The Quest for Visual Interest

Mohammad Soleymani
Swiss Center for Affective Sciences
University of Geneva
Switzerland
mohammad.soleymani@unige.ch

ABSTRACT

In this paper, we report on identifying the underlying factors that contribute to the visual interest in digital photos. A set of 1005 digital photos covering different topics and of different qualities was collected from Flickr. Images were annotated by a pool of diverse participants on a crowdsourcing platform. 12 bipolar ratings were collected for each photo on 7-point semantic differential scale, including dimensions related to interest, emotions and image quality. Every image received 20 annotations from unique participants. The most important appraisals and visual attributes for visual interest in photos was identified. We found that intrinsic pleasantness, arousal, visual quality and coping potential are the most important factors contributing to visual interest in digital photos. We developed a system that automatically detects the important visual attributes from low level visual features and demonstrated their significance in predicting interest at individual level.

Categories and Subject Descriptors

H5.5 [**Information storage and retrieval**]: Content Analysis and Indexing

Keywords

Interest; emotion; appraisal; crowdsourcing; computer vision.

1. INTRODUCTION

Deciphering why a user is interested in one image over another can benefit image retrieval and recommendation engines. Users actively create, look at and share millions of images on platforms such as Flickr[1] and Pinterest[2] on a daily basis. Both users and content repositories will benefit from novel methods for both indexing the content and better understanding users' desires while interacting with the content delivery platforms. Interest is an affective state that drives users' attention, and in combination with users' intent, it constructs users' preference and shapes their behavior on multimedia delivery platforms.

[1]http://flickr.com
[2]http://www.pinterest.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
MM'15, October 26–30, 2015, Brisbane, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806364.

1.1 Related work

In the last decades, psychologists increasingly supported the idea that interest is an emotion [14]. Appraisal theory is one of the most widely-accepted theories that explains the development of emotional experience. According to this theory, cognitive judgment about, or appraisal of, a situation is a key factor in the emergence of emotions. When an agent, in our case a person, faces a stimulus a series of concurrent evaluations or appraisals results in an emotional episode. The appraisals start with goal-relevance. If the stimulus is not relevant to the person, there is no further evaluation and therefore no emotion. If the stimulus is relevant to her, then further evaluations regarding its intrinsic pleasantness, the person's coping potential, novelty-complexity and its position to the norms and values of the person impacts the type of emotion she feels [13]. Silvia [14] has studied the appraisal mechanism of interest. He found novelty-complexity and coping potential to be the most important appraisals in the process of feeling interest. He has also identified that people with a higher level of familiarity with the subject have a higher level of interest in more complex forms of the stimuli. Therefore, comprehensibility and novelty both contribute to the feeling of interest. He later found that people can be categorized into different groups according to how much interest they feel towards an object or situation [14]. The first group, higher on curiosity and openness personality trait, is more likely to be interested by novel and more complex stimuli. For the second group, however, coping potential and comprehensibility was more important.

Gygli et al. and Grabner et al. [4, 3] demonstrated how the visual content features related to unusualness, aesthetics and general preference in images are important in predicting visual. However, the predicted interest does not take the personal differences into account and these different attributes are not separately learned. As a result their best performing results were obtained by combining all the visual features related to different attributes. Halonen et al. [5] identified a set of characteristics that are related to visual interestingness, including aesthetics, affect, colors, composition, genre, and personal connection.

1.2 Main contributions

In this paper, we are reporting a work in progress on identifying the underlying visual and personal factors contributing to the construction of visual interest as an emotion. To

this end, following the work of Silvia and colleagues [14], we first identified the appraisals that are important for interest. Then, a set of ratings on 12 different scales was collected to identify the visual attributes relevant to image interestingness, e.g., quality and complexity [14, 5]. The visual attributes that are important for visual interest are identified and learned from low level visual features. An automatic person-specific interest detection is also implemented and presented. The major contributions of this work are as follows: first, the appraisal structure of interest is studied using a set of real-world digital photos; second, for the first time, the effect of visual attributes and appraisals on visual interest in photos are investigated; and third, a content-based visual interest prediction from visual attributes is presented and evaluated.

