DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-119

Evaluation of aesthetic appeal with regard of user's knowledge

Pierre Lebreton, Technische universität Ilmenau, Audio Visual technology, Ilmenau/Germany; Alexander Raake, Audio Visual technology, Technische universität Ilmenau, Ilmenau/Germany, Marcus Barkowsky, LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, Rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France

Abstract

Perception of aesthetic appeal in images depends on image features and internal factors of the observer. Previous work has shown that depending on the type of images: art, architecture, faces, or landscape, etc. the inter-variation of the perception of aesthetic between participants is not equal. The evaluation of Art and architecture are found less consistent than for faces or landscape. It has been theorized that these differences of consistency between persons are due to their lack of expertise into the field of art and architecture compared to recognizing beautiful landscapes and faces. Therefore, this study analyses the rating behaviour of the participants based on their knowledge of photography. Two different types of subjective data were collected: one using social media, and the other one based on a crowdsourcing experiment. In both cases, different groups of users have been identified and differences of ratings between these groups are found. Therefore, it seems to be important to consider who is considered as a test participant in a study targeting evaluating aesthetic appeal in photography.

Introduction

In this paper the evaluation of aesthetic appeal of pictures is considered. The perception of aesthetic appeal is considered as highly depending on people's perception, nevertheless a large amount of research has been performed to find invariant among people enabling the characterization of intrinsic properties of aesthetically appealing pictures. When considering aesthetic appeal, two main aspects can be considered: one is the picture's features and the other one is the internal factors of the observer.
The identification of features enabling the characterization of whether a picture is appealing has been widely studied. These features can be technically oriented and take into account factors such as the noise, the color contrast, the brightness contrast, the realism, the composition, and the simplicity of the picture [1]. These studies have been performed using both evaluation involving test participants and prediction algorithms. Going beyond technical factors, the question of novelty of the picture to the user and how it makes him feel compared to his previous experience is also of relevance and strongly affects the appeal scores [2]. To address this, work has been done on the prediction of the novelty of a picture by comparing the distance between the image under evaluation to a large database [3] (The distance between images was evaluated using the method described in [4]).

Additionally, it was found that specific content characteristics can be identified enabling higher agreement across participants. These for example include facial attractiveness and the presence of the opposite sex [5]. In case of landscape pictures, the agreement between participants can be found affected by factors such as naturalness, complexity, mystery, contrast, coherence [6]. These results have been the motivation the work performed on aesthetic appeal prediction using algorithms.

However, with artwork and architecture pictures a lower agreement between participants can be found [7]. The lower agreement between participants for the artwork pictures show that preference cannot be universally explained via visual features [7] and therefore there are internal factors which affect the judgments of the participants. A hypothesized explanation is people rate the pictures based on the novelty and how these pictures make them feel compared to previous experiences [2]. Therefore, as in the case of artwork or architecture, the participants do not have the same expertise this results in differences in their availability to perceive the novelty of a new photo [7]. This results in larger variation between the participants ratings. Pursuing these results, the comparison of the agreement between participants ratings for real-world images and abstract images shows that there is a higher agreement between participants who are asked to rate real-world images than abstract images [8]. This is due to the availability of a semantic interpretation of the picture which is shared by the participants and is not available in the abstract images. Therefore, the authors conclude that without semantic preference for a picture over another one is highly individual.

In parallel, one very important aspect is the evaluation methods and the participants used in the tests. In the context of real-world images (with semantic content), aesthetic appeal of pictures have been evaluated using different approaches and different types of crowds which have different photography skills. In laboratory tests, images were evaluated using many factors such as "image quality", "how the picture moves the viewer", the "appeal", the "aesthetics", the "originality", the "image harmony", the "imaginability" of the picture, the "pleasantness", or the "preference" in a pairwise experiment. Additionally, it has been shown that these scores can be obtained through crowdsourcing experiments using platforms such as Amazon Mechanical Turk or Microworker [9]. Another approach is to use annotations provided by photographers during a picture shooting contests such as "DPChallenge.com". These platforms provide statistical measures about the number of times a picture was viewed, preferred, its average rank within the contest, the number of comments, the number of times it was judged as favorite, etc. [1, 10]. Similarly, previous research used the community website "photo.net" which provides subjective scores by users on different scales including "aesthetics" and "originality" [3]. The number of times a picture was viewed and the number of ratings was also considered as in the previously mentioned study. One strong asset of such an approach is it provides annotations using a large panel of test participants having different background and

