Unveiling the Multimedia Unconscious: Implicit Cognitive Processes and Multimedia Content Analysis

Marco Cristani[1]
Alessandro Vinciarelli[2,3]
[1]University of Verona (Italy)
[2]University of Glasgow (UK)
marco.cristani@univr.it
vincia@dcs.gla.ac.uk

Cristina Segalin[1]
Alessandro Perina[4]
[3]Idiap Research Institute (Switzerland)
[4]Microsoft Research (USA)
cristina.segalin@univr.it
alperina@microsoft.com

ABSTRACT

One of the main findings of cognitive sciences is that automatic processes of which we are unaware shape, to a significant extent, our perception of the environment. The phenomenon applies not only to the real world, but also to multimedia data we consume every day. Whenever we look at pictures, watch a video or listen to audio recordings, our conscious attention efforts focus on the observable content, but our cognition spontaneously perceives intentions, beliefs, values, attitudes and other constructs that, while being outside of our conscious awareness, still shape our reactions and behavior. So far, multimedia technologies have neglected such a phenomenon to a large extent. This paper argues that taking into account cognitive effects is possible and it can also improve multimedia approaches. As a supporting proof-of-concept, the paper shows not only that there are visual patterns correlated with the personality traits of 300 Flickr users to a statistically significant extent, but also that the personality traits (both self-assessed and attributed by others) of those users can be inferred from the images these latter post as "favourite".

1. INTRODUCTION

Until a few years ago, production and diffusion of multimedia data required skills and infrastructure that were the privilege of a few individuals and organizations (archives, digital libraries, online repositories, etc.) [39]. Nowadays, technologies as ubiquitous and user-friendly as smartphones and tablets allow one to easily create multimedia material (pictures, videos, soundbites, text and their combinations) and share it with others - typically through social media or other online technologies - by simply pushing a button. In such a technological landscape, multimedia data is not just a way to transmit knowledge and information - as it used to be traditionally for any type of data [6, 42] - but one of the channels through which we interact with others.

The core idea of this article is that, in such an unprecedented scenario, *the exchange of multimedia data has become a form of human-human communication and, therefore, it should involve the cognitive phenomena typically observed in human-human interactions.* This applies in particular to implicit cognitive processes that take place outside our conscious awareness, but still shape to a large extent our perception of the world and our behavior [24], namely the tendency to express and attribute to others goals, values, intentions, traits, beliefs and any other type of socially relevant characteristics [50].

To have a measure of how much multimedia data have become a means of communication between people, it is sufficient to consider a few statistics available on Youtube at the moment this article is being written[1]: while uploading every day 12 years of video material, Youtube users access the popular on-line platform one trillion times per year, an average of 140 visits per person on Earth (the figure refers to 2011). In other words, there seems to be no multimedia sample produced by one person that is not consumed by someone else. Unlike a mere few years ago, creation, diffusion and sharing of multimedia data is no longer the exclusive prerogative of skilled professionals, but the everyday practice of the lay person. Multimedia data are no longer, or no longer exclusively, the carefully crafted product of creativity and communication skills, but the spontaneous expression of common individuals involved in everyday social interactions.

The problem is that our cognitive processes are the result of a long evolutionary history and cannot change at the pace of technology. Therefore, our cognition keeps following patterns that were shaped during times when technology was far from existing [33]. In particular, a large body of evidence shows that our cognition constantly works to make sense of the world around us and that this happens, to a large extent, "*effortlessly, and even unintentionally*" [49]. This means that the information we gather and process through our conscious attention - the typical realm of current multimedia technologies - is only one of the factors that drive our reactions towards the environment, the others being "*implicit, even automatic processes: implicit attitudes, inferences, goals and theories, and the affect and behaviors they produce*" [50], where the word "*implicit*" means outside our conscious awareness.

To the best of our knowledge, multimedia approaches neglected so far to a large, if not full, extent the phenomena above. Most of the current technologies take into account only information that can be automatically detected in the data (*e.g.*, objects in pictures) or inferred from it (*e.g.*, genre

[1]`http://www.youtube.com/yt/press/statistics.html`

MM'13, October 21–25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10.
http://dx.doi.org/10.1145/2502081.2502280.

from music). The few attempts to take into account implicit cognitive processes focused on observable effects, including emotional, behavioral and physiological reactions, used, *e.g.*, for retrieval [1] and tagging purposes [37]. However, such reactions might be difficult to detect, especially in settings where social norms impose behavioral limitations (*e.g.*, public spaces). Furthermore, observable reactions are nothing else than effects that follow the actual changes in the user, namely those that concern "*implicit attitudes, inferences, goals and theories*" (see above).

The possible solution to such a state of affairs is that cognitive changes behind observable reactions are not random, but tend to follow, according to the *Brunswik Lens* [5], stable and predictable patterns. The Brunswik Lens, one of the most effective models developed in cognitive psychology, provides a framework suitable for investigating how multimedia data can be adopted as an observable evidence of "attitudes, inferences, goals and theories" (see above) of data producers. Symmetrically, the model helps to explain how data consumers attribute "attitudes, inferences, goals and theories" to data producers.

This paper shows that cognitive effects are detectable at least in the case of the interplay between Flickr pictures and personality traits of Flickr users. The results, obtained over PsychoFlickr (a novel image dataset of 60,000 images posted by 300 individuals), show not only that there is a statistically significant correlation between personality traits of the users and features extracted from the images they post, but also that the same features are correlated with the traits that picture observers assign to the users, even if observers and users have never been in contact. Furthermore, both associations are sufficiently stable to be learned by supervised statistical classifiers. This opens up to a set of applications like, *e.g.*, automatic attribute prediction: given a pool of images, the goal is to infer personal characteristics of its owner. This goal is performed here by projecting images on low-dimensional manifolds and exploiting sparse regression.

The rest of this paper is organized as follows: Section2 surveys research trends relevant to this work, Section 3 describes the Brunswik Lens model, Section 4 presents the dataset used for the experiments, Section 5 reports on experimental evidence supporting the core-idea of this paper, Section 6 presents application domains that can benefit from this work and the final Section 7 draws some conclusions.