2. DATA COLLECTION

2.1 Image selection

We aimed to create a diverse set of images both in terms of topic and aesthetics. A set of queries was constructed from image titles in International Affective Picture System [9] and Kodak Lossless True Color Image Suite[3]. Chu et al. [1] showed person familiarity in images depicting people to be an important factor in their interestingness. Therefore, to add a set of familiar faces, the 20 most famous celebrities according to Forbes[4] were also added to the query terms. In total, 139 queries were chosen for image selection. To ensure the diversity of images in terms of quality and aesthetics, we tried to diversify the photos based on the cameras that captured them. Flickr API was used to collect images published under Creative Commons (CC) license[5]. First a list of all camera brands and their models were retrieved. At every step of querying Flickr, 15 different models of each brand were randomly selected and the top 4 images according to relevance and date-posted (ascending) were retrieved. We decided to avoid the Flickr interestingness sort option due to its bias towards more popular images and its undisclosed technical definition. To include the photos without EXIF metadata and known cameras, for each query, top 20 photos were also retrieved regardless of the cameras capturing them. To avoid bias from the shape of the image and its details during the annotations, only images with landscape orientation with the aspect ratio 4:3 were kept. Black and white images, photos with watermarks and photos depicting extreme nudity (depicting genitals) or extremely unpleasant images were removed from the dataset. The black and white images were discarded to have a more homogenous set. A significant number of queries, including most of the queries related to celebrities did not yield any relevant results. Therefore, a second round of photo collection was performed. In the second round, at each step 200 CC licensed images with the permission for derivation and no requirement regarding the camera were collected for the missing queries. In the second round, images that were not very far in shape from landscape orientation with 4:3 aspect ratio were cropped and added to the dataset after manual selection. We randomly subsampled the collected photos from the first round considering the presence of the results from all the queries. The photos relevant to the queries with a small number or no results in the first round were manually selected from the results of the second round. In total 1005 photos were selected and resized to 600×400 pixels.

[3]http://r0k.us/graphics/kodak/

[4]http://www.forbes.com/celebrities/list/

[5]http://creativecommons.org/licenses/

(a) Most interesting

(b) Least interesting

Figure 1: The most and least interesting images according to the average ratings given by workers.

2.2 Labeling via crowdsourcing

In order to asses the Big-Five personality traits, we used the fake-proof personality questionnaire proposed by Hirsh and Peterson [6] in addition to a short version of IPIP test [2]. Similar to the work of Silvia et al. [14], the same personality questionnaires were used to assess the following traits: openness scale, sensation seeking, the tendency to seek new and varied experiences, curiosity, epistemic curiosity and perceptual curiosity. General interest in arts, sports, people, cars, celebrities, animals, scenery and food was also assessed using a 7-point Likert scale.

Crowdsourcing provides the scale and efficiency required for large scale data collection. Therefore, to collect annotations, we turned to Amazon Mechanical Turk[6]. It is essential to ensure the quality of labels collected through crowdsourcing. We followed a two-step strategy that we designed based on many current state-of-the-art crowdsourcing approaches [8]. The first step was a recruitment phase which included the personality, demography and general preference questionnaires in addition to annotating and describing one sample image. We were able to efficiently monitor the consistency of the responses by calculating the correlation between the Big-Five results of the IPIP test and the fake-proof test. Workers who randomly selected the responses were easily identified. We could also verify that the workers had a good knowledge of English by reading their description of the sample image. Moreover, we could verify that the scripts were properly working and compatible with the workers' browser. Two sets of qualifications were given to the workers who provided satisfactory responses to the qualification task. One set was given to the workers with a lower openness to experience score and one to the ones with the higher scores. For the second phase, image labeling, two identical batches that required 10 unique workers were published for these two groups. This way, we were sure to receive the same amount of labels from workers with different degree of openness to experience. Each Human Intelligence Task (HIT) in the second phase, consisted of providing 12 ratings on a 7-point semantic differential scale to 5 images. These scales were selected to identify the appraisals and visual attributes relevant to visual interest [14, 5]. A list of

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

all the scales is available in Table 1. This dataset is freely available for academic research[7].

Table 1: The 12 bipolar scales and their inter-rater agreements are given; for α the higher the better.

Scale	Krippendorf's α
Complex - Simple	0.13
Low quality - High quality	0.23
Appealing - Unappealing	0.29
Natural - Staged	0.30
Pleasant - Unpleasant	0.24
Arousing - Soothing	0.31
Familiar - Unfamiliar	0.14
Easy - Hard (to understand)	0.18
Comprehensible - Incomprehensible	0.13
Coherent - Incoherent	0.11
Boring - Exciting	0.24
Interesting - Uninteresting	0.20

2.3 Analysis of annotations

The qualification HIT was initially published for 150 workers. No requirement such as HIT acceptance rate or the number of completed HITs was required to avoid shrinking the pool of eligible workers. We further extended the qualification HIT for 150 more workers due to the lack of workers with lower openness trait. In total, there were 300 workers who performed the qualification HIT. We extended the qualification of high and low openness to 52 and 43 workers respectively. The qualified workers were invited to participate in the main image labeling HITs via MTurk messaging. Out of the total of 95 qualified workers, 66 workers (69.5% of the invited workers) performed at least one of the main HITs. From 66 workers who participated in our main HITs, 39 were male and 27 were female and they were mostly from the USA. They were in average 36.0 ± 11.4 years old. The workers on average spent 11.49 seconds annotating every image on 12 different scales. The whole crowdsourcing campaign cost about $2'400.00 USD for performing 20'100 image labeling tasks and took 11 days to complete. The average effective hourly rate for the image labeling HITs was $5.74 USD. Krippendorf's alpha on ordinal scale was calculated for each scale and is listed in Table 1.