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-119

expertise.
However, it is important to mention that special care should be taken when using data from social media as other factors than aesthetics appeal and internal factors are involved. These other factors include peer effect and more generally social relationship between persons. Differences of understanding of the different scales can also be possible issue. Therefore, the picture properties may even less contributes to the obtained votes. Finally, physiological measurement and their relation with aesthetics were also considered by measuring the brain activity after viewing different stimuli [11], or looking at factors such as heart rate or skin conductance [12].

In conclusion to these studies, one aspect which appears until now weakly studied is the expertise of the participants and its link with the inter-participant agreement in ratings. Although it was found that pictures of artwork and architecture showed lower agreement between participants because of their lack of expertise, past research have mainly focused on finding intrinsic features enabling agreement between all kinds of participants (naive, and expert) resulting in high variation across participants. It is believed that expert observer may be more consistent in their way of rating as they may share the same perception of the technical properties of the pictures. Nevertheless, it is still expected that they will differ due to their internal factors. For expert observers, the contribution of visual features may not be as weak as indicated in previous research. In this paper, it is proposed to study the differences of ratings between persons and their consistency as a function of the participant's expertise. The second main objective is to study the relationship between the photography knowledge of the participants and the differences of ratings: how expert observers differ from naive ones in terms of consistency and scores.

To this aim, two different sources of data is used. First, the user behavior in social media is analyzed, and secondly a crowdsourcing experiment is conducted. The paper will address how features can be extracted from social media enabling to study the relationship between user ratings. In a second step a crowdsourcing experiment is described targeting the evaluation of the participant's expertise and aesthetic appeal ratings.

Section 2 describes the work conducted in analyzing social media to study the relation between rating behavior and photographer expertise. In section 3, a crowdsourcing experiment aiming to study the relation between user expertise and ratings will be described. And section 4 concludes this paper.

Analyzing social media

Photography-sharing websites such as Flickr or 500px have a large number of images of various appeals and also have a large number of users with different expertise and skills. Getting valuable information to study aesthetic appeal from this data is not easy as social components between users adds to the traditional factors affecting user's behavior faced to an image. This section will describe two indicators for picture's aesthetic appeal prediction based on social media analysis.

Identification of Indicator

To study picture aesthetic appeal, the 500px website[1] was considered. This website provides to many users the ability to share pictures, indicates whether they like it, and comment them. This social media has the advantage to consider photography as an art, therefore users do not exchange photos to illustrate their daily life but aims to show their skills in photography. Using the API of 500px it is possible to retrieve different kinds of information on the pictures and the photographers. These indicators include ratings, the highest rating received by the picture, number of views, timestamps, number of votes, the number of comments, ... It also provides technical parameters such as the size of the image, and which camera and lenses were used to take the picture. Social information includes the relation between photographers: who is following who, what pictures does someone like, and so on.

It should be stated that the meta-information "rating", and "highest rating" are internal features computed by the site 500px using proprietary algorithm. It does not relate to the image appeal, but to the current popularity of the picture and how long the picture has been posted on the website.

A first naïve indicator is defined by computing per-picture the ratio between the number of likes a picture received and the number of views (Equation 1). This characterize how many people "liked" a picture i (L_i) compared to the number of people who have seen this picture (V_i).

$$E_i = \frac{L_i}{V_i} \tag{1}$$

Research questions

Based on this measure, different questions were addressed:

1. Is there a difference between what photographers like in terms of quality?
2. Is there a relationship between the quality of their production and what they like?

To answer the first research question, data were collected using the 500px API to obtain 36 different photographers who have indicated liking at least 100 pictures. Meta-information for the pictures were retrieved as well. For each of the pictures liked by one of these photographers, the score using the equation 1 was computed. It should be mentioned that the 100 pictures liked by each participant are not the same. The first analysis evaluates if the distribution of the images' score is different between participants. A MannWhitney U test was used to explain the pictures' score as a function of the different users. It can be seen that the variable photographer has a significant effect on the scores computed based on the number of likes and views ($F = 31.48$, $p < 0.01$). Therefore, the research question 1), can be answered as: yes, according to the quality indicator described in Equation 1, the images liked by the photographers have different distribution of quality.