2. NEIGHBORING AREAS

The key-idea of this article is that the exchange of multimedia data has become a form of human-human communication and, therefore, it should give rise to the same cognitive phenomena (*e.g.*, see [49, 50]) typically observed in any human-human interaction. To the best of our knowledge, this article is the first attempt to adopt such a perspective in multimedia technologies. However, several domains consider neighboring issues that, while being different from the ones proposed in this article, still include aspects relevant to this work.

The application of the *sociotechnical* perspective in studying the use of digital libraries - until a few years ago the most common infrastructure for the exchange of multimedia data - is one of the earliest attempts to take into account social issues in technological applications: "*To understand, use, plan for and evaluate digital libraries, we need to attend to social practice, which we define as people's routine activities that are learned, shaped, and performed individually and together*" [39]. The main difference between the sociotechnical perspective and the research direction proposed in this work is that the former focuses on use and usability issues (especially in professional and institutional settings) while the latter targets the communication between individuals, a step made possible only by recent technologies (social media, mobile devices, etc.).

In parallel, several efforts were done to improve multimedia technologies by automatically detecting and understanding emotional, behavioral and physiological reactions of data consumers (*e.g.*, if a person watching a video laughs, then the video can be tagged as "*funny*") [1, 26, 37]. The core-idea of these trends is that the content of the data produces observable changes in data consumers, then the observation of these latter provides information about the data. The main difference with respect to this article is that the accent is on the data content, like in most of the multimedia technologies, and not on the communication process underlying the data exchange between individuals.

More recently, some works investigated the interplay between observable characteristics of multimedia data and cognition [27, 55]. The first work [27] considers images tagged as "*favourite*" by a certain person as an expression of her aesthetic preferences and shows that, given a certain amount of pictures tagged as "*favourite*" by a certain individual, it is possible to predict whether the same will happen for another picture or set of pictures. The second work [55] investigates the characteristics of abstract paintings that stimulate certain emotional reactions rather than others. Both works shift the attention from the bare content of images to their potential role in a communication process, namely a person expressing aesthetic preferences in [27] and a painter eliciting emotions in [55]. However, unlike the perspective advocated in this article, both works take into account only one of the parties involved in the communication process.

To the best of our knowledge, the only two works that seem to consider multimedia data as a form of communication are in [12, 15]. The work in [15] studies the perception of profile pictures on social media and, in particular, the agreement between the actual personality traits of profile holders and traits attributed by others based on the profile picture. The work in [12], does a similar analysis, but it considers all elements that can appear in a profile. Not surprisingly, these works focus on social media, an interaction-oriented technology that allow users to use multimedia material to communicate with others. However early, the approaches in [12, 15] seem to confirm the action of implicit cognitive processes when using multimedia data in a communication scenario, the key-idea advocated in this article. Still, both works focus on a specific case and do not try to identify the underlying perspective that can be applied to many different cases.

3. THE MULTIMEDIA LENS MODEL

This section provides a conceptual framework that illustrates the perspective proposed in this article, namely a simplified version of the *Brunswik's Lens* (see Figure 1), the model originally proposed in [5] and successively modified to investigate, among other interaction phenomena, the influence of nonverbal behavior in face-to-face interactions [44] or the judgment of rapport [3].

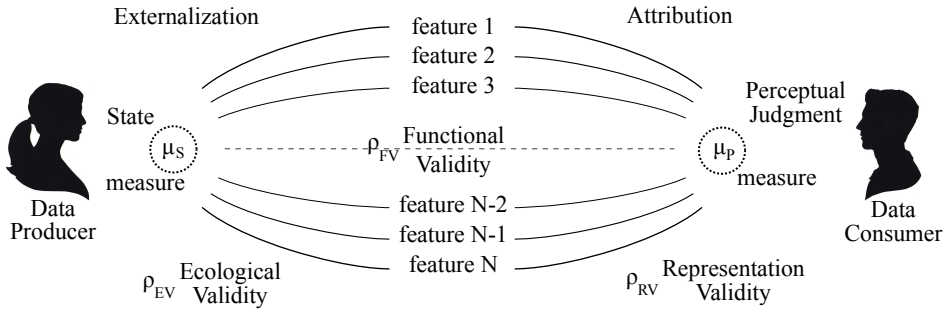

Figure 1: The picture shows a simplified version of the Brunswik Lens Model adapted to the exchange of multimedia data between a Data Producer and a Data Consumer.

In the model of Figure 1, the multimedia data is considered a form of communication between "*Data Producers*" (DP) and "*Data Consumers*" (DC). The key-idea of this article is that the process includes not only the exchange of content, a problem that the multimedia indexing and retrieval community has extensively investigated for at least two decades, but also implicit cognitive processes typical of any human-human interaction like, *e.g.*, the spontaneous attribution of socially relevant characteristics (attractiveness, trustworthiness, etc.) or the development of impressions.

The DP is always assumed to be in a certain *state* that can be either *transient* (*e.g.*, emotions, attitudes, goals, physiological conditions etc.) or *stable* (*e.g.*, personality traits, values, social status, etc.). In operational terms, the states are defined as quantitative measures (identified as μ_S in Figure 1) to be obtained via objective processes depending on the particular case under observation. For example, in the case of the social status, the measure can be the yearly income of the DP, while for the physiological condition it can be the heart rate or the galvanic skin conductance. In many cases, the states correspond to psychological constructs (*e.g.*, personality traits or interpersonal attractiveness) and the measures are the outcome of psychometric questionnaires. These latter are typically administered to the DPs and include questions associated to *Likert* scales (see Section 5 for an example).

According to the model, the multimedia data are an *externalization* of the DP state, i.e. an observable effect of it. Furthermore, the data is all the DCs know about a DP. From an operational point of view, the data correspond not only to the actual multimedia material (*e.g.*, pictures, video, soundbites, etc.), but also to any *feature* that can be extracted, manually or automatically, from the material itself. The empirical covariation of state measures and features quantifies the *ecological validity* of these latter, i.e. their effectiveness in accounting for the DP state. In Figure 1, the ecological validity is indicated with ρ_{EV} and, typically, it corresponds to the correlation or the Spearman coefficient between features and μ_S.