3. METHODS

3.1 Appraisal structure and relevant attributes

To study the effect of different attributes, including appraisals on interest, we used a mixed-effect linear model estimating the interest from the other ratings. Appraisal of coping potential was constructed by averaging the scores given to "easy to understand - hard to understand", "comprehensible - incomprehensible" and "coherent - incoherent" scales. Appraisal of novelty is related to both complexity and familiarity; however due to their low correlation ($\rho = -0.23$), we decided to keep them separate. Appraisal of intrinsic pleasantness was constructed by averaging pleasantness and aesthetics scores. Arousal, naturalness and quality were also added as fixed effect independent variables. Other ratings which can be associated with appraisals were set as fixed effect independent variables and images as a random effect covariate. A mixed-effect linear model enables us to model the between image variations in a random effect term while studying the effect of the independent variables on the dependent variable, i.e., interest. Interest scores were calculated by averaging the interest scale and reversed boring-exciting scale.

3.2 Attribute learning and interest prediction

A set of attributes important to visual interest was identified. An automatic system that can identify these attributes will be useful for an interest-centered indexing scenario.

Gygli et al. [4] identified a set of features that are related to visual quality, aesthetics and general preference. Inspired by their findings, the following features were extracted from images: spatial pyramids of sift histograms [10] (as a general scene descriptor), jpeg compression rate from an uncompressed image (as an indicator of image complexity), the color histogram, contrast, naïve arousal score [11] and edge distribution [4]. Khosla et al. [7] identified a set of features that were effective in predicting memorability. Even though memorability and interestingness have inverse correlation [4], the features can be effective in capturing the semantic content of the image. The following features were extracted based on [7]: histogram of oriented gradients (HOG) at 2×2 and 3×3 cells which were max-pooled at 2 spatial pyramid levels; color name feature; local binary patterns (LBP) which is reduced by max pooling; and a bag of word representation of GIST descriptors using the dictionary size of 512. The features that are taken from [7] are extracted using the feature extraction toolbox[8].

Given the nature of the attributes, we used a regression with sparse approximation of data [12], which performed slightly better than Support Vector Regressor (SVR) with an RBF kernel with much lower processing time. The sparse representation of the dictionary (here the normalized training set) was calculated using the Spectral Projected Gradient Method for ℓ_1-minimization (SPGL1) [15]. A Principal Component Analysis (PCA) was used for dimensionality reduction. PCA was applied on the training set in each iteration of the cross-validation. After calculating the number of principal components that carry 95% of the variance, the corresponding mapping was used to reduce the dimensionality of the test set. A leave-one-out cross-validation strategy was used for the evaluation.

The most straightforward way of predicting users' interest given a set of images with ratings is to use the labeled images of each user as the training set. We implemented this baseline method using the same regressor explained in the previous section. In order to demonstrate the significance of the attributes, a second method is proposed that learns individual interest from the visual attributes learned from the ratings given by a larger population. To this end, we trained a linear regression that combines the visual attributes for each image and each user. Only the data from the 22 workers who annotated at least 400 images were processed for individual interest prediction.

4. EXPERIMENTAL RESULTS

4.1 Appraisal structure

The statistical analysis on the effect of appraisals and attributes was performed using a mixed-effect linear model. It was found that the most important attributes were intrinsic pleasantness and arousal which further justifies the

[7]http://cvml.unige.ch/databases/visInterest

[8]https://github.com/adikhosla/feature-extraction

affective dimension of interest. Complexity and familiarity both contributed positively to higher interest whereas coping potential had a negative effect on average. Naturalness had the smallest effect in presence of the other factors. The coefficients of the fixed effects were as follows: arousal: 0.29, intrinsic pleasantness: 0.11, familiarity: 0.01, quality: 0.07, complexity: 0.02, coping potential: -0.08, and naturalness: 0.01.

The traits that are related to curiosity and openness were averaged to generate an openness score. The workers were divided into two groups of high and low openness and the previous model was fit on their annotations. Although the results were similar, for the groups with a higher openness personality trait, complexity coefficient was higher (0.17 vs. 0.13) and coping potential had a more negative effect (-0.09 vs. -0.04). This is in agreement with the findings of Silvia and colleagues [14] who observed that people with higher openness are interested in stimuli that is more difficult to understand and is more complex.