To answer the second research question, data were collected using the 500px API to obtain for different photographers the images they like and the images they have produced. For higher re-

[1] https://500px.com

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-119

liability on the collected data, only photographers who have pro duced more than 50 pictures and liked more than 50 pictures wer selected. This resulted in a selection of 175 photographers. Fig ure 1 depicts the relationship between the mean values of at leas 50 scores computed as described in Equation 1, of the picture produced by a photographer compared to the mean values of th scores of the images this same photographer liked. It can be ob served that the photographers who produced the highest qualit pictures (quality according to the metric Equation 1) likes onl pictures of higher quality. On the contrary, the photographers wh produced lower quality images like a wide range of image qual ity. This indicates that expert photographers are more selectiv in what kind of image they like, whereas less advanced photographers will like a larger span of image quality. It can be noted that in Figure 1 different spans of indicator values can be seen between produced an liked images due to the difference in popularity between the considered photographers.

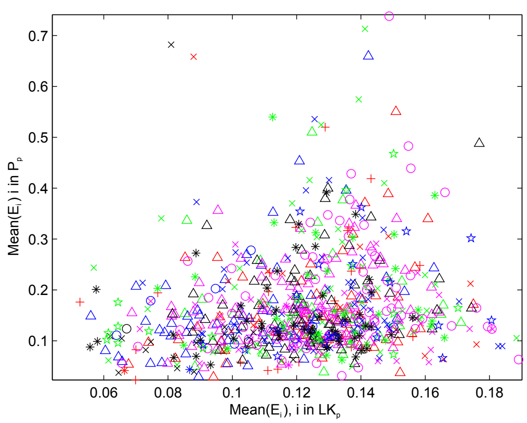

Figure 1. *Relationship between the production of a photographer and his own selection. LK_p is the set of images liked by the photographer p. P_p is the set of images produced by the photographer p.*

Discussion on the indicator

The limitations of the proposed indicator is revealed by the distribution of the indicator's value as a function of image category. A MannWhitney U test to explain the scores provided by the indicator depending on the picture category shows that picture categories have a significant impact on the scores (F=61.71, $p < 0.01$). For example, it can be seen that pictures belonging to the category nude are rated differently. This is due to the semantic contained in these pictures.

A second limit of the proposed indicator can be found in the temporal evolution of the number of views and votes along time (see Figure 2). Indeed, the number of views increases at a different speed than the number of likes. Lots of users are acting silently and do not report any kind of preference. Another critical aspect is all user are not acting fully honestly: from the comments and analysis of the like evolution, it can be identified that many users only rate and comment other profiles to advertise their own profile. This also affects the performance of the proposed indicator.

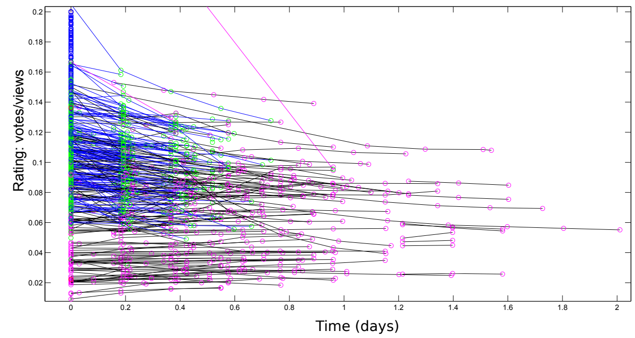

Figure 2. *Evolution of rating across time: many users are acting silently and do not report "likes". The colors encode the category to which the picture belong in the website 500px. The blue and black lines correspond respectively to the categories "popular" and "incoming".*

Alternative indicator

To overcome the limits of the previous indicator: the fact that the user does not necessarily report their preference, and the fact that users may not act fully honestly, it is proposed to consider the temporal evolution of the "likes". The general idea behind this measure is only highly aesthetically appealing picture will be able to keep receiving "likes" after a certain amount of time. Whereas less appealing pictures will be forgotten and lost in the high amount of new content put on the website and will not be consulted/liked anymore. This also enables to address the issue of users advertising for their own work as a high activity will only be observed after the publication of the picture, but will not be maintained after a larger amount of time. Figure 3 depicts the relationship between the number of likes as a function of the time since the pictures were inserted into the website on a logarithm scale. From this Figure, it can be seen that different categories of pictures can be identified: the pictures whose number of likes keep growing, and the pictures whose number of likes converge quickly. It can also be observed that only a limited number of images will achieve high scores.