When the DCs consume the data, they attribute the DP a state of measure μ_P. The process is called *attribution* and μ_P is referred to as *perceptual judgment*. For example, the DCs can attribute a certain yearly income to the DP based on the pictures and videos this latter shows. In the case of the psychological constructs (see Section 5), the attribution process is typically unconscious and it takes place spontaneously, whether the DC needs it (wants it) or not [49, 50]. In principle, μ_S and μ_P should have the same value (or at least similar values), but communication processes are always noisy, especially when the communication takes place through ambiguous channels like multimedia data are. The empirical covariation between features and perceptual judgments accounts for the *representation validity* of the features (identified as ρ_{RV}), i.e. for the influence these latter have on the attribution process. Like in the case of ρ_{EV}, the most common measurements of ρ_{RV} are correlation and Spearman coefficient.

The cognitive processes this paper focuses on are active in particular at the perceptual judgment stage, when DCs unconsciously develop an impression about the DP even if all they know about this latter is the multimedia data they are consuming. However, the processes are important for the DP as well because, in a communication scenario, there is no data production without an attempt to convey an impression, i.e. to ensure that μ_S and μ_P are close to each other. The empirical covariation of μ_S and μ_P (identified as ρ_{FV} in Figure 1) accounts for the latter aspect and it is called *functional validity*. Furthermore, the full version of the Brunswik Lens [5] allows one to include contextual information. This makes it possible the integration of the technologies proposed in this article into context aware approaches [2].

4. THE PSYCHOFLICKR DATASET

We experimented the core-idea of this paper on Flickr[2], one of the most popular online photo-sharing platforms. To this purpose we collected a corpus, dubbed *PsychoFlickr*, that reflects the Lens Model and includes both pictures and personality assessments. The corpus is publicly available at [*the dataset will be made available in case of acceptance*] and was collected as follows: we contacted 300 random "*pro*" users, i.e. individuals that pay a yearly fee in order to access privileged Flickr functionalities. These users are expected to be, on average, more adept than others to photography language and techniques. For each of these 300 users, we collected the 200 latest pictures made by others they labeled as "*favourite*", for a total of 60,000 images. Furthermore, each user filled the BFI-10 (Big Five Inventory 10) [40], a personality questionnaire aimed at measuring the personality of an individual in terms of the *Big Five*, five broad dimensions shown to capture most of the individual differences [43]. The outcome of the BFI-10 is a five dimensional

[2]`www.flickr.com`.

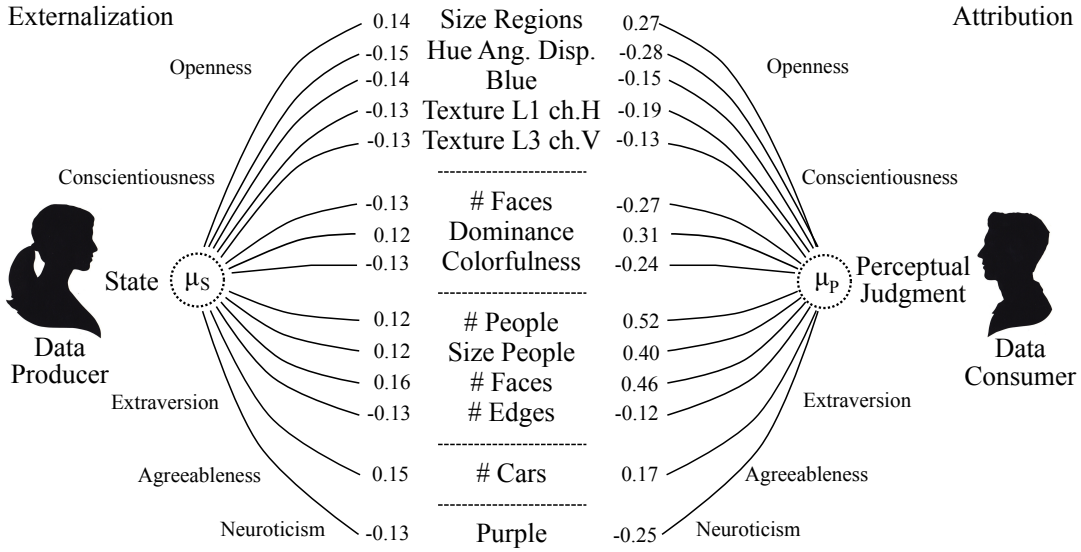

Figure 2: The picture shows the Brunswik Lens model for the PsychoFlickr dataset, where the state corresponds to the Big Five traits (as per assessed with the BFI-10). Ecological and Representation validities are measured with the Spearman Coefficient and the picture shows (for each trait) features for which both values are statistically significant ($p < 5\%$).

vector where each component measures how high an individual is with respect to each of the Big Five traits, namely *Openness* (tendency to be intellectually open, curious and have wide interests), *Conscientiousness* (tendency to be responsible, reliable and trustworthy), *Extraversion* (tendency to interact and spend time with others), *Agreeableness* (tendency to be kind, generous, etc.) and *Neuroticism* (tendency to experience the negative aspects of life, to be anxious, sensitive, etc.). In particular, for each trait, we have an integer which goes from -4 (low tendency) to 4 (high tendency).

Finally, we hired eight *assessors* that looked at the set of 200 favorite images provided by each of the users and, for each of them, filled the BFI-10 questionnaire. However, while the Flickr users adopted the self-assessment version of the BFI-10, the assessors used the other-assessment version. In other words, the users rated statements like "*I am a reserved person*", while the assessors rated statements like "*This person is reserved*", where by "person" is meant the user that has labeled the images under examination as "*favourite*". Each of the 8 assessors filled the personality questionnaires for each of the 300 users. The users and the assessors were never in contact and, furthermore, the images were the only information the assessors had at disposition about the users under exam. The 8 personality ratings produced by the different assessors about the same user were averaged to obtain the *perceptual judgment* (according to experimental psychology practices [40]).

