DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Experiencing the interestingness concept within and between pictures

Christel Chamaret, Claire-Hélène Demarty, Vincent Demoulin, Gwenaëlle Marquant; Technicolor; Cesson-Sévigné, France

Abstract

Interestingness is the quantification of the ability of an image to induce interest in a user. Because defining and interpreting interestingness remain unclear in the literature, we introduce in this paper two new notions, intra- and inter-interestingness, and investigate a novel set of dedicated experiments.

More specifically, we propose four experimental protocols: 1/ object ranking with a pre-defined word list, 2/ pair-wise comparison, 3/ image ranking and 4/ eye-tracking. We take advantage of experimenting on the same dataset to draw potential links between the collected data and to state on the agreement between subjects. While we do not evidence a relationship between the local (intra) and global (inter) notions of interestingness, we do observe correlated outputs throughout the different protocols. Beyond the low or moderate values obtained from inter-rater agreement metrics, we point out the experimental reproducibility to argue about the universal nature of the interestingness notions.

In addition, we bring deep insights on the relationships between interestingness and 7 other criteria, some of them already pointed out in the literature as being linked with interestingness. Unusualness and emotion seem to be the strongest enablers for interestingness. These insights are highly relevant for future work on modeling.

Introduction

Understanding what makes something interesting is a challenging task for scientists. Classic theories, coming from the psychology community, preferably attribute the interest elicitation to objective features related to the considered stimulus, e.g., *novelty, complexity, uncertainty and conflict* as mentioned by Berlyne in the 1960s [1]. Recent progress in emotion psychology revisited this approach: the interest is rather due to the people's appraisal(s) of an event, as demonstrated also for emotion [2, 3]. Thus, not only the stimulus is responsible for the interest elicitation, but additionally and mainly, the way the event is received, understood and a source of motivation [4] for each person. It questions the universality of the interest concept, that we are going to discuss in this paper.

Going deeper into this way, Silvia [5, 6] proposed two distinct appraisals to explain the origins of interest: 1/ the evaluation of an event's *novelty and complexity* (transmitting the original idea of classic theories), 2/ the evaluation of an event's *comprehensibility*. Basically, if people interpret a stimulus as being new and understandable, this latter induces a high interest. In other words, the novelty combined with a slight complexity piques people's curiosity and elicits an interest.

In the image processing community also is the interest concept addressed, but rather under the name Interestingness [14, 15] or Importance [10, 13]. Halonen *et al.*[24] define the interestingness as the "power of an object to awaken responses other than those called forth by its aesthetic form. [...] interestingness gives emotional or conceptual meaningfulness".

Whatever its denomination, interestingness prediction has regained importance nowadays, since it could highly serve today's applications, such as search engines, recommendation tools, advertising, e-learning, etc... One indicator of such trend may be observed also in Flickr website which provides a tag, associated to each image, named Interestingness (purely inferred from user's annotation).

Because interestingness is a high semantic concept, learning a computational model of how people perceive interestingness is very challenging due to the diversity of stimuli and to the differences in human perception. Until now, this topic has always been considered in the image processing community regarding two independent axes: 1/ Relative importance of objects in images [10, 13], 2/ Interestingness-based ranking of images [14].

Because defining and interpreting interestingness remain unclear in the literature, whatever the community (psychology or image processing), we claim that we need a novel set of specific experiments to confront the local and global views, i.e., the main axes mentioned previously. We also have the intuition that the universality of interestingness should be further explored and validated, as questioned by the psychology community. More generally, studying interestingness may allow to go one step further towards increasing the semantic understanding of content.

In this paper, our objectives are threefold. First, we investigate the universality of interestingness under a new experimental view. Demonstrating the reproducibility of results through our experiments would allow us to design confidently an efficient computational model of interestingness as a future work. Several experiments will help us measure two different notions of interestingness, which we formalize as:

- Intra-interestingness (local notion): the relative interest of the elements of an image,
- Inter-interestingness (global notion): the interest of an image when compared to others.

Second, thanks to the creation of a common dataset, our challenge will be to gauge the correlation between these two notions. Finally, targeting a modeling purpose, we will also evaluate which criteria, among aesthetics, emotion, unusualness,... are linked to interestingness.

For the sake of clarity, in the rest of the paper, we will call *concepts* things that are not physically *palpable*, i.e., that cannot be objectively segmented or partitioned, e.g., a color or a mood, contrary to an object or a person.

The paper is organized as follows. It will first draw the picture of previous works. Second, the proposed experiments will

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

be detailed, followed by a presentation of statistical results we obtained from these experiments. Finally, we will discuss successively potential links between the intra and inter-interestingness aspects, the image criteria responsible for inducing interestingness and its universality, before proposing some perspectives for future work.

Related works

As introduced previously, some efforts from the psychology community have been devoted to characterize the interest elicitation. Using either simple stimuli (e.g., polygons) or other modalities (e.g., poems), Silvia [5] demonstrated the involvement of novelty, complexity and comprehensibility for eliciting interestingness.

Closer to our intent, Gygli *et al.* [14] estimate numerically the contribution of an extensive set of criteria in the context of interestingness-based ranking of images. They evidence three groups of criteria as being highly relevant to the characterization of inter-interestingness: aesthetics, unusualness and scene types. In addition, they argue for the existence of a universal concept related to inter-interestingness, based on the high values of agreement obtained between observers.

Moreover, the affective track was also addressed to validate the original intuition that pleasure [16], high pleasantness [17], instant enjoyment [18] are a source for creating high interest. Nonetheless, this has been invalidated in other works [4] where paintings rated as interesting were also objectively assessed negative, disturbing, complex... Recently, inspired by [14, 26, 27], Soleymani [25] investigates how some factors such as affective content, quality, coping potential and complexity are correlated with visual interest in images. He concludes that the most important attributes were intrinsic pleasantness and arousal, justifying the affective dimension of interest. He also points out the links that exist between curiosity, openness and personality traits of people, and interestingness.

Focusing now on intra-interestingness, some studies get interested in local features and their role in assessing the relative importance of objects in images. In this vein, [10, 13] explore the influence of object sizes, locations, categories or context. More precisely, Berg *et al.* [13] makes a link between object importance and user descriptions of images. Thus, they are able to point out various factors related to perceived importance, such as image composition, content semantics (category of object or scene), and context, i.e., common vs unusual. For example, they conclude that larger objects and/or closer to the center of the image are more likely to be described. Moreover, they hypothesize that people usually perceive the animated objects as the main subjects of a picture, whereas common objects are identified as background content elements. Indoor scenes are also much more likely to be mentioned than outdoor scenes. Last, objects in an unusual setting are more likely to be described than when in a common setting.