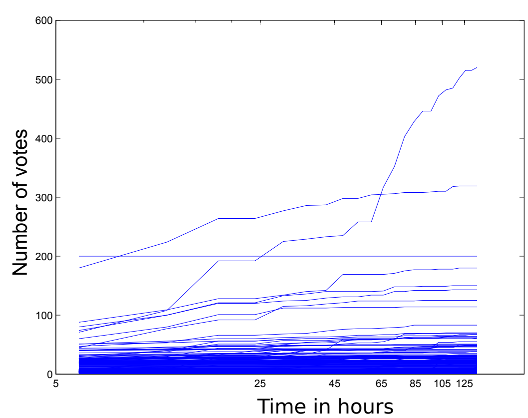

Figure 3. *Evolution of likes along the time for different images.*

Using linear regression between the logarithm of the time since the picture has been available on the site and the number of likes, it is possible to derive an indicator of how the picture

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-119

has been received by the community. Equation 2 provides the relationship between these features. $L_{i,o}$ is the number of likes the picture i have received at the time of the observation o. The value $T_{i,o}$ corresponds to the amount of time the picture i was available on the site at the time of the observation o. The couples of values (α_i, β_i) describes an image i on two criteria. α_i indicates how the new image was perceived, and its evolution on the website. β_i indicates the number of likes the picture has reached when the measurement started to be performed. The couple of values (α_i and β_i) need to be considered since even highly appealing pictures receive less likes after a certain amount of time. Therefore, the images can be classified into four categories as described in Table 1. Experimental results have shown that a value of α_i larger than 20 enables to identify highly appealing pictures, and a value of α_i lower than 0.6 will identifies less appealing pictures. This result from the fact that aesthetically appealing pictures will still receive attention from the community after a large amount of time, while less appealing picture tends to be forgotten.

$$L_{i,o} = \alpha_i \times log(T_{i,o}) + \beta_i \quad (2)$$

Discussion on the second indicator

One factor which can be discussed is the amount of time required to perform the evaluation. Current results are based on the observation of the evolution of the pictures' meta-information every 6 hours over one week. This has shown to be sufficient considering the number of images incoming within 500px as a clear distinction between appealing and non-appealing picture is quickly drawn using the data.

A limitation of the proposed method is the effect of the visibility of the pictures on 500px. Indeed, what makes a picture receive a "like", is of course its properties and the internal factor of the observer, but it is not limited to this. To "like" a picture, a user needs at first to be aware that the picture exists. This can be due to social connection or thanks to visibility of the picture on the website 500px. A picture put in front will be seen more times and will receive more "likes". In addition, peer effect and knowing who liked and produce the picture can affects the user decision. To further study the different quality indicators and the relationship between expertise and ratings a crowdsourcing experiment was conducted and is described in the next section.

Crowdsourcing study

To study the relation between the participant's knowledge of photography and his preferences, a crowdsourcing experiment was conducted.

Subjective experiment

The experiment was composed of three main tasks. The first one is a demographic, photography habit, and camera equipment questionnaire. The second task is a pairwise comparison experiment. Participants were asked to report their preferences for 7 different pairs of images. Finally, the third task evaluates participant knowledge in photography by asking technical questions. Five different images with specific flaws were presented to the user, and he was asked to identify the flaws by answering 5 different questions.

Demographic and habit questionnaire

The preliminary questionnaire was aimed to identify the participants' habit with regards to photography. After a demographic questions, the participants were asked about their equipment, their habit of sharing pictures: do they share their pictures, and if so how (via social networks, email, media storage), how they look for new ideas, do they study photography (in photo clubs, forums, read books, ...).

