The Pictures we Like are our Image: Continuous Mapping of Favorite Pictures into Self-Assessed and Attributed Personality Traits

Crisitina Segalin[1], *Student Member, IEEE*, Alessandro Perina[2], Marco Cristani[1], *Member, IEEE*, and Alessandro Vinciarelli[3], *Member, IEEE*

***Abstract*—Flickr allows its users to tag the pictures they like as "favorite". As a result, many users of the popular photo-sharing platform produce galleries of favorite pictures. This article proposes new approaches, based on Computational Aesthetics, capable to infer the personality traits of Flickr users from the galleries above. In particular, the approaches map low-level features extracted from the pictures into numerical scores corresponding to the Big-Five Traits, both self-assessed and attributed. The experiments were performed over 60,000 pictures tagged as favorite by 300 users (the PsychoFlickr Corpus). The results show that it is possible to predict beyond chance both self-assessed and attributed traits. In line with the state-of-the-art of Personality Computing, these latter are predicted with higher effectiveness (correlation up to 0.68 between actual and predicted traits).**

***Index Terms*—Computational Aesthetics; Personality Computing, Big Five Personality Traits, Automatic Personality Perception, Automatic Personality Recognition.**

I. Introduction

IS a picture worth a thousand words? It seems to be so when it comes to mobile technologies and social networking platforms: taking pictures is the action most commonly performed with mobile phones (82% of the American users), followed by exchanging text messages (80% of the users) and accessing the Internet (56% of the users) [1]. Furthermore, 56% of the American Internet users either post online original pictures and videos (46% of the total Internet users) or share and redistribute similar material posted by others (41% of the total Internet users). In other words, "*photos and videos have become key social currencies online*" [2].

Pictures streams are "*often seen as a substitute for more direct forms of interaction like email*" [3] and interacting with connected individuals appears to be one of the main motivations behind the use of online photo-sharing platforms [4]. The pervasive use of *liking* mechanisms, online actions allowing users to publicly express appreciation for a given online item, further confirms the adoption of pictures as a social glue [5]: on Flickr, *likes* between connected users are roughly 10^5 times more frequent than those between non-connected ones [6].

Besides helping to maintain connections and express affiliation, liking mechanisms are a powerful means of possibly involuntary self-disclosure: statistical approaches can infer personality traits and hidden, privacy sensitive information (e.g., political views, sexual orientation, alcohol consumption habits, etc.) from likes and Facebook profiles [7]. Furthermore, online expressions of aesthetic preferences convey an impression in terms of characteristics like prestige, differentiation or authenticity [8]. For this reason, this article proposes new approaches, based on Computational Aesthetics, capable to infer the personality traits of Flickr users, both self-assessed and attributed by others, from the pictures they tag as favorite. In other words, this article proposes an approach aimed at mapping the pictures people like into personality traits.

Regression analysis appears to be the most suitable computational framework for the problem. This applies in particular to Multiple Instance Regression (MIR) [9], [10] because Flickr users typically tag several pictures as favorite. Therefore, there are multiple instances (the favorite pictures) in a *bag* (the user that tags the pictures as *favorite*) associated with a single value (a personality trait of the user). Previous approaches show that it is possible to infer the aesthetic preferences of people through an appropriate weighting of their favorite pictures [11]. This work extends such a principle to the inference of personality traits and addresses the problem with a multiple instance strategy. In particular, this article proposes a set of novel methods that build an intermediate representation of the pictures - using topic models - and then perform regression in the resulting space, thus improving the performance of standard MIR approaches operating on the raw features extracted from the pictures.

The experiments have been performed over *PsychoFlickr*, a corpus of 60,000 pictures tagged as favorite by 300 *Pro* Flickr users[1] (200 randomly selected favorite pictures per user). For each user, the corpus includes two personality assessments. The first has been obtained by asking the users to self-assess their own Big-Five Traits, the second has been obtained by asking 12 independent assessors to rate the Big-Five Traits of the users (see Section III for details). In this way, according to the terminology introduced in [12], it is possible to perform both Automatic Personality Recognition (APR) and Automatic Personality Perception (APP), i.e. the prediction of self-assessed and attributed traits, respectively.

The reason for addressing both APR and APP is that self-assessed and attributed traits tend to relate differently to different aspects of an individual [13] and, therefore, both

[1]University of Verona (Italy), [2]Italian Institute of Technology (Italy), [3]University of Glasgow (UK).

[1] At the moment of the data collection, *Pro* users were individuals paying a yearly fee in order to access privileged Flickr functionalities.

need to be investigated. In the case of this work, the results suggest that favorite pictures account only to a limited extent for self-assessed traits (the best APR result is a correlation 0.26 between actual and predicted traits) while they have a major impact on attributed ones (the best APP result is a correlation 0.68 between actual and predicted traits). To the best of our knowledge, this is one of the few works where APP and APR have been compared over the same data (see [12] for an extensive survey). This is an important advantage because it allows one to assess the effectiveness of a given type of behavioural evidence (the favorite pictures in this case) in conveying information about personality.

The results above show that the proposed approach is more effective in the case of the attributed traits, i.e. in the case of APP. While not necessarily corresponding to the actual traits of people, attributed traits are still predictive of important aspects of social life [13]. In particular, attributed traits determine, to a significant extent, the way others behave towards a given individual, especially in the earliest stages of an interaction [14]. Furthermore, sociologists have observed that the social identity of an individual does not result only from her actual characteristics, but also from the characteristics attributed by others: "*We need to recognise that identification is often most consequential as the categorisation of others, rather than as self-identification*" [15]. For this reason, the literature proposes approaches aimed at predicting both self-assessed and attributed traits [12] and this work addresses both problems.

The rest of this paper is organised as follows: Section II surveys previous work, Section III presents the PsychoFlickr Corpus, Section IV introduces the low-level features extracted from the pictures, Section V describes the new MIR approaches developed for this work, Section VI reports on experiments and results and the final Section VII draws some conclusions.

II. Previous Work

The experiments of this work lie at the crossroad between Computational Aesthetics and Personality Computing (see [12] and [16] for an extensive surveys). To the best of our knowledge, no other works have addressed the problem of mapping favorite pictures into personality traits. However, several works have addressed separately the inference of aesthetic preferences from pictures and the inference of personality traits (both assessed and self-assessed) from social media material. The rest of this section proposes a survey of the main works presented in both areas.

A. Computational Aesthetics

Computational Aesthetics (CA) target "*[...] computational methods that can make applicable aesthetic decisions in a similar fashion as humans can*" [17]. In the particular case of pictures, the goal of CA is typically to predict automatically whether human observers like a given picture or not (for a more general discussion on the relation between technology and aesthetics, see [18]). In most cases, the task corresponds to a binary classification, i.e. to predict automatically whether a picture has been rated high or low in terms of visual pleasantness [19], [20], [21], [22], [23]. Unlike Implicit Tagging [24], CA does not try to measure or detect the reaction of people to get an indication of what the content can be (e.g., by tagging as "*funny*" an image when a person laughs at it). The sole target of CA is to identify image properties that discriminate between appealing pictures and the others.

The experiments of [19] aim at predicting whether a picture has been rated as visually pleasant or not (the two classes correspond to top and bottom pleasantness ratings assigned by human observers, respectively). The experiments are performed over a set of 1,664 images downloaded from the web. The features extracted from the images account for the properties of color, composition and texture. The classification accuracy achieved with Support Vector Machines is higher than 70%. Similar experiments are proposed in [20] over 6,000 images (3,000 per class). The features account for composition, lighting, focus controlling and color. The main difference with respect to the other approaches presented in this section is that the processing focuses on the subject region and on its difference with respect to the background. The classification accuracy is higher than 90%.

In the case of [21], the experiments are performed over digital images of paintings and the task is the discrimination between high and low quality paintings. The features include color distribution, brightness properties (accounting for the use of light), use of blurring, edge distribution, shape of picture segments, color properties of segments, contrast between segments, and focus region. In this case as well, the task is a binary classification and the experiments are performed over 100 images. The best error rate is around 35%. In a similar vein, the approach proposed in [22] detects the subject of a picture first and then it extracts features that account for the difference between foreground and background. The features account for sharpness, contrast and exposure and the experiments are performed over a subset of the pictures used in [19]. Like in the other works presented so far, the task is a binary classification and the accuracy is 78.5%. The approach proposed in [23] adopts an alternative approach for what concerns the features. Rather than using features inspired by good practices in photography, like the other works presented so far, it uses features like SIFT and descriptors like the Bag of Words or the Fisher Vector. While being general purpose, these are expected to encode the properties that distinguish between pleasant and non-pleasant images. The experiments are performed over 12,000 pictures and the accuracy in a binary classification task (6,000 pictures per class) is close to 90%.

B. Personality and Social Media

The literature proposes several works investigating the interplay between the traces that people leave on social media (posts, pictures, profiles, likes, etc.) and personality traits, both self-assessed [25], [26], [32], [27], [28], [29] and attributed [29], [30], [31]. Table I contains a synopsis of the

Ref.	Subj.	Samples	Features	Task	Ext.	Agr.	Con.	Neu.	Ope.	Other
[25]	167	167 Facebook Profiles	profile info., egocentric networks, LIWC	R	0.12 MAE	0.10 MAE	0.10 MAE	0.11 MAE	0.10 MAE	

[26]	209	209 RenRen profiles	Profile info., usage statistics, emotional states	C(2)
				C(3)
					83.8
					71.7
					F
						69.7
						72.3
						F
							82.4
							70.1
							F
								74.9
								71.0
								F
									81.1
									69.5
									F	

[27]	156	473 posts on FriendFeed	Some LIWC categories	U						average accuracy 63.1
[28]	10000	10000 blog posts	LIWC	C(2)	80.0 ACC					
[29]	300	60, 000 favorite pictures	visual patterns, aesthetic preferences	R	0.19 ρ	0.17 ρ	0.22 ρ	0.12 ρ	0.17 ρ	
[30]	440	440 pictures	photo content, appearance	CA						see text
[31]	5216	5216 social media profiles	presence of personal information	CA						see text

TABLE I: The table reports, from left to right, the number of subjects involved in the experiments, number and type of behavioral samples, main cues, type of task and performance over different traits. LIWC stands for Linguistic Inquiry Word Count (a psychologically oriented text analysis approach). The column "Other" refers to works using models different from the Big-Five. R stands for regression, U stands for unsupervised classification, C(n) for classification with n classes, and CA for correlational analysis. The performance for the classification tasks is reported in terms of Mean Absolute Error (MAE), F-Measure (F), accuracy (ACC) and correlation (ρ). The performances are not reported for comparison purposes (the results have been obtained over different data), but to provide full information about the works described.

works presented in this section [2].