The agreement among the assessors was measured with the Krippendorff's α [23], a reliability coefficient suitable for a wide variety of assessments (binary, nominal, ordinal, interval etc.), and robust to small sample sizes. Table 1 reports the α values for the different traits. The values are statistically significant and comparable to those observed in the literature for zero acquaintance scenarios [4], i.e. situations where assessors and subjects being rated do not have any personal contact (like it in the *PsychoFlickr* corpus).

	O	C	E	A	N
α	0.07	0.16	0.23	0.19	0.20

Table 1: Krippendorff's α for the Big Five traits.

In terms of the Lens Model, the "*pro*" users are the Data Producers, the assessors are the Data Consumers, the personality is the state and the outcome of the BFI questionnaire is the state measure (see Section 3 for more details).

5. MULTIMEDIA LENS AND FLICKR

This section measures, in quantitative terms, whether implicit cognitive processes are actually at work in the scenario underlying the PsychoFlickr corpus or not. In particular, the section addresses the following questions:

- Is there a consistent relation between features extracted from sets of "favourite" images and personality traits of Flickr users (both self-assessed and attributed)?
- If yes, is the relation sufficiently stable to automatically predict the personality traits of Flickr users (both self-assessed and attributed) based on sets of "favourite" images?

If the key-idea of this work holds, and implicit cognitive processes influence the exchange of multimedia data (according to the Lens Model), then the answer should be positive in both cases.

5.1 Features and Personality

We adopted a wide, though not exhaustive, spectrum of features, here grouped into two families (see Table 2). On

Category	Name	L	Short Description
Aesthetics	Use of light	1	Average pixel intensity of V channel [10]
	HSV statistics	4	Mean of S channel and standard deviation of S, V channels [28]; *Hue angular dispersion* in IHLS color space [31]
	Emotion-based	3	Amount of *Pleasure*, *Arousal*, *Dominance* [28, 51]
	Colorfulness	1	Colorfulness measure based on Earth Mover's Distance (EMD) [10, 28]
	Color Name	11	Amount of *Black*, *Blue*, *Brown*, *Green*, *Gray*, *Orange*, *Pink*, *Purple*, *Red*, *White*, *Yellow* [28]
	Entropy	1	Image entropy [27]
	Wavelet textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on the HSV channels [10]
	Tamura	3	Amount of *Coarseness*, *Contrast*, *Directionality* [47]
	GLCM-features	12	Amount of *Contrast*, *Correlation*, *Energy*, *Homogeneity* for each HSV channel [28]
	Edges	1	Total number (#) of edge points, extracted with Canny [27]
	Level of detail	1	Number of regions (after mean shift segmentation) [7, 17]
	Regions	1	Average *size* of the regions (after mean shift segmentation) [7, 17]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [10, 28]
	Rule of thirds	2	Mean of S,V channels in the inner rectangle of the image [10, 28]
	Image parameters	2	Size of the image [10, 27]
Content	Objects	28	Objects detectors [13]: we kept the number of instances (#) and their average bounding box *size*
	Faces	2	Number (#) and *size* of faces after Viola-Jones face detection algorithm [52]
	GIST descriptors	24	Level of openness, ruggedness, roughness and expansion for scene recognition [36].

Table 2: Summary of all features. The column 'L' indicates the feature vector length for each type of feature.

(a) (b)

Figure 3: Collage (a) and (b) are a random selection of "favourite" pictures of subjects high and low in Extraversion (as per attributed by the assessors), respectively. The most important difference is that extrovert individuals show a preference for pictures portraying people (80% of the samples in the collage) while introvert show the opposite preference (17% of the pictures in the collage).

one side, we have the cues that focus on aesthetic aspects [10, 28]: the reason is that the PsychoFlickr corpus includes pictures posted as "*favourite*", i.e. likely to represent the aesthetic preferences of the users under examination. On the other side, we focused on the content of the images; to this end, we employed robust probabilistic object detectors [13] (for a complete list of all detectable objects see [13]); we also retained the average area (the algorithm gives also the bounding box of the detected objects). In addition, we focused on the faces, adopting the standard Viola-Jones face detection algorithm [52] implemented in the OpenCV libraries. Finally, we adopted the GIST scene descriptors, which amounts to apply a set of oriented band-pass filters.

Figure 2 shows the features with higher ecological (covariation of self-assessment and features) and representation (covariation of features and perceptual judgment) validity with respect to the Big Five traits. The covariations, measured with the Spearman Coefficient, are statistically significant ($p < 5\%$). Therefore, *implicit cognitive processes seem to be actually at work when Flickr users share their set of favourite images*. The answer to the first question at the beginning of this section is positive. Further confirmation comes from Figure 3, showing a random selection of images labeled as "favourite" by extravert (collage "a") and introvert (collage "b") subjects. The former appear to picture people way more frequently than introvert ones (80% and 17% of the images in the collage, respectively).

5.2 Personality Prediction

Ecological and representation validity values of Figure 2 seem to suggest that predicting personality traits (both self-assessed and attributed) using "favourite" images is possible. The problem was cast as a regression instance on the traits of the users, considering users as the sets of their preferred images (see appendix A for a description of the regression approach). The performance was measured with the Spearman correlation coefficient between actual and predicted personality traits, the higher the coefficient, the closer the prediction to the true value. The results are reported in Table 3 where the first column shows the trait, the second explains

Trait	Label	Max ρ	Mean (Std) ρ	% s.s.
O	Self	0.25	0.17 (0.04)	100%
	Attributed	0.35	0.32 (0.04)	100%
C	Self	0.24	0.22 (0.03)	44%
	Attributed	0.57	0.49 (0.05)	100%
E	Self	0.28	0.19 (0.05)	88%
	Attributed	0.62	0.55 (0.03)	100%
A	Self	0.20	0.17 (0.03)	55%
	Attributed	0.52	0.45 (0.05)	100%
N	Self	0.14	0.12 (0.07)	7%
	Attributed	0.60	0.54 (0.04)	100%

Table 3: Prediction Results. ρ is the Spearman Correlation Coefficient

whether the predicted trait is self-assessed or attributed by the assessors, "Max ρ" is the maximum correlation found across the tests (i.e. the various configurations of the regression approach), "Mean (Std)" are mean value and standard deviation computed on correlations with p-values $< 5\%$, and "% s.s" is the percentage of different configurations of the regression approach that resulted in a statistically significant result.