Spain *et al.* [10, 11] notice that when describing images, humans give different priorities to different objects in images. Based on this assumption, they define a function to model the level of priority of each object. Because objects were named quasi-independently they argue that the process of naming objects in an image is akin to drawing balls from an urn without replacement. Hence, this function predicts the relative importance of each object directly from a segmented image while combining several object-related and image-related features. Out of the 46 tested features, many of them were found to be redundant. The authors also conclude that some features such as object position and size are informative whereas saliency is not. This work is relevant in the sense that their importance estimator works without awareness of what the object is (object identity). However, the way they catch the importance notion is questionable. The authors ask observers to "name 10 objects [they] see in the picture", assuming that they are also the most interesting ones. This experiment also restrains its conclusion to object importance, as no concepts were used by the observers.

Other alternative protocols for the annotation of interestingness have been investigated. The ESP game by Ahn *et al.* [7] presents the same image to two players. As it is currently implemented, players are encouraged to produce a matching word as quickly as possible with their partner. When multiple games are played iteratively on the same image, the named words form an ordered list. The implicit idea is that words associated with more important objects will tend to appear earlier. However, some bias is introduced as players use strategies to reach consensus as quickly as possible.

Elazary *et al.* [8] base their approach to measuring interestingness on the LabelMe [9] database. LabelMe is a communal online annotation tool to build image databases for computer vision research: people are asked to recognize and segment as many object categories as they can in images. Elazary's assumption is that the order in which objects are named during the annotation process reflects how interesting these objects are. However, annotations from past users are visible. Consequently, an object can only be segmented once, producing a single list of present objects in each picture for all users. The ease to outline a given object also influences the order of drawing, and consequently introduces some bias in the obtained importance ranking.

Even though most cited literature converges on the universality of interestingness, Grabner *et al.* [12] mention that human interest is also raised from the context and from personal experiences. For example, someone will intuitively show more interest in pictures with his/her own children than in pictures with some people he/she does not know. Hence, authors argue that the interestingness process is not universal: general consensus appears but this processing is also strongly affected by task instruction, individual attentional resources, prior knowledge, and personal motivations. Chu *et al.* [23] explore the link between people, object or scene familiarity and interestingness. They conclude that familiar people are preferred for faces, whereas novel objects or places are the right indicators in natural scenes, challenging the universal nature of interestingness.

As a conclusion, the presented works are dealing with the topic of interestingness at either a local or a global scale. None did try to state on both notions together. Moreover, they do not agree on which criteria are significantly related to the elicitation of interest. A questionable point is also about the universality of the interestingness concept. In this paper, we attempt to go one step further on all these subjects by proposing and interpreting a new set of experiments.

Experiments

To investigate the two notions of intra and inter-interestingness and also potential links with some other criteria,

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Figure 1. *Transportation dataset - 49 images, 900x600 pixels each. Red boxes correspond to the '25 images' subset used in some of the experiments.*

we conducted several experiments, with different protocols, to assess the reproducibility. Because we expect to draw conclusions about potential correlations between these two notions, we also propose a new dataset which will serve as a common basis for all experiments.

Table 1 summarizes all the experiments that are described in details below. We invite the reader to refer to this table to ease the understanding of our approach. Experiments A, B and C share the full image dataset and focus only on interestingness (either intra or inter) as single criterion, but address different sources of populations and protocols. Experiments D_{PWC} and D_{TIR} are more dedicated to pointing out links that may exist with six other criteria: aesthetics, emotion, unusualness, complexity, comprehensibility and information, i.e., the capability of an image to deliver a message. They were simultaneously conducted with the same population using two different protocols and numbers of images. All protocols were designed jointly by four experts in the field.

Table 1 - Summary of the different experiments conducted for the intra- and inter-interestingness assessement. Second column gives the targetted interestingness notion (intra / inter) and the number of tested criteria (mono: interestingness only / multi: all 7 criteria).

Expe.	Type, Crit.	Protocols	Popul.	Image nb.
A	Intra, Mono	Ranking of at most 5 words from a list of 10.	15	49
B	Inter, Mono	Pair-wise comparison (PWC)	34	49
C	Intra, Inter, Mono	Eye-tracking condition. Intra task: Mention the first five most interesting elements. Inter task: Decide on whether each picture is interesting.	10+10	49
D_{PWC}	Inter, Multi	Pair-wise comparison (PWC)	12 to 15	25
D_{TIR}	Inter, Multi	Ten Images Ranking (TIR)	51	10

Transportation dataset

Figure 1 shows the Transportation dataset: 49 natural images in the field of transportation. Each image of size 900x600 pixels, contains several objects and concepts (at least 5). More precisely, the vehicle type is diversified and the indoor or outdoor scenes take place in cities or in the countryside. A six experts team was involved in the picture selection. This may reduce subjective aspects in the dataset creation.

In order to study the links that may exist between the six criteria cited above and interestingness, we try to balance as much as possible the dataset characteristics in terms of each of them. This selection remains purely qualitative, no specific model is used to eventually quantify the aforementioned features.

On Figure 1, images with red boxes belong to a 25 images subset which were used for some of the conducted experiments.

Experiment A

This first experiment involves 15 naive observers (10 males, 5 females, age average = 40.6, $stdev = 9.9$), and focuses on intra-interestingness. Individual observer is shown successively each of the Transportation dataset's image (the observer's distance to the screen being 2-3 times the screen height, experiment is conducted in a dark room, free of noise). Each image is displayed for 15 seconds. Observers have then 20 seconds to identify and rank up to 5 words out of a predefined list of 10 to answer the question : "According to you, which elements make this image interesting?". Word lists have been chosen by the same team of six experts that created the dataset. The per-image word list was created randomly once and then proposed to each participant, with the only constraint of keeping concepts all together at the end of the list. For each image, the proposed list contains mostly objects (7 in average), while average number per image is 1 for living beings and 2 for concepts. Table 2 gives statistics about the lists' composition. 1 to 4 concepts are proposed for 96% of the images; 1 to 2 living beings are proposed in 78% of the images and each list contains from 4 to 9 objects.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Table 2 - Statistics on the proposed image word lists. Each line gives respectively the percentage of images in the dataset with N objects, N living beings, or N concepts.

Nb of	Objects	Living beings	Concepts
0	-	22%	5%
1	-	61%	16%
2	-	17%	46%
3	-	-	26%
4	2%	-	7%
5	4%	-	-
6	22%	-	-
7	46%	-	-
8	21%	-	-
9	5%	-	-
10	-	-	-

Experiment B

34 naive observers are involved, through a pair-wise comparison (PWC) protocol. Observers were nearly equally balanced between males (19) and females (15); their age average is 38.7 ($stdev = 14.4$). Again, observers are situated at the distance of 2-3 times the screen height. For half of them, the experiment was conducted in a dark room, with no environmental noise. The other half conducted the experiment in a familiar environment (e.g., their own place). This difference of (un)controlled protocol did not show any influence on the experimental results afterward.

To avoid fatigue, each observer is only shown a subset of 294 pairs among the $C_{49}^2 = 1176$ possible pairs, as described in [19]. For each pair, observers have to answer the question "Which of these two images is the most interesting?". Thus, only the interestingness criterion is tested in this experiment. Image pairs, extracted from the Transportation dataset, are displayed for 8 seconds, but no time limit is fixed to record the observers' votes.