The approach proposed in [25] infers the self-assessed personality traits of 167 Facebook users from the absence or presence of certain items (e.g., political orientation, religion,etc.) in the profile. The results, obtained with regression approaches based on Gaussian Processes and M5 algorithm, correspond to a mean absolute error lower than 0.15. Given that personality scores are defined along a 5 points scale, such an average error can be considered low, but it is unclear whether it is roughly the same for all subjects or it tends to be low for people on the extremes and high for those in the middle of the scales. Furthermore, the low number of trainig items (roughly 150) and the high number of features might have led to overfitting. In a similar way, the experiments presented in [26] predict whether 209 users of *Ren Ren* (a popular Chinese social networking platform) are in the lowest, middle or highest third of the observed personality scores. The features adopted in such a work include usage measures such as the post frequency, the number of uploads, etc. The results show an F-measure up to 72% depending on the trait. However, the performance seems to be higher for those traits where one of the three classes is more represented than the others, then the improvement is low with respect to a basic approach always giving as output the most represented class. APR on Facebook profiles was the subject of an international benchmarking campaign [3] the results of which appear in [32]. The main indication of this initiative is that selection techniques applied to large sets of initial features lead to the highest performances. However, the experimental setup adopted for the challenge (participants have all the data at disposition since the beginning) cannot exclude overfitting. In particular, it is unclear whether feature selection techniques have been applied only to the training set or to the entire corpus (if this is the case, features have been selected using information from the test set), thus overestimating the performance.

Given the difficulty in collecting large amounts of self-assessments, two approaches propose to use measures like the number of connections or the lexical choices in posts as a criterion to assign personality scores to social media users [27], [28]. In the case of [27], the proposed methodology measures first whether features like the use of punctuation (e.g., exclamation marks) or emoticons is stable for a given user, then it uses the most stable features to assign personality traits. The results show an accuracy of 63% in predicting the actual self-assessed traits of 156 users of *FriendFeed*, an Italian social network. The most interesting novelty of this work is the attempt to avoid the collection of assessments, an expensive and time-consuming, process, through unsupervised approaches trying to assign similar personality profiles to user similars in terms of online activities. However, the performances are not sufficient to actually replace the collection of self-assessed traits. Similar considerations apply to the case of [28], the authors simply label the users as extravert or introvert depending on how many connections they have. The resulting labels are then predicted automatically using lexical choices, i.e. calculating how frequently people use words falling in the different categories of the Linguistic Inquiry Word Count.

While the works presented so far address the APR problem, the experiments presented in [30], [31] target APP. In particular, both works investigate the agreement between self-assessed traits of social media users and traits that these are attributed by observers after posting material online. In the case of [30], the focus is on profile pictures and the experiments show that the agreement is higher when the picture subjects smile and do not wear hats. The other work [31] performs a similar analysis over profiles showing personal information and the results show that the agreement improves when people post information about their spiritual and religious beliefs, their most important sources of satisfaction, and material they consider to be funny. However, in both these works the ratings were made by one assessor only and, therefore, it is unclear

[2] Section III-B provides an introduction to personality and its measurement. Should the reader be unfamiliar with these concepts, reading Section III-B can help to better understand Section II-B and the content of Table I.

[3] `http://mypersonality.org/wiki/doku.php?id=wcpr13`

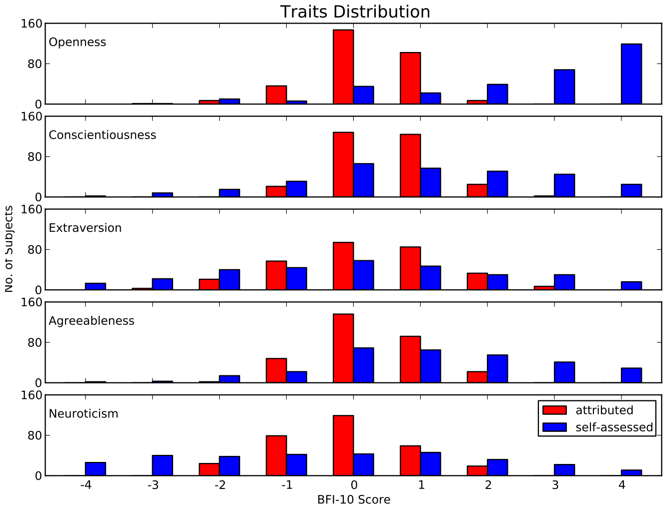

Fig. 1: The figure shows the distribution of self-assessed and attributed traits.

whether they can account for the average impression a subject conveys.

To the best of our knowledge, the work in [29] is the only one that addresses both APR and APP using the same data. The experiments are performed over the PsychoFlickr Corpus, the same dataset used in the experiments of this article (see Section III). The results show that the agreement between self-assessed and attributed traits range between 0.32 and 0.55 depending on the trait. Furthermore, experiments based on Counting Grid models and Lasso regression approaches lead to a correlation up to 0.62 for the attributed traits and up to 0.22 for self-assessed ones. Compared to the work in [29], the experiments of this article propose entirely different approaches, include the comparison between seven different methodologies based on Multiple Instance Regression, present a more extensive correlational analysis and provide a full description of the features adopted. In other words, this work includes substantial novelties and differences with respect to the results of [29].

Overall, the best performances tend to be observed for Extraversion and Conscientiousness (see Table I). This is not surprising because personality psychologists have observed that these are the traits that human observers tend to perceive more clearly as well [33]. However, there are specific contexts where traits typically difficult to observe become more *available*, i.e. more accessible to human observers and, therefore, easier to predict [13]. This is the case, e.g., of the higher performances obtained for Openness in [29], [31].

III. PsychoFlickr: Pictures and Personality

The experiments of this work are performed over *PsychoFlickr* [4], a corpus designed to investigate the interplay between aesthetic preferences and personality traits. The corpus includes pictures that 300 Flickr users have tagged as favorite (200 pictures per user for a total of 60,000 samples). Furthermore, for each user, the corpus includes both self-assessed and attributed traits (see Section III-B). Therefore, PsychoFlickr allows one to perform both APP and APR experiments.

A. The Subjects

The subjects included in the corpus were recruited through a *word-of-mouth* process. A few Flickr Pro users were contacted personally and asked to involve other Pro users in the experiment (typically through the social networking facilities available on Flickr). The process was stopped once the first 300 individuals answered positively. The resulting pool of users includes 214 men (71.3% of the total) and 86 women (28.7% of the total). The age at the moment of the data collection is available only for 44 subjects (14.7% of the total)[5]. These participants are between 20 and 62 years old and the average age is 39. However, it is not possible to know whether this is representative of the entire pool. The nationality is available for 288 users (96.0% of the total) that come from 37 different countries. The most represented ones are Italy (153 subjects, 51% of the total), United Kingdom (31 subjects, 10.3% of the total), United States (28 subjects, 9.3% of the total), and France (13 participants, 4.3% of the total).

B. Personality and its Measurement

Personality is the latent construct that accounts for "*individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns*" [34]. The literature proposes a large number of personality models and PsychoFlickr adopts the Big-Five (BF) Traits, five broad dimensions that have been shown to capture most individual differences [35]. The reason behind this choice is twofold: On the one hand the BF is the model most commonly applied in both personality computing [12] and personality science [13]. On the other hand, the BF model represents personality in terms of five numerical scores, a form particularly suitable for computer processing.

The scores of the BF model account for how well the behavior of an individual fits the tendencies associated to the BF Traits, i.e. *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 the BF framework, assessing the personality of an individual means to calculate the five scores corresponding to the traits above. The literature proposes several questionnaires designed for such a task (see [12] for a list of the most important ones). The personality assessments of PsychoFlickr have been obtained with the BFI-10 [36], a list of ten items

[4] The corpus is available at `http://vips.sci.univr.it/dataset/psychoflickr/PsychoFlickr.rar`

[5] Personal information is extracted from the Flickr profiles where the users are allowed to hide the details they prefer to keep private.

Self-Assessment	Attribution	Trait
I am reserved	The user is reserved	Ext
I am generally trusting	The user is generally trusting	Agr
I tend to be lazy	The user tends to be lazy	Con
I am relaxed, handle stress well	The user is relaxed, handles stress well	Neu
I have few artistic interests	The user has few artistic interests	Ope
I am outgoing, sociable	The user is outgoing, sociable	Ext
I tend to find fault with others	The user tends to find fault with others	Agr
I do a thorough job	The user does a thorough job	Con
I get nervous easily	The user gets nervous easily	Neu
I have an active imagination	The user has an active imagination	Ope

TABLE II: The BFI-10 [36] is the short version of the Big-Five Inventory. Each Item is associated to a Likert scale, from -2 ("*Strongly disagree*") to 2 ("*Strongly agree*") and contributes to the integer score (in the interval $[-4, 4]$) of a particular trait, see the third column. The answers are mapped into numbers (e.g., from -2 to 2). The table shows the questionnaire in both self-assessment and attribution version.

associated to 5-points Likert scales ranging from "*Strongly Disagree*" to "*Strongly Agree*". The main advantage of the BFI-10 is that it can be filled in less than one minute while still providing reliable results. Table II shows the BFI-10 for both self-assessment and attribution of personality traits. In the self-assessment case, people fill the questionnaire about themselves and the result is a personality self-assessment (necessary to perform APR experiments). In the attribution case, people fill the questionnaire about others and the result is a personality attribution (necessary to perform APP experiments).

The 300 Flickr users included in the corpus were asked to fill the self-assessment version of the BFI-10 and were offered a short analysis of the outcome as a reward for the participation. The chart of Figure 1 shows the distribution of the scores for each trait. In line with the observations of the literature, the self-assessments tend to be biased towards socially desirable characteristics (e.g., high Conscientiousness and low Neuroticism) [37]. In the case of PsychoFlickr, this applies in particular to Openness, the trait of intellectual curiosity and artistic inclinations. A possible explanation is that the pool of subjects includes individuals that tend to consider photography as a form of artistic expression. However, no information is available about this aspect of the users.

In parallel, 12 independent judges were hired to attribute personality traits to the 300 subjects of the corpus. The judges are fully unacquainted with the users and they are all from the same country (Italy) to ensure cultural homogeneity. They were asked to watch the 200 pictures tagged as favorite by each user and, immediately after, to fill the attribution version of the BFI-10 ("Attribution" column of Table II). Each judge has assessed all 300 users and the 12 assessments available for each user were averaged to obtain the attributed traits. The judges were paid 95 Euros for their work. The chart of Figure 1 shows the resulting distribution of scores. Since the judges are fully unacquainted with the users and all they know about the people they assess are the favorite pictures, the ratings tend to peak around 0, the score associated to the expression "*Neither agree nor disagree*".

