the answer, the image associated and a flag indicating whether it is the correct answer or not.
- **Table category** stands for the 16 categories discussed in the first experiment Chapter 3 and each category has several images. Although, not necessary in this phase of our work we decided to include this information only for future reference.
- **Table filter** holds the three Instagram filters chosen for this experiment and each filter has several images.
- **Table participant** characterizes all the participants and, thus, holds information on each of them. Each participant or row has a Facebook identification number, a Facebook username, a hometown, a picture, a game score, the user's answer to the first verification question and a flag for the completion of the training. If accessing the game through Microworkers platform, the respective participant row will store his or her worker ID as well.
- **Table participant_image** holds the relationship between an image and a participant in this context to avoid redundancy in the database. This table stores the observation time and the score given by each participant to an image.
- **Table payment** stores the payment codes generated for each participant in each campaign.

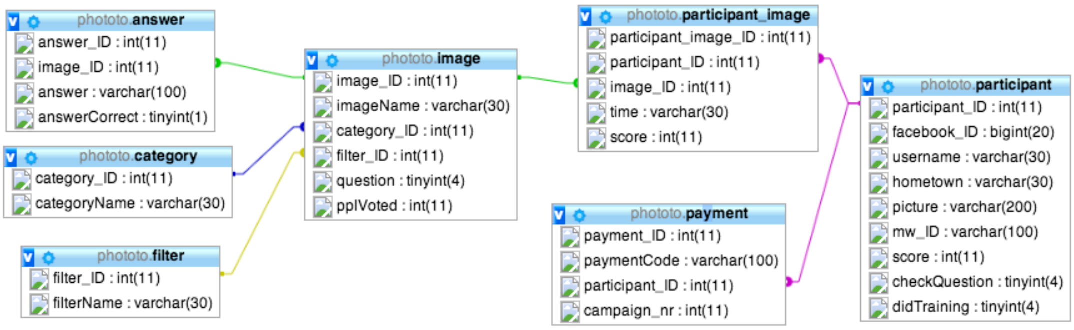


Figure 35: Database relational diagram.

4.3.3. Coupling with Microworkers and Facebook

When Phototo showed to be stable enough in local development, it was made online and only accessible for Facebook users. As a consequence, if one tries to follow the link of Phototo[35] will be prompt with a request to login in Facebook. If logged in Facebook, the app will ask for permission to access the participant's personal data and to post on their behalf. If granted, the web-app will fetch the participant's hometown and profile picture. The latter permission allows the participant to share the app in his or hers timeline. The app was first shared and used by our personal network to easily detect possible bugs.

After coupling with Facebook, we linked the application to the Crowdsourcing platform Microworkers. Microworkers offers two types of campaigns: Basic and Hire Group. In the Basic type, the employer only has to choose where the workers should be from, while in the Hire Group, the employer could hire a specific group of workers previously defined by the system of by the employer himself.

[35] http://apps.facebook.com/phototo_tud/

To follow the two-step design recommended in (Hossfeld et al 2013) and in (Soleymani et al 2010), we first identified a pseudo-reliable group of participants, who will be later performing the actual experiment. Accordingly, we first created two Basic type campaigns, one for North America and major English speaking countries, such as USA, UK, Canada, and Australia, and a second for Western Europe, including workers from France, Germany, Italy, Ireland, the Netherlands, and Sweden. We chose to limit ourselves to those two groups of countries because in (Redi et al 2013) those were showed to provide the most reliable scores (and in higher accordance with the laboratory experiment outcomes).

Of the participants to those two Basic type campaigns, those who correctly answered the verification question and had a standard deviation of the image observation time lower than 20 s (procedure used in (Redi et al 2013)) were grouped into a pseudo-reliable group created in Microworkers. Afterwards, 17 Hire Group campaigns were created so that this pseudo-reliable group could score all the 360 images.

Furthermore, in a campaign, workers need to submit a payment code in the task and the employer has to verify it and pay the worker if satisfied, otherwise give an explanation on why not satisfied. Each payment code needs to be unique for each participant and so it was generated by calculating the sha1 hash (US Secure Hash Algorithm 1) of the concatenation of 3 strings: worker ID, campaign number and step, where step represents the shared secret for all the campaigns. Worker would get their payment code after completing the first set of 4x5 images. Also a flexible threshold was created to limit the amount of workers that can be working at the same time in the campaign to give time to the slow participants to finish their task and to avoid a possible denial of service attack.

4.4. Data analysis and reliability

In total, this web application registered 672 visitors from October 2nd, 2013 till December 8th, 2013: 258 visitors from Microworkers and 414 from Facebook. Figure 36 and Figure 37, reports some usage statistic after approximately one week of usage (2nd to14th October 2013, the experiment was coupled with Microworkers on the 8th October 2013).

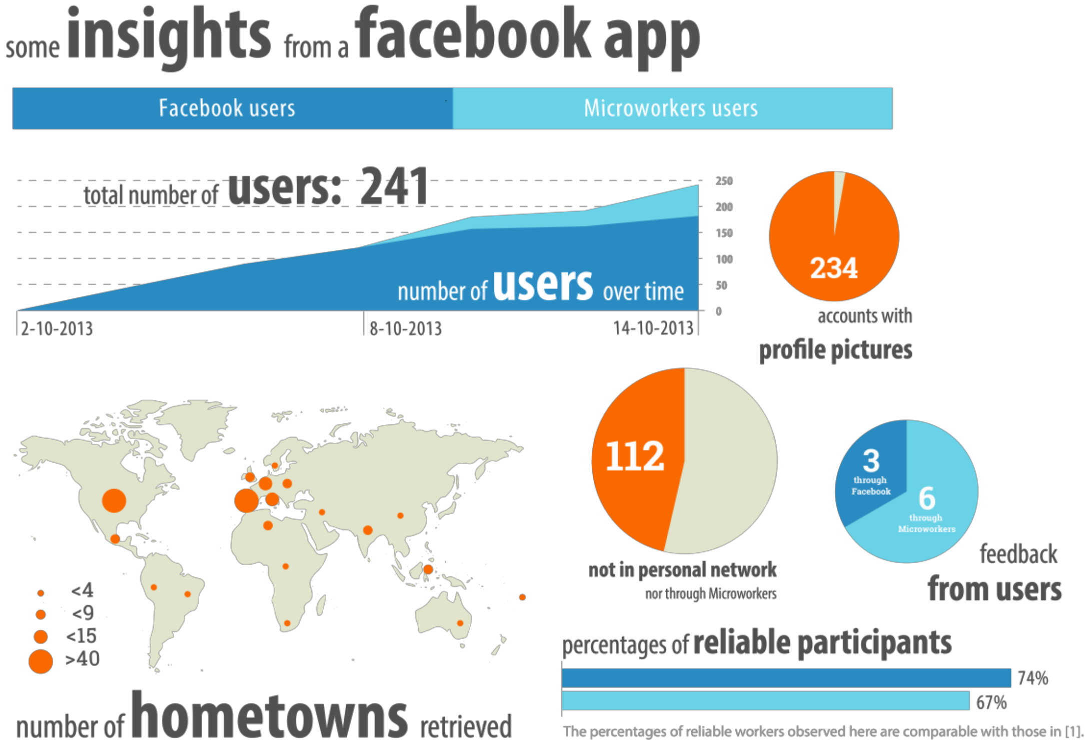


Figure 36: Preliminary analysis and statistics of the collected data on the 14th October 2013.

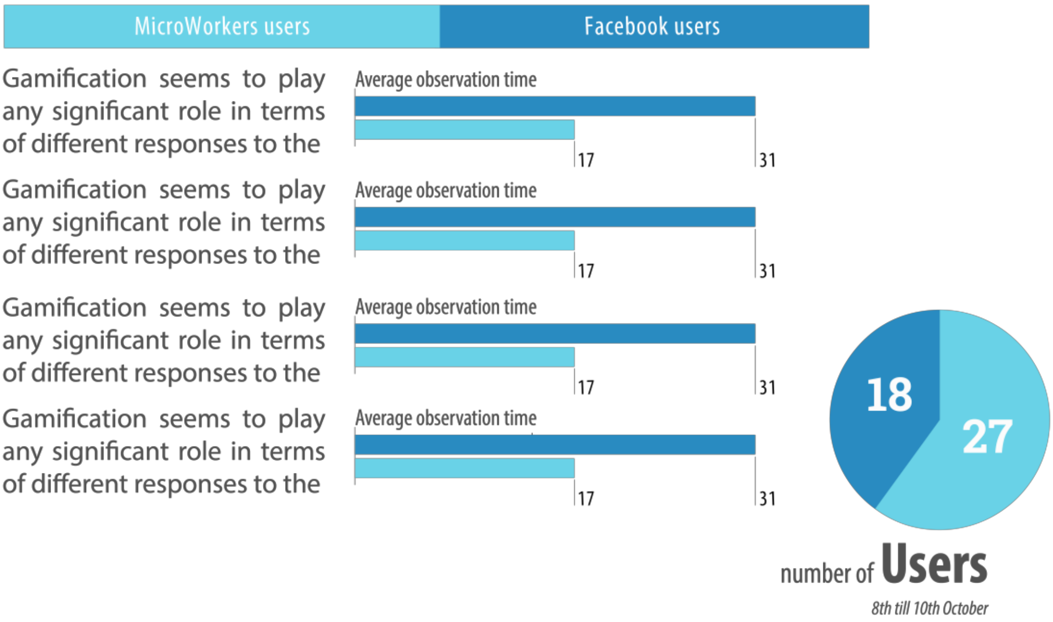


Figure 37: Preliminary comparative analysis between Microworkers (paid) and Facebook (volunteer) users on the 14th October 2013.

Of the eventual 672 visitors, 157 (121 volunteers and 36 Microworkers) did not authorize our app to access their personal data (profile picture, user name and hometown), which was made compulsory and thus, did not access our application. Further, 188 did not perform or complete the training session (142 volunteers and 46 Microworkers) and 22 participants did not rate any image: 18 Facebook users and 4 Microworkers users. This points out that, although Facebook attracted many users, about half of them did not complete the experiment successfully. As a result, from the 305 participants completed at least one session of the experiment, 133 were Facebook volunteers and 172 were Microworkers.

	Facebook Participants	**Microworkers Participants**
Participants not getting to the verification question	29%	14%
Participants not completing the training	34%	17%
Participants not rating any image	5%	2%
Participants completing the scoring of the first set of images	32%	67%
Average number of images rated by participants	18.49	45.69
Total number of ratings	6658	16388

Table 8: Comparison between Facebook (volunteer) users and Microworkers (paid) users.

From Table 8, one can see that Facebook users were twice less eager to complete the game than Microworkers users, which might be due to the money as incentive. Also, while Microworkers users knew the reasoning behind the application from the task description, for Facebook users was curiosity that got them there, but the app may have not matched their interests.

In total, we got 23046 scores, in average per image 64.183 ratings, ranging from a minimum of 48 to a maximum of 78. In this final analysis, we first filtered out unreliable users, then we compared the data collected with those obtained in the previous experiment (Chapter 3), to look for (in)consistencies between crowdsourcing and lab-based experiments. Finally, we looked at the impact of digital filters on aesthetic appeal to answer the core questions of this work.

4.4.1. Reliability

If identifying outliers in a crowdsourcing-based experiment is already tricky, doing so for a subjective evaluation adds an extra layer of complexity to the task (Kittur et al 2008). That is because for judgment measurements, there is no right or wrong answer, e.g. if an image of a cat is pretty for ten people except for one, that does not mean that the person is an outlier, nor that he did not complete the task in a rigorous way. Therefore, to filter out the outliers we adopted a multi-fold technique (see Figure 38):

1. First, we excluded the 12 participants that did not answer the verification question correctly. As mentioned before, this question had been incorporated at the beginning of the application as extra reliability measure (see Figure 29 in section 4.2.3.) to create a pseudo-reliable panel of participants, following the recommendations of (Hossfeld et al 2013): 6 were Facebook (volunteer) users and the remaining 6 were Microworkers' users.
2. Second, we rejected all the 42 participants not rating seriously and being distracted during the subjective test, whose observation time standard deviation exceeded 20 seconds. This value was chosen to accommodate possible variations in download speeds of different participants but reject participants with significantly high variations in completion times (Redi et al 2013): 20 were Facebook (volunteer) users and the remaining 22 were Microworkers' users.
3. Content questions about test contents were added to check reliability as well. These questions (18 in total) would appear after the last four-image presentation in each task, but had limited time to be answered (timer), which may have prevented serious participants to provide the right answer. Thus, we decided to only exclude the 39 participants that answered more than one question wrong.: 11 were Facebook (volunteer) users and the remaining 28 were Microworkers' users.
4. Due to the difficulty of discriminating a right answer from a wrong answer in preference judgments, we decided instead to calculate the amount of shared taste (Carbon et al 2011) between participants. For that reason, we computed the correlation between each participant's raw scores for each image and the average of the remaining participants. Through this analysis, we identified 3 extra participants who had suspiciously rated all images with the same score: 3 Microworkers' users.

On this last criteria, the minimum, maximum and average correlation can be found in Table 9, excluding the participants filtered out by these last stage. In order to compare with the laboratory data, we have also repeated the same correlation procedure for the participants of the laboratory experiment of Chapter 3. As one can see, even though the maximum and minimum existent correlations differ, the average correlation is comparable for both set ups.

	Laboratory experiment	**Crowdsourcing experiment**
Minimum correlation found	0.3485	-0.4071
Maximum correlation found	0.5913	0.8944
Average of the correlations	0.4796	0.4074

Table 9: Participant rating agreement from our previous (laboratory) experiment and from the crowdsourcing experiment.

In total, through the four abovementioned steps, the number of Microworkers and volunteer participants was reduced by about 34.3% and 27.8% respectively. Namely, in Figure 38, we can see the proportion of participants that were excluded in each phase. Further, the final number of each was 113 Microworkers and 96 volunteer participants.

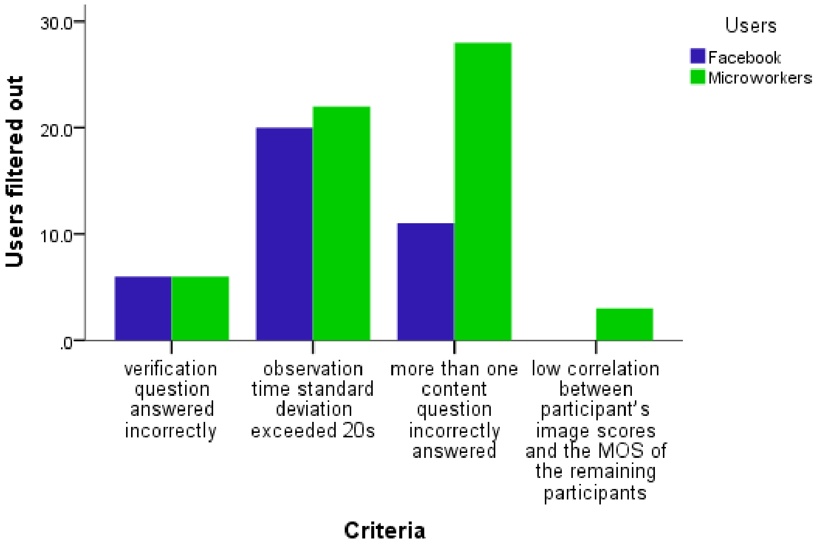


Figure 38: Comparison between the number of Facebook and Microworkers users filtered out through the adopted multi-fold technique.

Eventually, 79 workers from region 1 and 34 from region 2 participated and for every 20 images rated they earned 0.30 USD. The total cost of the experiment was approximately 254 USD. By comparing with (Redi et al 2013), we had a similar number of participants from region 1 and 2, a higher number of images to rate and we spent less. Also, we filtered out 31% of our total participants, less than in (Redi et al 2013). In fact, even if we do not consider the workers from region 3 in (Redi et al 2013), we have filtered approximately the same as in (Redi et al 2013). A possible reason for these results is the added value of gamification, which confirm the findings reported in (Hossfeld et al 2013).

The ratio of unreliable volunteers was smaller with respect to Microworkers, nevertheless, a large amount of Facebook users did not get through with the first session of the experiment. We can conclude here that, of those volunteer Facebook participants that completed at least a series of 4x5 ratings, they took the task more seriously than the Microworkers. However, we had a large dropout rate. The reason for this might be that we did not promote the app enough or that the gamification needs to be improved. Also, the high number of visitors that haven't played the game shows that Facebook policies might have scared the users, with the permissions asked. Further research should be considered on that fact.

Finally, we considered the SOS hypothesis presented in (Hossfeld et al 2011) and previously described in equation (1) (see section 2.3.3.). For that purpose, we used the raw opinion scores on aesthetic appeal provided by the crowdsourcing experiment to compute the MOS and SOS according to equation (2) and (3), respectively (also in section 2.3.3.)..

The SOS hypothesis was employed to check the remaining 209 participants inter-rater consistency. We used a non-linear least squared function to obtain an SOS parameter (a) of 0.30. According to (Hossfeld et al 2011), the SOS parameter a is about 0.30 for web surfing and about 0.17 for image quality studies. According to the criteria used in (Hossfeld et al 2011), that considered online experiments closely related to web surfing, we can then claim our filtered experimental data to be valid suitable for the subsequent analysis.

4.4.2. Comparison of crowdsourcing-based and laboratory-based results

In our previous work of (Redi et al 2013), we replicated the experiment in Chapter 3 but in a crowdsourcing setting and we only collected ratings on images aesthetic appeal and content recognisability . The crowdsourcing-based experiment in (Redi et al 2013) was implemented using the QualityCrowd framework coupled with the Microworkers platform. Moreover, for this experiment three different regions were selected based on their countries' adequate expertise in the English language: region 1 (CS-R1) covered North America and major English speaking countries, such as USA, UK, Canada, and Australia, region 2 (CS-R2) covered Western Europe, including France, Germany, Italy, Ireland, the Netherlands, and Sweden, and region 3 (CS-R3) covered Asia, with Bangladesh, India, Pakistan, Philippines, Singapore, and Thailand. Our aim in this work was to compare laboratory and crowdsourcing evaluations, as well as the performance between the three regions. We found that crowdsourcing-based evaluations can be rather consistent with laboratory-based evaluations when scoring recognisability , but not so much for aesthetic appeal. In general, our results showed the performance of region CS-R1 and region CS-R2 to be comparable, which was not the case for region 3.

It was then interesting to look whether there was some consistency in how people would rate aesthetic appeal of images, even though the different methodology (Single Stimulus for the first two experiments and 4-wise comparison for the last) and the level of control of environmental conditions (low for CS experiments and high for the lab experiment). Thus, at this point, we compared the outcomes of this last experiment with those of the lab-based experiment described in Chapter 3 and those of our previous Crowdsourcing experiment (Redi et al 2013).