From the resulting pair comparisons, a ranking of the images is obtained thanks to the Bradley-Terry-Luce (BTL) model [28].

Experiment C

We conducted an eye-tracking campaign composed of three different tasks. 30 participants (20 males, 10 females, age average = 35.2, $stdev = 10.2$) observed the same 49 pictures and were confronted to one specific task. This latter goal is to mimic the behavior of a visual analysis when assessing the local or global aspects of interestingness. For each task, the participants were located in a dark room, free of noise. Before each trial, the subject's head was correctly positioned so that his/her chin pressed on a chin-rest. The SMI RED IView X system with a 50 Hertz sampling has been employed. Each subject watched randomly the complete dataset: the presentation time was 10 seconds for each image and a grey/neutral image with a randomly located cross was presented for 2 seconds between each image to minimize a potential centered bias. Participants were also informed that questions can be asked after the presentation of a group of stimuli. The questions were randomized and only asked in order to keep the subject concentrated on the task.

We divided all participants over the three tasks. 10 participants were recorded in a free viewing condition, then they were instructed to explore freely the pictures. 10 others had to rank at most 5 interesting elements on their own (no predefined list was proposed) (Experiment C-Intra) and the last 10 should decide on whether the picture is interesting or not (Experiment C-Inter). Note that we are considering on this paper only the users' answers (not the eye fixations) of these two latter tasks, that were collected manually in parallel of the eye movements.

Experiments D

Two complementary experiments aim at identifying the causes while inferring interestingness. Hence, they were conducted after a first analysis of the results of experiment B. 25 images only were chosen from the transportation dataset, again with a view of balancing as much as possible representatives from the 7 criteria but also with the constraint to be distributed as much as possible within the ranking resulting of experiment B. This choice was done empirically by the authors. Resulting 25 images are shown with a red box in Figure 1.

For these experiments a same population of 51 observers was tested, with following statistics: 34 males, 17 females, age average = 38.0, $stdev = 11.0$. Viewing conditions were the same as for the experiment B. This time though, all seven criteria were proposed for comparison.

These two experiments differ in their protocol. While first experiment, D_{PWC}, follows a Pair-Wise Comparison (PWC) protocol, as in experiment B, second experiment, D_{TIR}, is based on a Ten Images Ranking (TIR) protocol. Each tested person is asked to successively follow one first experiment of type D_{PWC}, 7 consecutive experiments of type D_{TIR}, and a last experiment of type D_{PWC}, as described below.

- Experiment D_{PWC}: Each user involved in this experiment follows the exact pair-wise protocol as for experiment B, with a task related to one criterion among the 7 input criteria. For a given user, the pair-wise task was conducted twice, on two different criteria. These 2 criteria were chosen to balance as much as possible the number of responses among all 7 criteria after each user, therefore leading to a population of 12 to 15 users per criterion.
- Experiment D_{TIR}: This experiment follows an image ranking protocol. For each user, a randomly selected set of 10 images from the 25 input images, is displayed in two lines on the screen. The user is then asked to rank them in increasing order according to one criterion, e.g., for aesthetics, from the least to the most aesthetics image. Viewing conditions are identical to what is used for experiment B, but all 10 images are presented at once to the user, and no duration limitation is applied for the ranking task. Each user proceeds with the ranking of the same 10 images for each of the 7 criteria. Criteria are proposed randomly to each user and for each of them, the 10 images are laid randomly in the two lines at initialization time. Finally, all 7 criteria were assessed on different sets of 10 images by 50 different users. For the sake of comparison with the other experiments, rankings that are obtained for each person for a given criterion are converted into pair comparisons over the entire set of 25 images. BTL model is then used to obtain a global ranking of the 25 input images, for all users.

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Evaluation metrics

In this section, more details are provided about two metrics employed in the evaluation of collected data.

Inter-rater Agreement

Any experiment involving a task of data collection requires a metric to assess the consistency between annotators, also called raters or coders. Plenty of metrics (e.g., Percent Agreement, Scott's Pi, Cohen's Kappa, Fleiss's K, Cronbach's Alpha and so on) flourished in the literature during the last decades [22]. Consequently, the experimenters face a critical choice: how to pick up a relevant metric with the suitable properties adapted to their data.

In this study, we employ the Krippendorff's alpha (α) metric because it satisfies many crucial conditions to evaluate annotated data [22]. First, it is neither sensitive to the number of collected data nor to the rater's order. Second, it raises a minimum and maximum values which allow us to quantify some effects and to draw a conclusion. Thus, 1 represents a perfect agreement, while 0 reflects the absence of agreement. Third, it is applicable to different measurements, such as nominal (specific categories), ordinal (data ordering), interval and ratio data. Also, it is independent from the format of collected data (e.g., number of categories, range of data). Finally, one useful property is its ability to cope with missing samples which realistically occur in most experiments.

Interestingly, Krippendorff's α encompasses several known metrics targeting specific cases. Thus, Spearman's rank correlation coefficient ρ is equivalent to Krippendorff's α between two raters for the ordinal data case. In the same vein, Pearson's intraclass correlation coefficient matches Krippendorff's α between two raters for the interval data case. Both correspondences are valid when considering a large set of units or samples.

In its general form, Krippendorff's α can be written as:

$$\alpha = 1 - \frac{D_o}{D_c} \tag{1}$$

where the disagreement accountable to chance is expressed by D_c, while D_o is the disagreement observed among the collected data. Coincidence matrices are built to express D_o and D_c regarding pair (m,n):

$$D_o = \frac{1}{l}\sum_m\sum_n o_{mn}\delta^2_{mn} \tag{2}$$

$$D_c = \frac{1}{l(l-1)}\sum_m\sum_n l_m \cdot l_n \delta^2_{mn}, \tag{3}$$

where l is the total number of annotated pairs, such as $l = \sum_m\sum_n o_{mn}$. Then, l_m (or l_n) represents the number of times the element (often called unit) m (respectively n) is compared with all other elements. o_{mn} is the collected observation when comparing the units m and n.

Also, δ^2_{mn} stands for the metric difference which determines the data type and measured reliability (nominal, ordinal, interval, ratio). In this paper, we address only the nominal and ordinal cases, for which the corresponding metrics are computed as follow:

$$\delta^2_{nominal,mn} = \begin{cases} 0, \; if \; m = n \\ 1, \; else \end{cases} \tag{4}$$

$$\delta^2_{ordinal,mn} = \left(\sum_{g=m}^{g=n} l_g - \frac{l_m + l_n}{2}\right)^2 \tag{5}$$

For more details on inter-rater agreement and concrete examples, the reader may refer to [21, 22].

Even though the author mentioned that *there are no magical numbers*, he recommended $\alpha \geq 0.8$ to ensure the data reliability and conceded a potential agreement if $\alpha \geq 0.66$ [21]. However, the design of the metric penalizes a lot the agreement as soon as one observer deviates from the others. As an example, let us define a pair-wise comparison experiment during which ten raters have expressed their opinion. If only one observer over ten disagrees, α drops from 1 to 0.6 (when 1 corresponds to full agreement). Knowing the subjectivity associated to our task, we argue that it may be acceptable to have lower Krippendorff's alpha values than the known cutoffs.