Research on consensus in the perception of the Big-Five suggests to measure the agreement between raters in terms of percentage of ratings variance shared across judges [38], [39]. The percentage can be measured by performing a two-way Analysis of Variance of the ratings [40]. If r_{ij} is the score that judge j assigns to subject i for a particular trait, then the grand mean of the scores $\bar{r}$ is

$$\bar{r} = \frac{1}{SR} \sum_{i=1,j=1}^{i=S,j=R} r_{ij}, \tag{1}$$

where S is the total number of subjects and R is the total number of raters. Correspondingly, it is possible to define the mean squares for subjects, raters, cells and interaction between raters and subjects as follows [40]:

$$\begin{array}{c} \mu_s^2 = \frac{R}{S-1} \cdot \sum_{i=1}^{S} (\sum_{j=1}^{R} r_{ij} - \bar{r})^2 \\ \mu_r^2 = \frac{S}{R-1} \cdot \sum_{j=1}^{R} (\sum_{i=1}^{S} r_{ij} - \bar{r})^2 \\ \mu_c^2 = \frac{1}{RS-1} \sum_{i=1,j=1}^{i=S,j=R} (r_{ij} - \bar{r})^2 \\ \mu_{s \times r}^2 = \mu_c^2 - \mu_r^2 - \mu_s^2. \end{array} \tag{2}$$

The expressions above allow one to define the percentage α of variance shared across raters as follows:

$$\alpha = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_{s \times r}^2}, \tag{3}$$

where $\sigma_s^2 = (\mu_s^2 - \mu_{s \times r}^2)/R$ is the estimate of the subject variance, $\sigma_r^2 = (\mu_r^2 - \mu_{s \times r}^2)/S$ is the estimate of the raters' variance and $\sigma_{s \times r}^2 = \mu_{s \times r}^2$ is the estimate of the variance of the interaction between raters and subjects. This latter is the component of the variance that cannot be associated only to raters or only to subjects and, hence, it is associated to the interaction between the two. In the case of the data used in this work, the α values are as follows: 0.08 for Openness, 0.14 for Conscientiousness, 0.28 for Extraversion, 0.19 for Agreeableness and 0.24 for Neuroticism.

The problem left open is whether the α values above can be considered acceptable or not. According to a study presented in [38] - to the best of our knowledge, the most extensive investigation of the problem so far - the median values of α over 9 articles on the perception of the Big-Five at zero acquaintance (the same situation as this article) are as follows: 0.07 for Openness, 0.13 for Conscientiousness, 0.32 for Extraversion, 0.03 for Agreeableness and 0.07 for Neuroticism. The comparison with the α values for this article (see above) confirms that the agreement level observed in this work is compatible with the agreement levels observed and accepted in the zero acquaintance personality perception literature [38], [39].

Category	**Name**	**d**	**Short Description**
Color	HSV statistics	5	Average of S channel and standard deviation of S, V channels [41]; *circular variance* in HSV color space [42]; *use of light* as the average pixel intensity of V channel [43]
	Emotion-based	3	Measurement of *valence*, *arousal*, *dominance* [41], [44]
	Color diversity	1	Distance w.r.t a uniform color histogram, by Earth Mover's Distance (EMD) [43], [41]
	Color name	11	Amount of *black*, *blue*, *brown*, *green*, *gray*, *orange*, *pink*, *purple*, *red*, *white*, *yellow* [41]
Composition	Edge pixels	1	Percentage of pixels classed as edge by the Canny detector [45]
	Level of detail	1	Number of regions (after mean shift segmentation) [46], [47]
	Average region size	1	Average *size* of the regions (after mean shift segmentation) [46]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [43], [41]
	Rule of thirds	2	Average of S,V channels over inner rectangle [43], [41]
	Image size	1	Size of the image [43], [45], [48]
Textural Properties	Gray distribution entropy	1	Image entropy [45]
	Wavelet based textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on HSV channels [43]
	Tamura	3	Amount of *coarseness*, *contrast*, *directionality* [49]
	GLCM - features	12	Amount of *contrast*, *correlation*, *energy*, *homogeneousness* for each HSV channel [41]
	GIST descriptors	24	Output of GIST filters for scene recognition [50].
Faces	Faces	1	Number of faces (extracted manually)

TABLE III: Synopsis of the features. Every image is represented with 82 features split in four major categories: Color, Composition, Textural Properties, and Faces. This latter is the only feature that takes into account the picture's content.

IV. Feature Extraction

The goal of this work is to map the favorite pictures of Flickr users into personality traits. The features adopted in this work focus on the contributions of *Computational Aesthetics* (see Section II) because these have been designed to account for the properties that make pictures visually appealing. The main assumption behind this choice is that pictures are often tagged as favorite for personal reasons (e.g., they show friends and relatives or are related to fond memories) [6], but the raters cannot access these motivations and can only access the appearance of the pictures. Popular feature extraction techniques like, e.g., SIFT and HOG have not been considered because they were originally conceived for other purposes, even if they have been shown to be effective in some tasks related to CA (see Section II). Furthermore, one of the main goals of this work is to show the very feasibility of a task like mapping favorite pictures into attributed traits. In this respect, the exploration of a wider spectrum of features can come at a later stage of the investigation.

A synopsis of the features adopted in this work is available in Table III. The features cover a wide, though not exhaustive, spectrum of visual characteristics and are grouped into three main categories: *color*, *composition* and *textural properties*. This follows the taxonomy proposed in [41], but it excludes the *content* category to make the process more robust with respect to the wide semantic variability of Flickr images. The only exception is a feature that counts the number of faces because these are ubiquitous in the images and, furthermore, the human brain is tuned to their detection in the environment [51].

A. Color

The feature extraction process represents colors with the HSV model, from the initials of *Hue*, *Saturation* and *Value* (this latter is often referred to as *Brightness*). This section describes features related to colors and their use.

HSV statistics: These features account for the use of colors and are based on statistics collected over H,S and V pixel values observed in a picture (see Figure 2). The H channel provides information about color diversity through its *circular variance* R [42]:

$$\begin{aligned} A &= \sum_{k=1}^{K}\sum_{l=1}^{L}\cos H_{kl}, \quad B = \sum_{k=1}^{K}\sum_{l=1}^{L}\sin H_{kl} \\ R &= 1 - \frac{1}{KL}\sqrt{A^2 + B^2} \end{aligned}$$

where H_{kl} is the Hue of pixel (k, l), K is the image height and L is the image width. On the S and V channels we compute the average and standard deviation; in particular, the average Saturation indicates chromatic purity, while the average over the V channel is called *use of light* and corresponds to a fundamental observation of image aesthetics, i.e. that underexposed or overexposed pictures are usually not considered aesthetically appealing [43]. The average of the Hue was not calculated because it cannot be associated to an intensity attribute (low, high), being an angular measure. Figure 2 provides examples of how the pictures change according to HSV statistics.

Emotion-based: Saturation and Brightness can elicit emotions according to the following equations resulting from psychological studies (*Valence*, *Arousal* and *Dominance* are dimensions commonly adopted to represent emotions) [44]:

$$\text{Valence} = 0.69 \cdot \bar{V} + 0.22 \cdot \bar{S} \quad (4)$$

$$\text{Arousal} = -0.31 \cdot \bar{V} + 0.60 \cdot \bar{S} \quad (5)$$

$$\text{Dominance} = -0.76 \cdot \bar{V} + 0.32 \cdot \bar{S} \quad (6)$$

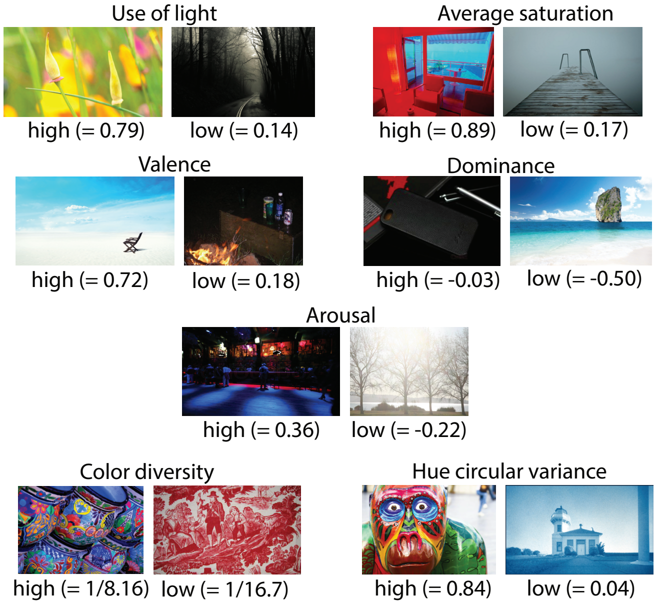

Fig. 2: The figure shows examples of how the visual properties of a picture change according to several color-related features.

where $\bar{V}$ and $\bar{S}$ are the averages of V and S over an image, respectively. See Figure 2 for examples of pictures with different levels of Valence, Arousal and Dominance.

Color diversity (Colorfulness): The feature distinguishes multi-colored images from monochromatic, sepia or low-contrast pictures. Following the approach in [43], the image under analysis is converted in the CIELUV color space, and its color histogram is computed; this representation is compared (in terms of Earth Mover's Distance) with the histogram of an ideal image where the distribution over the colors is uniform, that is, an histogram where all the bins have the same value (Figure 2 shows examples of pictures with different colorfulness).

Color name: Every pixel of an image can be assigned to one of the following classes identified in [52]: *black*, *blue*, *brown*, *grey*, *green*, *orange*, *pink*, *purple*, *red*, *white* and *yellow*. The sum of pixels across the classes above accounts not only for how frequently the colors appear in an image, but also for the style of a photographer. The classification of the pixels is performed using the algorithm proposed in [53] and it mimics the way humans label chromatic information in an image. The fractions of pixels belonging to each of the classes above are used as features.

B. *Composition*

The term *composition* refers to the organization of visual elements across an image regardless of its actual content. The features described in this section aim at capturing such an aspect of the pictures.

Edge pixels: The structure of an image depends, to a significant extent, on the edges, i.e. on those points where the image brightness shows discontinuities. Therefore, the feature extraction process adopts the Canny detector [45] to identify the edges and calculate the fraction of pixels in an image that lie on an edge (see Figure 3 for an example of edge extraction in an image).

Level of detail: images can be partitioned or *segmented* into multiple *regions*, i.e. sets of pixels that share common visual characteristics. The feature extraction process segments the images using the EDISON implementation [46] of the mean shift algorithm [47], and provides two features: i) the number of segments, accounting for the fragmentation of the image, and ii) the normalized average extension of the regions, that is, the mean area of the regions divided by the area of the whole image. On average, the more the details, the more the segments (see Figure 3).