For this comparison, we will be using only the aesthetic appeal scores related to the 90 images shared by these three experiments (i.e., those evaluated in the present, Facebook-based experiment). In the following analysis, we will adopt CS-R1, CS-R2, CS-R3 to denote the crowdsourcing data from the 3 regions examined in (Redi et al 2013), Lab to denote the data from the laboratory experiment of Chapter 3 and FB to denote the data from this last experiment.

When comparing opinion ratings across different studies, user studies typically recommend usage of the MOS and standard deviation of the obtained scores for comparisons (Hossfeld et al 2011). As in (Redi et al 2013), we first compared how well did the participants understood the task by calculating the level of inter-participant consistency for aesthetics. In order to do so, we averaged the normalized single scores' standard deviation assigned to each image by all participants in each of the experiments. This method is connected to the SOS hypothesis analysis did before. While the SOS hypothesis method yields an alternative procedure to compare different quality rating scales due to evidence found that the above-mentioned measures do not only depend on the underlying technical conditions of the system under test but are also affected by the rating scales used. In our case, we kept a 5 discrete rating scale in the 3 experimental set-ups, which allow us to compare standard deviations of the opinion scores without the need to worry about discretization.

Table 10 shows that participants were able to score images with a rather similar degree of consistency across all experiments, which allows for further comparison of the MAOZ collected in the current experiment.

	Lab	**FB**	**CS-R1**	**CS-R2**	**CS-R3**
Inter-observer consistency	0.8205	0.8484	0.6947	0.7127	0.7958

Table 10: Averaged the scores' standard deviation assigned to each image by all participants in each of the experiments.

To find out whether the way people rated image aesthetic appeal differed due to the different methodology and the level of control on environmental conditions from Lab to FB, we performed a Mann-Whitney U test on the collected MAOZ from both FB and Lab. We chose this nonparametric test because our data showed not to be normally distributed. Both MAOZ distributions showed to be significantly different (U = 1545, p = 0.000), specifically the MAOZ collected from FB showed to have highest aesthetic scores. This is a known effect in literature, for example study in (Jiang et al 2008) also reports that crowdsourcing scores are typically skewed towards the top end of the scale.

Our next step was to measure the strength of association between all the MAOZ collected in all the different experiments. To do so, we computed the Pearson correlation between the MAOZ values obtained in each experiment. All Pearson correlation coefficients showed to be statistically significant and are reported in Table 11. Nevertheless, it is visible that the Lab MAOZ have a generally weak correlation with Crowdsourcing scores, which is weakest for the CS-R3 experiment (in which, though, participants were found to be the least reliable (Redi et al 2013). Higher correlations can be found between the scores obtained in the Facebook experiment and those from the experiment in (Redi et al 2013), despite still lower than those found within the different regions in which experiment (Redi et al 2013) was performed. This fact is interesting as the experimental methodology (same for the Lab and the (Redi et al 2013) experiment seems to influence less the consistency of the scores across experiments than the environment in which the experiment is performed does (laboratory versus crowdsourcing). Results are more consistent within crowdsourcing experiments than between lab and crowdsourcing perhaps due to the similar level of control on environmental conditions. An explanation might have to do with the participant pool shared by both crowdsourcing setups. An alternative explanation might be due to high pool of participants used in both setups which leads to a more representative sample of the population than in a lab setup.

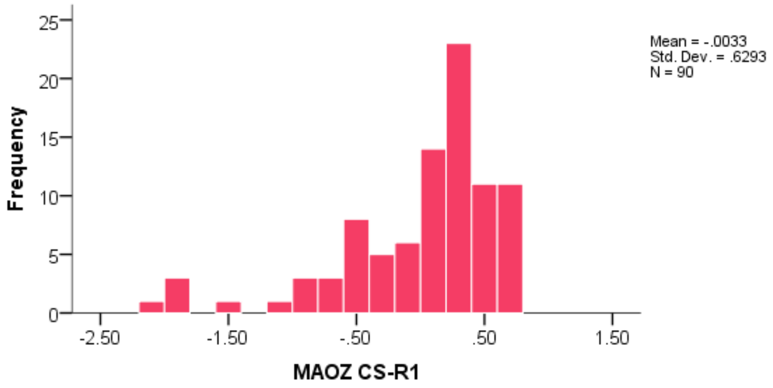
	FB	CS-R1	CS-R2	CS-R3
Lab	0.336	0.469	0.505	0.273
FB		0.664	0.668	0.613
CS-R1			0.938	0.805
CS-R2				0.808

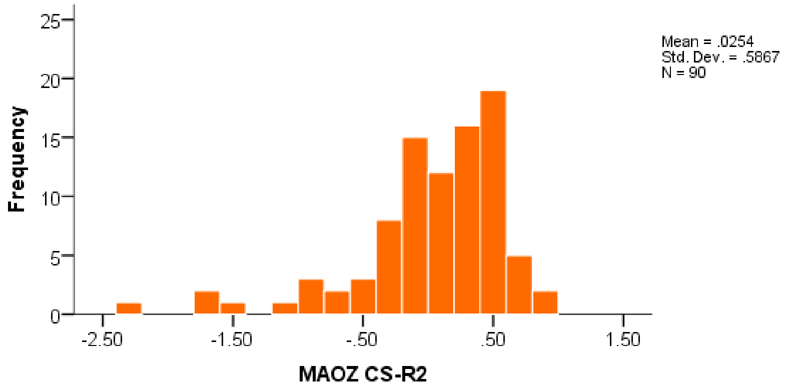
Table 11: All Pearson correlation coefficients showed to be statistically significant.

Finally, to be able to compare scores from these different experiments, we plotted each experiment aesthetic appeal MAOZ in a histogram with a common comparable scale in Figure 39. As one can see, crowdsourcing score distributions appear to be quite similar, skewed towards the top right end of the axis MAOZ scale. Lab scores instead seem more uniformly distributed across the whole scale. Further, Table 12 reports the skewness of each distribution. Despite of the fact that all the distributions share a negative skewness, the crowdsourcing distributions share a comparable value of skewness superior than the lab distribution, which is close to a zero skewness of a normal distribution.

	FB	CS-R1	CS-R2	CS-R3	Lab
Skewness	-1.177	-1.531	-1.577	-1.925	-.401

Table 12: Computed skewness of each distribution.





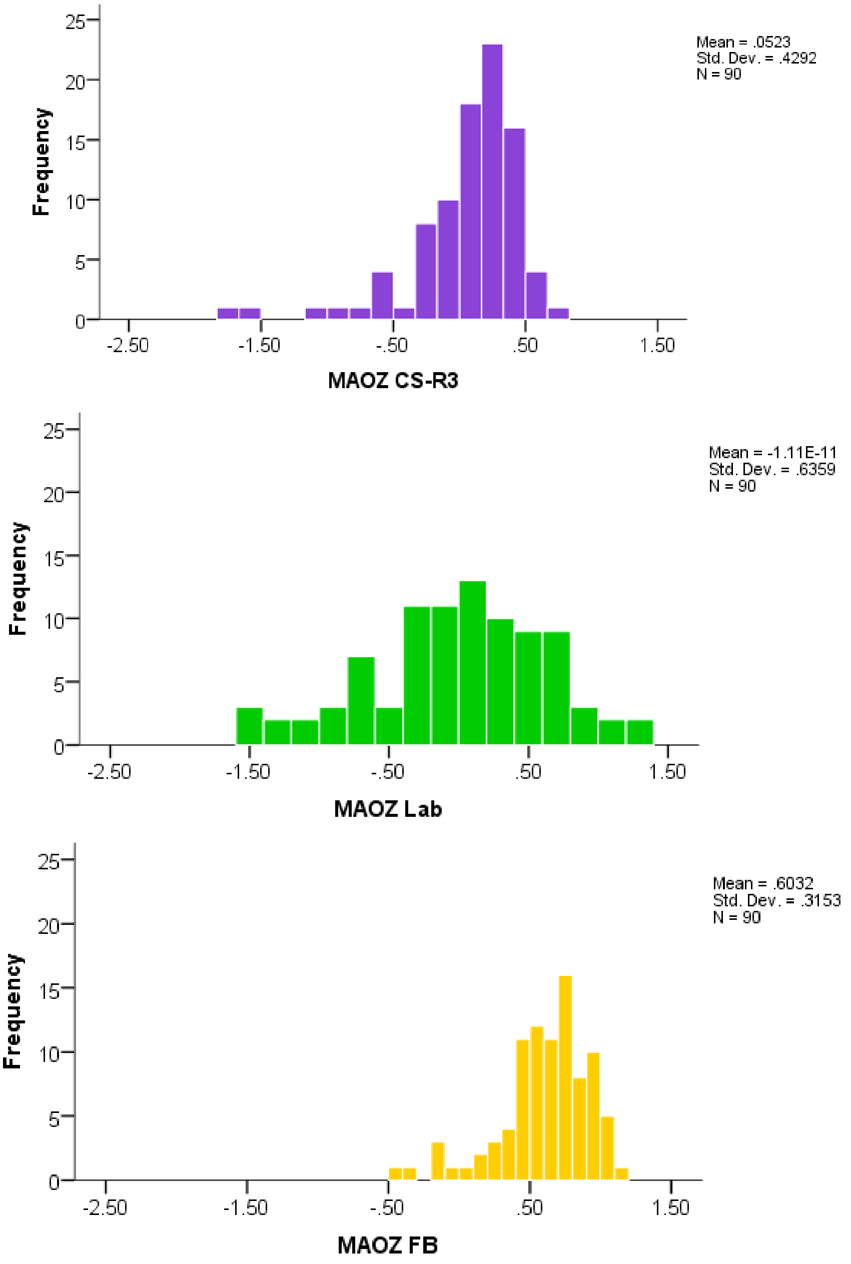


Figure 39: Histograms of the distributions of the MOAZ from Lab, FB , CS-R1, CS-R2 and CS-R3.

To appreciate better to what extent aesthetic appeal scores of one experiment could predict those of another experiment, at least at a coarse level, we assess the quality of the crowdsourcing outcomes from FB, CS-R1, CS-R2 and CS-R3 relatively to the Lab scores, we generated from each distribution of MAOZ three binned categories with an equal number of cases, based on equal percentiles, each containing 33.3% of the cases. Images in the first category (lower 33% of the scores) were considered as characterized by "low aesthetic appeal", those with scores within the 33% and 66% percent of the overall aesthetic appeal distribution were considered of "medium aesthetic appeal, and the remaining ones were included in the "high aesthetic appeal" category. We were then interested in checking whether images would be assigned to the same category across different experiments. To verify this, we built the confusion matrixes (Kohavi et al 1998) shown in Table 13. A confusion matrix, also called contingency table, usually contains information about the expected and predicted categories for classifier models. In this case, we set as expected categories the aesthetic appeal categories assigned in our present (FB) experiment. The confusion matrix shows how many of the images were classified in the same category in the crowdsourcing experiments and how many where judged as having a different aesthetic appeal level. The diagonal elements represent how many times the crowdsourcing-based aesthetic appeal category overlapped the laboratory-based category, while the off-diagonal elements

represent those that were distinct. The higher the diagonal values the best, and the accuracy of the matching is measured as the sum of the diagonal elements over the total number of images.
The accuracy is computed in Table 14. It can be seen that in this case the similarity in evaluations is comparable throughout the three experiments. This implies that the scores from the experiment FB did a good of a job and that the difference between the current experimental data and the reference data from (Redi et al 2013) and from Chapter 3. Interestingly, the accuracy was higher for CS-R1, and as mentioned previously, our participants from coming Microworkers were in their majority also from region 1.

		Predicted (FB)		
		1	2	3
Expected (Lab)	1	13	9	8
	2	11	11	8
	3	7	9	14

		Predicted (FB)		
		1	2	3
Expected (CS-R1)	1	18	8	4
	2	7	12	11
	3	6	9	15

		Predicted (FB)		
		1	2	3
Expected (CS-R2)	1	19	8	3
	2	7	13	10
	3	5	8	17

		Predicted (FB)		
		2	3	4
Expected (CS-R3)	2	18	8	4
	3	7	10	13
	4	6	11	13

Table 13: Confusion matrix of the predicted scores from FB on the expected scores of Lab, CS-R1, CS-R2 and CS-3, respectively.

	Predicted (FB)
Expected (CS-R1)	0.500
Expected (CS-R2)	0.544
Expected (CS-R3)	0.456
Expected (Lab)	0.422

Table 14: Computed accuracy for FB scores predictions on the experimental data from (Redi et al 2013) and Chapter 3.

4.4.3. Impact of digital filters on aesthetic appeal of images

As a first step, we checked whether the remaining image rations would follow a normal distribution to properly setup the following analysis. In this case, we used the Kolmogorov-Sminnov test normality of our data due to its applicability to large datasets. We found our data not to be normally distributed, which indicates the need for using non-parametric tests in the following.
We then addressed the core questions of our research:

- Do filters improve the aesthetic appeal of an image?
- What relations can be established between the usage of the most popular filters and the visible attributes and computed features of a photograph?

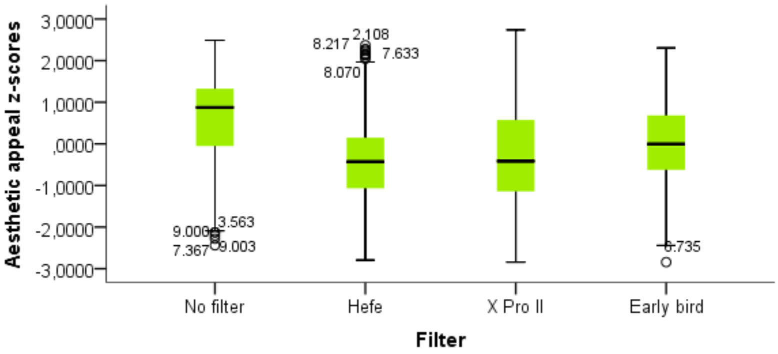


Figure 40: Graphically representation of the distribution of aesthetic scores for each filter type.

To answer both questions, we first of all ran a Kruskal-Wallis H test to check whether the distributions of the scores obtained for the normal images and the images filtered with Hefe, Xpro and Early bird separately had the same median. We can report that there was a statistically significant difference between the different types of filter (chi = 197.793, df = 3, p = 0.000). To complement this test, we used the Mann-Whitney U test. It was then clear that the absence of filter group scored significantly higher in aesthetics compared to the images with the filter Hefe (U = 195.5, p = 0.000), X Pro II (U = 335.0, p = 0.000) and Early bird (U = 699.5, p = 0.000). In respect to the type of filter, images with the filter Early bird revealed to have significantly higher aesthetic appeal scores than the ones with Hefe (U = 1458.5, p = 0.000) and X Pro II (U = 2332.0, p = 0.000). These results can also be visualized in the boxplot in Figure 40. As a result, we could answer to our questions that 1) at least for the consumer images included our dataset, there was no added value in terms of aesthetic appeal when applying a filter; rather, the aesthetic appeal decreased and 2) the early bird filter seems to produce more aesthetically pleasing results than the other two. As mentioned before, in terms of complexity, Early bird modifies more an image than Hefe and X Pro II. Besides altering brightness, contrast and fill the image with a different colour that contrasts with the given image as the other filters, Early bird alters hue and saturation as well as adjusts intensity levels of image shadows, mid-tones, and highlights. Further, Early bird refines the edges of an image to make them finer. The high level of manipulation that the filter does might be behind the high aesthetic appeal scores, although, to be certain, further research is necessary. We can then agree that the popularity associated to this filter (see section 4.1.1. Figure 20) is related to the aesthetic appeal created by its employment.

To investigate better the impact of the filters on the image features and, in turn, their impact on the aesthetic appeal, we again referred to the features measured in Chapter 3: the adapted versions of the colourfulness feature from (Hasler et al 2003) and the contrast feature in (Matkovic 2004). In Chapter 3, we had already computed these two features for our unfiltered images, therefore, we ran again both algorithms in our filtered images.

Afterwards, we tested whether, per filter, the aesthetic MAOZ per image, the colourfulness value and the contrast value were linearly correlated. The statistically significant ($p < 0.05$) Pearson correlation coefficients can be seen in Table 15 with a green cell background, while the not statistically significant ($p > 0.05$) coefficients are represented with a cell red background. Despite the fact that these results did not show any interesting relation, we decided to study these effects graphically and use boxplots to depict the aesthetic appeal sing z-scores in terms of colourfulness and contrast, as well as to depict the different types of filters in terms of colourfulness and contrast (see Figure 41 and Figure 42). As expected, these did not show any trend between aesthetic appeal and colourfulness or contrast but allowed us to quickly examine these distributions explicitly. On what concerns to the different types of filters, images with X Pro II have in general higher colourfulness but share the same average value as for images with Hefe, which means that images Hefe are more concentrated near the mean value (level 4). Compared to no filter, Early bird images' colourfulness has the same average value (level 3) but are more concentrated around this level while no filter has a higher dispersion towards the high values of colourfulness. Instead, for contrast, all the different versions of images have the same average value of contrast, although no filter and Hefe have a higher dispersion towards lower values of contrast while in contrast X Pro II images have a higher dispersion towards higher values of contrast. Then, Early bird images are distributed evenly towards higher and lower values. Although, due to the correlations between these features and the aesthetic scores of images with Early bird, we cannot relate these differences with the higher aesthetic appeal scores of this type of filter compared with the remaining ones.

	Colourfulness	No filter
Contrast	-0.297	0.245
Colourfulness		-0.176
	Colourfulness	**Hefe**
Contrast	-0.117	0.130
Colourfulness		-0.167
	Colourfulness	**X Pro II**
Contrast	-0.302	-0.152
Colourfulness		0.197
	Colourfulness	**Early bird**
Contrast	-0.178	0.026
Colourfulness		-0.003

Table 15: Pearson correlations between the features designed in Chapter 3 and each sample with different filters of aesthetic appeal scores.

Furthermore, we looked for a possible effect of each of different levels of colourfulness and contrast in the aesthetic appeal MAOZ, i.e. to understand whether people's aesthetic rating differed based on the level of contrast or colourfulness amongst images. For that purpose, we ran a Kruskal-Wallis H-test between the aesthetic MAOZ distribution and each feature. Both test results between the aesthetic MAOZ and the colourfulness levels and between the aesthetic MAOZ and the contrast levels did not show a statistically significant difference between the different colourfulness levels or the contrast levels in aesthetic appeal. This means that when people rated aesthetic appeal, their rating behaviour did not changed based on the 6 levels of colourfulness neither based on the5 levels of contrast.