Spain and Perona

Interestingly, one can, in a straightforward way, derive an estimation of the Krippendorf alpha agreement metric from the probabilistic model of Spain and Perona, which was mentioned as a key contribution for the intra-interestingness assessment in [10] and [11]. This is what we establish in this section.

Spain and Perona define an object's importance in a particular image as the probability that a human observer naming objects will name it first. By modeling the process that generates an observer's sequence, they enable the measure of an object importance from relatively few observers. The process of drawing balls with different sizes from an urn without replacement mimics the process of naming objects in an image. The size of the balls are the parameters and model the objects' importance. Maximum likelihood estimation is used to fit the model to a set of object sequences obtained from human observers' annotations.

Still fitting to the analogy with the urn model, the probability of the n^{th} word in the list to be cited at rank r+1, $p(n, r+1)$, can be derived from the Spain and Perona importance $p(n,1)$ in the following way:

$$p(n, r+1) = \sum_{\sigma \in S_r^{W_n}} p(\sigma)\frac{p(n,1)}{\sum_{j \notin \sigma} p(j,1)} \tag{6}$$

where $S_r^{W_n}$ is the set of all ordered sequences containing r words but not the word W_n. If the sequence σ contains r words indexed by i_j, $p(\sigma)$ is given by:

$$p(\sigma) = \frac{\prod_{k=1}^{r} p(i_k,1)}{\prod_{l=1}^{r-1}(1 - \sum_{k=1}^{l} p(i_k,1))} \tag{7}$$

From this, we can estimate the coefficients of the reliability and coincidence matrices detailed by Krippendorf in [20]. By considering C independent ordered sequences obtained from the urn model with the above parameters, we match the case of a complete reliability matrix (no missing data) for C observers where the units are the balls. For a given sequence, the code for a given ball is the rank at which the ball has been drawn. Each code occurs exactly once per sequence and every unit is valued by exactly C observers. The probability $p(n,k,r)$ to obtain exactly k times the ball numbered n at rank r is given by the binomial law:

$$p(n,k,r) = \binom{C}{k} p(n,r)^k (1 - p(n,r))^{C-k} \tag{8}$$

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Hence, coefficients o_{nn} located on the diagonal of the coincidence matrix are given by:

$$o_{nn} = \frac{1}{C-1}\sum_r\sum_k p(n,r,k)k(k-1) \quad (9)$$

We can also compute the probability $p(n1,k1,n2,k2,r)$ to get exactly k_1 times the ball n_1 and k_2 times the ball n_2 at rank r:

$$p(n_1,k_1,n_2,k_2,r) = \binom{C}{k_1}\binom{C-k_1}{k_2} p(n_1,r)^{k_1} p(n_2,r)^{k_2}(1-p(n_1,r)-p(n_2,r))^{C-k_1-k_2} \quad (10)$$

The coefficient $o_{n_1n_2}$ is obtained by:

$$o_{n_1n_2} = \frac{1}{C-1}\sum_r\sum_{k_1}\sum_{k_2} p(n_1,k_1,n_2,k_2,r)k_1k_2 \quad (11)$$

From the general form given in [20] and considering the ordinal metric, we obtain the following estimation of the Krippendorf's alpha coefficient:

$$\alpha = 1-(NC-1)\frac{\sum_c\sum_k o_{ck}(k-c)^2}{\sum_c\sum_k (k-c)^2} \quad (12)$$

where N stands for the total number of words, leading to the conclusion that, from the Spain and Perona probabilistic model, it is possible to derive an estimation of the Krippendorf alpha metric.

Correlation and causality

In order to study the relationship that may exist between the data of our experiments, we provide the $R-square$ or R^2 measurement, while fitting a regression line (or any other more adapted model) between the data of the different experiments. Interestingly, R^2 provides an indication about how well a simple relationship can be found between two independent data sets. However, if it does testify of a link, this is not necessarily a link of causality.

The higher the R^2 (maximum is 1), the better the model fits and the better the correlation. The R^2 values must nevertheless be assessed in accordance with the number of samples. Indeed, it may happen that for a large sample set, the correlation ρ between two variables is equal to 0, while for a smaller subset it exhibits a high correlation. The significance of R^2 values from observed samples has to be checked in order to test the null hypothesis, i.e., the hypothesis that the observed value comes from a population for which $\rho = 0$. The level of significance test (above which we reject the null hypothesis) is set to 5% in this paper, meaning that the measured R^2 values have less than 5% likelihood of occurring by chance. Table 3 recalls, for different numbers of samples, corresponding cutoffs for significant values of R^2.

Results

In this section, we first present our experimental results for each of the two main notions independently, then jointly. For the reasons described previously, we employ the Krippendorff's alpha reliability (KAR) [21] as agreement metric. This will allow to compare agreement levels from the different tasks.

Table 3 - Significant values of R^2 for a level of significance (p-value) of 5%.

Samples#	R^2
10	≥ 0.30
25	≥ 0.11
49	≥ 0.05

Intra-interestingness notion

Experiment A

In this experiment, 3040 words out of 3675 possible at most were ranked by 15 observers, representing about 4 words per image per observer. Table 4 presents statistics on the ranked-words categories (objects, living beings and concepts) in experiment A. These figures confirm that living beings attract viewers' attention more than objects and indicate a preference for concepts to point out interestingness. This is clearly emphasized when focusing only on top ranked words.

For the 49 images contained in the dataset we computed two values of the Krippendorff's alpha coefficient. The first one, directly from the 15 rankings and the second one derived from the words' importance obtained by using the Spain and Perona model. A high correlation value is obtained between those two results as shown on Figure 2. From the collected data of Experiment A, we therefore establish empirically the validity of the theoretical demonstration developed in section *Spain and Perona*.

Table 4 - Intra A - Statistics on ranked word categories.

	Words in database	All ranked words	Words in Top 3	Words ranked 1st
Number	490 = 49 x 10	3040	2173	735 = 49 x 15
Object	70%	63%	60%	51%
Concepts	21%	26%	29%	39%
Living beings	9%	11%	11%	10%

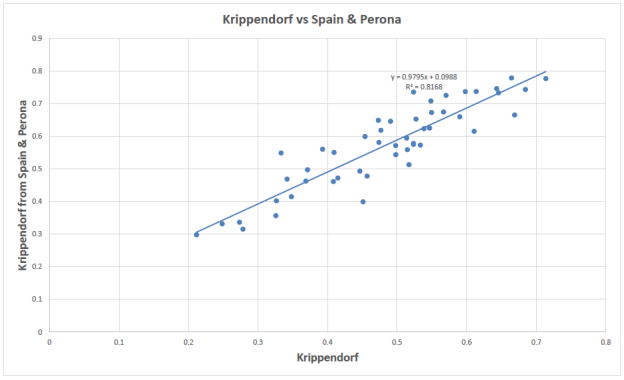


Figure 2. *High correlation between the Krippendorf's alpha coefficient directly computed from data and the one estimated from the Spain & Perona model.*

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Experiments A and C-Intra

For the word ranking task, in both experiments A and C-Intra, we computed the importance of all words (490) according to the probabilistic model proposed by Spain and Perona in [10] from the collected data. Since in experiment C-Intra, there was no list of preselected words, we discarded the cited words that were not in the predefined word list of experiment A. Globally they represent 15% of the cited words. Among all first words proposed, none is discarded, and only 3 were discarded among 343 at the second rank. Hence, this reinforces the chosen list proposed by the expert annotators in experiment A, and it also ensures a possible comparison between experiments A and C-Intra, as, at least for the first two words, we came up with quite the same word list from the observers themselves.