Low depth of field (DOF) indicator: An image with low depth of field corresponds to a shot where the object of interest is sharper than the background, drawing the attention of the observer [41], [43] (see Figure 3). To detect low DOF, it is assumed that the object of interest is central; the image is thus decomposed into wavelet coefficients (see the next section), which measure the frequency content of a picture: in particular, high frequency coefficients (formally, level 3 as used in the notation of Eq. (10)) encode fine visual details. The low DOF indicator calculates the ratio of the high frequency wavelet coefficients of the inner part of the image against the whole image is calculated. In specific, the image is divided into 16 equal rectangular blocks $M1, \ldots M16$, numbered in row-major order. Let $w_3 = w_3^{HL=v}, w_3^{LH=h}, w_3^{HH=d}$ denote the set of wavelet coefficients in the high frequency of the hue image I_H. The low DOF indicator DOF$_H$ for hue is computed as follows,

$$\text{DOF}_H = \frac{\sum_{(k,l)\in M_6\cup M_7\cup M_{10}\cup M_{11}} w_3(k,l)}{\sum_{i=1}^{16}\sum_{(k,l)\in M_i} w_3(k,l)} \quad (7)$$

High DOF$_H$ indicates an apparent low depth of field in the image. The low depth of field indicator for the Saturation and the Brightness channels is computed similarly on the correspondent image channels. Figure 3 shows the difference between images with different Depth of Field.

The rule of thirds: Any image can be ideally divided into nine blocks - arranged in a 3×3 grid - by two equally-spaced horizontal lines and two equally-spaced vertical lines. The *rule of thirds* is a photography composition guideline that suggests to position the important visual elements of a picture along such lines or at their intersections. In other words, it suggests where the most salient objects should lie in the image. The rule of thirds feature in image aesthetics simplifies the photographic technique, analyzing the central block of the image by keeping the average values of Saturation and Brightness [41], [43]:

$$f_S = \frac{9}{KL}\sum_{k=K/3}^{2K/3}\sum_{l=L/3}^{2L/3} S_{kl} \quad (8)$$

where K is the image height, L is the image width and S_{kl} is the Saturation at pixel (k,l). A similar feature f_V can be calculated for the Brightness.

Image size: the size of the image is calculated as the total number of pixels and used as feature.

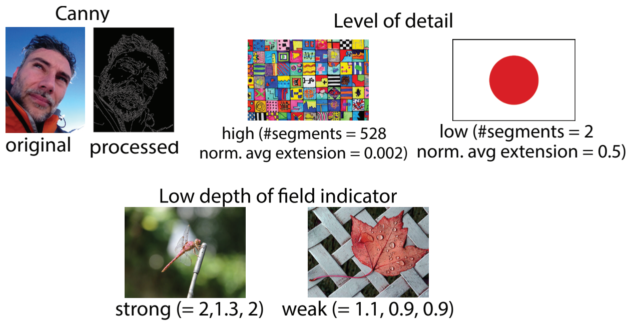

Fig. 3: The figure shows the effect of the Canny algorithm and an example of the visual properties associated to Level of Detail and Low Depth of Field.

C. Textural Properties

A texture is the spatial arrangement of intensity and colors in an image or in an image region. Textures capture perceptual aspects (e.g., they are more evident in sharp images than in blurred ones) and provide information about the subject of an image (e.g., textures tend to be more regular in pictures of artificial objects than in those of natural landscapes). The features described in this section aim at capturing textural properties.

Entropy: The entropy serves as a feature to measure the homogeneousness of an image. The image is first converted into gray levels; then, for each pixel, the distribution of the gray values in a neighborhood of 9×9 pixels is calculated (that is, the gray level histogram of the patch) and the entropy of the distribution is computed. Finally, all the entropy values are summed, and divided by the size of the image. The more the intensity tends to be uniform across the image, the lower will be the entropy (see Figure 4 for the impact of Entropy on visual characteristics). In the below expression, P_i is the probability that the difference between two adjacent pixels is equal to i, and Log_2 is the base 2 logarithm.

$$E = -\sum_i P_i Log_2 P_i \tag{9}$$

Wavelet textures: Daubechies wavelet transform can measure the spatial smoothness/graininess in images [43], [41]. The 2D Discrete Wavelet Transform (2D-DWT) of an image aims at analyzing its frequency content, where high frequency can be associated intuitively to high edge density. The output of a 2D-DWT can be visualized as a multilevel organization of square patches (see Figure 5). Each level corresponds to a given frequency analysis of the original image. In the first level of decomposition, the image is separated into four parts. Each of them has a quarter size of the original image, and a label. The upper left part is labeled LL (LowLow) and is a low-pass version of the original image. The vertical LH (LowHigh), horizontal HL (HighLow) and diagonal HH (HighHigh) parts can be assumed as images where vertical, horizontal and diagonal edges at the finest scale are highlighted. We can call them *edge images* at level 1. The subdivision can be further applied to find coarser edges, as the figure shows, performing again the wavelet transform to the coarser (LL coefficients) version at half the resolution, recursively, in order to further decorrelate neighboring pixels of the input image. The Daubechies wavelet transform is a particular kind of wavelet transform, explicitly suited for compression and denoising of images.

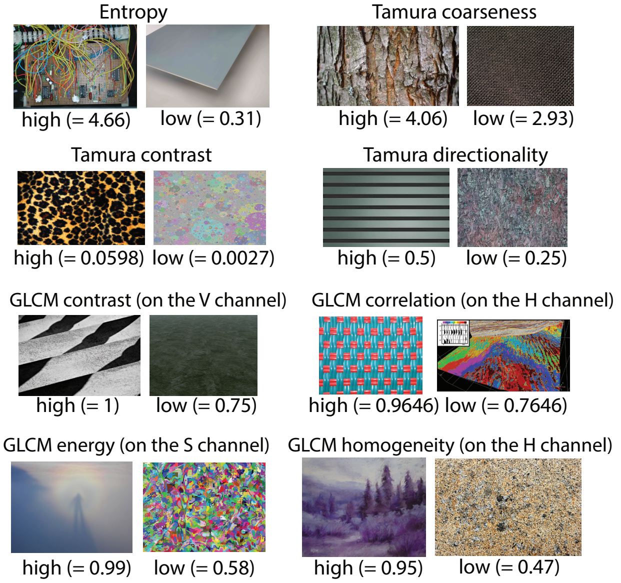

Fig. 4: The figure shows examples of pictures where the value of the textural features is high and low.

The feature extraction process computes a three-level wavelet transform on H, S and V channels separately. At each level, we have three parts which represent the edge images, called w_i^h, w_i^v and w_i^d, where $i \in 1, 2, 3$, $d = HH$, $h = HL$ and $v = LH$, to resemble the kind of edges that are highlighted (diagonal, horizontal, vertical, respectively). The wavelet features are defined as follows:

$$wf_i = \frac{\sum_{k,l} w_i^h(k,l) + \sum_{k,l} w_i^v(k,l) + \sum_{k,l} w_i^d(k,l)}{(|w_i^h| + |w_i^v| + |w_i^d|}, \tag{10}$$

for a total of 9 features (three levels for each of the three channels). The values k, l span over the spatial domain of the single w taken into account, and the operator $|\cdot|$ accounts for the spatial area of the single w. The corresponding wavelet features of saturation and brightness images have been computed similarly. In other words, for each color space channel and wavelet transform level, we average the values of the high frequency coefficients. We extracted three more features by computing the sum of the average wavelet coefficients over all three frequency level for each HSV channel (see Figure 5).

Tamura: In [49], six texture features corresponding to human visual perception have been proposed: coarseness, contrast, directionality, line-likeness, regularity and roughness. The first three have been found particularly important, since they are tightly correlated with human perception, and have been considered in this work. They are extracted from gray level images.

Coarseness: The feature gives information about the size of texture elements. A coarse texture contains a small number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2516994, IEEE Transactions on Affective Computing

of large texels, while a fine texture contains a large number of small texels. (see Figure 4). The coarseness measure is computed as follows. Let X an $I \times J$ matrix of values $X(i,j)$ that can for instance be interpreted as gray values:

1) For every point (i,j) calculate the average over neighbourhoods. The size of the neighbourhoods are powers of two, e.g.: $1 \times 1, 2 \times 2, 4 \times 4, \ldots, 32 \times 32$:

$$A_k(i,j) = \frac{1}{2^{2k}} \sum_{n=1}^{2^{2k}} \sum_{m=1}^{2^{2k}} X(i - 2^{k-1} + n, j - 2^{k-1} + m) \tag{11}$$

2) For every point (i,j) calculate the difference between the not overlapping neighbourhoods on opposite sides of the point in horizontal and vertical direction:

$$E_k^h(i,j) = |A_k(i + 2^{k+1}, j) - A_k(i - 2^{k-1}, j)| \tag{12}$$

and

$$E_k^v(i,j) = |A_k(i, j + 2^{k+1}) - A_k(i, j - 2^{k-1})| \tag{13}$$

3) At each point (i,j) select the size leading to the highest difference value:

$$S(i,j) = \arg\max_{k=1\ldots5} \max_{d=h,v} E_k^d(i,j) \tag{14}$$

4) Finally take the average over 2^S as a coarseness measure for the image:

$$F_{crs} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} 2^{S(i,j)} \tag{15}$$

Contrast: It stands, in rough words, for texture quality. It is calculated by

$$F_{con} = \frac{\sigma}{\alpha_4^z} \quad with \quad \alpha_4 = \frac{\mu_4}{\sigma_4} \tag{16}$$

where $\mu_4 = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} (X(k,l) - \mu^4)$ is the fourth moment about the mean μ, σ^2 is the variance of the gray values of the image, and z has experimentally been determined to be $\frac{1}{4}$. In practice, contrast is influenced by the following two factors: range of gray-levels (large for high contrast), polarization of the distribution of black and white on the gray-level histogram (polarized histogram for high contrast). For an example, see Figure 4.

Directionality: It models how polarized is the distribution of edge orientations. High directionality indicates a texture where the edges are homogeneously oriented, and conversely. Given the directions of all the edge pixels, the entropy E of their distribution is calculated; the directionality becomes then $1/(E+1)$. Textures with edges oriented along a single direction will be distributed as a single peak, thus $E = 0$, and maximal directionality (=1). Conversely, pictures with edges whose orientation is distributed in a uniform manner will have low directionality (~ 0). For an example, see Figure 4.

Gray-Level Co-occurrence Matrix (GLCM) features: The GLCM is a matrix where the element (i,j) is the probability $p(i,j)$ of observing values i and j for a given channel (H, S or V) in the pixels of the same region W. In the feature extraction process, W includes a pixel and its right neighbor and, therefore, the GLCM includes the probabilities of observing one pixel where the value j is at the right of a pixel where the value is i. The GLCM serves as a basis for calculating several features, each obtained separately over the H, S and V channels [54]:

Contrast: It is the average value of $(i-j)^2$, the square difference of values observed in neighboring pixels: $C = \sum_{i,j=0}^{L-1} (i-j)^2 p(i,j)$, where L is the number of possible values in a pixel. The value of C ranges between 0 (uniform image) and $(L-1)^2$ (see Figure 4 for examples of pictures with high and low contrast).