Once they finished filling up the questionnaire, they could switch to the second task.

Design of the pairwise comparison test

49 images were used in this test. They were based on a selection of pictures taken from 500px. It is worth mentioning that the images were properly referenced at the end of the test enabling the participant to find their respective authors. The selection of the images was done using the analysis presented in section . Images with different slopes ranging from 0.01 until 500 according to the previously presented indicator was selected. This enable to balance the test with both highly and lowly appealing pictures.
The design of the pairwise comparison was done using the optimized square design [13], the images were ordered in the square matrix based on the indicator computed in section , enabling to maximize the comparison between similarly aesthetically appealing images. Thanks to the square design, it is possible to estimate the Bradley-Terry [14] scores for each image with a limited number of pairwise comparisons. Considering that 49 images was used, a total of $49 \cdot (\sqrt{49} - 1) = 294$ comparisons are needed to apply the Bradley-Terry model. Each comparison was repeated 10 times. Resulting in 2940 comparison required. These comparisons were distributed over 420 participants from the Microworker platform. Each of them had to report their preference for 7 pairs of images.

Image flaws evaluation

In the last part of the test, participants were asked to vote on six different criteria: how did they like the picture, the composition, the exposure, the color, the sharpness of the picture, and whether a clear subject can be identified in the picture. The first five scales were rated on the 5 grade scale (with the labels "excellent", "good", "fair", "poor", "bad"). The presence of a subject was a binary reply (yes/no). Except for the question to evaluate if they liked the picture or not, an option "I do not understand the scale" was added. This enabled participants to report if they do not understand the proposed evaluation concept. Finally, the participants were also provided with the ability to add comments on each picture. Comments were not mandatory. The goal of this part of the test was to evaluate their ability to criticize a picture, and if they understand the technical terms such as "composition", "exposition", "identified subject".

Result

The first main result is the comparison between the indicator presented in section , and the subjective data from the pairwise comparison test. The Bradley-Terry (BT) model was applied to the pairwise comparison data. Figure 4 depicts the relationship between the BT-Scores and the proposed indicator. The indica-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-119

α_i	β_i	Description
Low	Low	New image i on 500px, with low aesthetic appeal.
High	Low	New image i on 500px, growing rapidly and having high aesthetic appeal.
Low	High	Old image i on 500px, growing slowly. It is a highly aesthetic appealing image, but have been on the site for a long time.
High	High	Old image i on 500px, growing rapidly. It is a highly aesthetic appealing image. This case is unlikely, but would indicate a highly appealing picture.