These levels of importance are strongly correlated ($R^2 = 0.46$), advocating for a weak dependence to the protocols in the case of the word ranking task.

Inter-interestingness notion

Experiment B

Thanks to BTL modeling, we end up with an interestingness ranking of all images (see Figure 3), which shows promising qualitative characteristics: most interesting images seem to be mostly aesthetics and/or unusual, whereas images drawing negative emotions seem to be pushed to the bottom of the scale. These observations are partly confirmed later with complementary experiments and objective statistical data.

Nevertheless, inter-observers' agreement per image, computed using the Krippendorff's Alpha-Reliability, shows small values as illustrated in Figure 4. As a conclusion, three scenarii can be envisioned: either the universality of inter-interestingness can be questioned, the task protocol should be reviewed, or the chosen metric (KAR) is not applicable. However, agreement values seem to be higher at both ends of the interestingness scale than for middle values, as confirmed by $R^2 = 0.29$ when fitting a polynomial curve to the data, which is intuitively understandable: users agree more on what is most or least interesting.

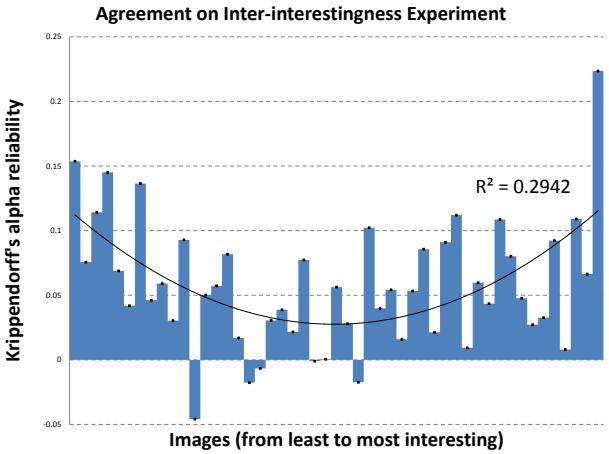


Figure 4. *Experiment B - Inter-observers' agreement (agreement values are ranked according to interestingness).*

Experiments B and C-Inter

In this section, we aim at finding a correlation between the data collected in experiment B and in experiment C-Inter. While a global ranking of all pictures on a scale is obtained in experiment B, binary answers are aggregated in experiment C-Inter (interesting $= 1$; not interesting $= -1$) for each participant and each picture. Note that the two experiments do not have not a similar task, in the sense that experiment B is a comparison task, while in experiment C-Inter, an absolute assessment of interestingness is collected. However, the global notion of inter-interestingness is caught by both experiments.

For the sake of comparison, we need to derive a ranking with all pictures from experiment C-Inter. In Figure 4, we demonstrated that the inter-rater agreement is encoded along the ranking scale, meaning that the agreement appeared higher for the most and least interesting pictures. We assume the same behavior for experiment C-Inter and translate the binary answers into a scale dependent on the data agreement. To do so, we derive this scale s as the ratio of collected answers regarding the interestingness of picture i, expressed as $s_i = \frac{1}{N}(2n_i - N)$, where N is the total number of observers and n_i the number of observers answering positively to picture i's interestingness.

We compare the obtained ranking to the one derived from experiment B. Interestingly, global image interestingness values measured from both experiments exhibit a significant correlation, with a R^2 value of 0.36.

However, per-picture agreements of both experiments show no correlation. We also note that KAR values are globally higher for experiment C-Inter than for experiment B (Figure 5).

Thus, we conclude that experiment C-Inter reproduces experiment B's ranking which can be then considered meaningful.

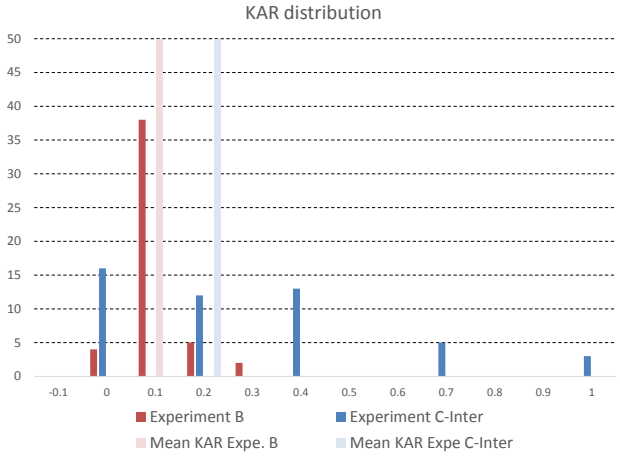


Figure 5. *Experiments B and C-Inter - Distribution of KAR values for all pictures for both experiments.*

Experiment D_{PWC}

Similarly to what was proposed in Figure 4 for experiment B, we were able to fit a polynomial curve, between inter-observers agreement and ranking for each of the 7 criteria. For 7 criteria out of 7, obtained R^2 values (see Table 5) are significative enough to demonstrate that users agree more on both ends of the rankings than on middle range, which is, once again, intuitively un-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Figure 3. *Experiment B - Ranking of images per interestingness: from lowest (left-hand side) to highest (right-hand side) values. Right end of the scale: mostly aesthetics and unusual images. Left end of the scale: mostly images with negative emotions.*

derstandable. Nevertheless, as for experiment B, the small values (mostly under 0.3) of inter-observers' agreements whatever the criterion should also be noted.

Table 5 - Experiment D_{PWC} - R^2 values obtained when fitting a polynomial curve between inter-observers' agreement and ranking for each of the 7 criteria.

	R^2
Interestingness	0.2574
Unusualness	0.6572
Emotion	0.2741
Aesthetics	0.7349
Information	0.4034
Complexity	0.1912
Comprehensibility	0.515

In order to state on potential links between the 6 additional criteria and interestingness, we also computed the correlations between their rankings and the interestingness ranking. All R^2 values are proposed in first column of Table 6. Unusualness, emotion and, to a lesser extent, aesthetics, appear to be correlated with interestingness.

Table 6 - Experiments D_{PWC} and D_{TIR} - R^2 values for correlations between rankings of interestingness and of each other criterion.