Correlation: It is the coefficient that measures the covariation between neighboring pixels:

$$\sum_{i,j=0}^{L-1} \frac{(i-\mu)(j-\mu)p(i,j)}{\sigma^2}, \tag{17}$$

where $\mu = \sum_{i,j=0}^{L-1} ip(i,j)$, and $\sigma^2 = \sum_{i,j=0}^{L-1} p(i,j)(i-\mu)^2 + \sum_{i,j=0}^{L-1} p(i,j)(j-\mu)^2$. The correlation ranges in the interval $[-1, 1]$ (see lower part of Figure 4).

Energy is the sum of the square values of the GLCM elements: $\sum_{i,j=0}^{L-1} p(i,j)^2$. If an image is uniform, the energy is 1 (see Figure 4).

Homogeneity is a measure of how frequently neighboring pixels have the same value (see Figure 4):

$$H = \sum_{i,j=0}^{L-1} \frac{p(i,j)}{1 + |i-j|} \tag{18}$$

The feature tends to be higher when the elements on the diagonal of the GLCM are larger (see Figure 4).

Spatial Envelope (GIST): it is a low dimensional representation of a scene that relies on Gabor Filters to capture a set of perceptual dimensions, namely *naturalness*, *openness*, *roughness*, *expansion*, *ruggedness* [50]. The outputs of the GIST filters are used as features.

D. Number of Faces

All features presented so far are content independent, i.e. do not take into account what the images show. This feature is the only exception to such an approach because human faces are frequently portrayed in Flickr pictures and, furthermore, there are neural pathways that make the human brain particularly sensitive to faces [51]. In this work, the number of faces is calculated manually for each of the 60,000 pictures of the dataset. Every visible face was counted, irrespectively of its scale, pose, size and occlusion. Facial expressions were not taken into account. Automatic face detectors were avoided because they are not sufficiently robust to deal with the variability of favorite pictures. These often portray people in unusual poses and a preliminary analysis shows that the Viola-Jones detector [55] identifies only 70% of the faces in the corpus. This introduces noise difficult to model and quantify.

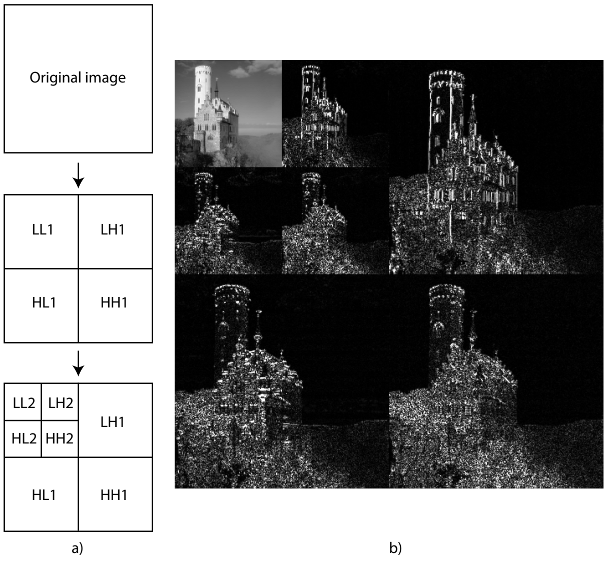

Fig. 5: The figure shows how the wavelet decomposition works.

V. Inference of Personality Traits

This section presents the regression approaches adopted to map the features described in Section IV into personality traits. The regression is performed separately for the Big-Five traits because these result from the application of Factor Analysis to behavioural data [56] and, therefore, they are indepedent. The goal of the experiments is to infer the personality traits - both self-assessed and attributed - from the multiple pictures that a user tags as favorite. Hence, Multiple Instance Regression (MIR) [9], [10] appears to be the most suitable computational framework because it addresses problems where there are multiple instances (the favorite pictures of a Flickr user) for a *bag* (the Flickr user) associated with one value (the score of the Flickr user for a particular trait). Furthermore, MIR approaches can deal with cases where only a subset of the bag instances actually account for the value to be predicted, or when all the bag instances have a role in its definition [57]. In this scenario the latter hypothesis could be the most reasonable, but no experiments have been carried out to actually pinpoint which image(s) is (are) more significant to determine the personality score.

Before applying the regression approaches (both the baselines and the proposed approaches), the features are discretized using $Q = 6$ quantization levels (Q values between 3 and 9 were tested, but no significant performance differences were observed: in any case, $Q = 6$ gave a slightly better performance). The intervals corresponding to the levels are obtained by splitting the range of each feature in the training set into Q uniform, non-overlapping intervals. In this way, the 82 features describing each picture (see Section IV) can be interepreted as counts. The main motivations for not clustering features individually are, on one hand, to remain dataset independent (different datasets result into different clusters) and, on the other hand, to limit computational costs when the number of features is high.

In the following, each Flickr user u corresponds to a bag of favorite pictures B^u ($u = 1, \dots, 300$) and five traits y_p^u (p = O, C, E, A, N). The trait values predicted by the MIR approaches are denoted with $\hat{y}_p^u$. The notation does not distinguish between self-assessed and attributed traits because the two cases are treated independently of each other. The element $\mathcal{C}_z^t$ of the feature matrix $\mathcal{C}$ is the value of feature z ($z = 1, \dots, 82$) for picture t ($t = 1, \dots, 6 \times 10^4$). Finally, $v(t)$ is a function that takes as input a picture index t and returns the corresponding user-index u.

A. Baseline Approaches

The general MIR formulation is NP-hard [9] and this requires the adoption of simplifying assumptions. The most common one is to consider that each bag includes a *primary instance* that is sufficient to predict correctly the bag label. In the experiments of this work, this means that each bag B^u includes only one picture t - with $v(t) = u$ - that should be fed to the regressor to obtain as output the trait score y_p^u (for a given p). However, the primary instance cannot be known a-priori for a test bag. Furthermore, the bags of this work include 200 pictures and, therefore, using only one of them means to neglect a large amount of information. For this reason, this work adopts different baseline MIR approaches, more suitable for the PsychoFlickr data. Essentially, these baselines assume that each instance carries a role in determining the value of the bag label, in line with the assumptions of [57]. Furthermore, the baseline approaches include a regressor that always predicts the average value of the traits as per estimated over the training set.

Baseline: The simplest baseline approach consists in predicting always the average of a given trait in the training set. The reason for using the average is that this is the value that minimizes the Root Mean Square Error when a regressor always predicts the same value.

Naive-MIR [10]: The simplest approach consists in giving each picture of a bag B^u as input to the regressor. As a result, there is a predicted score $\hat{y}_p^u(t)$ for each picture t such that $v(t) = u$. The final trait score prediction $\hat{y}_p^u$ is the average of the $\hat{y}_p^u(t)$ values. The main assumption behind the Naive-MIR is that all the pictures of a bag carry task-relevant information and, therefore, they must all influence the predicted score $\hat{y}_p^u$.

cit-kNN [58]: Given a test bag B^u, this methodology adopts the minimal Hausdorff distance [59] to identify, among the training bags, both its R nearest neighbors and its C-nearest citers (the training bags that have B^u among their C nearest neighbors). The predicted score $\hat{y}_p^u$ is then the average of the scores of both R nearest training bags and C nearest citers training bags. The approach does not include an actual regression step, but still maps a test bag B^u into a continuous predicted score $\hat{y}_p^u$.

Clust-Reg [60]: The Clust-Reg MIR includes three main steps. The first consists in clustering all the pictures of the training bags using a kmeans, thus obtaining C centroids c_j in the feature space ($j = 1, \dots, C$). The second step considers all the images of a bag B^u that belong to a cluster c_k and averages them to obtain a prototype k. The same task is performed for all training bags B^u. In this way each training bag is represented by C prototypes at most. The third step trains C regressors r_i $(i = 1, \ldots, C)$ - each obtained by training the model in Section V-C over all prototypes corresponding to one of the C clusters - and identifies r_k, the one that performs best on a validation set. At the moment of the test, r_k is applied to prototype k of a test bag B^u to obtain the predicted score $\hat{y}^u_p$. The regressor operates on a prototype k that represents only the test-bag pictures surrounding the centroid c_k. Therefore, the Clust-Reg implements the assumption that only a fraction of the pictures carry task-relevant information.

B. Latent representation-based methods

The baseline approaches presented in Section V-A operate in the feature space where the pictures are represented. Such a scheme is not suitable when the number of instances per bag is large - like in the experiments of this work - because it is not possible to know a-priori what are the samples that carry information relevant to the task. In this respect, the main novelty and advantage of the approaches presented in this section consist in mapping the pictures of the training bags onto an intermediate latent space $\mathcal{Z}$. This latter is expected to capture most of the information necessary to perform the trait scores' prediction. While being proposed for the experiments of this work, the new MIR approaches can be applied to any problem where the number of instances per bag is large.

Topic-Sum: The main assumption of this approach is that the pictures of a test bag B^u distribute over topics, i.e. over frequent associations of features that can be learnt from the images of the training bags. Such an approach is possible because the features extracted from the pictures have been quantized and the feature vectors can then be considered as vectors of counts or Bags of Features (see beginning of Section V). The topic model adopted in this work is the Latent Dirichlet Annotation (LDA) [61]. The LDA expresses the topics as probability distributions of features $p(\mathcal{C}^u_z|k)$, with $k = 1, \ldots, K$ and $K << D$ (D is the dimension of the feature vectors). Once the topics are learnt from the training bags, a test bag B^u can be expressed as a mixture of topics:

$$p(B^u) = \sum_{k=1}^{K} p(\mathcal{C}^u_z|k)p(k|u), \tag{19}$$

where $\mathcal{C}^u_z$ is the feature matrix of the images belonging to B^u, and the coefficients $p(k|u)$, called *topic proportions*, measure how frequently the topics appear in test bag B^u. The regressor of Section V-C is trained over vectors where the components correspond to the topic proportions of the training bags. At the moment of the test, the topic proportions of a test bag B^u are fed to the resulting regressor to obtain $\hat{y}^u_p$.

Gen-LDA: This approach learns a LDA model [61] from the pictures of each training bag and then it fits a Dirichlet distribution $p(\cdot; \alpha^u)$ on the resulting topic proportions (see description of Topic-Sum above). The parameter vectors α^u are then used to train the regressor of Section V-C. At the test stage, the α^u parameters of the Dirichlet distribution corresponding to the topic proportions in test bag B^u are then given as input to the regressor to predict the trait scores.