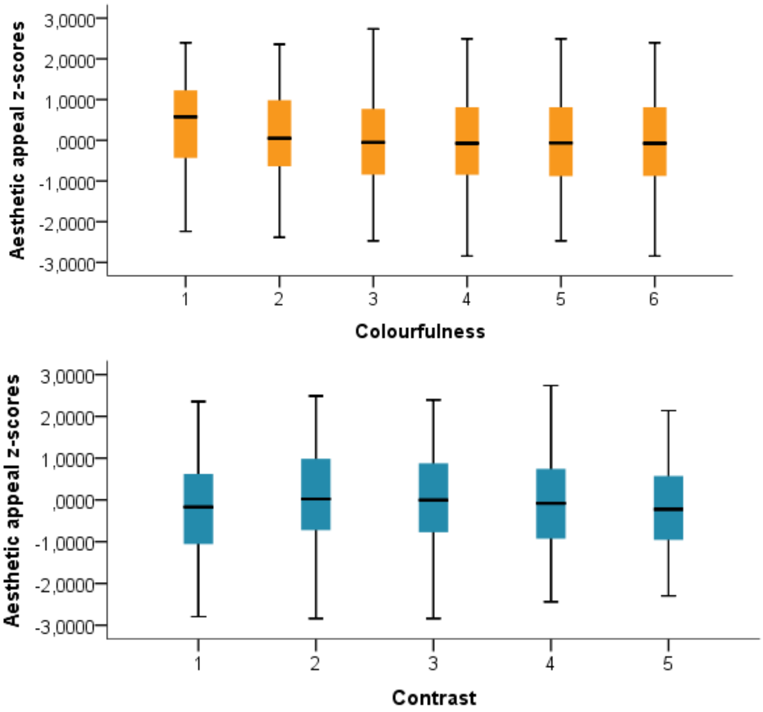


Figure 41: Boxplots on the distribution of aesthetic single z-scores for each level of the adopted features.

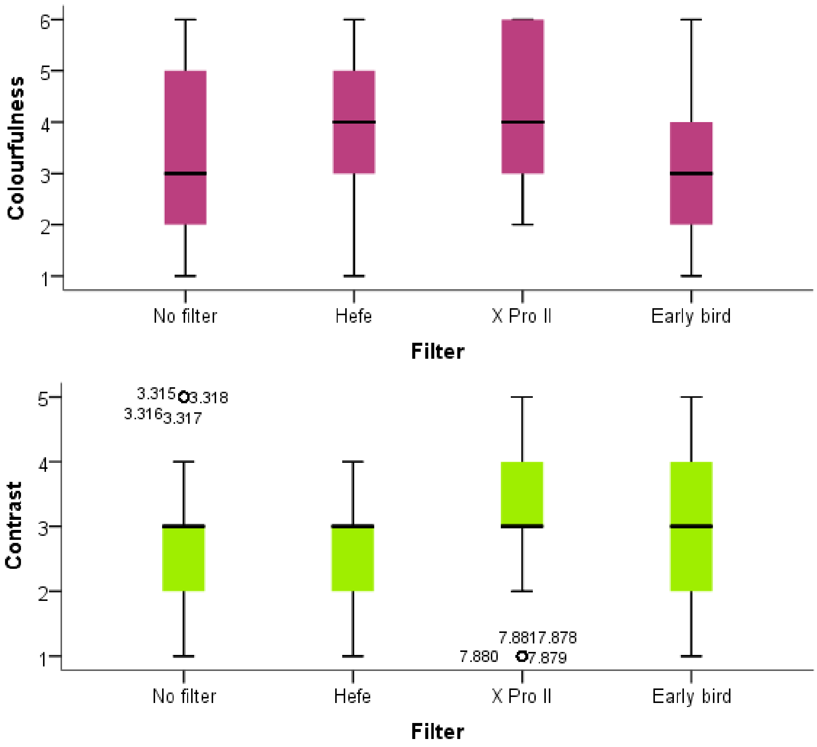


Figure 42: Boxplots on the distribution of the adopted features for each type of filter.

In addition, we were interested in finding out whether the application of a specific filter on an image, with a certain original level of colourfulness or contrast would help in terms of increased aesthetic appeal. Therefore to study how do the aesthetic appeal scores change when a filter is applied, given the contrast or colourfulness original values of the image, we plotted the line graphs in Figure 43.

Based on the original (unfiltered) contrast, in the case of the Early bird filter the aesthetic appeal increases slightly only for initial high contrast images but overall is constant for the other initial levels of contrast. A more noticeable increase of aesthetic appeal for initial high contrast images can be seen when using Hefe. Besides, there is only a slight increase for initial low-medium contrast images (levels 2 and 3). The opposite effect happens for X Pro II, in which aesthetic appeal increases slightly for low contrast images and decreases for high contrast images.

Regarding the original (unfiltered) colourfulness, in the case of the Early bird filter, a slight increase of aesthetic appeal can be seen for images with medium-high initial colourfulness (level 4). For the remaining images the aesthetic appeal seems to vary very little. For Hefe instead, there is a small aesthetic appeal increase for images with low-medium initial colourfulness (levels 2 and 3) and a decrease of aesthetic appeal for images with initial high colourfulness. In contrast, X Pro II usage decreases aesthetic appeal for initial low-medium colourfulness (level 3) and increases for initial high colourfulness. Additionally, it is clear that X Pro II increases the contrast more than the other two filters, which might be overused for the selected images (see Figure 42 section 4.4.3.).

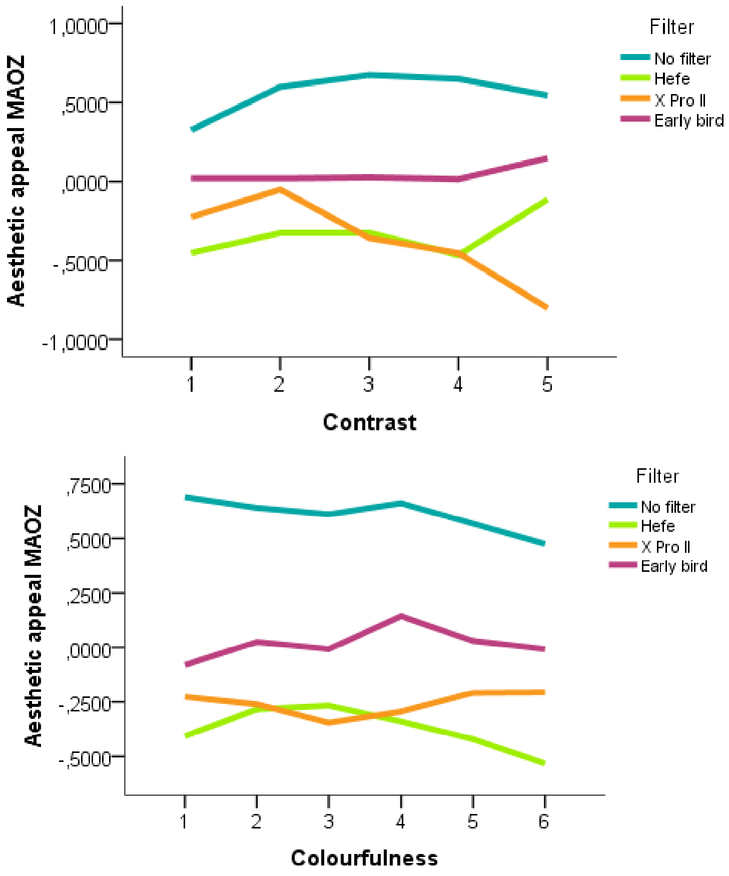


Figure 43: Line graphs to show how aesthetic appeal scores change when a filter is applied, given the contrast or colourfulness original values of the image.

To finalize our analysis, we looked at the recognisability data from the first experiment. As one might recall, recognisability had been defined as how clear the subject of the image is to the observer. For example, the content of images presenting some kind of distortion, such as blur, might be difficult to recognize. Then, what is familiar has to be recognizable, whereas what is recognizable may or not be familiar. Familiarity, instead was defined as how often has the participant seen the content of the image and that might change in terms of changes in colour, e.g. a cat with three yellow spots might be more familiar if one has a cat with three yellow spots at home but with the usage of filters colours can perhaps change the effect of familiarity. Additionally, from the attributes studied in Chapter 3, recognisability had shown a significant almost moderate correlation with aesthetic appeal. Nevertheless, on the assumption that content recognisability does not change with the filter usage, because a filter doesn't now distort an image at the point that a cat is not a cat anymore, we focused this analysis only on content recognisability . Therefore, we assigned the same recognisability scores collected on the images without filter to the ones with filters and plot these in Figure 44. Next, to assess how well recognisability scores can explain and predict aesthetic appeal, we determined the coefficients of determination (see Figure 44). The coefficients of determination is simply the squared value of the correlation coefficient. These represent the proportion of the variation from one attribute that is predictable from the other attribute, which, with the help of the graph, allow us to estimate to what extent aesthetic appeal can be predicted based on image recognisability . It appears that 36% of the total variation in aesthetic appeal can be explained by recognisability. In unfiltered images, when filters are applied the association between recognisability and aesthetic appeal reduces its strength significantly.

Additionally, the coefficient of determination is the square of the Pearson correlation coefficient. So the coefficient of determination 0.354 between recognisability and aesthetic appeal for images with no filter will result in a Pearson correlation coefficient of 0.595. Likewise, in Chapter 3 we also exposed a positive correlation between these attributes but much weaker. One possible reason might be due to the effect of the other attributes. We had shown that the colour likeability attribute was a suppressive variable in this relation. Another possible reason might be the small pool of participants used in the first experiment.

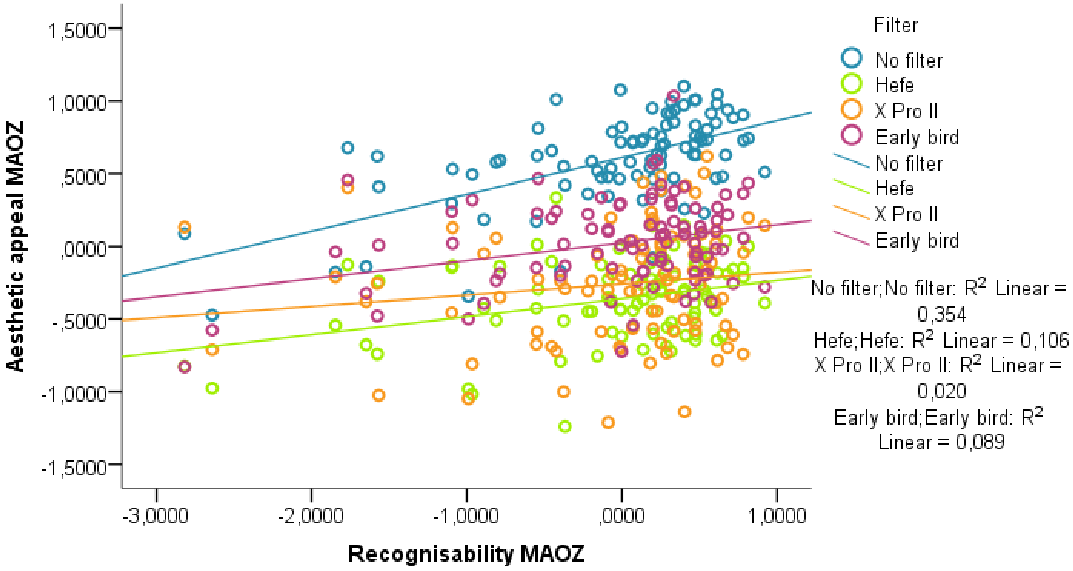


Figure 44: Dependency of aesthetic appeal MAOZ on recognisability MAOZ.

4.5. Conclusions

In this chapter we investigated the impact of digital filters on aesthetic appeal using a crowdsourcing approach in contrast to Chapter 3. For that purpose, we choose the most popular filters in Instagram: Early bird, X Pro II and Hefe.

We used the contrast and colourfulness features, ROI and aesthetic appeal MAOZ collected in the previous chapter to build a representative dataset for this experiment.

Our approach entailed the design of a web app with a playful interface hosted in the social network Facebook and coupled with the crowd-provider platform Microworkers. This latter point allowed us to elaborate on the importance of a good crowdsourcing design as well as to compare the two types of users: paid and volunteer. From the 672 total number of users registered, 367 only visited the app. These were mainly Facebook (volunteer) participants which might have been due to their curiosity, whereas Microworkers (paid) participants had a reward as second incentive. Another alternative explanations might have be because Microworkers participants were already acquainted with similar tasks or because these were given an explanation of the tasks beforehand, besides of the clarification in the app itself. That can also explain why Microworkers participants rated more images in average than Facebook participants.

The remaining 305 participants were then succumbed to a reliability data analysis which comprehended a multi-fold approach. The number of Facebook and Microworkers participants was reduced here by about 34.3% and 27.8%, respectively. By comparison with the crowdsourcing experiment in (Redi et al 2013), we had a similar number of participants, a higher number of images to rate and we spent less. Additionally, we filtered less participants (31%) than the mentioned study. A possible reason for these numbers might be due to the added value of gamification.

We then compared the collected data from the different types of setups: crowdsourcing from the current experiment and from (Redi et al 2013) and laboratory from Chapter 3. Participants showed to score images with a rather similar degree of consistency across the experiments in the crowdsourcing environment rather than when comparing with the laboratory setup. This can be either explained by the shared crowd of participants between this experiment and the one in (Redi et al 2013) or due to high pool of participants used in both setups which leads to a more representative sample of the population than in a lab setup.

In terms of score distributions, crowdsourcing experiments showed to share the same negative skewness, in accordance with what found in the literature (Jiang et al 2008).

Regarding the impact of digital filters in aesthetic appeal, we found in our dataset, there was no added value in terms of aesthetic appeal when applying one of the three most popular filters. Instead, amongst the filters, Early bird filter seemed to produce more aesthetically pleasing scores than Hefe and X Pro II. This seems to be in accordance with the correspondingly high popularity of this filter in user statistics (see section 4.1.1. Figure 20). The high level of manipulation behind this filter compared with the other two might be an reason for its high aesthetic appeal scores.

Furthermore, the features colourfulness and contrast described in Chapter 2 (see section 2.4.2.) did not show any interesting relation with the aesthetic scores of any of the filters. Therefore, we cannot relate the differences in colourfulness or contrast with the high or low aesthetic scores of any type of filter.

We also found that participant’s rating behaviour while rating aesthetic appeal did not changed based on the different levels of colourfulness neither on the different levels of contrast. Then, we checked aesthetic preference based on the original (unfiltered) contrast and colourfulness. We showed that if the image has originally a low level of contrast the application of X Pro II will help in terms of increased aesthetic appeal. In contrast, if the image has originally a high level of contrast the application of Hefe or Early bird will help in terms of increased aesthetic appeal. Given the original values colourfulness, the aesthetic appeal scores will change in an approximate inverse way when each filter is applied. For originally low colourfulness the aesthetic appeal scores will be higher upon the usage of Hefe and Early bird and for originally high colourfulness the aesthetic appeal scores will be higher upon the usage of X Pro II. Additionally, it is clear that X Pro II increases the contrast more than the other two filters, which might be overused for the selected images (see Figure 42 section 4.4.3.).

Finally, we looked at relation between recognisability scores collected in Chapter 3 and the aesthetic scores collected in this experiment and we unveiled a higher correlation between recognisability and aesthetic appeal than the one reported in Chapter 3. A possible explanation might be due to the effect of the other attributes studied in Chapter 3. Accordingly, we had shown that the attribute colour likeability was a suppressive variable in this relation. Other possible explanations might be caused by the small pool of participants used in the first experiment or by the subset of images that we selected.

5. Conclusions

In this work, we have investigated the added value of digital filters on aesthetic appeal of images using a crowdsourcing approach. Recently, a big hype on social media has started with a new social network called Instagram. Instagram is a smartphone application as well as a photo-based social network, which provides users with digital filters giving a vintage antique look to your photos. Since its appearance, more and more people have become a digital filter user. The motivation behind this study was to address the question whether images with Instagram like filters are more aesthetically pleasant as well as adding knowledge to field of Computational Aesthetics on the understanding the processes of an aesthetic appreciation of a photo.
To reach this understanding, an empirical study of user preferences is necessary. This entails collecting users opinions on large sets of photographic material, which is costly and time consuming when performed through traditional methods in controllable laboratory settings. Crowdsourcing offers an alternative, granting access to a diverse and numerous set of users that for a small compensation are willing to perform tasks such as rating the aesthetic appeal of images. Nevertheless, crowdsourcing exposes to high risk of collecting unreliable ratings. In this thesis, we tackled this risk by enhancing crowdsourcing tasks with gamification, as well as determining a rigorous method to establish the reliability of the ratings handed in by the test participants.

Our work contributes for the computational aesthetic community with the added knowledge on relationship between filters, features attributes and aesthetic appeal. Besides, we have also proven the added value of gamification in crowdsourcing. Further, we have reported on the differences in reliability between Microworkers (paid) and Facebook (volunteer) users and on the reproduce of laboratory results in a crowdsourcing environment. We also have shown that digital filters do not enhance aesthetic appeal of pictures as we expected.

For that purpose, as a first step towards our understanding of the added value of digital filters in image aesthetic appeal, we run a lab-based, pilot scaling study to collect information on the objective (features) and subjective (attributes) image properties that impact aesthetic appeal. We collected subjective judgments in terms of recognisability, familiarity, colour likeability and aesthetic appeal on a big dataset previously selected and categorized in terms of content.

We selected then three Instagram like filters in terms of their popularity in Instagram: Early bird, Hefe and X Pro II. Our dataset was then reduced so that we could apply these filters to a representative dataset in terms of aesthetic appeal, salient regions and the two features studied: colourfulness and contrast. Our following step was then materialized by the design of our main experiment as a Facebook playful web application, later coupled with the crowd-provider Microworkers. In this latter experiment, we have only collected aesthetic appeal scores, in contrast with the first experiment. To address reliability we spend some effort designing a playful interface, which had shown to lead to successful results in the crowdsourcing literature. Besides, we added in our implementation many elements recommended in crowdsourcing user studies. It should be noted that the present work did not follow the traditional technique used in other fields of making use of a crowdsourcing platform with a simple web-based interface.

Our initial collected data has revealed a strong correlation between colour likeability and aesthetic appeal as well as between familiarity and recognisability. In relation to the features contrast and colourfulness, we have unveiled that colour likeability decreases after a contrast increase and increases when colourfulness increase, which is in agreement with psychological studies. Also, it seems that low contrast images rank higher in aesthetics, which goes against what has been showed in the literature. Moreover, low contrast images might take longer to observe because its content is harder to distinguish. Our analysis on the impact of the image content on the aesthetic attributes and aesthetic appeal revealed interesting results, namely confirmed people low aesthetic appeal towards abstract photography and that in contrast to the literature, images with people scored low in aesthetics.
On what concerns to our second experiment, we had two types of participants: Facebook (volunteer) and Microworkers (paid) participants. Our analyse revealed that many Facebook users only visited the application developed perhaps dragged by curiosity. In contrast, the small number of Microworkers that have not actively participated might be explained by the reward as an incentive or because these were already acquainted with similar tasks or even because they were given an explanation of the task beforehand, besides of the clarification in the app itself. These reasons can also explain why Microworkers participants rated more images in average than Facebook participants.