Interestingness and ...	Exp. D_{PWC}	Exp. D_{TIR}
...Unusualness	**0.308**	**0.6461**
...Emotion	**0.2365**	**0.6433**
...Aesthetics	**0.1594**	**0.1301**
...Information	0.064	**0.2177**
...Complexity	0.0624	**0.2772**
...Comprehensibility	0.0086	**0.2827**

Experiment D_{TIR}

Again, we computed correlations between interestingness and the 6 other criteria and show resulting R^2 values in second column of Table 6. Compared to experiment D_{PWC}, this exhibits higher correlations for all criteria. Unusualness and emotion are especially well correlated, and to a lesser extent, so are information, complexity and comprehensibility.The comprehensibility curve, not shown in this paper, also exhibits some interesting negative correlation.

Experiments D_{PWC} and D_{TIR}

We present in Table 7, the correlations that exist between experiments D_{PWC} and D_{TIR}, for all 7 criteria. In all cases, despite the involvement of the same population, high R^2 values are obtained, voting for an equivalence of results between two different protocols (pair-wise comparisons and image rankings), whatever the criterion. Furthermore, two different groups of features can be isolated: Group 1: unusualness, aesthetics, emotion, informativeness, interestingness and complexity ($R^2 > 0.6$) and Group 2: comprehensibility (R^2 0.16).

To reinforce this result, we also computed the Spearman coefficients *rho* for each criterion either globally (see column 2 of Table 7) or per user. For the later, for a given criterion, a *rho* value per user was computed between the corresponding ranking from experiment D_{PWC} and the ranking given by chosen user in experiment D_{TIR}. In these distributions (see Figure 6), peaks for high *rho* values mean that most of the users' rankings are highly correlated with the ranking obtained by pair-wise comparisons. It should be noted that the criteria can be grouped in the same two groups as what we obtained through ranking correlations.

Experiments B, D_{PWC} and D_{TIR}

Going further to what was presented in previous paragraph, we concentrated on interestingness only and computed potential correlations between all experiments: B, D_{PWC} and D_{TIR}.

Table 8 shows the obtained R^2 values between interestingness rankings for all 3 experiments. Overall, all conducted experiments show interesting correlations, as R^2 values are high enough to be significantly representative. Higher correlations are obtained

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

Table 7 - Experiments D_{PWC} and D_{TIR} - Correlations between both rankings and Spearman coefficients for each of the 7 criteria.

	R^2	Spearman ρ
Interestingness	0.69	0.79
Unusualness	0.85	0.92
Emotion	0.81	0.87
Aesthetics	0.82	0.95
Information	0.91	0.92
Complexity	0.67	0.85
Comprehensibility	0.17	0.24

either for identical protocols (experiments B and D_{PWC}) or for the same population (experiments D_{PWC} and D_{TIR}).

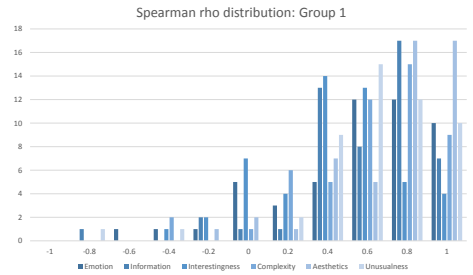
Intra- and Inter-interestingnesses

In this section, we attempt to find some link or common mechanisms between the intra- and inter-interestingness experiments.

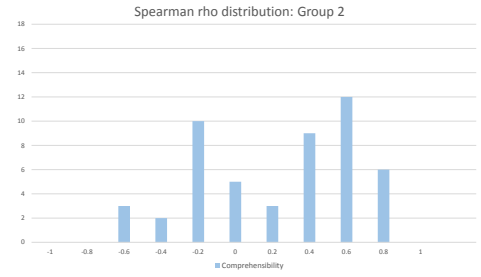
Agreement

Since these two types of experiments produce incomparable outputs, we focus on the per-picture agreement and composition (proportion of objects, living beings and concepts).

First, we state on agreement values for both the intra and inter tasks and observe that images reaching high, resp. low, agreements in each task are not the same. In other words, no correlation appears between the distribution of agreements collected for each experiment A, and B or D_{PWC}. It seems that the mechanisms involved for assessing the global and local interest of a picture



(a)



(b)

Figure 6. *Experiments D_{PWC} and D_{TIR} - Spearman rho distributions for all criteria between both experiments, following the groups' notations elaborated on ranking correlation.*

Table 8 - Experiments B, D_{PWC} and D_{TIR} - Correlations of interestingness rankings.

	Exp. D_{PWC}	Exp. D_{TIR}
Exp. B	0.5769	0.5101
Exp. D_{PWC}	na	0.6933

differ and conduct to different agreements. Thus, a picture with a low agreement in the intra-interestingness task may lead to a high consensus for the inter-interestingness task.

Second, we investigate the proportion of objects, living beings and concepts as a factor explaining the strong agreement of some pictures. We did not find any evidence of relationship between them, whatever the experiment.

Link with Spain & Perona

For every image in our dataset, and for Intra A, we observe a very good fitting of the distribution of objects' importance obtained with the Spain & Perona modelization to a single parameter exponential model in the log-log domain. On the dataset, the average of the squared Pearson coefficient R^2 is higher than 0.88 with a minimum value at 0.66.

Comparing the image ranks obtained in experiment D_{PWC} with the parameters of the exponential model, we did not success in pointing out any significant correlation level. When focusing more on the particular importance of the three word categories, objects, concepts and living beings in every image, we did not observe any particular distribution correlated with the image rank.

In other words, if concepts and living beings seem to be the most interesting elements within an image, they do not seem to impact the interestingness of the image when it is compared to some others.

Discussion

This paper presented an original work on the concept of interestingness within and between pictures. First, we introduced the concept of intra-interestingness (local or within) and of inter-interestingness (global or between). Second, we designed a dedicated dataset employed in all experiments. Third, we designed, set up and discussed five complementary experiments. In the following discussion, several key aspects are developed: from the data reproducibility to the concept universality via the potential underlying or correlated criteria related to interestingness.

Linking the local and global aspects of interestingness

This first investigation of the intra- and inter-interestingness on the same dataset did not conduct to find any evidence of a relationship between these two notions. The task of intra-interestingness conducted to a higher agreement between annotators than the one of inter-interestingness. We could not find any correlation between the two tasks, even when considering subgroups of pictures with specific features (repartition of concepts/living beings/objects), top ranked words and so on.

Also, the inter-interestingness final ranking cannot be linked to the Spain & Perona distributions, obtained from the word rankings in the intra-interestingness experiment.

As a conclusion, we were not able to draw a picture for mod-

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

eling conjointly the two notions. Finding a link would have allowed to design a two-stage model, which could benefit from one notion to enrich the modeling of the second one. However, our experiments do not allow to state on the fact that the brain mechanisms for each task correlate and require a common architecture. This point clearly needs more investigations.

Reproducibility of interestingness tasks

Intra aspect

By performing twice the task of intra-interestingness with two different protocols, we aim at finding correlated results. While in experiment A the subjects had to select the first five most interesting elements within a list of ten proposed words, another group of participants were placed in a context of eye movement measurement in experiment C with a similar task of citing the five most interesting elements but without a predefined list. By applying the model of Spain and Perona, we could estimate the importance of each word with respect to the original list.