Gen-MoG: This approach learns a Mixture of Gaussians (with diagonal covariances) with C components from the pictures of the training bags, then it considers all the images of a test bag B^u to estimate the following for $c = 1, \ldots, C$:

$$Z^u(c) = \sum_{t:v(t)=u} p(c|t) \tag{20}$$

where $p(c|t)$ is the *a-posteriori* probability of component c in the mixture when the picture is t. The values $Z^u(c)$ are given as input to the regressor to predict the trait scores. Compared to the *Multiple Instance Cluster Regression* [60], a similar methodology, the main difference of the Gen-MoG is that an instance is softly attributed to all the components of the Mixture of Gaussians through the probabilities $p(c|t)$.

CG: This approach is based on the *Counting Grid* (CG) [62], a recent generative model which embeds BoF representations like those used in this work in d-dimensional manifolds. The CG allows one to map each picture t of the training bags onto a 2-dimensional grid lying on a smooth manifold, i.e. a manifold where close positions correspond to close images in the original feature space. The grid has E_1 rows and E_2 columns and, typically, $E_1 \times E_2 << N$, where N is the number of pictures in the bag. After such a training step, every test bag B^u is projected onto the same manifold and becomes a set of locations $L^u = \{\ell^t\}$ on the 2-dimensional grid, i.e. a distribution of the test bag pictures over the grid. The distribution - an $E_1 \times E_2$ dimensional vector - is first smoothed by averaging over a 5×5 window and then given as input to the regressor of Section V-C to predict the trait scores.

C. Regression

All the methods above, except cit-kNN, require a regressor to predict the trait scores. The one adopted in the experiments of this work has the following form:

$$\hat{y}^u_p = \sum_{k=1}^{K} \beta_k x^u_k, \tag{21}$$

where the $\beta = (\beta_1, \ldots, \beta_K)$ are the regressor parameters and the values x^u_k are the parameters that, according to the different methods presented above, represent a test bag B^u (e.g., the Dirichlet distribution parameters in Gen-LDA). The β_k are estimated by minimising the Mean Square Error $E(\beta)$:

$$E(\beta) = \sum_{u=1}^{U} \left(y^u_p - \sum_{k=1}^{K} \beta_k x^u_k\right)^2, \tag{22}$$

where U is the number of training bags. The problem was regularized using LASSO [63], an approach which constrains the L_1 norm of the least squares solution, thus acting as model selection method, by enforcing the sparsity on coefficients β. The regularizer in the Lasso estimate is simply expressed as a threshold on the L1-norm of the weight β:

$$\sum_k |\beta_k| \leq t \tag{23}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2516994, IEEE Transactions on Affective Computing

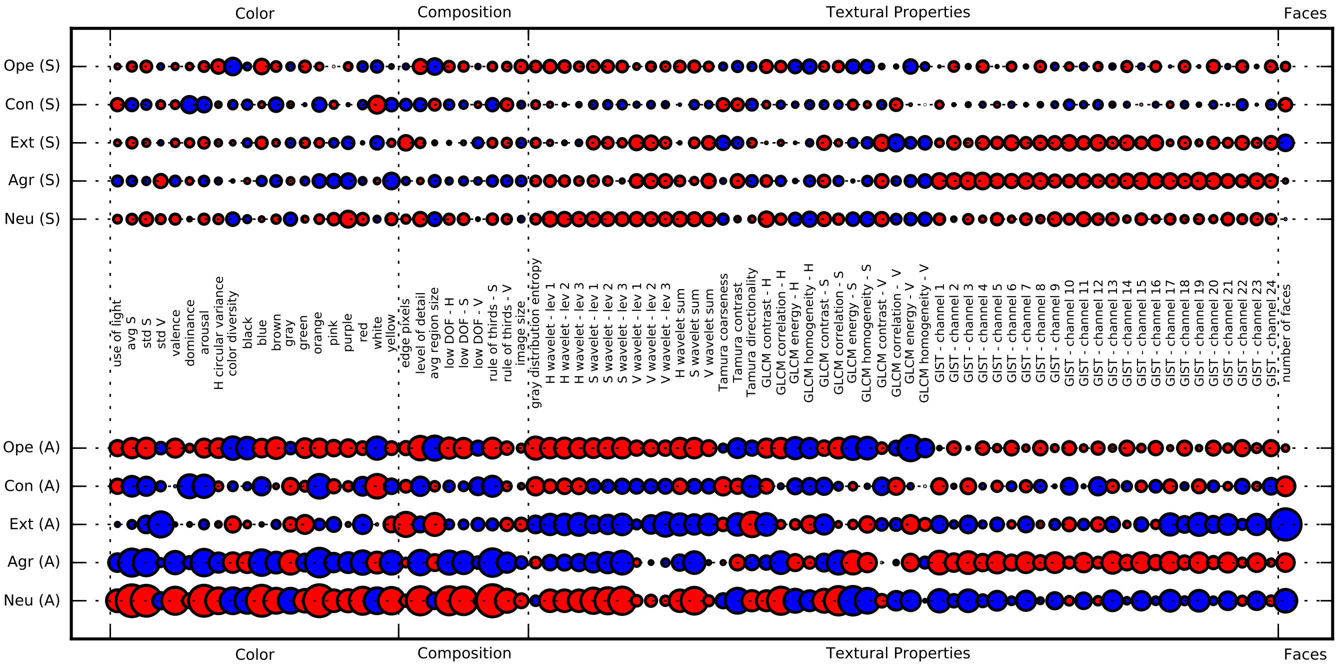

Fig. 6: The plot shows the Spearman Correlation Coefficients ρ between features and traits, both self assessed (upper part) and attributed (lower part). The bubbles are blue when the correlation is positive and red when it is negative. The largest bubbles correspond to $|\rho| = 0.55$ while the smallest ones correspond to $|\rho| \simeq 0.00$. Correlations for which $|\rho| \geq 0.12$ are statistically significant at level 0.05.

This term acts as a constraint that has to be taken into account when minimizing the error function. By doing so, it has been proved that (depending on the parameter t), many of the coefficients β_j become exactly zero [63]. This is particularly relevant to the problem of this work, because it allows one to model the fact that not all the features are correlated with a given trait.

VI. Experiments and Results

This section presents first a correlational analysis aimed at showing the relationship between features and traits and then the regression experiments performed in this work.

A. Correlational Analysis

Given the vectors $\vec{x}_i^{(u)}$ ($i = 1, \ldots, 200$) extracted from the 200 images favored by a user u, it is possible to use their average $\vec{x}^{(u)}$ as a representative of the corresponding bag. Figure 6 shows the covariation, measured with the Spearman Coefficient ρ, of $\vec{x}^{(u)}$ components and traits, both self-assessed and attributed (blue and red bubbles account for positive and negative values of ρ, respectively). The size of the bubbles, proportional to the absolute value of ρ, represents the strength of the relationship between a given feature and a trait: " *[...] the sign of the correlation coefficient has no meaning other than to denote the direction of the relationship. Correlations of* 0.75 *and* −0.75 *signify exactly the same degree of relationship. It is only the direction of that relationship that is different*" [40].

The covariation is high for the attributed traits, but limited for the self-assessments. In particular, ρ is statistically significant (at level 0.05) for 48.5% of the features in the case of the attributed traits and only for 8.3% of the features in the case of self-assessments. This suggests that the visual properties of the images covariate with the impression that the judges develop about the Flickr users, but do not account for the self-assessments that the users provide. For this reason, the rest of this section focuses on the attributed traits.

Color properties (see Section IV-A) covariate to a significant extent with all traits and, in particular, with Agreeableness and Neuroticism. However, the properties that are positively correlated with one trait tend to be correlated negatively with the other and conversely. In other words, Agreeableness and Neuroticism seem to be perceived as complementary with respect to color characteristics. This applies, e.g., to average saturation ($\rho = 0.40$ for Agreeableness and $\rho = -0.55$ for Neuroticism), percentage of orange ($\rho = 0.45$ and $\rho = -0.56$), blue ($\rho = 0.36$ and $\rho = -0.52$) and red ($\rho = 0.30$ and $\rho = -0.40$) pixels, arousal ($\rho = 0.38$ and $\rho = -0.52$) and valence ($\rho = 0.27$ and $\rho = -0.40$). Overall, the judges appear to assess as high in Agreeableness users that like images eliciting pleasant emotions and showing pure colors. Conversely, the judges consider high in Neuroticism people that like images stimulating intense, unpleasant emotions and contain colors with low saturation.

Complementary assessments can be observed for Openness and Conscientiousness as well when it comes to the relationship with compositional properties (see Section IV-B). The features of this category that covariate most with the two traits