In line with the literature on personality computing [29], the performances achieved over self-assessed traits are lower than those obtained over attributed ones. The reason is that the former depend on an individual assessment and tend to be more noisy while the latter, resulting from the consensus among different assessors, tend to correlate better with measurable characteristics of the data. In particular, for the attributed traits, all configurations of the regression approach tested in the experiments led to statistically significant results, while for the self-assessed traits this happens only for Openness. The best performance is achieved, for the attributed assessments, for Extraversion and Conscientiousness. The same applies to most of the works on personality computing presented in the literature and the reason is that Extraversion and Conscientiousness are the two traits people perceive more quickly and effectively [20]. In the case of this work, the performance is satisfactory for Neuroticism and Agreeableness as well. The trait for which the performance is lower is Openness. The probable explanation is that the distribution of the users along such trait is peaked around high values (Openness is the trait of creativity and "pro" Flickr users are, not surprisingly, higher than the average along the trait).

The results presented in Table 3 are statistically significant ($p < 5\%$) and, therefore, the answer to the second question of the beginning of this section is positive. In other words, implicit cognitive processes seem to influence the attribution of personality traits in Data Consumers watching "favourite" pictures on Flickr.

6. POTENTIAL APPLICATIONS

Section 2 surveys areas that share some aspects with the perspective proposed in this article while still showing substantial differences. This section focuses on application domains that involve the exchange of multimedia data and, therefore, might benefit from taking into account the cognitive processes that, according to the results presented above, seem to influence such type of process.

Understanding and modeling of cognitive processes involved in multimedia data consumption are likely to be beneficial for Human Information Interaction (HII), the domain studying the "*relationship between people and information*" [14]. HII researchers are particularly interested in modelling people *proflections*, i.e. individual's conscious and unconscious projections on information objects (*e.g.*, pictures) and the reflections that other people and machines create to those projections (*e.g.*, links and annotations) [30]. This applies in particular to multimedia retrieval technologies that might be enhanced by taking into account not only the data content (like most of current technologies do [26]), but also the interplay between content and perceptual judgments (personality, values, goals, intentions, etc.).

The role of cognitive biases can be of interest for Digital Humanities as well, especially for what concerns the effort towards "*new modes of knowledge formation enabled by networked, digital environments*" and the focus on "*distinctive modes of producing knowledge and distinctive models of knowledge itself*" [6]. In particular, Digital Humanities investigate the impact of media authoring technologies on the transmission of knowledge and information, a phenomenon likely to involve implicit cognitive processes like those described in this work. In a similar vein, the perspective adopted in this article can be useful in Big Data Analytics - the domain aimed at making sense of large amounts of unstructured data [32] - one of the most important challenges technology faces today. In particular, there is consensus among Big Data experts that no useful information can be extracted from large databases without associating automatic mining approaches and human interpretation [35]. This latter is likely to be influenced by cognitive processes similar to those illustrated in the experiments of this work.

Viral marketing, the "*diffusion of information about the product and its adoption over the network*" [25], is an advertisement technique aimed at spreading information as widely as possible through (mostly online) *word of mouth* mechanisms. The previous part of this paper has shown that the exchange of multimedia data, being a form of human-human communication, can be thought of as a form of word of mouth. Therefore, implicit cognitive processes might contribute to explain and enhance virality. In the same vein, communication strategies based on social media can benefit from the prediction of perceptual judgments likely to be attributed to a given multimedia message diffused through online social platforms [21].

According to the European Consumer Commissioneer, "*Personal data is the new oil of the internet and the new currency of the digital world*" [16]. The "states" of data producers (see Section 3) often correspond to personal characteristics of potential interest for different bodies (*e.g.*, companies trying to model their customers or governments interested in gathering information about the population). Approaches like those presented in this work can help to obtain such information by analyzing publicly available data that people usually post on personal home pages, Youtube, Facebook, etc. [22]. In parallel, the development of technologies capable of going beyond the mere content and infer personal characteristics of data producers require a redefinition of the concept of privacy and a careful analysis of ethical issues [9].

A peculiar form of communication through multimedia material is the participation in online games where several participants interact via avatars or animated characters.

The choice of a particular character or particular gaming strategies and options is likely to convey information about the player "states" (see, *e.g.*, the approaches in [18, 54, 56] for the case of personality). In a similar way, computer mediated communication can be influenced by implicit cognitive processes via interface characteristics like the profile picture of Skype users.

Creating and viewing photographs as a process of self-insight and personal change is the main principle of *phototherapy* and *therapeutic photography* [53, 46], two recent psychology perspectives; for the therapists, "*Images provide an undercurrent of emotion and ideas that enrich interpersonal dynamics, often on a level that is not fully conscious or capable of being verbalized*". Of particular interest for these fields is how the language of composition and visual design intersects with the language of unconscious primary cognitive processes, including emotional/ideational association. Our study suggests that answers to these questions may be found with the help of computers.

Last, but not least, it is apparent the crucial role that image processing and machine learning would have; at the same time, our study delineates new challenges for these areas; for example, discovering visual patterns that correlate with personal traits in a stronger way than ordinary features or fit better cognition processes could be a research mission for the field of deep learning and feature learning [45]. Generative modeling can also be involved, looking for new models that mimic the way diverse visual features should be combined together to communicate a certain personal trait. An immediate example applies to the Counting Grid used in Section A.1: in this paper, we employed CG as a mere dimensionality reduction strategy, without accounting the traits label. Including this information may lead to a low-dimensional embedding where nearby images exhibit similar features *and* personal traits.