We then developed a thorough data analysis approach to identify unreliable users. The number of participants filtered out showed to be lower than in our previous work where we used a Microworkers coupled with a simple web-based interface. The added value of gamification might explain these outcomes.

Further, Facebook users have proven to be more reliable in general than paid users by 6.5%. This emphasizes the fact that if a user study is implemented in an joyful manner, Facebook presents a good platform to get data faster and cheaply with higher degree of reliability.

Participants showed to score images with a rather similar degree of consistency across the experiments in the crowdsourcing environment rather than when comparing with the laboratory setup. This can either be explained by the shared crowd of participants between the crowdsourcing experiments or due to the high pool of participants used in both setups which leads to a more representative sample of the population than in a lab setup. Furthermore, in terms of score distributions, crowdsourcing experiments showed to share the same negative skewness, in accordance with what found in the literature.

Regarding the impact of digital filters in aesthetic appeal, we found in our dataset, there was no added value in terms of aesthetic appeal when applying one of the three most popular filters. That result is obviously surprising if the popularity of Instagram is taken into account although it can be supported by the user statistics on the filter usage. Nevertheless it is possible to speculate on the reasons of such outcome. One possible reason might be because we did not taken into account the screen size. Instagram is only available in mobile phones and so it only allows the user to observe an image in a phone screen size. So it might be that if one takes into account the effect of this condition (the small screen size of a mobile phone), the impact of digital filters on aesthetic appeal can instead be higher. Another reason might be because we used a 4-wise comparison as methodology instead of single stimulus. Instagram limits its users to observe one image with a filter (or without) at a time, in a single stimulus condition. Thus, it might be that the users are affected by a possible shift in their internal reference when applying a filter, forgetting the original unfiltered. And so, with this approach the impact of digital filters on aesthetic appeal would instead be higher. Another possible reason maybe that Instagram filter users represent a particular segment of aesthetic preference underrepresented in the experimental sample. Within the three most popular filters, Early bird showed to have higher concentration of high aesthetic scores which agrees with its popularity in the social application. The high level of manipulation behind this filter compared with the other two might be an reason for its high aesthetic appeal scores.

As far as the relationship filter-features-attributes-appeal is concerned, in this second analysis, we did not find any significant correlation between the aesthetic scores of filtered images and the features considered of interest in the lab experiment. Though, we studied how aesthetic preference was enhanced based on the original (unfiltered) contrast and colourfulness. We showed that if the image has originally a low level of contrast the application of X Pro II will help in terms of increased aesthetic appeal. In contrast, if the image has originally a high level of contrast the application of Hefe or Early bird will help in terms of increased aesthetic appeal. Given the original values colourfulness, the aesthetic appeal scores will change in an approximate inverse way when each filter is applied. For originally low colourfulness the aesthetic appeal scores will be higher upon the usage of Hefe and Early bird and for originally high colourfulness the aesthetic appeal scores will be higher upon the usage of X Pro II. Additionally, it is clear that X Pro II increases the contrast more than the other two filters, which might be overused for the selected images.

Moreover we exposed a stronger correlation between aesthetic appeal and recognisability, which was supressed in the first experiment due to the effect of the other attributes. Other possible explanations might be caused by the small pool of participants used in the first experiment or by the subset of images that we selected.

To sum up, we have shown that crowdsourcing can be valuable solution to address the drawbacks associated with the laboratory studies. In addition, our approach presents a valuable solution to address the crowdsourcing reliability problem. The success of our approach in the present work came out from combining a crowdsourcing platform (Microworkers) with gamified experiment based in a web-application hosted in Facebook. We have developed a methodology to test the reliability of the outcomes of a crowdsourcing experiment in a robust way.

Besides a careful planning, explained in detail in this work, precautions were also introduced to filter out suspicious participants when processing the data. An appropriate planning and an adequate safeguard are essential for an ease filtering of suspicious participants.

Nevertheless, more research needs to be done on how to detect outliers and how to fully captivate participants.

6. Recommendations for future research

After presenting the conclusions some recommendations are given for further research in the scope of Computational Aesthetics using crowdsourcing:

- **Development of a crowd-testing framework** – In Computational Aesthetics the added value of this framework would save time and effort when designing the experiment and would allow researchers to focus only in their research question.
- **Consider the mobile screen size** – Since Instagram is only mobile, an interesting research would be on the effect of screen size in the impact of digital filters in aesthetic appeal.
- **Consider a Single Stimulus method –** The Instagram app constrains the user to observe an image with a filter at time (see Figure 45). Therefore, would be interesting to use a similar approach and compare the results of the two experiments. Trying other methodology instead such as paired comparison might also be bring new knowledge concerning digital filters.

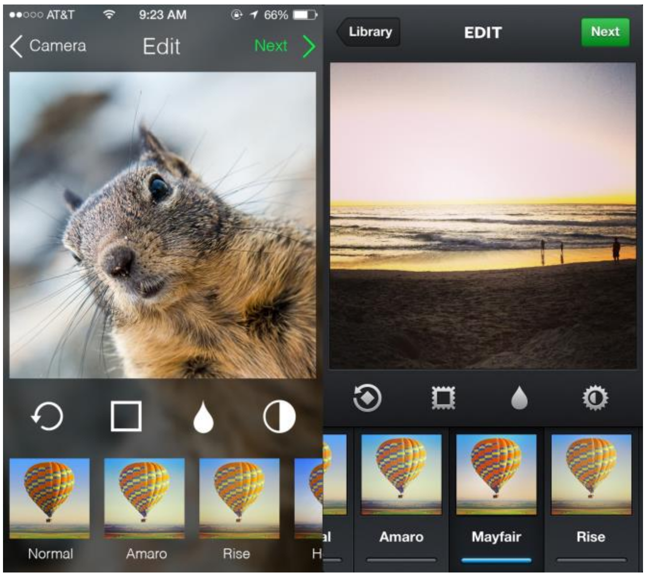


Figure 45: Example of Instagram usage.

- **Consider more usability tests** – In web development, before implementing an online application, one should test his interface with a set of participants to understand how usable it is. This approach might improve considerably the efficiency of crowdsourcing users.
- **Improvement of the converging lines metric** – We believe that the metric implemented needs some improvement in terms of algorithm. Further, the dataset to investigate this metric needs also to comprise this rule, i.e. one needs to select a representative number of images that represent examples of this rule.
- **Improvement of the simplicity of region metric** – The same as above but for this metric instead.
- **Consider sharing the application in other social networks** – In order to attract more users, one can try to share the application link in other social networks like dig, twitter, etc.
- **Improvement of playful factor** – A big point of improvement would be the playful factor of our design. Though playful, we believe that more elements could be added from game design studies.
- **Consider the classic image quality –** It would be interesting to look at the relationship between the attributes gathered (e.g. recognisability) and a no-reference metric such as BLIINDS or BRISQUE from (Sheikh et al 2006)(Wang et al 2004)(Sheikh et al 2005).
- **Consider using the demographic information –** Facebook profiles provide personal information on the users that can be extracted with authorization from them. One could use this information together with the filters usage.
- **Implementation of a computational model –** The development of a develop a predictive model with the collected data on the links between features and attributes and attributes and aesthetic appeal.

- **Consider a deeper analysis –** The preference data collected in this experiment could also be used for further analysis on the impact of the digital filters on the features of an image.

7. Bibliography

[1] **(Abraham et al 2009)** Abraham, A., Hassanien, A.E., Snasel, V. (2009). *Computational Social Network Analysis – Trends, Tools and Research Advances*. Springer.

[2] **(Alers et al 2010)** Alers, H., Liu, H., Redi, J., Heynderickx, I. (2010). Studying the risk of optimizing the image quality in saliency regions at the expense of background content. *Proceedings of IS&T/SPIE Human Vision and Electronic Imaging*.

[3] **(Axelsson 2007)** Axelsson, Ö. (2007). Towards a Psychology of Photography: Dimensions Underlying Aesthetic Appeal of Photographs 1, 2, 3. *Perceptual and Motor Skills, 105(2),* 411-434.

[4] **(Bench et a1 1996)** Bech, S., Hamberg, R., Nijenhuis, M., Teunissen, K., de Jong, H.L., Houben, P., Pramanik, S.K. (1996). Rapid perceptual image description (RaPID) method. *Proceedings of SPIE 2657 – Human Vision and Electronic Imaging,* 317.

[5] **(Berdan 2004)** Berdan, R. (2004). Composition and the elements of visual design. Retrieved January, 31 2014, from: http://photoinf.com/General/Robert_Berdan/Composition_and_the_Elements_of_Visual_Design.htm

[6] **(Bhattacharya et al 2010)** Bhattacharya, S., Sukthankar, R., Mubarak, S. (2010). A Framework for Photo-Quality Assessment and Enhancement based on Visual Aesthetics. *MM '10 – Proceedings of the international conference on Multimedia*, 271-280.

[7] **(Bonanos 2012)** Bonanos, C. (2012). *Instant: The Story of Polaroid*. Princeton Architectural Press.

[8] **(Box 2011)** Box, Daniel. (2011). *Instagram Filters as Photoshop Actions*. Retrieved January, 31 2014, from: http://dbox.tumblr.com/post/5426249009/instagram-filters-as-photoshop-actions

[9] **(Carbon 2011)** Carbon, C.C., (2011). Cognitive mechanisms for explaining dynamics of aesthetic appreciation. *i-Perception, 2(7),* 708 – 719.

[10] **(Cerosaletti et al 2009)** Cerosaletti, C.D., Loui, A.C. (2009). Measuring the Perceived Aesthetic Quality of Photographic Images. *QoMEx 2009 – International Workshop on Quality of Multimedia Experience*, 47-52.

[11] **(Cerosaletti et al 2011)** Cerosaletti, C.D., Loui, A.C., Gallagher, A.C. (2011). Investigating Two Features of Aesthetic Perception in Consumer Photographic Images: Clutter and Center. *Proceedings of SPIE 7865 – Human Vision and Electronic Imaging XVI,* 786507.

[12] **(Congcong et al 2009)** Congcong Li, Tsuhan Chen (2009). Aesthetic Visual Quality Assessment of Paintings. *IEEE Journal of Selected Topics in Signal Processing (Volume:3, Issue: 2)*, 236 – 252.

[13] **(Datta et al 2006)** Datta, R., Joshi, D.., Jia Li, Wang, J.Z. (2006). Studying Aesthetics in Photographic Images Using a Computational Approach. *Computer Vision – ECCV 2006 – Lecture Notes in Computer Science, Volume 3953*, 288-301.

[14] **(Datta et al 2007)** Datta, R., Jia Li, Wang, J. Z. (2007). Learning the consensus on visual quality for next-generation image management. *Proceedings of the 15th international conference on Multimedia – ACM*, 533-536.

[15] **(Dowling 2012)** Dowling, Stephen. (2012). *Did the Lomo camera save film photography?* Retrieved January, 31 2014, from BBC: http://www.bbc.com/news/magazine-20434270

[16] **(Eickho et al 2012)** Eickho, C., Harris, C.G., de Vries, A.P., Srinivasan, P. (2012). Quality Through Flow And Immersion: Gamifying Crowdsourced Relevance Assessments. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,* 871-880.

[17] **(Engeldrum 2000)** Engeldrum, P.G. (2000). *Psychometric Scaling: a toolkit for imaging systems development*. Imcotek Press.

[18] **(Engelke et al 2011)** Engelke, U., Kaprykowsky, H., Zepernick, H., Ndjiki-Nya, P. (2011). Visual Attention in Quality Assessment. *Signal Processing Magazine – IEEE, 28(6),* 50-59.

[19] **(Everingham et al 2010)** Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A. (2010). The PASCAL visual object classes (voc) challenge. *International journal of computer vision, 88(2)*, 303 – 338.

[20] **(Faerber et al 2010)** Faerber, S. J., Leder H., Gerger G., Carbon C.C. (2010). Priming semantic concepts affects the dynamics of aesthetic appreciation. *Acta Psychologica, 135(2),* 191 – 200.

[21] **(Fedorovskaya et al 2013)** Fedorovskaya, E.A., de Ridder, H. (2013). Subjective matters: from image quality to image psychology. *Proceedings of SPIE 8651 – Human Vision and Electronic Imaging XVIII,* 86510O.

[22] **(Freeman 2013)** Freeman, M. (2007). *The Photographer's Eye: Composition and Design for Better Digital Photos*. Focal Press.

[23] **(Griffin et al 2007)** Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset*.

[24] **(Hartmann et al 2008)** Hartmann, J., Sutcliffe, A., De Angeli, A. (2008). Towards a Theory of User Judgment of Aesthetics and User Interface Quality. *Journal ACM Transactions on Computer-Human Interaction (TOCHI), Volume 15, Issue 4,* article No. 15.

[25] **(Hasler et al 2003)** Hasler, D., Suesstrunk, S.E. (2003). Measuring colourfulness in natural images. *Proceedings of SPIE 5007 – Human Vision and Electronic Imaging VIII*, 87.

[26] **(Hoenig 2005)** Hoenig, F. (2005) Defining Computational Aesthetics. *Computational Aesthetics'05 – Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 13-18.

[27] **(Hoffman 2005)** Hoffman, Gretchen. (2013). *How Twitter, Facebook and Blogs Influence Fashion Shoppers*. Retrieved January, 31 2014, from: www.netbase.com/all-posts/how-twitter-facebook-and-blogs-influence-fashion-shoppers/

[28] **(Honan 2005)** Honan, Mathew (2007). *Apple unveils iPhone*. Retrieved January, 31 2014, from: http://macworld.com/article/1054769/iphone.html

[29] **(Hossfeld et al 2011)** Hossfeld, T., Schatz, R., Egger, S. (2011). SOS: The MOS is not enough! Third International Workshop on Quality of Multimedia Experience (QoMEX), 131 – 136.

[30] **(Hossfeld et al 2011)** Hossfeld, T., Schatz, R., Zinner, T., Seufert, M., Tran-Gia. P. (2011). Transport Protocol Influences on YouTube Video streaming QoE. *Technical Report 482, University of Würzburg*.

[31] **(Hossfeld et al 2013)** Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia. P. (2013). CrowdTesting: A Novel Methodology for Subjective User Studies and QoE Evaluation. *Technical Report 486, University of Würzburg*.

[32] **(Ishai et al 2007)** Ishai, A., Fairhall, S.L., Pepperell R. (2007). Perception, memory and aesthetics of indeterminate art. *Brain Research Bulletin, 73(4),* 319-324.

[33] **(Isola et al 2011)** Isola, P., Jianxiong Xiao, Torralba, A., Oliva, A. (2011). What makes an image memorable? *CVPR – IEEE Conference on Computer Vision and Pattern Recognition,* 145 – 152.

[34] **(Itti et al 2001)** Itti, L., Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2,* 194 – 203.

[35] **(ITU 2012)** ITU (2012). *Recommendation ITU-R BT.500-13 (01/2012) – Methodology for the subjective assessment of the quality of television pictures*. International Telecommunication Union.

[36] **(Jiang et al 2008)** Jiang Yang, Adamic, L.A., Ackerman, M.S. (2008). Crowdsourcing and knowledge sharing: strategic user behavior on taskcn. *EC '08 Proceedings of the 9th ACM conference on Electronic commerce,* 246 – 255.

[37] **(Jiang et al 2010)** Jiang, Wei, Loui, A.C., Cerosaletti, C.D. (2010). Automatic Aesthetic Value Assessment in Photographic Images. *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME)*, 920 – 925.

[38] **(Joshi et al 2011)** Joshi, D., Datta, R., Fedorovskaya, E., Quang-Tuan Luong, Wang, J.Z., Jia Li, Jiebo Luo (2011). Aesthetics and Emotions in Images. *IEEE Signal Processing Magazine (Volume: 28 , Issue:5),* 94 – 115.

[39] **(Kittur et al 2008)** Kittur, A., Chi, E.H., Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *CHI' 08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 453-456

[40] **(Koch et al 1987)** Koch, C., Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. *Matters of Intelligence, Synthese Library Volume 188*, 115-141.

[41] **(Kohavi et al 1998)** Kohavi, R., Provost, F. (1998). Glossary of terms. *Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, 30(2–3)*.

[42] **(Komar et al 1997)** Komar, V., Wypijewski, J., Melamid, A. (1997). *Painting by Numbers: Komar and Melamid's Scientific Guide to Art*. University of California Press.

[43] **(Lai-Kuan et al 2009)** Lai-Kuan Wong, Kok-Lim Low (2009). Saliency-enhanced image aesthetics class prediction. *ICIP 2009 – 16th IEEE International Conference on Image Processing*, 997 – 1000.

[44] **(Lassalle et al 2012)** Lassalle, J., Gros, L., Morineau, T., Coppin, G. (2012). Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception? *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 1 – 6.

[45] **(Le Callet et al 2012)** Le Callet, P., Möller, S., and Perkis, A. (2012). Qualinet White Paper on Definitions of Quality of Experience. *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).*

[46] **(Le Callet et al 2013)** Le Callet, P., Niebur, E. (2013). Visual Attention and Applications in Multimedia Technologies. *Proceedings of the IEEE, 101(9),* 2058 – 2067.

[47] **(Leder et al 2004)** Leder, H., Belke, B., Oeberst, A., Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95(4)*, 489 – 508.

[48] **(Lind 1980)** Lind, R. (1980). Attention and the Aesthetic Object. *The Journal of Aesthetics and Art Criticism, Vol. 39, No. 2,* 131-142.

[49] **(Locher 2011)** Locher, P. (2011). Contemporary experimental aesthetics: State of the art technology. *i-Perception, 2(7),* 697–707.

[50] **(Mansilla et al 2011)** Mansilla, W.A., Perkis, A., Ebrahimi, T. (2011). Implicit experiences as a determinant of perceptual quality and aesthetic appreciation. *Proceedings of the 19th ACM international conference on Multimedia*, 153-162.

[51] **(Mantel et al 2013)** Mantel, C., Guyader, N., Ladret, P., Ionescu, G., Kunlin, T. (2013). Characterizing eye movements during temporal and global quality assessment of h.264 compressed video sequences. *Proceedings of SPIE 8291 – Human Vision and Electronic Imaging XVII*, 82910Y.

[52] **(Matkovic et al 2005)** Matkovic, K., Neumann, L., Neumann, A., Psik, T., Purgathofer, W. (2005). Global contrast factor-a new approach to image contrast. *Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 159 – 167.