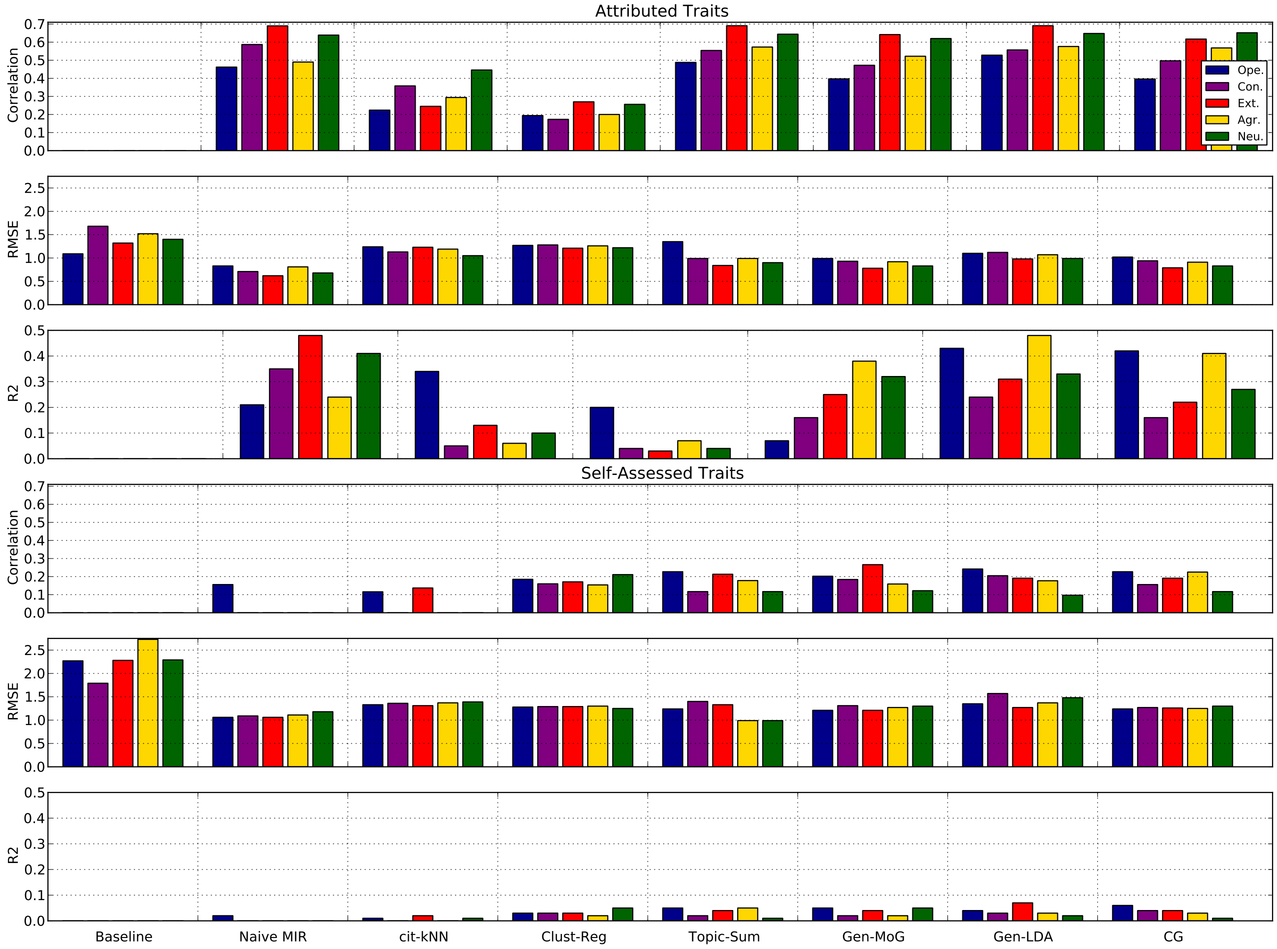

Fig. 7: The figure shows APP (upper charts) and APR (lower charts) performances in terms of Spearman Correlation Coefficient between actual and predicted traits, Root Mean Square Error (RMSE) and R^2 metric. In the case of the correlations, missing bars correspond to non statistically significant values.

are rule of thirds ($\rho = -0.21$ for Openness and $\rho = 0.22$ for Conscientiousness) and level of detail ($\rho = -0.30$ and $\rho = 0.19$). Therefore, unconventional compositions displaying a few details tend to be associated with high Openness (the trait of creativity and artistic inclinations) while conventional compositions with many details tend to be associated with high Conscientiousness (the trait of reliability and thoroughness).

Textural features (see Section IV-C) appear to covariate with the perception of most traits, especially when it comes to the properties of the *Gray-Level Co-occurrence Matrix* (see Section IV-C). In the case of Openness, the highest correlations are observed for exposure (measured in terms of brightness energy) and image homogeneousness (measured in terms of gray distribution entropy). The covariation is positive for the former ($\rho = 0.35$) and negative for the latter ($\rho = -0.27$). Therefore, people that like pictures with homogeneous illumination and uniform textural properties tend to be perceived as higher in Openness. For Conscientiousness, the covariation ($\rho = 0.23$) is significant only for the Tamura directionality. Hence, there seems to be no relationship between the trait and textural properties. In contrast, several textural properties covariate with the attribution of Extraversion. High contrast in hue, meaning large color differences in neighboring pixels, and saturation, meaning chromatic purity, are associated with high Extraversion scores ($\rho = 0.26$ and $\rho = 0.21$, respectively). The same applies to Tamura contrast ($\rho = 0.25$) and directionality ($\rho = -0.33$) of the images.

The value of ρ for the number of faces, the only content related feature considered in this work, is statistically significant at 0.01 confidence level for all traits except Openness. The ρ value is negative for Conscientiousness ($\rho = -0.2$) and Agreeableness ($\rho = -0.17$) and positive for the other traits. Not surprisingly, the absolute covariation is particularly high ($\rho = 0.53$) for Extraversion, the trait of sociability and interest for others, and Neuroticism ($\rho = -0.28$), the trait of the difficulties in dealing with social interactions.

According to personality psychologists, the traits that people tend to perceive more clearly are Extraversion and Conscientiousness [33]. However, different data can make different traits more or less *available*, i.e. more or less accessible to human observers [13]. The correlational analysis shows that favorite pictures convey impressions more effectively for Agreeableness and Neuroticism than for the other traits. This seems to suggest that the raters develop an impression in terms of whether a person is overall nice (a typical characteristic of people high in Agreeableness) or not (a typical characteristic of people high in Neuroticism). This appears to be confirmed by the fact that the correlations for the two traits have often opposite sign, meaning that a person perceived to be neurotic is not perceived to be agreeable and conversely.

B. Experimental Setup

All the experiments of this work have been performed using a Leave-One-User-Out approach: the models are trained over all the pictures of PsychoFlickr except those tagged as favorite by one of the Flickr users included in the corpus (see Section III). The traits of these latter are then predicted using the excluded pictures as test set. The process is then iterated and, at each iteration, a different user is left out. The hyper-parameters of the methods introduced in Section V-A and Section V-B have been set through cross-validation: all parameter values in a search range were tested over a subset of the training set and the configurations leading to the highest performance were retained for the test. The main advantage of the setup above is that it allows the use of the entire corpus to measure the performance of the inference approaches while still preserving a rigorous separation between training and test set.

Hyper-parameters and search ranges for the methods described above are as follows: for the cit-kNN, number of nearest citers C and number of nearest neighbors R were searched in the ranges $[4, 10]$ and $[2, 8]$, respectively; for Clust-Reg and Gen-MoG, the number C of clusters was searched in the set $\{5, 10, 20, \ldots, 100\}$; for Topic-Sum and Gen-LDA, the number of topics K was searched in the set $\{50, 70, 90, 110, 130, 150\}$; for CG, the grid sizes were searched in the set $\{20 \times 20, 25 \times 25, \ldots, 65 \times 65\}$. Whenever there was no risk of confusion, the same symbol has been used for different hyper-parameters in different models. These ranges have been set by considering common works on object recognition for what concerns number of topics, number of clusters, and grid size [62], [61], and the original papers of cit-kNN [58] for what concerns the number of citers and neighbors.

C. Prediction Results

Figure 7 reports the results obtained with the regression methods described in Section V for both self-assessed and attributed traits. The performance is assessed with three different metrics, namely Spearman correlation coefficient between scores predicted automatically and scores resulting from the BFI-10 questionnaire (see Section III), Root Mean Square Error (RMSE) and R^2. The reason is that different performance metrics account for different aspects and only the combination of multiple metrics can provide a complete description of the results.

In line with the state-of-the-art of Personality Computing [12], APP results tend to be more satisfactory than APR ones. In the case of this work, the reason is that the judges are unacquainted with the users. Therefore, the pictures dominate the personality impressions that the judges develop and, as a result, the correlation between visual features and trait scores is higher (see Figure 6). Furthermore, the consensus across the judges is statistically significant (see Section III-B). These two conditions help the regression approaches to achieve higher performances. When the users self-assess their personality, they take into account information that is not available in the favorite pictures like, e.g., personal history, inner state, education, etc. Therefore, the correlation between visual features and trait scores is low. This does not allow the regression approaches to achieve high performances.

APP and APR performances are similar in terms of RMSE, but correlation and R^2 are better for APP than for APR. The probable reason is that, in the case of APP, the regressor tends to maintain the mutual relationships between personality scores, i.e., the regressor tends to predict higher scores for those subjects that tend to be rated higher by the assessors. This explains why the correlations are statistically significant (actual and predicted traits covariate to a statistically significant extent) and more satisfactory than R^2 and RMSE results.

According to personality psychology, "*[...] a compelling argument can be made for emphasizing comparisons among individuals, which we do in everyday life [...] and which is useful for practical purposes*" [64]. This means that what is important is not to predict the actual personality scores that individuals have been attributed, but to ensure that the subjects that have been attributed higher scores by the raters tend to be assigned higher scores by the regressor as well. In this respect, the Spearman Correlation Coefficient appears to be the performance metric that better fits the indications of personality psychology.

All approaches have been compared with a baseline that simply predicts the average of the trait values observed in the training set. The performance of the baseline is lower than the performance of all other approaches to a statistically significant extent (see Figure 7). The weakest approaches (cit-kNN and Clust-Reg) are those that make hard decisions to exclude part of the pictures in a test bag. This seems to suggest that all pictures carry task-relevant information and the most effective approach is to make soft decisions by combining complex generative models (e.g., LDA and CG) and sparsity control regressors (see Section V-C). This is the case of the best performing methods, namely Topic-Sum, Gen-MoG, CG and Gen-LDA (this latter has the best overall performance). The good performance of the Naive MIR further confirms that all images in a test bag contribute to influence the attributed traits and, hence, must be used for the regression. This observation has two possible explanations. The first is that all pictures influence the impression that each judge develops about the Flickr users. The second is that each judge is influenced by a different subset of pictures and the attributed traits - the average over the traits attributed individually by each judge - are therefore influenced by all pictures in a test bag.

For every trait, it is possible to split the range of the observed scores into quartiles. The performance of the regressors has been measured separately over subjects that fall in the top and bottom 25% of the observed scores and over the remaining subjects. Overall, the performance tends to be higher for

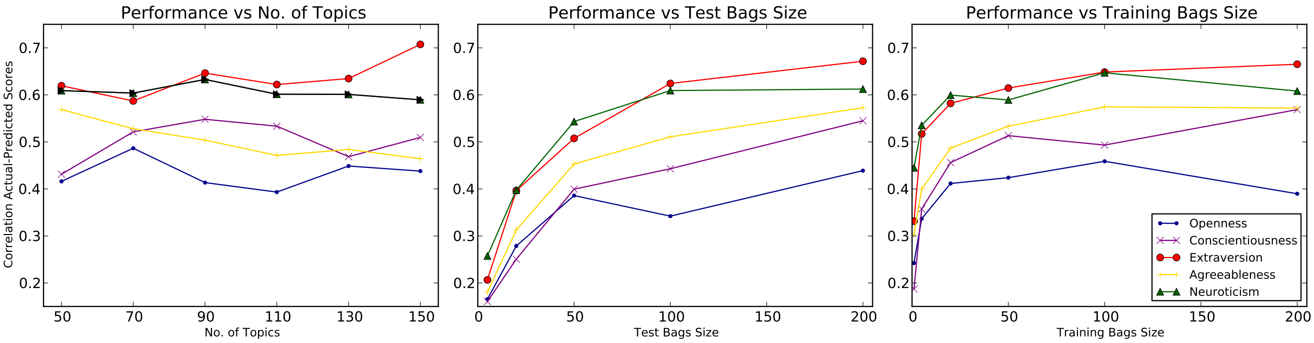

Fig. 8: From left to right, the first plot shows the performance as a function of the number of topics, the second and the third report the performance as a function of test and training bags size, respectively (the plots correspond to the APP results).

subjects that are closer to the extremes of the scales because these can be reached only when there is higher agreement between the raters, i.e., when the relationship between visual features and traits is more consistent. This means that the approach tends to be more effective when an individual is far from the average along one of the Big Five dimensions.