The list presented in this section is far from being exhaustive, but it is representative of the scenarios where the investigations proposed in this article can be relevant, namely those where individuals produce, exchange and consume (possibily multimedia) data.

7. CONCLUSIONS

This article advocates the idea that the exchange of multimedia data has become a human-human communication scenario and, therefore, it involves the same cognitive phenomena of any other form of interaction between people, especially when it comes to expression and mutual attribution of socially relevant characteristics (attractiveness, social status, personality, goals, values, intentions, etc.). As a supporting evidence, the paper proposes experiments on the interplay between personality traits and Flickr pictures. The results show that the personality of an individual can be predicted, to a statistically significant extent, through the pictures she labels as "*favourite*". Furthermore, the experiments show that the images can be used to predict the traits that others attribute to such an individual. Therefore, at least for what concerns personality, the exchange of images via Flickr seems to work according to the *Brunswik Lens* (see Section 3), the cognitive model underlying social interactions. In other words, the key-idea proposed in this work appears to hold.

To the best of our knowledge, such a perspective has never been adopted in a multimedia technology context before. The probable reason is that multimedia data became an interaction channel only recently, when the diffusion of appropriate technologies for data production (cameras, smartphones, tablets, etc.) and consumption (social media, digital libraries, etc.) made it possible to exchange multimedia data as easily as we previously exchanged written material (letters, messages, etc.) [6].

This new scenario opens several research questions (the list is not exhaustive):

- Is it possible to improve multimedia technologies by taking into account implicit cognitive processes?
- Do implicit cognitive processes influence our behavior as multimedia technology users?
- Does multimedia technology need to change to accomodate implicit cognitive processes? If yes, how?
- Can traditional multimedia technologies contribute to understand the influence of cognitive processes in data production and consumption? If yes, how?
- What do we reveal about ourselves when we share multimedia data?
- What is the effect of the multimedia data we share on the impression others develop about us?

It can be expected that the perceptual judgments we make about those who produce the data we consume end up influencing our perception of the data. For example, we might tend to like more or to find more relevant data produced by people we perceive as more similar to us. If actually observed, such an effect (known in psychology as "*similarity-attraction*" [8]), might not only improve retrieval technologies, but also contribute to explain our behavior as users and, in ultimate analysis, lead to higher technology usability and effectiveness. Similar considerations apply to any technology that involves the consumption of data.

Symmetrically, the increasing amount of multimedia information we produce and share (Instagram pictures, Tweets including pointers to video and audio data, etc.) is probably contributing to a larger and larger extent to our "*appearance*", one of the characteristics that influence most the impression others develop about us (an effect known as the "*halo-effect*" psychology [34]). However, while we know how to manage our appearance in face-to-face interactions, in most cases we are still not aware of the way others see us through the lens of the multimedia data we produce.

The two examples above show how cognitive and technological issues are tightly intertwined in everyday scenarios involving production and consumption of multimedia data. The two cases focus on specific aspects, but the perspective proposed in this work might show that the full range of phenomena taking place in face-to-face interactions (see [11] for a monograph) take place through multimedia data as well. If true, the door would be open towards new multimedia applications as well as novel findings in cognitive sciences.

APPENDIX

A. THE REGRESSION APPROACH

To apply a standard regression approach is problematic because there are multiple images associated to the same target. Straightforward algorithms like, *e.g.*, summing all the image descriptors of each user, and then perform regression, does not work because such process adds noise to a

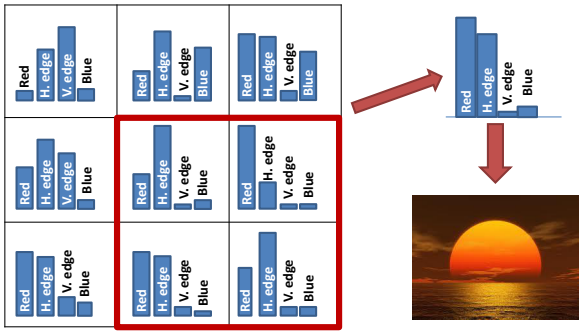

Figure 4: Generating an image from a simple 3×3 counting grid: given a 2×2 window on the grid, we average the feature counts, obtaining a bag of features which corresponds to the final image. V. edge and H. edge are toy features meaning vertical and horizontal edges, respectively.

weak signal. Multiple instance regression [41] is also unadvisable because of its high computational complexity, especially when the number of images for each user is large.

Therefore, we propose an alternative approach composed by the following three steps:

1. **Feature Extraction and Normalization**. We first extract from all the images the set of features listed in Section 5.1; since each z-th cue expresses the level of presence of a given quantity, i.e. a count c_z, we can think each image as an histogram of counts $\{c_z\}$, or bag-of-features (BoF). After that, we normalize each c_z to ensure that each feature takes values in the same range. This avoids some features (*e.g.*, number of edges) to overcome others (*e.g.*, GIST, amount of coarseness)

2. **Clustering**. After dividing the users in training and testing users, we consider all the images of the training users. By means of a clustering algorithm, we learn a low-dimensional representation that maps each t-th image (i.e. its BoF) in a 2-dimensional location ℓ^t, lying on a smooth manifold. As clustering method we employed the *Counting Grid* [38], a recent generative model which embeds BoF representations in N-dimensional manifolds[3]. This way, each user u becomes a set of locations $L^u = \{\ell^t\}$ on the manifold.

3. **Regression and Trait Prediction**. Considering the training users, we train a regressor to the personality traits. In specific, for each user u we have a five-dimensional target that characterizes the Big Five personality traits $p \in \{O, C, E, A, N\}$, where each trait is described by a value $y_p^u \in [-4, 4]$. As regression method, we used Lasso [48]. Trait prediction amounts to test the regressor on the test users.

In the following, we will detail the latter two steps of the process.

A.1 Clustering: the Counting Grid Model

[3]Here we decide $N = 2$ for the sake of clarity; other dimensions can be explored. In addition, we tried different dimensionality reduction approaches (Mixtures of Dirichlet distributions), leading to inferior performances.