[53] **(McGarvey et al 2004)** McGarvey, Jim, (2004). *The DCS Story – 17 years of Kodak Professional digital camera systems – 1987-2004*. Retrieved January, 31 2014, from: http://nikonweb.com/files/DCS_Story.pdf

[54] **(Merriam-Webster 2014)** Merriam-Webster. *Aesthetics*. Retrieved January, 31 2014, from: http://merriam-webster.com/dictionary/aesthetics

[55] **(Norusis 1990)** Norusis, M-J. (1990). *SPSS base system user's guide*. SPSS

[56] **(Obrador et al 2010)** Obrador, P., Schmidt-Hackenberg, L., Oliver, N. (2010). The role of image composition in image aesthetics *ICIP 2010 – 17th IEEE International Conference on Image Processing,* 3185 – 3188.

[57] **(Peres 2007)** Peres, M. (2007). *The Focal Encyclopedia of Photography – 4th Edition*. Focal Press.

[58] **(Peters et al 2007)** Peters, G., Aesthetic Primitives of Images for Visualization (2007). *IV '07 – Proceedings of the 11th International Conference Information Visualization*, 316 – 325.

[59] **(Prakel 2009)** Prakel, D. (2009). *The Visual Dictionary of Photography*. AVA Publishing.

[60] **(Redi 2013)** Redi, J.A. (2013). Visual quality beyond artifact visibility. *Proceedings of SPIE 8651 – Human Vision and Electronic Imaging XVIII,* 86510N.

[61] **(Redi et al 2011)** Redi, J., Liu, H., Zunino, R., Heynderickx, I. (2011). Interactions of visual attention and quality perception. *Proceedings of SPIE 7865 – Human Vision and Electronic Imaging XVI*, 78650S.

[62] **(Redi et al 2011)** Redi, J.A., Hantao Liu, Zunino, R., Heynderickx, I. (2011). Interactions of visual attention and quality perception. *Proceedings of SPIE 7865 – Human Vision and Electronic Imaging XVI,* 78650S.

[63] **(Redi et al 2011)** Redi, J.A., Heynderickx, I. (2011). *TUD Image Quality Database: Interactions*. http://mmi.tudelft.nl/iqlab/interactions.html.

[64] **(Redi et al 2011)** Redi, J.A., Heynderickx, I., (2011). Image quality and visual attention interactions: Towards a more reliable analysis in the saliency space. *QoMEX 2011*, 201-206

[65] **(Redi et al 2012)** Redi, J.A., Heynderickx, I. (2012). Image integrity and aesthetics: towards a more encompassing definition of visual quality. *Proceedings of SPIE 8291 – Human Vision and Electronic Imaging XVII*, 829115.

[66] **(Redi et al 2013)** Redi, J.A., Hossfeld, T., Korshunov, P., Mazza, F., Povoa, I., Keimel, C. (2013). Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognisability and Aesthetic Appeal. *CrowdMM '13 – Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 29 – 34.

[67] **(Redi et al 2013)** Redi, J.A., Povoa, I. (2013). The Role of Visual Attention in the Aesthetic Appeal of Consumer Images: A Preliminary Study. *Proceedings of Visual Communications and Image Processing (VCIP)*, 1 – 6.

[68] **(Rhodes et al 2001)** Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in non-western cultures: In search of biologically based standards of beauty. *Perception, 30(5),* 611 – 625.

[69] **(Riegler et al 2013)** Riegler, M., Lux, M., Kofler, C. (2013). Frame the Crowd: Global Visual Features Labeling boosted with Crowdsourcing Information. *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop*, Volume 1043.

[70] **(Robson 2002)** Robson, C. (2002). *Real world research: A resource for social scientists and practitioner-researchers*. Wiley.

[71] **(Shapiro et al 2001)** Shapiro, L., Stockman, G. (2001). Computer Vision. *Prentice-Hall.*

[72] **(Sheikh et al 2005)** Sheikh, H. R., Wang, Z., Cormack, L., Bovik, A. C. (2005). LIVE image quality assessment database release 2.

[73] **(Sheikh et al 2006)** Sheikh, H.R., Sabir, M.F., Bovik, A.C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing, 15(11),* 3440 – 3451.

[74] **(Smith 2013)** Smith, Chris. (2013). *Highlight and Auto Enhance launched at Google I/O as powerful new Google+ photo tools*. Retrieved January, 31 2014, from: http://trustedreviews.com/news/highlight-and-auto-enhance-launched-at-google-i-o-as-powerful-new-google-photo-tools

[75] **(Soleymani et al 2010)** Soleymani, M., Larson, M. (2010). Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation,* 4 – 8.

[76] **(Szechter et al 2007)** Szechter, L.E., Liben, L.S. (2007). Children’s Aesthetic Understanding of Photographic Art and the Quality of Art-Related Parent – Child Interactions. *Child Development, Volume 78, Issue 3*, 879–894.

[77] **(The Economist 2013)** The Economist (2013). *Facebook is bad for you: Get a life! – Using the social network seems to make people more miserable no. 8849, vol. 408.*

[78] **(Thomas 2013)** Thomas, Owen. (2013). *Here's A Delicious Way Facebook Could Profit From Instagram—And Destroy OpenTable And Yelp*. Retrieved January, 31 2014, from: http://businessinsider.com/instagram-food-photos-are-a-phenomenon-2013-1

[79] **(Truta 2013)** Truta, Filip. (2013). *The Top 3 Most Popular Cameras on Flickr Are iPhones*. Retrieved January, 31 2014, from: http://news.softpedia.com/news/The-Top-3-Most-Popular-Cameras-on-Flickr-Are-iPhones-337949.shtml

[80] **(Vliegendhart et al 2012)** Vliegendhart, R., Larson, M., Pouwelse, J. (2012). Discovering User Perceptions of Semantic Similarity in Near-duplicate Multimedia Files. *Proceedings of the First International Workshop on Crowdsourcing Web Search,* 54-58

[81] **(Wagemans 2011)** Wagemans, J. (2011). Towards a new kind of experimental psycho-aesthetics? Reflections on the Parallellepipeda project. *i-Perception, 2(6),* 648-678.

[82] **(Wallraven et al 2009)** Wallraven, C., Cunningham, D., Rigau, J., Feixas, M., Sbert, M. (2009). Aesthetic appraisal of art - from eye movements to computers. *Computational Aesthetics'09 – Proceedings of the Fifth Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 137-144.

[83] **(Wang et al 2004)** Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing, 13(4),* 600 – 612.

[84] **(Wang et al 2010)** Wang, J., Chandler, D.M., Le Callet, P. (2010). Quantifying the Relationship between Visual Salience and Visual Importance. *Proceedings of SPIE 7527 – Human Vision and Electronic Imaging XV*, 75270K .

[85] **(Xiao et al 2010)** Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., Torralba, A. (2010). SUN Database: Large-scale Scene Recognition from Abbey to Zoo. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3485 – 3492.

[86] **(Yiwen et al 2008)** Yiwen Luo, Xiaoou Tang (2008). Photo and Video Quality Evaluation: Focusing on the Subject. *ECCV ’08 – Proceedings of the 10th European Conference on Computer Vision: Part III*. 386-399

8. Appendixes

A. **THE ROLE OF VISUAL ATTENTION IN THE AESTHETIC APPEAL OF CONSUMER IMAGES: A PRELIMINARY STUDY**

B. **Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognisability and Aesthetic Appeal**

C. **Poster used for recruitment of laboratory subjects**

D. **Phototo user interface design**

Appendix A :

THE ROLE OF VISUAL ATTENTION IN THE AESTHETIC APPEAL OF CONSUMER IMAGES: A PRELIMINARY STUDY

THE ROLE OF VISUAL ATTENTION IN THE AESTHETIC APPEAL OF CONSUMER IMAGES: A PRELIMINARY STUDY

Judith A. Redi, Isabel Povoa

Intelligent Systems, Delft University of Technology, The Netherlands
j.a.redi@tudelft.nl, misabelpovoa@gmail.com

ABSTRACT

Predicting the aesthetic appeal of images is of great interest for a number of applications, from image retrieval to visual quality optimization. In this paper, we report a preliminary study on the relationship between visual attention deployment and aesthetic appeal judgment. In particular, we seek to validate through a scientific approach those simplicity and compositional rules of thumb that have been applied by photographers and modeled by computer vision scientists in computational aesthetics algorithms. Our results provide a confirmation that both simplicity and composition matter for aesthetic appeal of images, and indicate effective ways to compute them directly from the saliency distribution of an image.

Index Terms— Aesthetic appeal, visual attention, visual quality, image saliency

1. INTRODUCTION

The possibility to predict the aesthetic appeal of images has recently attracted a lot of interest from the multimedia community, being it crucial for a number of applications, from multimedia information retrieval to computer graphics [1]. Recent research has also shown that aesthetic appeal of images plays a role in the tolerance that users have to visual distortions [2]. Augmenting objective quality metrics [3] with a prediction of aesthetic appeal could therefore significantly improve their ability to assess the overall pleasantness of images towards a finer optimization of multimedia delivery systems.

Computational aesthetics models [1] have attempted to mimic processes underlying the human appreciation for image aesthetics. Factors such as color rendering [4], semantic content [5], familiarity [6], image simplicity [7] and compliance to compositional rules [8] have been modeled through computer vision techniques towards a reliable estimation of aesthetic appeal. Existing models have achieved good performance, but the room for improvement is still large. This might be due to the fact that most of these models are often inspired by photographers' rules-of-thumb [6], which have not been validated in a scientific way. Both image simplicity (i.e., clarity of the subject [7]) and compositional rules (e.g., the well-known rule of thirds [8]), for example, are tools used by photographers to guide the observer's visual attention towards the image subject and ease perceptual fluency. Very few studies have attempted at checking their validity in a systematic way, e.g. by verifying with empirical measurements the existence of a relationship between the deployment of visual attention, image simplicity/compliance to compositional rules and aesthetic appeal. In fact, several studies have looked into the relationship between visual attention and art [9, 10]; however, they mostly analyzed visual scan paths, in relation to either viewing task [9] or painting genre [10]. To the best of the authors' knowledge, no studies so far have inquired the role of visual attention in aesthetic appeal in relation to composition or image simplicity for regular consumer images.

On the other hand, studying visual attention in relation to image preferences has been shown to have a high added value in the contingent field of objective image quality assessment [11]. The deployment of visual attention was shown to play a major role in quality appreciation: artifacts visible in the region of interest of an image are more likely to be noticed and therefore more annoying for observers [12]. As a result, modulating the distortion visibility measurements with saliency information was shown to be beneficial for objective metrics' accuracy. Also, it has been shown that in some cases, it might be sufficient to compute distortion visibility only in the region of interest of the image, implying significant savings in terms of computational complexity of the metrics [11]. Similar principles could be applied to computational aesthetics metrics; however, until now little work has been done in this direction, besides several, remarkable attempts at estimating compliance to compositional rules through the use of visual attention models [13, 14].

In this work, we present the (preliminary) findings of a large study involving 200 consumer images and 14 participants, whose eye movements were tracked while judging the aesthetic appeal of the images. We analyze to what extent the way visual attention is deployed during the image evaluation is related to simplicity, composition, and eventual aesthetic appeal. To achieve this, we define several indicators of attention deployment based on fixation and saliency [15] information. We confirm that simplicity (in

Fig. 1. Samples from the image database used in the experiment, along with their categories.

terms of low image clutter [7]) is positively correlated to aesthetic appeal, and that compositional rules such as the rule of thirds can be validated through saliency analysis. Furthermore, the indicators we define for simplicity and composition analysis can be easily implemented in computational aesthetics metrics starting from the output of visual attention models (e.g., [16]).

2. AN EYE TRACKING STUDY FOR UNDERSTANDING AESTHETIC APPEAL

We designed a within-subjects experiment, in which fourteen participants were asked to judge the aesthetic appeal of images while their eye movements were being tracked. The number of participants was chosen in line with what advised in [17, 18]. Previous studies based on eye-tracking also have shown that a number around 15 is sufficient to guarantee stable results [2,12].

2.1. Image material

A set of 200 images was included in the experiment. Of these, 56 corresponded to those already included in study [2], 26 were chosen from images freely available online, and 118 were taken from the private collection of an amateur photographer.

Images were selected to cover a wide range of subject categories, keeping the sample as representative as possible of a general image population. The dataset was labeled based on 16 categories from the website 500px.com, for both expert and amateur photography (see Fig. 1). The following criteria were considered when selecting the categories:

- Compliance to categories used in computer vision literature (e.g., the LHI dataset [19]), as in the case of *Landscapes*, *People* and *Sport*.
- Frequent occurrence in social networks, as in the case of *Food* and *Fashion*.
- Need to encompass different levels of familiarity [6], as in the case of *Abstract* and *Celebrities*.

2.2. Apparatus

The experiment was performed in a room with constant illumination at approximately 70 lux, in an environment compliant to ITU recommendations [18]. A 23" LED backlight monitor having a resolution of 1360x768 was used to display the stimuli. Participant's face movements were constrained by a chinrest at a distance of 0.7 meters from the display. A *SensoMotoric Instruments* GmbH Eye Tracker with a sampling rate of 50/60 was used to track the participants' eye movements during the image viewing. The instrument has a pupil tracking resolution of 0.1° and a gaze position accuracy of 0.5 to 1.

2.3. Methodology

For each image in the database, participants were asked to score its aesthetic appeal in a Single Stimulus setup [18], using a 5-point discrete scale ranging between very low (1) and very high aesthetic appeal (5).

Because of the large number of images involved, fatigue and memory effects might have affected the data collection (as revealed by a pilot experiment). As a consequence, participants were asked to score images in two sessions, involving 100 images each, to be performed in different days. Each session lasted on average 40 minutes, including a short break after scoring the first 50 images.

All participants were first briefed about the general setup of the experiment and their task. A short training session (consisting in rating 3 images) was performed to allow participants to familiarize with their task. The images provided in the training were not intended to be anchoring stimuli for the scoring scale, as we did not want to prime participants with specific criteria for judging images. Participants had no time constraints in observing the images prior to scoring (both in the training and in the actual experiment). Before each image, participants' initial fixation point was forced to be in the center of the image by displaying a white cross in the middle of the screen (with a neutral background). The scoring scale was accessible only after completing the viewing of an image, in order to avoid distraction during the image observation. Images were presented in a randomized order for every participant.

At the beginning of every session (and after every break) the eye-tracker was calibrated on the participant's gaze based on a 13-points grid.

3. DATA ANALYSIS

Individual Aesthetic Appeal Scores were processed according to the procedure recommended in [18], which pointed out the presence of one outlier participant, then excluded from the analysis. Scores were then normalized per participant and transformed into individual Aesthetic Appeal Z-Scores (AAZ), eventually ranging between -3.01 and 3.47. To verify inter-observer consistency in scoring, we computed the standard deviation across the scores given to the same image by the participants, and then averaged it across all image. This resulted into a value of 0.82, corresponding to 12% of the aesthetic appeal range covered by the AAZ, in line with previous results in the field [2].

3.1. Eye tracking data analysis

We processed eye-tracking recordings in order to collect information on both eye movements and attention deployment. With respect to the latter, we processed fixation data according to [20] to obtain, per each image, visual importance information in the form of saliency maps. Saliency maps [15] represent the probability, pixel per pixel, for a location in the image to be attended by the (average) observer. As such, they outline the areas in the image which attract most attention. We believe this information can be helpful in our analysis for two main reasons. First, they may provide a powerful tool to estimate simplicity, in terms of how visually crowded (how many areas of the image attract attention, as a measure of clutter, or low simplicity) is the image. Second, it is commonly assumed that highly salient areas correspond to the most important elements in the image. Photographers intentionally compose images so that visual attention is driven to these elements; an analysis of salience could therefore reveal the compliance of an image to compositional rules-of thumb, to be later matched to an actual benefit in terms of aesthetic appeal.

The following steps were performed to create saliency maps from raw eye-tracking data:

1. All fixations lasting less than 100 ms were discarded from the recordings;
2. For each image I of size W_I x H_I, locations fixated by every observer were identified and added to a fixation map $FM^{(I)}(x,y)$, eventually gathering all fixation points from all observers;
3. $FM^{(I)}(x,y)$, was then smoothed by applying a grey scale patch with Gaussian intensity distribution whose variance (σ) was approximating the size of the fovea (~2° of visual angle). The resulting saliency map element $SM^{(I)}(k,l)$, at location (k,l) was therefore computed as:

$$SM^{(I)}(k,l) = \sum_{f=1}^{N_f} \exp\left[-\frac{(x_f - k)^2 + (y_f - l)^2}{\sigma^2}\right] \qquad (1)$$

with (x_j, y_j) being the pixel coordinates of the fth fixation ($f=1...N_f$) in $FM^{(I)}(x,y)$, and $k \in [1, W_I]$, $l \in [1, H_I]$.

We also produced binary versions of the saliency maps *SM*, in order to isolate the Region(s) of Interest (ROI) of the image. To compute our Binary Maps (*BM*) we performed the following extra steps:

4. A saliency threshold th^S was determined, common for all maps, as one third of the maximum saliency value across all maps. A threshold for saliency was preferred over a threshold for the size of the ROI area (as used in other works, e.g., [12]), in order to isolate areas that were equally salient across all images. Of course, the value of the threshold was established in a somewhat arbitrary way and changes in the threshold may affect the results reported in the following section. We delegate to future studies further investigations on these aspects.
5. For each image I, its binary map $BM^{(I)}$ was determined as:

$$BM^{(I)}(x,y) = \begin{cases} 1 & if \quad SM^{(I)} > th^S \\ 0 & otherwise \end{cases} \qquad (2)$$

4. RESULTS

4.1. Analysis of viewing strategy

As a first step, we investigated possible relationships between eye movements' characteristics and aesthetic appeal z-scores (AAZ). In particular, per each subject and image, we considered the number of fixations and saccades, their average duration, and the amplitude and velocity of saccades. These indicators are often used to describe visual strategy [21] and were found to be related to both viewing task [22] and perceived quality [23]. We report in Table 1 their mean across all observers and images, and related Standard Error. The average number and duration of fixations was found to be comparable to that obtained for other studies in the field [10] and slightly lower than that found, e.g., for technical quality scoring [21].

To check whether a relationship existed between viewing strategy parameters and judgments of aesthetic appeal, we also computed the linear correlation coefficient between these quantities and the corresponding AAZ. As shown in Table 1, none of the parameters was found to be a predictor to aesthetic appeal, as instead was found in, e.g., [22]. In

Table 1. Statistics of eye movements and correlation with aesthetic appeal judgments.

	Mean		Correlation with AAZ
	Statistic	Std. Error	
Number of fixations	20.17	0.288	-0.017
Duration of fixations	381,79	2.522	0.019
Number of saccades	16,55	0.251	-0.021
Duration of saccades	30,73	0.276	0.012
Amplitude of saccades	2,11	0.054	-0.050
Velocity of saccades	63,57	1.052	0.040

that case, the duration of fixations was negatively correlated with video quality, perhaps because the sudden appearance of visual artifacts would capture and hold attention in an unnatural way. In the case of static images, such surprise effect does not apply. This, along with difference in viewing task [22] could partially explain this discrepancy in viewing behavior.