We evidence a significant correlation between the two experiments. Since the same word ranking is reproduced in two different protocols, we do believe the intra-interestingness task is doable, the concept is understood by naive subjects. Consequently, the agreement is strong in that sense.

Inter aspect

Regarding the global aspect of interestingness, we set up three different protocols to derive a potential correlation between the picture rankings. Hence, a pair-wise protocol (Experiments B and D_{PWC}), an eye-tracking protocol with a task about the global interestingness nature of viewed pictures (Experiment C-Inter) and a simultaneous ranking of 10 pictures (Experiment D_{TIR}) are used.

In all cases, we clearly evidence important correlations between the three resulting rankings for interestingness, as stated in the *Results* section. These correlations may nonetheless be somewhat debated by a few arguments related to the experiments conditions.

Hence, considering the experiments B and C-Inter, the ranking scores of C-Inter are not directly collected but derive from the agreement values per picture. This is potentially questionable in terms of methodology. Also, a smaller correlation is obtained between experiments B and D_{TIR}, for which both the population and protocol are different. Indeed, experiments B and D_{PWC} share the same protocol, whereas experiments D_{PWC} and D_{TIR} share the same population of observers, and both reach higher correlation values. Nevertheless, all these correlation values are still meaningful, even for a significant demanding level (p-value ≤ 0.001).

Considering again the experiments B and C-Inter, we evidence a higher agreement for the absolute rating of interestingness versus the ranking through pair comparison. Thus, we conclude that it is surprisingly easier and more universal to assess the global interestingness, absolutely and without a reference, than by comparison. Indeed, the pair-wise comparison protocol is known in self-reporting to ease the observers' assessment, since making a decision by preference is simpler than absolute rating [29]. One explanation could be the long experiment duration and associated lassitude in the case of experiment B.

Consistent results are obtained for the 6 additional tested criteria when comparing rankings for experiments D_{PWC} and D_{TIR}, reinforcing our conclusion that whatever the protocol, and to a lesser extent whatever the population, we were able to apprehend the global characteristics of such criteria.

As mentioned for intra-interestingness, this proves the validity of measuring and further modeling such criteria. Once again, the reproducibility of results legitimates the strength of the observed agreements, despite their moderate value.

Correlated criteria

Experiments D_{PWC} and D_{TIR} aim at quantifying the role of other criteria in the assessment of interestingness. Some of them are well studied in the literature, e.g., emotion, aesthetics, complexity, unusualness while others remain unexplored in terms of annotation and modeling, e.g., comprehensibility, information.

Experiments D_{PWC} and D_{TIR} both exhibit correlations between interestingness and the 6 other studied criteria. Both unusualness and emotion happen to be highly correlated to interestingness, whatever the experiment D_{PWC} and D_{TIR}. Experiment D_{TIR} also shows correlations for the other 4 criteria, which in decreasing order are comprehensibility, complexity, information and aesthetics. These results reinforce what could have been intuitively deduced and what was sometimes also said in the literature [14], at least for unusualness and emotion. It is also interesting to note that for both experiments, comprehensibility was inversely correlated to interestingness, driving the conclusion that an easily understandable image tends to be less interesting, contrary to what was said in [5, 6].

Apart from aesthetics, correlations with the 3 other criteria, comprehensibility, complexity and information were not present in experiment D_{PWC}. We may wonder if these differences do come from the bigger size of population for experiment D_{TIR}, or from the protocol. It may also reflect the difficulty that people had to assess some of the criteria. For instance, complexity was apparently not well understood by observers during experiments, if we consider their spontaneous questions on how complexity was defined. Another reason of this difference may also be the subjective interpretation of each criterion depending on the observers.

In any case, these correlations are a first answer towards a feasible modeling of the interestingness notion, although correlation does not necessarily means causality.

Agreement and universality

This section aims at discussing the universality of interestingness. Some of our first results highlighting strong correlations between interestingness rankings for different protocols and populations clearly allow to establish, to a certain extent, the existence of interestingness universality. In our experiments, we further showed that intra and inter interestingness notions exist and are understood by people. They can also be measured on average. Of course, additional experiments with totally different populations, from different cultural backgrounds for example, should be conducted, to verify that universality goes beyond cultural differences. This last point may highlight differences, especially if unusualness causes interestingness, as we tended to conclude. Indeed, what is unusual for some population may be perceived differently for some other population; this is the cultural heritage (e.g., 2^{nd} row, 6^{th} image in Figure 1, Asian train).

Nevertheless, our conclusion that extrem pictures get more agreement than neutral or middle pictures in the interestingness

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

scale is also leveraging interestingness universality at least for some categories of pictures. Moreover, one should look at the low values of agreements reached by our experiments against what is generally obtained in the literature for some other criteria such as aesthetics, emotions [30]. For these criteria, agreement values are in the same range, although it has been proven that modeling was still a difficult task, but feasible. This tends to vote for a possible modeling of interestingness for an average observer, with the underlying idea that this notion reaches at least a certain degree of universality.

This paper also brings a highlight on inter-rater agreement metrics. While these latter have been discussed in the literature and argued to be universal or rather well designed for some specific cases, most of them suffer from proposing a meaningful scale, leaving any experimenters into deep consideration about its exploitation. This is reinforced when considering subjective tasks. In this paper, we favor the reproducibility of experiments as a major argument to the agreement and provide the related scores from the Krippendorff metric. Most of these computed values were low and very far away from 0.66, such as arbitrarily recommended by Krippendorff [21]. Nevertheless, this paper set up some subjective and difficult tasks and could serve as proposing reference values for future studies.

Future directions

Regarding the observations on the distribution of agreements as a function of ranking, we recommend to restrict any model's outputs to three categories/classes, i.e., low, neutral and high interestingness levels. This would encompass the potential uncertainty on annotations and seems more realistic to ensure exploitable outputs from the model.

In terms of protocols, we evaluated four of them to measure the global interestingness of a picture. It seems that they are all reliable in the sense that they provide meaningfull rankings. However, we could advice to employ an absolute rating of interestingness, since this protocol provides a higher agreement (in our case) than pair-wise comparison and allows the annotation of a larger amount of stimuli at constant duration. Thus, it paves the way to the very popular large scale annotation process through crowdsourcing.

Beside the question of protocols for collecting data, the employed stimuli remain a tricky aspect of this kind of study. As a strategy to create our dataset, we purposely restrict our study to a specific category of pictures, but without controlling finely the low-level stimuli features (spatial frequencies, color...) and their semantic composition. This could be a way to refine the dataset or create a new one more objectively.

Finally, since we did not find any evidence of links between the local and global aspects of interestingness with the presented data, we plan to later exploit the eye movements collected for these two tasks to attempt to draw a conclusion about this point.