The counting grid (CG) is a generative model recently introduced in [38] for analyzing images collections. It assumes that images are represented as histograms $\{c_z\}$ or bags of features, where c_z counts the occurrences of feature z.

Considering its two-dimensional version, a CG is a 2D finite discrete grid where each location $\mathbf{i} = (x, y)$ contains a normalized count of features $\pi_{\mathbf{i},z}$. Under this model, an image (i.e. its BoF $\{c_z\}$) could be thought as produced by the following generative process: a small window is located in the grid, averaging the feature counts within it to obtain a local probability mass function over the features, and then generating from it an appropriate number of features in the bag (see Fig. 4). In other words, unlike a straightforward embedding (e.g. PCA) that links an image with a point location, the counting grid forces the image to link with a small window of locations. Given that the size $E_1 \times E_2$ of a counting grid is usually small compared to the number of images, this also forces windows linked to different images to overlap, and to co-exist by finding a shared compromise in the feature counts located in their intersection. The overall effect of these constraints is to produce locally smooth transitions between strongly different feature counts by gradually phasing features in/out in the intermediate locations.

In practice, local neighborhoods in the grid represent similar concepts and images mapped in close locations are somehow similar.

Formally, the counting grid $\pi_{\mathbf{i},z}$ is a 2D finite discrete grid, spatially indexed by $\mathbf{i} = (x, y) \in [1 \dots E_1] \times [1 \dots E_2]$, and containing normalized counts of features indexed by z. Thus, we have $\sum_z \pi_{\mathbf{i},z} = 1$ everywhere on the grid. A given BoF $\{c_z\}$ is generated by selecting a certain location $\mathbf{k}$, calculating the distribution $h_{\mathbf{k},z} = \frac{1}{W_n} \sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z}$ by averaging all the words counts within the window $W_\mathbf{k}$ (with area W_n) that starts at $\mathbf{k}$, and then drawing features counts from this distribution.

In other words, the position of the window $\mathbf{k}$ in the grid is a latent variable; given $\mathbf{k}$, the likelihood of $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{c_z} = \alpha \prod_z \Big(\sum_{\mathbf{i} \in W_\mathbf{k}} \pi_{\mathbf{i},z} \Big)^{c_z}, \qquad (1)$$

where α is a fixed normalization factor.

To learn a counting grid, we need to maximize the likelihood over all training images T, that can be written as

$$p(\{\{c_z^t\}, \mathbf{k}^t\}_{t=1}^T) \propto \prod_t \prod_z \Big(\sum_{\mathbf{i} \in W_{\mathbf{k}^t}} \pi_{\mathbf{i},z} \Big)^{c_z^t}, \qquad (2)$$

which is intractable, much like in mixtures; therefore, it is necessary to employ an iterative EM algorithm. Starting from a random initialization of the counting grid π, the E-step aligns all bags of features to grid windows, to match the bags' histograms, inferring where each bag maps on the grid, i.e.

$$q^t(\mathbf{i}) \propto \exp \sum_z c_z^t \cdot \log h_{\mathbf{i},z} \qquad (3)$$

In the M-step the model parameter, i.e. the counting grid π, is re-estimated. For details on the learning algorithm and on its efficiency, the reader can refer to the original papers [38, 19]. For our purposes, the most interesting outputs are the posterior probabilities q^t's, the position in the grid of each image. Summing over the entire grid the contributes $q^t(\mathbf{i})$, which are due to the images of a user, provides a signature L^u. Essentially, it is a 2D matrix, of the same dimension of

the grid, where some locations $\{\mathbf{i}\}$ are weighted by the q^t's, indicating that in such locations there are some images of the user u.

A.2 Regression and Trait Prediction

To assess the validity of our prediction method, we used the Leave-One-User-Out paradigm. We considered CGs of various complexities with size $E = [20 \times 20, 25 \times 25, \ldots 65 \times 65]$ and window $W = [5 \times 5]$ and we learnt a model with all the images belonging to the training users. Then, we computed L^u for each user, and used this representation to regress on the profile y_p^u. We learned the regression weight vector $\mathbf{w}$ by minimizing the error function

$$E(\mathbf{w}) = \sum_{\mathbf{i}=1}^{L} \left(\mathbf{y_p} - \mathbf{w}^T \cdot L^u(\mathbf{i}) \right)^2 \quad (4)$$

where L indicates the number positions in the grid. We solved the problem using Lasso [48], a shrinkage and selection method for linear regression which enforces the sparsity on coefficients $\mathbf{w}$ by bounding the sum of the absolute values of the coefficients. The bound is a constraint that has to be taken into account when minimizing the error function. At this point the training phase is completed and to predict the personality train of the held out (test) user, we 1) infer the mappings of its images on the counting grid (Eq.3), 2) compute is latent description $L^{test}(\mathbf{i})$ and 3) by calculating

$$\hat{y}_p^{test} = \mathbf{w}^T \cdot L^{test}, \quad (5)$$

we predict the trait.

B. REFERENCES

[1] I. Arapakis, J.M. Jose, and P.D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 395–402. ACM, 2008.

[2] M. Baldauf, S. Dustdar, and F. Rosenberg. A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing*, 2(4):263–277, 2007.

[3] F.J. Bernieri and J.S. Gillis. Judging rapport: Employing brunswik's lens model to study interpersonal sensitivity. In J.A. Hall and F.J. Bernieri, editors, *Interpersonal Sensitivity. Theory and Measurement*. Lawrence Erlbaum, 2001.

[4] J.C. Biesanz and S.G. West. Personality coherence: Moderating self–other profile agreement and profile consensus. *Journal of Personality and Social Psychology*, 79(3):425–437, 2000.

[5] E. Brunswik. *Perception and the representative design of psychological experiments*. University of California Press, 1956.

[6] A. Burdick, J. Drucker, P. Lunenfelds, T. Presner, and J. Schnapp. *Digitial Humanities*. MIT Press, 2012.

[7] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603 – 619, 2002.