4.2. Analysis of visual attention deployment

As mentioned in section 3.1, visual saliency can reveal important properties of the image, in particular related to visual clutter [7] and composition [8].

In the following, we describe a set of indicators that we designed to characterize both elements starting from Fixation and Saliency information and we check their relationship with aesthetic appeal.

4.2.1. Simplicity and clutter indicators

The following indicators were designed to attempt at estimating visual clutter from saliency information:

Peak Saliency: The peak value of a saliency distribution represents the location of the image that is more likely to attract the attention of an (average) observer. A high peak value indicates that in the image there is one location (i.e., an image element, possibly the subject) that is highly attractive. Lower values would instead indicate poor attractiveness, perhaps because of the presence of multiple attractive elements in the image (visual clutter). We calculate this quantity as $Peak\_S^{(I)} = max(SM^{(I)})$, and we expect it to be positively correlated to aesthetic appeal.

Saliency Spread: the spread of saliency values across the image measures whether the attention was directed towards a concentrated area (low clutter) or was instead distributed throughout the image (high clutter). We measure it by computing the standard deviation of the saliency distribution of each image: $Spread\_S^{(I)} = stdev(SM^{(I)})$, and we expect it to be negatively correlated to aesthetic appeal.

Number of fixations within the ROI: The less fixations are scattered in the background of the image, the more it is likely that there is a single object attracting the viewer's attention, which implies visual simplicity and low clutter. We compute this feature as:

$$nFix\_ROI^{(I)} = \sum_{x=1}^{W_I}\sum_{y=1}^{H_I} FM^{(I)}(x,y) \times BM^{(I)}(x,y) \qquad (3)$$

with $FM^{(I)}(x,y)$ being the fixation maps, and $BM^{(I)}(x,y)$ the binary map for each image I.

Dispersion of the fixations: Introduced in [23], this indicator intends to measure the spread of the fixations during the observation of an image. $Disp\_Fix^{(I)}$ is computed as the average Euclidean distance between each fixation in $FM^{(I)}$ and the centroid of fixations.

Number of distinct ROIs: When attention is divided over different elements in the image, there might be multiple peaks in the saliency distribution, and, as a result of the thresholding procedure described in section 3.1, this may originate multiple Regions of Interest in the binary maps. We define $No\_ROI^{(I)}$ the number of distinct ROIs retrievable in $BM^{(I)}$, and we expect this indicator to be negatively correlated with aesthetic appeal.

4.2.2. Indicators of compliance to composition rules

Several studies have attempted at using saliency information generated by visual attention models [13, 14] in order to predict the compliance of the image content to composition rules such as the *rule of thirds*. Such rule, often used by professional photographers, states that to ensure ease of view, the center of the main object should be located along the intersections of the lines that divide the image in thirds (see figure 2).

We replicate here a set of indicators that have been previously used for computational aesthetic models, attempting at verifying the compliance of the image to the Rule of Thirds:

- **Minimum Euclidean Distance** ($Dist\_thirds^{(I)}$) between the centroid of the (largest) ROI and the intersections of the line of thirds, normalized by the size of the image (as per [4])
- **Minimum Distance from the horizontal lines of thirds** ($Dist\_thirds\_h^{(I)}$) of the centroid of the ROI, normalized by the height of the image H_I
- **Minimum Distance from the vertical lines of thirds** ($Dist\_thirds\_v^{(I)}$) of the centroid of the ROI, normalized by the width of the image W_I

Furthermore, we compute the extent of the **area of the ROI** ($Area\_ROI^{(I)}$), normalized by the whole image area, to estimate the balance between main element and background.

4.2.3. Results

To better appreciate the impact of our indicators on aesthetic appeal, we first quantized all their values (except for indicator *No_ROI*) into three classes (low, medium and high indicator value). This was achieved by (1) detecting the 33rd and 66th percentiles of the distribution of the indicators throughout images and (2) assigning to all the images with

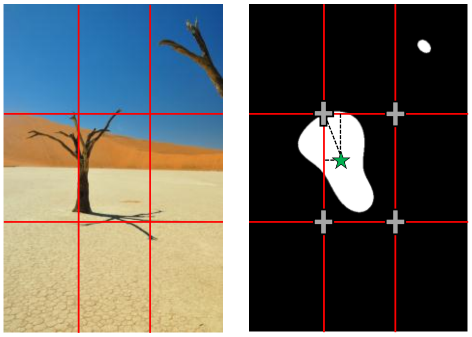


Fig. 2. Graphical explanation of the rule of thirds.

an indicator value lower than the 33rd percentile a value of 1 (low), to all images with an indicator value between the 33rd and 66th percentiles a value of 2 (medium) and a value of 3 (high) to all other images. Values of the percentiles are reported in table 2; impact of all indicators on aesthetic appeal can be visualized in fig. 3 and 4.

Figure 3 shows how our data confirm the negative effect of visual clutter on aesthetic appeal. Indicators *Peak_S* (df = 2, F = 20.88, sig = 0.000), *Spread_S* (df = 2, F = 7.38, sig = 0.001), and *No_ROI* (df = 4, F = 7.58, sig = 0.000) were found to have a significant effect on the aesthetic appeal judgments (AAZ). In particular, Figure 3.a confirms the expected relationship between *Peak_S* and AAZ: the higher the attractiveness of a single location in an image, the higher the aesthetic appeal. The relationship expected between the number of ROIs and AAZ is also confirmed (figure 3.e), with the aesthetic appeal decreasing with the increase in number of distinct ROIs (and consequent increase of clutter). An interesting effect is found for images for which no ROI was segmented (leftmost bar in Fig 3.e, *No_ROI* = 0): in this case the aesthetic appeal is also very low. This phenomenon is in line with what we expected: since we used a single threshold across all images, if no ROI was detected that was because no area in the image was sufficiently attractive to match the overall threshold. This might be due to the fact that attention was very spread across the image, which in turn could result from a high clutter in the image.

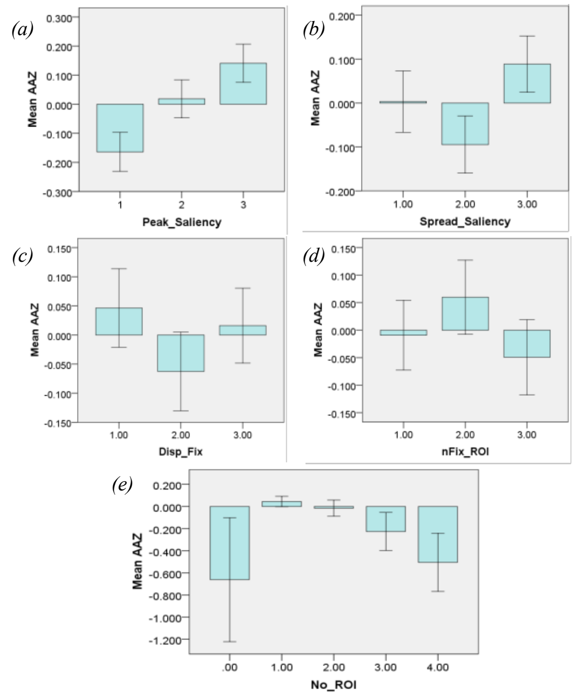


Fig. 3. Impact of clutter indicators on aesthetic appeal scores (AAZ). A level of 1 indicates low indicator values, 2 medium indicator values, and 3 high indicator values.

Figure 4 reports the relationship between the compositional features and the aesthetic appeal. We found no significant effect (df = 2, F = 2.38, sig = 0.093) of *dist_thirds* on AAZ. This is a quite interesting result, as it suggest that one of the most commonly used features to express image compliance to compositional rules might not properly reflect aesthetic appeal appreciation mechanisms. Interestingly, also features *dist_thirds_h* (df = 2, F = 1.728, sig = 0.178) and *area_ROI* (df = 2, F = 0.035, sig = 0.716) did not have a significant effect on aesthetic appeal. Conversely, the distance of the centroid of the ROI from the vertical lines of thirds *dist_thirds_v* has a significant effect on aesthetic appeal (df = 2, F = 17.32, sig = 0.000). It is also interesting to analyze the nature of this effect (figure 4.b). It seems that, for low values of *dist_thirds_v* (that is, the centroid of the ROI is close to the vertical lines of thirds) higher values of aesthetic appeal are obtained; for medium distances, aesthetic appeal significantly decreases, as expected; for high values of distance, the aesthetic appeal slightly increases again. This behavior can be explained by looking at the thresholds used to quantize the values of *dist_thirds_v* (table 2). These values are normalized by the width of the image; therefore, the maximum value that the indicator could assume is 1/3. As we can see, the maximum value found for indicator *dist_thirds_v* is ~1/6 = 0.17, which implies that in no case the centroid of the ROI is located in peripheral regions of the image. Furthermore, our analysis revealed that for the most part, ROI centroids are located in the central region of the image, i.e., that delimited by the four lines of thirds. As a result, we can assume that most of the images having a high distance of the ROI centroid from the vertical lines of thirds, are images whose ROI is located in the very center of the image. Centrality of the main subject has also been shown to be positively correlated to aesthetic appeal [7], which may partially explain our result.

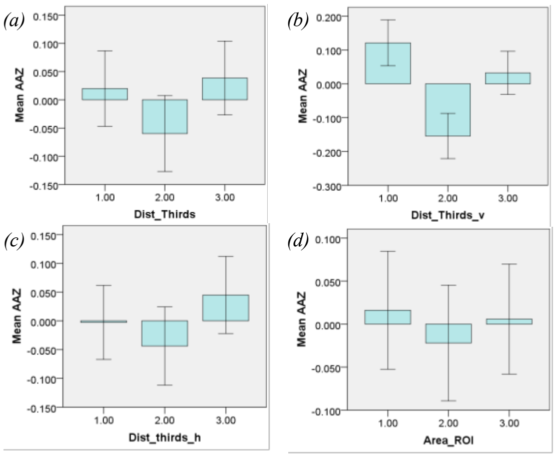


Fig. 4. Impact of image composition indicators on aesthetic appeal scores (AAZ). A level of 1 indicates low indicator values, 2 medium indicator values, and 3 high indicator values.

Table 2. Threshold used for the quantization of the indicators.

Indicator	Min	33rd Percentile	66th Percentile	Max
Peak_S	18,15	35,64	43,58	75,38
Spread_S	4,32	6,88	7,92	11,78
Area_ROI	0,00	0,05	0,08	0,28
nFix_ROI	0,17	0,52	0,64	1,00
Disp_Fix	0,27	0,41	0,50	0,77
Dist_Thirds	0,00	0,12	0,16	0,27
Dist_Thirds_v	0,00	0,10	0,14	0,17
Dist_Thirds_h	0,00	0,07	0,12	0,17

5. CONCLUSIONS

In this paper, we conducted a preliminary study on the role of visual attention in image aesthetic appeal appreciation. We tracked the eye movements of 14 subjects during the judgment of aesthetic appeal of a set of 200 consumer images and then analyzed the relationship between the attention deployment and the aesthetic appeal judgments. We designed a set of indicators extracted from human saliency (easily adaptable to saliency information gathered from computational models) that validated a negative correlation between image clutter (low simplicity) and aesthetic and a clear human preference for images having the most attractive object located either at the vertical line of thirds or the center of the image. It should also be mentioned that the influence on perceived quality of participant expertise as well as of the image content can be highly relevant, and will be investigated at a later stage.

No impact of the vertical placement of the ROI was found instead, which is useful information to simplify the computation of composition features in computational aesthetics model.

We intend to further investigate in the future on other relationships between attention deployment and aesthetics, features such as contrast of image, color and texture inside and outside the ROI. Furthermore, in the future developments of this study, we intend to validate the current findings into actual computational aesthetics models and to expand the pool of participants.

ACKNOWLEDGEMENTS

The authors are thankful to the amateur photographer Mark Dekker (http://gplus.to/markdekker), who donated most of the photographic material used in this study.

12. REFERENCES

[1] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *Sig. Proc. Magazine, IEEE*, vol. 28, 2011.

[2] J. A. Redi, "Visual quality beyond artifact visibility," in IS&T/SPIE Electronic Imaging, 2013, 86510N-86510N-11.

[3] S. S. Hemami and A. R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal Process-Image.* , vol. 25, no. 7, pp. 469-481, 2010

[4] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in Computer Vision–ECCV 2008, ed: Springer, 2008, pp. 386-399

[5] M. Redi and B. Merialdo, "Where is the interestingness?: retrieving appealing VideoScenes by learning Flickr-based graded judgments," in Proc. ACM Multimedia, 2012

[6] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in Computer Vision–ECCV 2006, pp. 288-301

[7] C. D. Cerosaletti, A. C. Loui, and A. C. Gallagher, "Investigating two features of aesthetic perception in consumer photographic images: clutter and center," SPIE Conference Series, 2011, p. 5

[8] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver, "The role of image composition in image aesthetics," in Proc. IEEE ICIP, 2010

[9] A. L. Yarbus and L. A. Riggs, Eye movements and vision vol. 2: Plenum press New York, 1967

[10] C. Wallraven, D. Cunningham, J. Rigau, M. Feixas, and M. Sbert, "Aesthetic appraisal of art-from eye movements to computers," Computational aesthetics, pp. 137-144, 2009

[11] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," Signal Processing Magazine, IEEE, vol. 28, pp. 50-59, 2011

[12] H. Alers, H. Liu, J. Redi and I. Heynderickx, “Studying the risk of optimizing the image quality in saliency regions at the expense of background content”, IS&T/SPIE HVEI, 2010

[13] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," In Proc. IEEE ICIP, 2009

[14] Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in Proc. ACM Multimedia, 2010

[15] C. Koch and S. Ullman, “Shifts in Selection in Visual Attention: Toward the Underlying Neural Circuitry,” Human Neurobiology, vol. 4, no. 4, pp. 219-27, 1985

[16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," Trans. Pattern Analysis and Machine Intelligence, vol. 20, 1998

[17] Engeldrum, Peter G., “Psychometric scaling: a toolkit for imaging systems development,” Imcotek Press, 2000

[18] ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” Geneva, (2002)

[19] B.Yao, X. Yang, and S. Zhu, “Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks.” EMMCVPR, 2007

[20] Redi, J., and Heynderickx, I., “Image Quality And Visual Attention Interactions: Towards A More Reliable Analysis In The Saliency Space,” Proceedings of QoMEX, (2011)

[21] R. J. Peters, A. Iyer, L. Itti, and C. Koch, “Components of bottom-up gaze allocation in natural images,” *Vision Research*, vol. 45, no. 18, pp. 2397–2416, August 2005

[22] Ninassi, O. Le Meur, P. L. Callet, D. Barba, and A. Tirel, “Task impact on the visual attention in subjective image quality assessment,” Proc. EUSIPCO-06, 2006

[23] C. Mantel, N. Guyader, P Ladret, G. Ionescu and T. Kunlin “Characterizing eye movements during temporal and global quality assessment of h.264 compressed video sequences,” in Proc. SPIE 8291, HVEI XVII, 2012

Appendix B :

Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognisability and Aesthetic Appeal

Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal

Judith Redi
INSY, TU Delft
j.a.redi@tudelft.nl

Tobias Hoßfeld
University of Würzburg
tobias.hossfeld@uni-wuerzburg.de

Pavel Korshunov
MMSPG, EPFL
pavel.korshunov@epfl.ch

Filippo Mazza
IRCCyN, ECN
filippo.mazza@irccyn.ec-nantes.fr

Isabel Povoa
INSY, TU Delft
misabelpovoa@gmail.com

Christian Keimel
Technische Universität München
christian.keimel@tum.de

ABSTRACT

Research on Quality of Experience (QoE) heavily relies on subjective evaluations of media. An important aspect of QoE concerns modeling and quantifying the subjective notions of 'beauty' (aesthetic appeal) and 'something well-known' (content recognizability), which are both subject to cultural and social effects. Crowdsourcing, which allows employing people worldwide to perform short and simple tasks via online platforms, can be a great tool for performing subjective studies in a time and cost-effective way. On the other hand, the crowdsourcing environment does not allow for the degree of experimental control which is necessary to guarantee reliable subjective data. To validate the use of crowdsourcing for QoE assessments, in this paper, we evaluate aesthetic appeal and recognizability of images using the Microworkers crowdsourcing platform and compare the outcomes with more conventional evaluations conducted in a controlled lab environment. We find high correlation between crowdsourcing and lab scores for recognizability but not for aesthetic appeal, indicating that crowdsourcing can be used for QoE subjective assessments as long as the workers' tasks are designed with extreme care to avoid misinterpretations.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*perceptual reasoning, representations, data structures, and transforms*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*evaluation/methodology, video*

Keywords

Crowdsourcing, Aesthetics, QoE, Subjective evaluations.

CrowdMM'13, October , 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2396-3/13/10 ...$15.00.

1. INTRODUCTION

Crowdsourcing (CS) is a powerful tool for gathering subjective ground truth for large multimedia collections. Big amounts of users (microworkers) can be reached to accomplish a set of small tasks in exchange for a symbolic payment, which is particularly convenient when large user studies have to be conducted. By designing appropriate micro-tasks, sufficiently reliable data can be gathered in an inexpensive and time-effective way. As a result, CS has become a popular tool for media tagging [4], investigation of cognitive responses to media fruition [8], evaluation of privacy filters [13], etc.

Research on Quality of Experience (QoE) [15] relies on understanding user preferences in terms of perceptual quality and overall enjoyment of multimedia. To this end, studies are conducted in a controlled Laboratory (Lab) environment, with fixed lighting and experimental conditions [19], since the goal is to collect information on the user sensitivity to impairments in the media signal [7] and the related quantification of their annoyance. In this context, CS has often been considered not appealing for QoE research, as it would not guarantee the necessary level of environmental control to provide reliable data. Lately, however, the sensitivity-centric definition of QoE has been challenged, and it was shown that QoE depends also on user preferences and personality, context of media usage, and quality of the interaction with the system [15, 20]. With the acceptance of this more encompassing definition of QoE, the interest in using CS for QoE research has grown significantly [2, 7, 11]. Nevertheless, some doubts remain regarding the extent to which CS can provide reliable QoE data.

To understand the benefits and limits of using crowdsourcing in QoE evaluations, we look into how well QoE ratings collected in a controlled lab environment can be replicated by a crowdsourcing experiment. Being QoE a multifaceted quantity [15], we focus specifically on aesthetic appeal, which has been recently shown to play an important role in QoE judgments [20]. Understanding aesthetic appeal of media is of major interest for the multimedia community, which has indeed devoted a lot of effort to it lately [10], although often based on ground truth collected through social media platforms. Interestingly, very few efforts have been made towards quantifying the aesthetic appeal in a more controlled way. In this study, we make a first attempt at collecting more rigorous ground truth on aesthetic appeal of consumer images in a lab environment, and we check to what extent CS can be used to collect the same type of information.