4.2 Attribute and interest detection

We attempted learning the attributes that are related to interest, namely, quality, familiarity, naturalness, coping potential, intrinsic pleasantness, arousal, complexity and general interest. All the labels were generated based on the averaged scores given by 20 participants. The scores were scaled between [-0.5 ,+0.5]. At every iteration of cross-validation, visual content features were normalized by subtracting the mean and dividing by the standard deviation of the feature values in the training set. The detection results are evaluated by r-squared (r^2) metric which adjusts for the random level (averaged score from the training set) and is roughly equivalent to the squared correlation. Root-mean-square error (RMSE) and Pearson correlation (ρ) are also reported. The results are shown in Table 2.

Table 2: Attribute detection results are given. The last two rows include the results for individual interest prediction. All attribute and interest scores are scaled between [-0.5 ,+0.5]; for RMSE $\in [0, 1]$ the lower the better; for r^2 and ρ the higher the better. Acronyms: Intr.: Intrinsic; Ind.: Individual, BL: Baseline, At. Attribute-based.

Attribute	r^2	ρ	$RMSE$
Quality	0.26	0.51	0.13
Coping potential	0.06	0.27	0.11
Naturalness	0.32	0.57	0.18
Intrinsic pleasantness	0.16	0.40	0.15
Familiarity	0.08	0.30	0.14
Arousal	0.37	0.14	0.09
Complexity	0.13	0.37	0.12
General interest	0.20	0.44	0.13
Ind. interest (At.)	**0.08**±0.00	0.27±0.08	0.24±0.07
Ind. interest (BL)	**0.03**±0.00	0.26±0.09	0.24±0.08

The evaluation based on r^2 demonstrates that for individual-specific interest prediction, the attribute-based method significantly outperforms the direct baseline approach (based on one tailed t-test on r^2; $p < 0.01$).

5. CONCLUSIONS

In this work, we investigated the underlying factors for visual interest. Affective content, quality, coping potential and complexity are shown to have a significant effect on visual interest in images. We demonstrated that using the visual attributes that are learned from a population, it is possible to detect individual interest. Content-based methods are however limited by their input. The first appraisal in an affective episode is goal-relevance which cannot be assessed only from the content without the knowledge of a user and the context. Therefore, having social information and personal connections, especially for photos depicting people, is very important. We also expect that some appraisal subcomponents that we did not consider in this study have to be identified; for example, there are different categories of intrinsic pleasantness and the role of aesthetics in cognitive appraisal theory is still unclear. In this paper, promising results are reported for detecting interest and its relevant attributes. In the future, the attributes can be learned from a larger set of images and public databases specific to image quality and aesthetics assessment.

6. ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation's Ambizione grant. We would like to thank Danny Dukes for helpful discussions.

7. REFERENCES

[1] S. L. Chu et al. The effect of familiarity on perceived interestingness of images. In *IS&T/SPIE EI*, 2013.

[2] L. R. Goldberg et al. The international personality item pool and the future of public-domain personality measures. *J. Res. Per.*, 40(1):84–96, 2006.

[3] H. Grabner et al. Visual interestingness in image sequences. In *ACM MM*, 2013.

[4] M. Gygli et al. The Interestingness of Images. In *IEEE ICCV*, 2013.

[5] R. Halonen et al. Naturalness and interestingness of test images for visual quality evaluation. In *SPIE IQSP*, 2011.

[6] J. B. Hirsh and J. B. Peterson. Predicting creativity and academic success with a "Fake-Proof" measure of the Big Five. *J. Res. Pers.*, 42(5):1323–1333, 2008.

[7] A. Khosla et al. Memorability of image regions. In *NIPS*, 2012.

[8] A. Kittur et al. Crowdsourcing user studies with Mechanical Turk. In *ACM CHI*, 2008.

[9] P. Lang et al. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, Univ. Florida, 2008.

[10] S. Lazebnik et al. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[11] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.

[12] P. Noorzad and B. L. Sturm. Regression with sparse approximations of data. In *EUSIPCO*, 2012.

[13] K. R. Scherer. Studying the emotion-antecedent appraisal process: An expert system approach. *Cogn Emot.*, 7(3):325–355, 1993.

[14] P. J. Silvia et al. Are the sources of interest the same for everyone? Using multilevel mixture models to explore individual differences in appraisal structures. *Cogn Emot.*, 23(7):1389–1406, 2009.

[15] E. van den Berg and M. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2009.