Images' aesthetic appeal and temporal evolution of likes

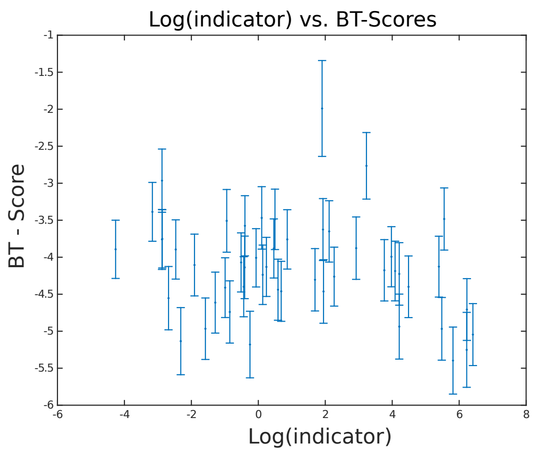

Figure 4. *Indicator vs. Bradley-Terry scores.*

tor cannot predict well the BT-Scores. There is a tendency of high values of the indicator to relate with highly preferred images. However, the reverse is not true and even images with low indicator values can be equally preferred as highly appealing pictures. An explanation for this result can be that the gradient of votes is not only due to aesthetic appeal of the pictures, and the popularity of the photographer affects both the visibility of the pictures and the decision to vote for the picture on the social media. Therefore it will be harder for less popular photographer to reach a similarly high gradient of votes. On the other side, participants from the crowdsourcing experiment were not aware of who produced the picture while rating and were not affected by a peer-effect or featuring effect by the website during the rating. In addition, the skill of each participant is not equal and adds another source of variation. The preliminary questionnaire and flaws evaluation tests were designed to study the knowledge of the user in photography and its link with the ratings behavior. In the image-flaw evaluation test, the participants were asked to report on a five grade scale to which extent they liked the presented pictures. Five different pictures were shown to them. A lilliefors test showed that the distribution of the liking rating across the participants is normal. Table 2 provides the results of non parametric Kruskal-Wallis test was applied to compare the distributions of the participants' ratings on the overall liking scale as function of the fact whether they have understood or not the technical scales. It can be seen that the evaluation concepts "exposition" and "presence of a main subject" are more discriminative, and tend to differentiate groups of participants having different ratings behavior. The different technical scales were found highly correlated as depicted in Figure 5. A PCA, showed that 75% of the variance is explained with two components. From the PCA (See Figure 6), it can be observed

Evaluation concept	Chi-sq	p
Composition	0.89	0.345
Exposition	3.74	0.0532
Color	0.05	0.824
Blurriness (Picture sharpness)	1.5	0.22
Main subject	5.56	0.0184

Kruskal-Wallis test comparing the understanding of scales, and rating behavior

that the axis like, composition, exposition, availability of a subject is closely related. Color and blurry being orthogonal to the other scales.

From the preliminary questionnaire, it was found that 70% of the users were male, 30% female. 37% of the users usually use Facebook to share a picture, followed by Instagram with 28% of the users. The other categories such as 500px, Flickr, Deviant Art was rather small with respectively 7%, 1% and 6%. 3% of the users reported using Snapchat, 12% do not use social media to share pictures and the remaining 6% reports using other ways. The camera equipment described by the users was mainly composed of smartphones and tablets, only 12% mentioned dedicated cameras. 6% reported having invested more than 600$ in photography equipment, 54% spent less than 200$ and 22% reported not having a camera. The population obtained seems then to be different than the one from the 500px website.

Considering the low number of participants using 500px, Flickr, Deviant Art, for the rest of the analysis these participants have been grouped in one category. Based non parametric Kruskal-Wallis test, it was found that depending on the main social media used by the participants the rating behavior in the last part of the test was different (Chi-sq=19.31, $p < 0.01$). Similarly, participants who spent more than 600$ in photography equipment rated significantly lower the images in the last part of the test (Chi-sq=15.9, $p < 0.01$).

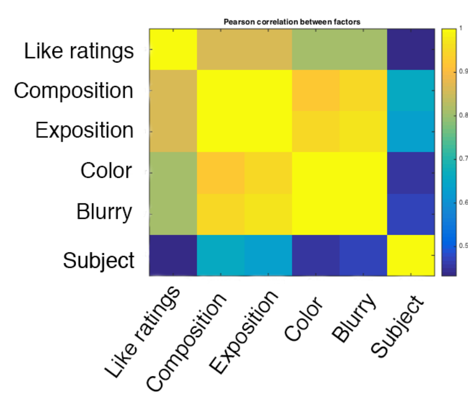

Figure 5. *Pearson correlation between scales.*

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-119

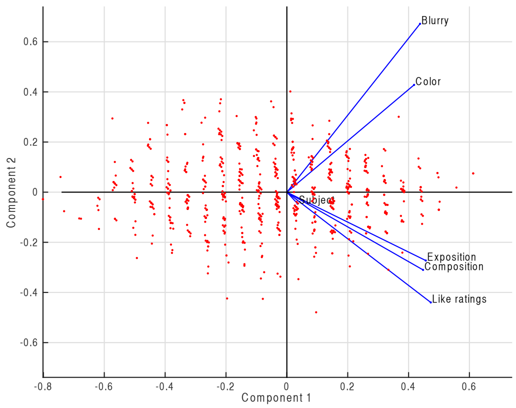

Figure 6. *Principal component analysis of the ratings.*

Conclusion and future work

This study aimed to analyze the relationship between the user's knowledge in photography and his rating behavior. Different approaches were employed, one is social media analysis and the other one is a crowdsourcing. The study based on social media described two different indicators based on the number of likes a picture received as well as the number of views and their respective temporal evolution. Based on the first indicator, a difference between images selected as favorite by photographers could be observed as a function of their ability to produced images widely accepted by other photographers. However, this first indicator is limited due to the content-dependency and social behavior of the photographers: users do not necessarily report liking pictures, and people only advertise their own work. A second indicator was established taking into account temporal evolution of the "likes" a picture received. This enable to overcome the "profile advertisement" aspects by hypothesizing that only highest appealing pictures will keep receiving ratings after a certain amount of time.