Conclusion

This paper exhibits several novelties compared to previous works on interestingness. First, we refine the state-of-the-art attempts by introducing two concepts: intra- and inter-interestingness. Following this formulation, we propose four experimental protocols: 1/ object ranking with a pre-defined word list, 2/ pair-wise comparison, 3/ eye-tracking recording under three conditions and 4/ image ranking. Some of them have never been addressed within the topic of interestingness.

We take advantage of experimenting on the same dataset to draw new crossed links between the local (intra) and global (inter) notions of interestingness and their characteristics. In addition, we bring deep insights related to the criteria responsible for assessing the picture's interest and go further towards the understanding of the universal character of interestingness.

The introduction of categories of observers could be further explored through additional experiments on user-related and context-related interestingness. Our conclusions will serve as a necessary and strong basis for a future computational modeling of content interestingness.

References

[1] Daniel E. Berlyne, Conflict, arousal and curiosity, McGraw-Hill, 1960.

[2] Richard S. Lazarus, Emotion and adaptation, New York: Oxford University Press. 1991.

[3] Paul J. Silvia, Appraisal components and emotion traits: Examining the appraisal basis of trait curiosity, Cognition and Emotion, 22(1), 94-113. (2008)

[4] Paul J. Silvia, Interest The curious emotion, Current Directions in Psychological Science, 17(1), 57-60. (2008)

[5] Paul J. Silvia, What is interesting? Exploring the appraisal structure of interest, Emotion, 5(1), 89.(2005).

[6] Paul J. Silvia, Exploring the psychology of interest, Oxford University Press, 2006.

[7] Luis Von Ahn and Laura Dabbish, Labeling images with a computer game, Proceedings of the SIGCHI conference on Human factors in computing systems, pg. 319-326. (2004)

[8] Lior Elazary and Laurent Itti, Interesting objects are visually salient, Journal of vision, 8(3), 3,(2008).

[9] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy and William T. Freeman, LabelMe: a database and web-based tool for image annotation, International journal of computer vision, 77(1-3), pg. 157-173, (2008)

[10] Merrielle Spain and Pietro Perona, Some objects are more equal than others: Measuring and predicting importance, ECCV 2008 (pp. 523-536). Springer Berlin Heidelberg (2008)

[11] Merrielle Spain and Pietro Perona, Measuring and predicting object importance, International Journal of Computer Vision 91.1 , pg. 59-76. (2011)

[12] Helmut Grabner, Fabian Nater, Michel Druey and Luc Van Gool, Visual interestingness in image sequences, Proceedings of the 21st ACM international conference on Multimedia, pg. 1017-1026, (2013)

[13] Alexander C. Berg, Tamara L. Berg, Hal Daume III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch and K. Yamaguchi, Understanding and predicting importance in images, Computer Vision and Pattern Recognition (CVPR), pg. 3562-3569, (2012).

[14] Michael Gygli, Herbert Grabner, Hayko Riemenschneider, Fabian Nater and Luc Van Gool, The interestingness of images, ICCV, pg. 1633-1640, (2013).

[15] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng and Hanfang Yang, Understanding and Predicting Interestingness of Videos, in AAAI, Vol. 1, No. 1, pg. 2, (2013).

[16] Irving Biederman and Edward Vessel, Perceptual Pleasure and the Brain A novel theory explains why the brain craves information and seeks it through the senses, American scientist, 94(3), pg. 247-253

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-139

(2006).

[17] Craig A. Smith and Phoebe C. Ellsworth, Patterns of cognitive appraisal in emotion, Journal of Personality and Social Psychology, 48(4), pg. 813, (1985).

[18] Ang Chen, Paul W. Darst and Robert P. Pangrazi, An examination of situational interest and its sources in physical education, British Journal of Educational Psychology, 71(3), pg. 383-400, (2001).

[19] Jing Li, Marcus Barkowsky and Patrick Le Callet, Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. SPIE Electronic Imaging, Stereoscopic Displays and Applications, Human Factors, 8648, pg.1-12, (2013).

[20] Klaus Krippendorff, Computing Krippendorf's Alpha-Reliability, Retrieved from $http://repository.upenn.edu/asc_papers/43$, (2011).

[21] Klaus Krippendorff, Content analysis: An introduction to its methodology, 3rd edition. Thousand Oaks, CA: Sage, 2013.

[22] Hayes, Andrew F., and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. Communication methods and measures 1.1 (2007): 77-89.

[23] Sharon Lynn Chu, Elena Fedorovskaya, Francis Quek and Jeffrey Snyder. The effect of familiarity on perceived interestingness of images. In IS&T/SPIE Electronic Imaging, HVEI, 2013.

[24] Raisa Halonen, Stina Westman and Pirkko Oittinen. Naturalness and interestingness of test images for visual quality evaluation. IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2011.

[25] Mohammad Soleymani, The quest for visual interest. In Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, 2015.

[26] Aditya Khosla, Jianxiong Xiao, Antonio Torralba and Aude Oliva. Memorability of image regions. In Advances in Neural Information Processing Systems, 2012.

[27] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In Proceedings of the international conference on Multimedia, 2010.

[28] RA Bradley, ME Terry. Rank Analysis of Incomplete Block Designs: The method of paired comparisons. Biometrika, 39 (3-4): 324-345, (1952).

[29] Yannakakis, Georgios N., and John Hallam. Ranking vs. preference: a comparative study of self-reporting. Affective computing and intelligent interaction. Springer Berlin Heidelberg, 437-446. 2011.

[30] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. Affective Computing, IEEE Transactions on 6.1 (2015): 43-55.

Author Biography

Christel Chamaret received two Master degrees from Polytech'Nantes and the University of Nantes in Electronic and Computer Engineering in 2003. She joined Technicolor in 2007 and became a senior scientist in 2013. She works in the field of human perception and contributed to the design of perceptual models (visual attention, harmony, aesthetic, emotion). She has also an expertize in user experiments (eye-tracking, side-by-side ranking). Previously, she has worked at INRIA Rennes (France) and Philips Research (The Netherlands).

Claire-Hélène Demarty graduated from Telecom ParisTech in 1994 and received a Ph.D. degree in Computer Science, Mathematical Morphology, from Mines ParisTech in 2000. She joined the Technicolor Research & Innovation Center in 2004 as a senior researcher and became senior scientist in 2010. Located in Rennes, France, she is working on scene understanding and multimedia indexing technologies. She is author or co-author of more than 20 papers and is holding several patents.

Vincent Demoulin graduated from the Institut de Formation Supérieur en Informatique et Communication (IFSIC) and received a DEA (Master) in Signal Processing and Telecommunication from Rennes I University (France) in 1994. He joined Technicolor in Rennes the same year. Since the beginning of 2010, he has been working on image processing as a senior scientist, with particular focus on local feature descriptors and image segmentation.

Gwenaëlle Marquant studied Physics and received the PhD degree in Signal Processing from Rennes University in 2000. In 2001-2002 she worked at Philips Research lab on exploratory video coding and perceptual coding. She joined Technicolor in 2003. She is actually working on scene understanding. Her expertise domain covers video compression as well as image/video processing and computer vision. She is author and co-author of several patents, contributions to the MPEG standard and papers.