[8] J.W. Condon and W.D. Crano. Inferred evaluation and the relation between attitude similarity and interpersonal attraction. *Journal of Personality and Social Psychology*, 54(5):789, 1988.

[9] R. Cowie. The good our field can hope to do, the harm it should avoid. *IEEE Transactions on Affective Computing (to appear)*, 2013.

[10] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision*, volume 3953 of *Lecture Notes in Computer Science*, pages 288–301. Springer Verlag, 2006.

[11] J. Elster. *Explaining Social Behavior*. Cambridge University Press, 1997.

[12] D.C. Evans, S. D. Gosling, and A. Carroll. What elements of an online social networking profile predict target-rater agreement in personality impressions. In *Proceedings of the International Conference on Weblogs and Social Media*, pages 45–50, 2008.

[13] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://www.cs.brown.edu/~pff/latent-release4/, 2010.

[14] R. Fidel. *Human Information Interaction*. MIT Press, 2012.

[15] S. Fitzgerald, D.C. Evans, and R.K. Green. Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement. In *Proceedings of AAAI International Conference on Weblogs and Social Media*, 2009.

[16] World Economic Forum. Personal data: the emergence of a new asset class. Technical report, World Economic Forum, 2011.

[17] C.M. Georgescu. Synergism in low level vision. In *Proceedings of the International Conference on Pattern Recognition*, pages 150–155, 2002.

[18] D. Johnson and J. Gardner. Personality, motivation and video games. In *Proceedings of the Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, pages 276–279, 2010.

[19] N. Jojic and A. Perina. Multidimensional counting grids: Inferring word order from disordered bags of words. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 547–556, 2011.

[20] C.M. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima. Fundamental dimensions of social judgment: Unrdestanding the relations bteween judgments of competence and warmth. *Journal of Personality and Social Psychology*, 89(6):899–913, 2005.

[21] A.M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

[22] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceeedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[23] K. Krippendorff. Reliability in content analysis. *Human Communication Research*, 30(3):411–433, 2004.

[24] Z. Kunda. *Social cognition: Making sense of people*. The MIT Press, 1999.

[25] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):5, 2007.

[26] M.S. Lew, N. Sebe, D. Chabane, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19, 2006.

[27] P. Lovato, A. Perina, N. Sebe, O. Zandoná, A. Montagnini, M. Bicego, and M. Cristani. Tell me what you like and I'll tell you what you are: discriminating visual preferences on Flickr data. In K.M. Lee, Y. Matsushita, J.M. Rehg, and Z. Hu, editors, *Proceedings of the Asian Conference on Computer Vision*, volume Lecture Notes in Computer Science 7724. Springer Verlag, 2012.

[28] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the ACM International Conference on Multimedia*, pages 83–92, 2010.

[29] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[30] G. Marchionini. Human–information interaction research and development. *Library & Information Science Research*, 30(3):165–174, 2008.

[31] K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2009.

[32] M. Minelli, M. Chambers, and A. Dhiraj. *Big Data, Big Analytics*. Wiley, 2013.

[33] C. Nass and S. Brave. *Wired for speech: How voice activates and advances the Human-Computer relationship*. The MIT Press, 2005.

[34] R.E. Nisbett and T.D. Wilson. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4):250–256, 1977.

[35] F.J. Olhorst. *Big Data Analytics*. Wiley, 2013.

[36] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[37] M. Pantic and A. Vinciarelli. Implicit Human-Centered Tagging. *IEEE Signal Processing Magazine*, 26(6):173–180, 2009.

[38] A. Perina and N. Jojic. Image analysis by counting on a grid. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 1985–1992, 2011.

[39] A. Peterson Bishop, N.A. van House, and B.P. Buttenfields, editors. *Digital Library Use*. MIT Press, 2003.

[40] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *in Journal of Research in Personality*, 41:203–212, 2007.

[41] S. Ray. Multiple instance regression. In *Proceedings of the International Conference on Machine Learning*, pages 425–432, 2001.

[42] D. Rosenberg. Data before the Fact. In L. Gitelman, editor, *Raw data is an oxymoron*, pages 15–40. MIT Press, 2013.

[43] G. Saucier and L.R. Goldberg. The language of personality: Lexical perspectives on the five-factor model. In J.S. Wiggins, editor, *The Five-Factor Model of Personality*. 1996.

[44] K.R. Scherer. Personality markers in speech. In *Social markers in speech*, pages 147–209. Cambridge University Press, Cambridge, 1979.

[45] K. Sohn, D.Y. Jung, H. Lee, and A.O. Hero. Efficient learning of sparse, distributed, convolutional feature representations for object recognition. In *IEEE International Conference on Computer Vision*, pages 2643–2650, 2011.

[46] J. Suler. The psychotherapeutics of online photosharing. *International Journal of Applied Psychoanalytic Studies*, 6(4):339–344, 2009.

[47] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6), 1978.

[48] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[49] J.S. Uleman, L.S. Newman, and G.B. Moskowitz. People as flexible interpreters: Evidence and issues from spontaneous trait inference. In M.P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 28, pages 211–279. Elsevier, 1996.

[50] J.S. Uleman, S.A. Saribay, and C.M. Gonzalez. Spontaneous inferences, implicit impressions, and implicit theories. *Annual Reviews of Psychology*, 59:329–360, 2008.

[51] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology General*, 123(4):394–409, 1994.

[52] P.A. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[53] J. Weiser. *Phototherapy techniques: Exploring the secrets of personal snapshots and family albums*. Jossey-Bass San Francisco, 1993.

[54] C.Y. Yaakub, N. Sulaiman, and C.W. Kim. A study on personality identification using game based theory. In *Proceedings of the International Conference on Computer Technology and Development*, pages 732–734, 2010.

[55] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe. In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. In *Proceedings of the ACM international conference on Multimedia*, pages 349–358, 2012.

[56] N. Yee, N. Ducheneaut, L. Nelson, and P. Likarish. Introverted elves & conscientious gnomes: The expression of personality in World of Warcraft. In *Proceedings of the Annual Conference on Human Factors in Computing Systems*, pages 753–762, 2011.