A crowdsourcing experiment was conducted, 49 images were compared using pairwise comparison. The temporal-based indicator was compared to the Bradley-Terry scores computed from the pairwise comparison experiment. Results show a high value of the temporal-based indicator will be associated with highly appealing pictures. But this observation is not reciprocal. The difference may be due to social aspects from the social network. The visibility of the picture on the website is also a main factor into the equation. All selected pictures were not equal in terms of visibility which affected the temporal evolution of likes, but not in the crowdsourced pairwise comparison tests. A second aspect is the difference in terms of population addressed in the experiments: one is composed of users with a high interest in photography, the others were less interested in photography as an art.

In the crowdsourcing experiment, a task was designed to evaluate the user-knowledge on photography. However, this task did not allow to determine categories of user ratings and the distribution of scores was found Gaussian. By looking into the type of social media people use, and the amount of money they spent in photography equipment it was possible to find significant differences of ratings between participants. However, the number of participants in the groups using sites oriented to photography as an art, and the users who spent a very high amount of money in photography equipment may be too low to be too conclusive.

Nevertheless, ratings behavior is different across participants and it is meaningful to consider it while evaluating aesthetic appeal in images. Further work will consider a larger crowd with users more interested into photography as an art to better identify the differences of ratings between groups. Regarding the social media analysis aspects, the visibility and interaction between user will be further analyzed to better differentiate pictures' characteristics and social aspects.

References

[1] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2006.

[2] I. Biederman and E. A. Vessel, "Perceptual pleasure and the brain," *American Scientist*, vol. 94, pp. 247253, 2006.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision (ECCV)*, 2006.

[4] J. Z.Wang, J. Li, and G.Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 947–963, 2001.

[5] D. E Berlyne, "Novelty, complexity, and hedonic value," *Perception & Psychophysics*, vol. 8, pp. 279286, 1970.

[6] R. Kaplan and S. Kaplan, *The experience of nature: A psychological perspective*, Michigan: Ulrichs Bookstore., 1995.

[7] Edward Vessel and Alexander H Denker, "Personalized visual aesthetics," in *Proceedings of SPIE - The international society for optical engineering*, 2014.

[8] Edward A. Vessel and Nava Rubin, "Beauty and the beholder: Highly individual taste for abstract, but not real-world images," *Journal of Vision*, vol. 10, no. 2, pp. 114, 2010.

[9] Judith Redi and Isabel Povoa, "Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd?," in *CrowdMM' 14 Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, 2014, pp. 25–30.

[10] Datta R., Li J., and Wang J. Z., "Learning the consensus on visual quality for next-generation image management," in *Proceedings of the ACM Multimedia Conference*, Augsburg, Germany, 2007.

[11] E.A Vessel, G. Starr, and N Rubin, "The brain on art: Intense aesthetic experience activates the default mode network," *Frontiers in Human Neuroscience*, vol. 6, no. 22, pp. 1–17, 2012.

[12] W. Tschacher, S. Greenwood, V. Kirchberg, S. Wintzerith, K van den Berg, and M. Trndle, "Physiological correlates of aesthetic perception in a museum," *Journal of Psychology of Aesthetics, Creativity, and the Arts*, vol. 6, pp. 96–103, 2012.

[13] Jing Li, Marcus Barkowsky, and Patrick Le Callet, "Subjective assessment methodology for preference of experience in 3DTV," in *11th IEEE IVMSP Workshop : 3D Image/Video Technologies and Applications*, Seoul, South Korea, Jun 2013, pp. 1–4.

[14] John C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller model of paired comparisons for image quality assessment," in *Proceedings IS&Ts Image Processing, Image Quality, Image Capture, Systems Conference*, April 2001.