We conducted an experiment in a controlled lab environment, in which the aesthetic appeal of 200 consumer images was rated by 14 paid participants in a single stimulus setup [19]. Along with this quantity, participants also rated the level of recognizability of the content of the image. This second quantity is related to perceptual fluency [17], which is known to have an effect on the aesthetic appeal of works of art. In this study, we wanted to check whether this effect was preserved also when judging the aesthetic appeal of consumer images.

We then replicated the same experiment in a crowdsourcing setting, by using the Microworkers[1] platform. About 390 workers from 16 countries evaluated (subsets of) our images, ensuring a variety in cultural and social backgrounds, which are known to impact aesthetic preferences. Adaptations to the protocol were needed to allow controlling the reliability of the workers, and checks were made using both control questions and timestamp information prior to analyzing the results and comparing them with the Lab data.

In the remainder of this paper, after a brief review of existing work on user studies on aesthetic appeal and crowdsourcing (Section 2), we describe the experimental protocol followed in the Lab and its adaptation to the crowdsourcing evaluation (Section 3). In Section 4, we analyze the reliability of the CS workers, and based on reliable workers only, in Section 5, we compare the outcomes of crowdsourcing experiment with lab experiment. We draw conclusions and possible future extensions of this study in Section 6.

2. BACKGROUND

Being able to model aesthetic preferences of users is a major concern for modern multimedia research. Information on image aesthetic appeal can help in retrieval and recommendation tasks, as well as in optimizing visual Quality of Experience [10, 20]. Before computational models can be created that reliably predict the aesthetic appeal of an image [9], in-depth knowledge is needed on actual user aesthetic preferences. This is not a trivial task, as aesthetic preferences are typically considered to be highly subjective and related to personal implicit experiences [17], cognitive biases, and personal opinions and memories [18]. Nevertheless, some research on the matter has been conducted by means of self assessments, eye tracking experiments and physiological measurements [22]. Color and saliency have been shown to play a major role in the aesthetic and emotional impact of an image [1, 23]. Furthermore, correlation between aesthetic ratings and familiarity has been reported in [3]. Content recognizability has been shown to have an influence on aesthetic appeal in [14, 17], and abstract paintings were found to be less likely appreciated by people with respect to immediate works of art [16].

Based on these studies and on classical geometrical canons (e.g., rule of thirds and golden ratio), researchers in computational aesthetics have proven to be able to capture useful information for the aesthetic assessment of images [9, 3, 16]. Nevertheless, reliable prediction of the aesthetic appeal of images is still to be achieved. To work towards that goal, computational aesthetic researchers need to rely on ground truth of how users judge the aesthetic appeal of large image collections. Since obtaining this sort of data from controlled experiments is expensive in time and cost [9], more and more researchers turn to community-contributed resources (i.e., from popular online image databases, such as *Photo.net* used in [3]) for data collection. These platforms, however, lack a strict protocol for image assessment and some users can create fraudulent accounts to increase their ratings, leading to unreliable evaluations. In this scenario, crowdsourcing seems to be an in-between solution, offering both the opportunity to reach out to large communities of users and controlling the aesthetic evaluation procedures.

Crowdsourcing is a further development of the outsourcing principle, where the granularity of work is reduced to small tasks that can be accomplished within a few minutes to a few hours and do not require a long-term employment. Tasks are often highly repetitive (e.g., image annotation) and are usually grouped in larger units, referred to as *campaigns*. Most *employers* submitting tasks to an anonymous crowd use a mediator in the form *crowdsourcing platforms* that maintains the crowd, manages the employers campaigns and handles the reimbursement of the workers on behalf of the employer after successful completion of the the tasks.

Amazon's Mechanical Turk (MTurk)[2] and Microworkers are typically used commercial Crowdsourcing platforms. MTurk is the largest crowdsourcing platform and is often used in research, as well as in commercial third-party applications; however, it allows only US residents or companies to submit tasks to the platform. The platform used in this contribution, Microworkers, allows not only international employers, but also worker diversity [5], whose geographic location can be chosen directly by the employer.

When it comes to subjective QoE evaluation tasks, Crowdsourcing tests require the presentation and assessment of different media in a suitable web-interface. Instead of implementing an appropriate interface separately for each QoE test, existing and publicly available frameworks as the *Qudrant of Euphoria* [2] and *QualityCrowd* [11] can be used. Chen's *Quadrant of Euphoria* provides an online service for the QoE evaluation of audio, visual, and audio-visual stimuli using pairwise comparison of two different stimuli in an interactive web-interface, where the worker can judge which of the two stimuli has a higher QoE. In contrast, the *QualityCrowd* framework is not an online service, but a complete open-source platform designed especially for QoE evaluation with crowdsourcing. It can be modified with relatively low effort for different assessment tasks (e.g., single or double stimulus) and provides a simple scripting language for creating campaigns including multi-modal stimuli, training sessions and control questions.

3. EVALUATION METHODOLOGY

We investigated aesthetic appeal and its relationship with some of the features analyzed in Section 2 by means of both a Lab-based and a Crowdsourcing-based experiment. To do so, we designed a within-subjects experiment, in which every participant had to evaluate several aspects of a set of images in a single stimulus setup [19]. Four quantities, namely aesthetic appeal, color likeability, familiarity, and recognizability were inspected in the lab environment. In the crowdsourcing setup only two quantities were inspected to simplify the task: recognizability ('how well can you understand what is represented in the image?') and aesthetic appeal ('how beautiful do you think is the image?').

3.1 Image material

We used a database of 200 images, out of which 56 corresponded to the ones used in [20], 26 were crawled from the web, and 118 were selected from the private collection of an amateur photographer. Images were chosen to encompass a wide range of image contents as generally available online, based on their classification into the categories used by *500.com*, an online database for both expert and amateur photography. As a result, images were chosen that could be classified into categories typically used in computer vision research (e.g., Landscapes and People), frequently occurring in social networks (e.g., Food and Fashion) and covering different

[1] http://microworkers.com/

[2] https://www.mturk.com/mturk/

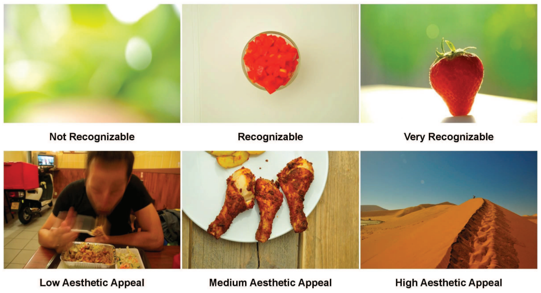


Figure 1: Example of images with different levels of recognizability and aesthetic appeal.

levels of familiarity and recognizability (e.g., Celebrities and Abstract). Images were also selected to roughly span a wide range of aesthetic appeal, based on the ratings already awarded to some of them on the website *500px.com*.

3.2 Lab-based Experiment

Fourteen paid participants took part in the Lab experiment, mostly originating from Europe. They were initially briefed about the general setup and their task. Then, they went through four short training sessions (each used 3 images, reflecting the evaluation scale) to ensure (1) the participant's acquaintance with the task and (2) the anchoring of the scoring scale for each quantity.

Each participant was then asked to assess color likeability, familiarity, recognizability and aesthetic appeal of each of the 200 images. They used four (one per quantity) 5-point discrete numerical scales, ranging from 1 being the lower score and 5 being the higher score. Semantic labels were added at the ends of each scale ("Bad Color" and "Excellent Color", "Not Familiar" and "Very Familiar", "Not Recognizable" and "Very Recognizable" or "Bad aesthetic appeal" and "Excellent aesthetic appeal", respectively). To avoid distraction during the image observation, these scales were kept in a follow-up separated screen.

To avoid fatigue effects that could harm the data collection procedure, due to the elevated number of images, the dataset was randomly split in two sets of 100 images each, to be evaluated by the same participant in two sessions, to be performed in different days. Each session lasted on average 40 minutes per participant, including a short break after scoring the first 50 images to minimize fatigue.

The experimental set-up followed the ITU-R BT.500 recommendation [19] and throughout the whole experiment, no time constraint was given for image observation and scoring.

3.3 Crowdsourcing Experiment

To repeat the experiment in a crowdsourcing environment, it was necessary to deal with two issues: (1) the fact that crowdsourcing tasks should not last longer than 5 to 10 minutes and (2) the risk of unreliable behavior of some of the workers, because of the distributed and remote nature of the test environment. Some adaptation in the experimental protocol was therefore needed to address these issues.

First of all, instead of two sessions with 100 images each, the crowd-based test consisted of 13 campaigns with 20 images each, where 5 of the images were the same for all campaigns to allow re-alignment and scale anchoring purposes. These 5 images corresponded to the 0th, 25th, 50th, 75th, and 100th percentiles of the distribution of all aesthetic quality scores as determined in the lab-based evaluation of the 200 images. The remaining 15 images per campaign were unique to each campaign. Due to this split, each worker in the crowd-based test only evaluated a subset of the original image set. Each worker could also participate in multiple campaigns.

To address the second issue, we implemented reliability control mechanisms to identify and filter out ratings from unreliable users or wrong test conditions [7]. Details on various reliability mechanisms for crowdsourcing experiments can be found in [6] and references therein. Unreliable user rating may be caused by language problems or wrong test conditions due to software errors or hardware incompatibilities, and need to be filtered out in order to avoid a falsification of QoE results. Additionally, there may also be *cheating* users who try to submit invalid or low quality work in order to reduce their effort while to maximizing their received payment, especially when this is very small [21]. We included therefore content questions [12, 7] in each campaign of 20 images after the 5th and 15th images. Furthermore, we targeted countries with an adequate proficiency in the English language, with an English speaking population larger than 10 million people or than 50% of the total population, as all test instructions were provided in English only. In order to limit the workers' participation to specific geographic regions, we used the Microworkers platform. We identified three regions in which workers could correspond to the above characteristics. Region 1 (CS-R1) corresponded to North America and major English speaking countries, such as USA, UK, Canada, and Australia, region 2 (CS-R2) corresponded to Western Europe, including workers from France, Germany, Italy, Ireland, the Netherlands, and Sweden, and region 3 (CS-R3) corresponded to Asia, including workers from Bangladesh, India, Pakistan, Philippines, Singapore, and Thailand. Each campaign was therefore replicated three times for each of the three geographic regions considered in this evaluation, resulting in a total of 39 campaigns.

We used the QualityCrowd [11] framework due to its flexibility and therefore easy adaptation to the task of aesthetics and recognizability evaluation. Similarly to the lab test, we also included a mandatory training to introduce the worker task and the same images used for the recognizability training in the lab experiment were used to allow workers practicing with the experimental interface. Each worker was presented with the image to be evaluated in a web interface that also provided two discrete five point scales to rate the content recognizability and aesthetic appeal of the shown image, similar to the computer-based interface used in the lab test. It is important to note that both questions were displayed on the same page as the corresponding evaluated images with recognizability question being on the left and aesthetic on the right, both below the image.

For each of the 39 campaigns, 30 different users participated and rated 20 images for 0.30 USD. In total, 28,080 images were rated consuming about 85 working hours at a total cost of 351 USD.

4. CROWDSOURCING RELIABILITY

Before comparing Lab and crowdsourcing results, the reliability of the crowdsourcing users has to be analyzed in order to identify and filter out unreliable user ratings. In the following, the results from the 13 different crowdsourcing campaigns are investigated. As mentioned in Section 3.3, each worker could participate in multiple campaigns. Figure 2 shows the histogram of the number of campaigns conducted by a single worker. It can be seen that regions CS-R1 and CS-R2 lead to similar results, while CS-R3 was significantly different. For CS-R1 and CS-R2, 6.61 and 7.36 campaigns were completed on average per user, respectively. Asian users (CS-R3) on average participated only in 2.47 campaigns. While at most $13 \cdot 30 = 390$ different workers could have participated per region, there were only 59 (CS-R1), 53 (CS-R2), and 158 (CS-R3) different workers, respectively. The higher user diversity in R3 may be caused by higher competition, as the workers are mainly located in Asia for Microworkers.com [5]. As a consequence, 14 and 15 workers from CS-R1 and CS-R2 are able to participate in all 13 campaigns, while no one from CS-R3 completes all campaigns.

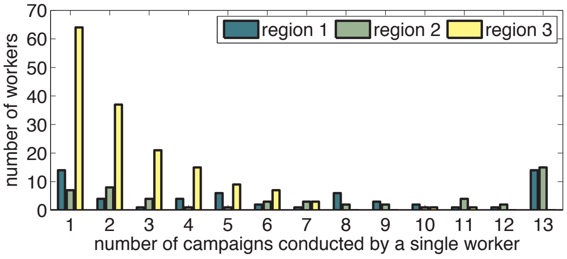


Figure 2: Number of campaigns conducted by a single worker

As a prerequisite to define a worker as ‘reliable’, all content questions about the images had to be answered correctly by an individual user. Figure 3 shows that the ratio of ‘reliable’ workers is similar for CS-R1 and CS-R2 with about 90 % over all campaigns. In contrast, only 70 % of workers from CS-R3 correctly answered all content questions. This discrepancy could be due to both language problems or cheating; either way, evaluations from these users could not be considered reliable, and were filtered out from the analysis presented in Section 5.

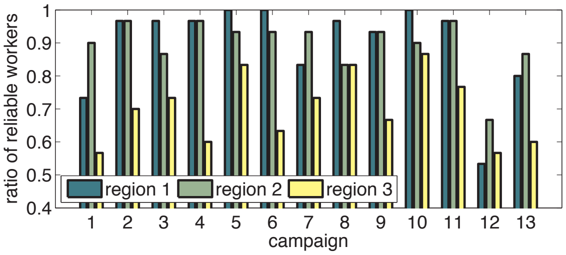


Figure 3: Ratio of workers who answered all questions correctly

The completion time per task was also considered for each user. A median task completion time of 3 min is observed for CS-R1 and CS-R2, while CS-R3 leads to 4 min. Taking a closer look at the mean task completion time reveals that for CS-R1 and CS-R2 the median task completion time is close to the mean completion time. However, for users in CS-R3, the average task completion time is significantly larger than the mean values. Thus, there are users with very large observation times for some images. The observation time per image is measured as the time from when the image is displayed until the time the user rating is given. Figure 4 shows the cumulative distribution function (CDF) of the standard deviation of the image observation duration per user in the different regions. Again, the curves from CS-R1 and CS-R2 overlap. However, the results for R3 are significantly different. In order to filter out users not rating seriously and being distracted during the subjective test, all users with a standard deviation of the image observation time larger than 20 s were rejected. This value was chosen to accommodate possible variations in download speeds of different users but reject users with significantly high variations in completion times.

Finally, unreliable participants were also identified as those rating images in a way that is significantly different with respect to the rest of the population. These outliers were also detected according to [19] and excluded from the subsequent analysis.

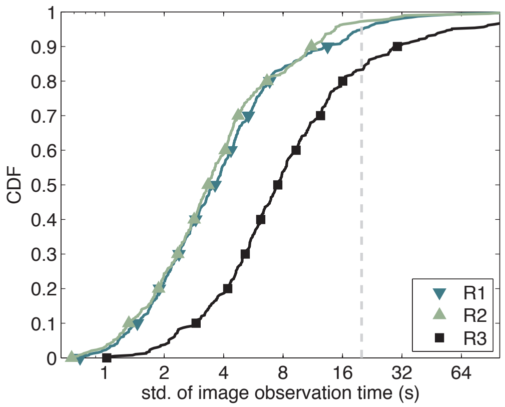


Figure 4: Cumulative distribution function (CDF) of the standard deviation (STD) of the image observation duration per user

Eventually, 14 % (CS-R1), 13 % (CS-R2), and 43 % (CS-R3) of the workers were filtered out, respectively.

5. LAB VS. CROWDSOURCING RESULTS

We computed normalized MOS (Mean Opinion Scores) from the lab and crowdsourcing experiments according to [19]. For lab-based scores, after rejection of one outlier participant, we normalized the ratings of each participant by subtracting from each individual score the mean score value for that participant and dividing it by the standard deviation of all the ratings of that same participant. Similarly, the scores were normalized for each crowdsourcing campaign separately. Although the original scores can also be compared, our experiments showed that the normalization allows for a better comparison, demonstrating all the disparities clearer. Normalized MOS for recognizability and aesthetics were computed as the mean values of normalized ratings given by all workers/participants who evaluated an image,but separately for each of the three regions (CS-R1, CS-R2, and CS-R3).

The primary goal of this study was to check whether subjective image judgments collected in a crowdsourcing and lab environments were consistent. As a starting point, we checked the degree of inter-participant consistency. We expected a similar level of inter-participant consistency across Lab and CS experiments to indicate a comparable level of understanding of the task and of the underlying image construct to be rated (either recognizability or aesthetics). Such similarity would in turn allow for a more fair comparison of the MOS. We computed thus the standard deviation of the scores assigned to the same image by all the participants evaluating it. High values of standard deviation for an image indicate high disagreement across participants on the judgment of that image. Table 1 shows the mean values of the standard deviation across all images in the database.

Participants were quite consistent in rating both recognizability and aesthetics. Furthermore, the degree of consistency is rather stable across Lab and CS conditions, with an exception for the crowd-

Table 1: Average standard deviation of individual scores across all images and participants

	Lab	CS-R1	CS-R2	CS-R3
recognizability	0.6590	0.6213	0.6430	0.7716
aesthetics	0.8164	0.7061	0.7198	0.7902

sourcing data obtained from CS-R3 in the recognizability scoring task. The results suggest that across all experiments participants were able to score images with an acceptable and similar degree of consistency, which allows for further comparison of the Mean Opinion Scores gathered in the experiment.

As a second step, we checked whether the MOS obtained from the Lab and CS experiment were similarly distributed. One way to test this is to check whether the MOS values for lab and CS originate from two distributions with the same median. We tested this by means of the Kruskal-Wallis test (MOS for Lab, CS-R1, CS-R2 and CS-R3 were found not to be normally distributed, hence the need for a non-parametric test). The test revealed that neither the recognizability MOS (df = 3, chi = 7.49, p = 0.0578) nor the aesthetics MOS (df = 3, chi = 1.37, p = 0.7126) had significantly different medians across Lab, CS-R1, CS-R2 and CS-R3. This can be visually inspected in Figure 5, where Lab and CS MOS distributions are shown to be spread around a similar range, without systematic scoring differences (e.g., aesthetics always scored lower in the lab experiment). Systematic differences were also excluded by running a Mann-Whitney U-test among all possible distribution pairs, which in all cases gave negative response.

From Figure 5 it is also noticeable that the distributions of Lab and CS scores do not always nicely overlap. To quantify this, we checked to what extent Lab and CS MOS were linearly correlated. The results are reported in Table 2. Interestingly, Lab and Crowdsourcing MOS are quite well correlated for the recognizability construct (above 0.8 except for R3, for which the correlation drops significantly); MOS obtained from CS participants originating from different geographical areas are also acceptably consistent. Consistency across geographic areas is maintained for the aesthetic scoring; however, this is not the case for the correlation between lab and crowdsourcing scores, for which a visible drop occurs. Especially CS-R3 MOS have little predictive power for the Lab scores of the same images (correlation coefficient of 0.23).

Table 2: Linear correlation between LAB MOS and CS MOS

	recognizability			aesthetics		
	CS-R1	CS-R2	CS-R3	CS-R1	CS-R2	CS-R3
Lab	0.869	0.856	0.652	0.398	0.418	0.228
CS-R1	-	0.956	0.752	-	0.932	0.750
CS-R2	-	-	0.791	-	-	0.794

To further investigate this mismatch between CS and Lab results, we checked whether the CS data would preserve the insights on the measured construct emerged from the obtained Lab data. To test this, we looked into the relationship between recognizability and aesthetic Lab scores. These two quantities were found to be not correlated (correlation coefficient of 0.19, Table 3). When computing the same quantity for the three CS experiments, we found instead recognizability scores to be highly correlated to aesthetic scores (above 0.85 for all regions). Again, we found a discrepancy between the Lab results and the CS results, probably due to the difference in scoring aesthetics.

The main surprise of the crowdsourcing experiments is that while recognizability shows high correlation with lab-based scores, aesthetics doesn't. There are a few ways to explain this phenomenon. In principle, the discrepancy between Lab and CS results could be due to a different interpretation of the aesthetic quality scoring task in the CS settings. However, participants were found to be equally consistent when scoring in Lab or CS (see Table 1), which suggests an equal clarity of the tasks. Another possible explanation is that in the lab test, participants had to evaluate four quantities, whereas in the CS experiment they focused only on recognizability and aesthetics. This may have primed participants, favoring an unconscious association of the two quantities. A third explanation could be that some microworkers are careless in the way they complete their task. If they may try to answer the first question (on recognizability) honestly, for the second question (on aesthetics) they could just replicate the judgment expressed for recognizability, to minimize their effort. This reasoning is supported by the fact that recognizability and aesthetics MOS in crowdsourcing tests are highly correlated, whereas this is not the case in the Lab.

Table 3: Correlation between recognizability and aesthetic MOS in lab and crowdsourcing experiments

	Lab	CS-R1	CS-R2	CS-R3
correlation	0.196	0.869	0.896	0.888

6. CONCLUSION

In this paper, we compared lab and crowdsourcing-based evaluations of image aesthetic appeal and content recognizability. We found that crowdsourcing workers can be quite consistent with lab participants in scoring recognizability, whereas this is not the case for aesthetic appeal. Further analysis of the results suggests that crowdsourcing can be used for this type of subjective assessments, but the evaluation methodology needs to be designed carefully to avoid misinterpretations or cheating by the online workers. In particular, priming, confusion or cheating effects may arise from the evaluation of two different quantities in the same task.

As current results do not indicate a clear cause for the discrepancy between lab and crowdsourcing scores, we intend to conduct another round of crowdsourcing experiments to clarify the matter further. To investigate confusion and cheating effects, a possibility would be to have workers repeating the same campaign with a reversed order of the questions (first aesthetics and then recognizability) or just one of the two questions at a time.

ACKNOWLEDGEMENTS

This work was conducted in the framework of COST Action IC1003 – Qualinet with partial support from the EC funded Network of Excellence VideoSense. Special thanks to photographer Mark Dekker (http://gplus.to/markdekker) for allowing to use his photos in the experiments and to Christopher Pangerl who helped organizing and conducting crowdsourcing evaluations.

7. REFERENCES

[1] O. Axelsson. Towards a psychology of photography: Dimensions underlying aesthetic appeal of photographs. *Perceptual and Motor Skills*, 105(August 2002):411–434, 2007.

[2] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. Quadrant of euphoria: a crowdsourcing platform for QoE assessment. *Network, IEEE*, 24(2):28–35, Mar. 2010.

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

[4] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, Apr. 2011.

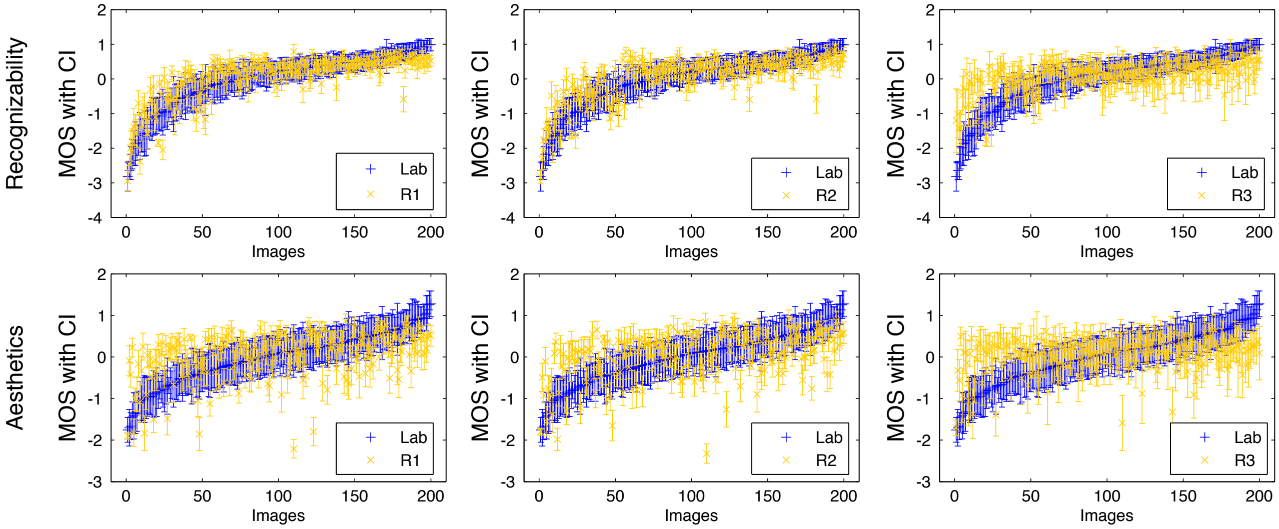


Figure 5: Comparison of Lab MOS with the CS MOS for the three scoring regions, for both recognizability (above) and aesthetics (below). Lab MOS (blue '+' markers) are sorted according to their magnitude.

[5] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com. In *Workshop on Future Internet and Next Generation Networks (FINGNet)*, pages 322–329, Seoul, Korea, June 2011.

[6] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. CrowdTesting: A Novel Methodology for Subjective User Studies and QoE Evaluation. Technical Report 486, University of Würzburg, Feb. 2013.

[7] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. Quantification of youtube qoe via crowdsourcing. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pages 494–499, Dec. 2011.

[8] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011.

[9] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[10] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[11] C. Keimel, J. Habigt, C. Horch, and K. Diepold. Qualitycrowd - a framework for crowd-based quality evaluation. In *Picture Coding Symposium (PCS), 2012*, pages 245–248, May 2012.

[12] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[13] P. Korshunov, S. Cai, and T. Ebrahimi. Crowdsourcing approach for evaluation of privacy filters in video surveillance. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia*, CrowdMM'12, pages 35–40, Nara, Japan, Oct. 2012.

[14] J. Lassalle, S. Member, L. Gros, T. Morineau, and G. Coppin. Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception? In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–6, 2012.

[15] P. Le Callet, S. Möller, and A. Perkis. Qualinet white paper on definitions of quality of experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Mar. 2013.

[16] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE journal*, 3(2):236–252, 2009.

[17] W. A. Mansilla, A. Perkis, and T. Ebrahimi. Implicit experiences as a determinant of perceptual quality and aesthetic appreciation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 153–162. ACM, 2011.

[18] S. Plous. *The Psychology of Judgment and Decision Making*. McGraw-Hill Series in Social Psychology. McGraw-Hill Education, 1993.

[19] Recommendation. ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland, 2012.

[20] J. A. Redi. Visual quality beyond artifact visibility. *Proc. SPIE*, 8651:86510N–86510N–11, Mar. 2013.

[21] S. Suri, D. Goldstein, and W. Mason. Honesty in an online labor market. In *Human Computation: Papers from the 2011 AAAI Workshop*, pages 61–66, 2011.

[22] W. Tschacher, S. Greenwood, V. Kirchberg, S. Wintzerith, K. van den Berg, and M. Tröndle. Physiological correlates of aesthetic perception of artworks in a museum. *Psychology of Aesthetics, Creativity, and the Arts*, 6(1):96–103, 2012.

[23] L.-k. Wong and K.-l. Low. Saliency-enhanced image aesthetics class prediction. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 997–1000, Cairo, Nov. 2009. Ieee.

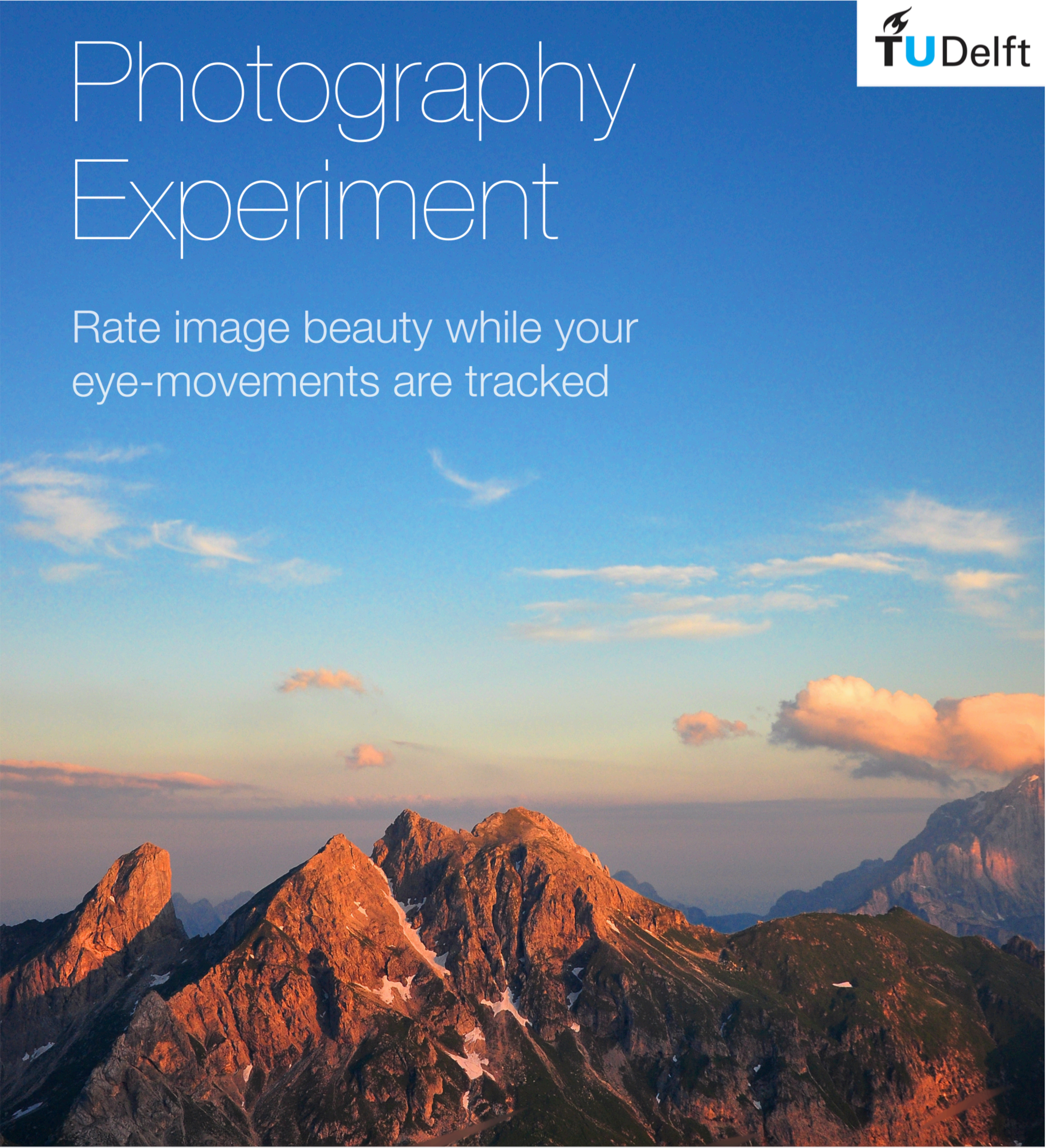
Appendix C :

Poster used for recruitment of laboratory subjects

TU Delft

Photography Experiment

Rate image beauty while your eye-movements are tracked



We are looking for amateur and expert photographers to take part in an experiment to evaluate image beauty.

Duration: 2 separate sessions ±40 min
Reward: €10 Gift Voucher for bol.com + refreshments

Interested?
Contact Isabel Povoa - misabelpovoa@gmail.com

Appendix D :
Phototo user interface design

Given the established protocol, the application requirements were to collect aesthetic appeal scores from participants in playful way. For that reason, we chose a flat design combined with a balanced use of light bright colours in the interface. Further, we chose to use only two typefaces: Dosis and Lobster from Pablo Impallari[36]. Whereas Dosis is a very simple, clean and rounded typeface perfect from short text, Lobster is a bold condensed typeface suitable for titles.

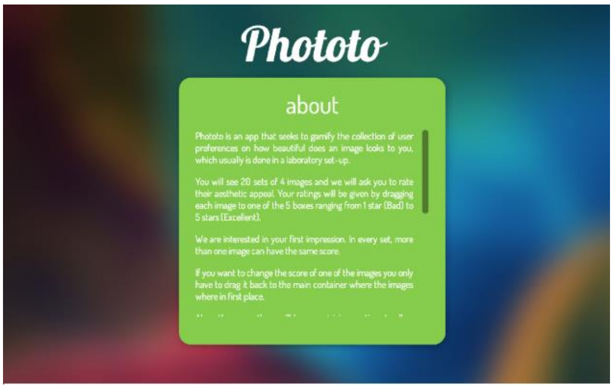
As mentioned before, we also used five containers with 1 to 5 stars for the scoring task. The action of sorting elements in boxes depending on their function/value is an intrinsic human value. Thus, this way of scoring would be naturally intuitive as well as playful.

Another element worth of mention is the arrows we used to guide a user in the right direction. The choice needed to match the playful design of the experiment and so, we used hand drawn, grunge style arrows.

To sum up, the main functionalities available for the user in the first screen are:

Participate in the experiment via the play button.

Get to know more about the web app (Figure 46) via the about button.



Phototo

about

Phototo is an app that seeks to gamify the collection of user preferences on how beautiful does an image looks to you, which usually is done in a laboratory set-up.

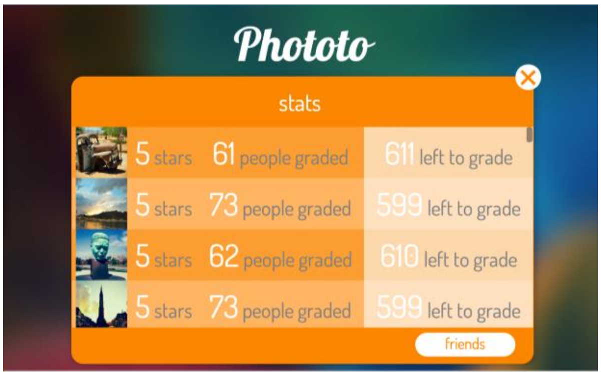
You will see 20 sets of 4 images and we will ask you to rate their aesthetic appeal. Your ratings will be given by dragging each image to one of the 5 boxes ranging from 1 star (Bad) to 5 stars (Excelient).

We are interested in your first impression. In every set, more than one image can have the same score.

If you want to change the score of one of the images you only have to drag it back to the main container where the images where in first place.

Figure 46: About explanation.

Get to know what is the average score so far for each of the images already rated by the participant, how many people have rated and how many people still need each of them as well as how many images each participant already rated, how many is missing rating and the participant's score (Figure 47) via the stats button.



Phototo

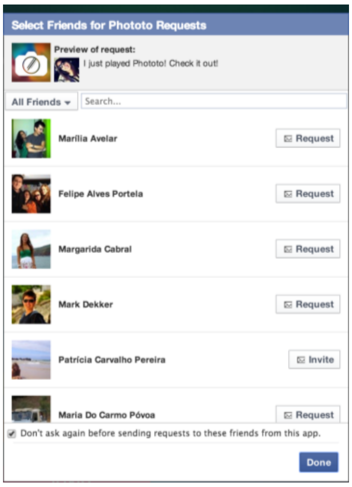
stats

5 stars	61 people graded	611 left to grade
5 stars	73 people graded	599 left to grade
5 stars	62 people graded	610 left to grade
5 stars	73 people graded	599 left to grade

Figure 47: Statistics screen.

Invite Facebook friends to participate (Figure 48) via the invite button.

[36] http://impallari.com/



I just played Phototo! Check it out!

Marília Avelar

Felipe Alves Portela

Margarida Cabral

Mark Dekker

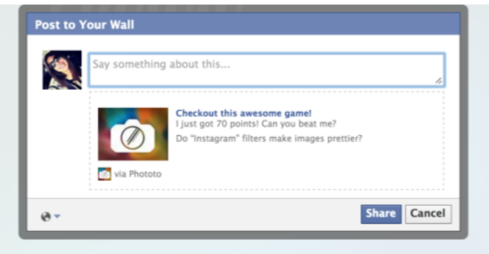
Patrícia Carvalho Pereira

Maria Do Carmo Póvoa

☑ Don't ask again before sending requests to these friends from this app.

Figure 48: Invitation screen.

Share the web app in your Facebook wall (Figure 49) via the share button.



Checkout this awesome game!

I just got 70 points! Can you beat me?

Do "Instagram" filters make images prettier?

Figure 49: Share screen.

To allow starting the experiment, we needed a play button to keep track of how many campaigns the user already had participated and to generate a payment code. We needed this type of control also to control how many and which images the participant had already scored. If a user was redirected from Microworkers, then we parsed the URL used to access the app in order to identify the worker ID and generate a payment code by the end of the rating task. We made use of one verification question, 18 content questions per participant for reliability checking purposes as well as tracked how long participants took to observe each 4-image presentation and their score. As a further matter, to ensure that the participant got acquainted with the rating task, we needed a training session and a simple instructions screen so that we could acknowledge that the participant was guided through the overall procedure.