On average, when taking into account only the subjects in the extreme quartiles (top and bottom 25% of the observed scores), the correlation increases by 99.1% for Openness, by 118.5% for Conscientiousness, by 176.0% for Extraversion, by 44.1% for Agreeableness and by 122.5% for Neuroticism. In the case of R^2 the same figures amount to 339.4% (Openness), 483.3% (Conscientiousness), 861.3% (Extraversion), 117.0% (Agreeableness) and 434.1% (Neuroticism). The performance improves in terms of RMSE as well and decreases by 4.0% for Openness, by 9.1% for Conscientiousness, by 25.68% for Extraversion, by 4.82% for Agreeableness and by 20.73% for Neuroticism. Similar effects are observed for APR, but the changes in performance are less significant (all improvements are lower than 50%).

Different types of data let different traits to emerge with more or less evidence [13]. This is the reason why not all the traits are predicted with the same effectiveness. In the case of the attributed traits, the values of the shared variance α (see Section III-B) provide a first indication of this phenomenon: There is higher agreement for traits that emerge more clearly or, at least, are perceived to do so by the judges. As a result, the performance tends to be better for traits where α is higher. Extraversion is the best predicted dimension for both attributed and self-assessed traits, in line with the results of both Personality Computing [12] and Personality Psychology [33]. The reason is that this trait is the most socially oriented and, therefore, it leaves more traces in observable behaviour [13]. In the case of attributed traits, the performance tends to be higher than average on Neuroticism. To the best of our knowledge, the literature does not provide indications about, but it seems to be the effect of the high correlation of the trait with the use of certain colors (orange, blue and red) and chromatic purity, as well as with the emotions elicited by the images. These effects are among the strongest observed in the PsychoFlickr corpus (see Figure 6). The lowest performance corresponds to Openness. The main reason is probably that the judges seem to manifest high uncertainty in assessing the trait. This is evident in Figure 1, where the distribution for attributed Openness shows the highest peak in correspondence of the bin centered around zero. Similarly, Openness is the trait that corresponds to the lowest α (see Section III-B).

D. Number of Topics, Bag Size and Performance

This section analyses in more detail the application of Gen-LDA to the prediction of attributed traits, the case for which the experiments above show the best overall performance. The leftmost plot of Figure 8 shows the performance as a function of the number of LDA topics. The range is $[50, 150]$ because outside this interval the performance falls rapidly: having less than 50 topics will fuse together into a single topics features that probably are highly uncorrelated, losing in expressive power. In the other case, after 150 topics the method starts producing results with high variance, much probably because of initialization issues. In any case, no value of the number of topics appears to be optimal for all traits. For the best predicted traits (Extraversion and Neuroticism), the performance remains roughly constant or even grows with the number of topics. For the other traits, the performance reaches its maximum in correspondence of different numbers of topics and then it falls before reaching the 150 limit. A possible explanation is that there is no advantage in increasing the number of topics when the covariation between features and traits is lower (Figure 6 shows that Openness, Agreeableness and Conscientiousness are more weakly correlated with the features than the other two traits). An interesting future work could be that of capturing the features that actually play a strong role for the traits inference, and in particular checking if the features that had strong correlation with the traits are also important for their prediction. This could not be the case, since topic models in general evaluate the role of different features when considered in a joint fashion (that is, those features which concur to a particular topic), and not taken independently as in the correlation analysis.

The central and rightmost plots of Figure 8 show the relationship between performance and size of test and training set bags, respectively (the ranges are driven here by the number of available images per user, that is, 200). In both cases, the performance grows with the number of pictures, but statistically significant performances can still be achieved with small bag sizes, i.e. 5 in the case of the training set and 1 in the case of the test set. This is particularly important in view of applications dealing with users that tag only a few images as favorite.

VII. Conclusions

This work has proposed an approach for mapping pictures tagged as favorite into personality traits, both self-assessed and attributed. The results show that the approach is particularly effective in the case of attributed traits, i.e., in the case of the personality impressions that the pictures convey. While not necessarily corresponding to the actual traits of an individual, attributed traits are still important because they are predictive of important aspects of social life, including attitude of others [14] and social identity [15].

The motivations for tagging a picture as favorite are multiple and include social and affective aspects like, e.g., positive memories related to the content and bonds with the people that have posted the picture (see [6] for an extensive introduction and analysis). However, features expected to account for how visually appealing a picture is appear to be effective in the case of attributed traits. One possible explanation is that the raters do not know the motivations for which a picture has been tagged as favorite, but can still make an aesthetic judgment. Therefore, it is possible that the traits are assigned on the basis of how visually appealing the favorite pictures are and not on the basis of the users' motivations. Such an effect has been extensively observed in face-to-face interactions where people tend to attribute socially desirable characteristics to individuals they find attractive, a phenomenon known as "*what is beautiful is good*" [65].

The performance of the approach proposed in the experiments tends to be higher when attributed traits are closer to the extremes of the scales (see Section VI-C), i.e., when the subjects are far from the average along a given trait. The main probable reason is that these are the cases for which there is higher agreement between the raters (the extremes can be reached only when most raters agree) and, hence, there is a more consistent relationship between physical characteristics of the data at disposition (the features extracted from the pictures in this work) and traits.

The above suggests that the performances of APP approaches can be expected to increase when there is high agreement between raters. However, the literature shows that this does not happen in zero-acquaintance personality assessment studies, where the best that can be expected is that the raters simply agree beyond chance [38], [39]. In particular, Section III-B shows that the agreement between raters observed in this work is in line with the Personality Psychology literature, where it is considered acceptable. In other words, the Personality Psychology literature [38], [39] suggests that low agreement between raters is a characteristic of the APP problem and not the result of poor data collection practices.

One of the main consequences is that personality assessments tend to peak in the central part of the scales [66]. Figure 1 shows that the PsychoFlickr Corpus is in line with the Personality Psychology literature from this point of view as well. In an APP perspective, one possible solution is to limit the experimental work to subjects that are at the extremes of the scales (several works in the Personality Computing literature adopt such an approach [12]). However, this might lead to overestimate the performances and, in any case, it is not possible to know whether a subject is at the extremes of a trait without having performed a prediction first. In this respect, it is an open research problem to develop techniques that discriminate between the subjects for which the agreement between raters is high and those for which it is low.

When it comes to the prediction of self-assessments, the possibility of achieving satisfactory performances depends on "*Relevance (i.e., the environment must allow the person to express the trait) and Availability (i.e., the trait must be perceptible to others)*" [13]. In other words, just because an individual holds a particular trait, that does not mean that the trait is manifested and perceptible in every possible situation. The results of this work suggest that the galleries of favorite pictures are an environment where self-assessed traits are neither available nor relevant. However, it is not possible to exclude that the low performance on self-assessed traits depends on the particular features adopted in this work. In fact, while features that capture visual appealing do not covariate with self-assessed traits, it is possible that other types of features do. Furthermore, the literature shows that the results achieved on self-assessed traits tend always to be lower than those obtained on assessed ones [12]. In the case of this work, the probable reason is that the Flickr users adopt information different from the pictures when they assess their own traits (e.g., their personal history and previous experiences) [13]. In other words, the pictures do not necessarily carry all the information that the subjects use when they perform a self-assessment.

The correlational analysis of Section VI shows that a large number of features covariate with the attributed traits to a statistically significant extent. The covariation is particulary high in the case of two traits - Neuroticism and Agreeableness - and the features related to the colors (see Figure 6). This can provide suggestions on how to manage online impressions using favorite pictures. For example, people that tag as favorite pictures where blue and warm colors (orange, brown, red and yellow) dominate tend to be perceived as more agreeable. In contrast, people that tag as favorite pictures where black and gray are frequent tend to be perceived as more neurotic. This is important because many "*use websites as a way to learn about someone they barely know*" [67] and, furthermore, the impressions conveyed through online activities have been shown to have an effect on important life issues like, e.g., the outcome of a job interview [68].

There are several directions for future work. The results of this article suggest cues that should be included in the feature set like, e.g., expression, gender, pose, scale and occlusion of human faces (if any). Similarly, the feature set might distinguish between indoor and outdoor pictures. However, these cues require the development of robust detectors because pictures posted on Flickr have quality and variability different from those observed in common literature benchmarks. Investigations in this sense can start with manual annotations to verify whether the cues actually have an impact. Other research efforts can focus on the regression approaches to be used. One possibility is to use deep learning strategies to design ad hoc features for the scenario at hand, thus discovering low level patterns of interest for trait prediction. Furthermore, LASSO could be substituted by non-linear or kernel regression approaches allowing one to take into account more complex relationships between features.

From an application point of view, this work contributes to recent multimedia trends trying to take into account the way people react to data they consume, whether this means to predict the emotions elicited by a painting [69] or to infer the content of videos and pictures from the behavioural reactions of people that watch them [24]. Furthermore, the results of this work seem to confirm the hypothesis that favorite pictures can work as social signals, i.e., as "*communicative or informative signals which [...] provide information about social facts*" [70]. This can possibly extend the scope of Social Signal Processing - the domain aimed at modeling, analysis and synthesis of social signals - to online interaction contexts [71].

REFERENCES

[1] M. Duggan and L. Rainie, "Cell phone activities 2012," Pew Research Center, Tech. Rep., 2012.

[2] L. Rainie, J. Brenner, and K. Purcell, "Photos and videos as social currency online," Pew Research Center, Tech. Rep., 2012.

[3] N. Van House, "Flickr and public image-sharing: distant closeness and photo exhibition," in *CHI 2007 Extended Abstracts on Human Factors in Computing Systems*, 2007, pp. 2717–2722.

[4] ——, "Personal photography, digital technologies and the uses of the visual," *Visual Studies*, vol. 26, no. 2, pp. 125–134, 2011.

[5] J. Suler, "Image, word, action: Interpersonal dynamics in a photo-sharing community," *CyberPsychology & Behavior*, vol. 11, no. 5, pp. 555–560, 2008.

[6] M. Lipczak, M. Trevisiol, and A. Jaimes, "Analyzing favorite behavior in Flickr," in *Proceedings of the International Conference on Multimedia Modeling*, S. Li, A. El Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, and C. Gurrin, Eds., vol. LNCS 7732. Springer Verlag, 2013, pp. 535–545.

[7] 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*, vol. 110, no. 15, pp. 5802–5805, 2013.

[8] H. Liu, "Social network profiles as taste performances," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 252–275, 2007.

[9] S. Ray and D. Page, "Multiple instance regression," in *Proceedings of the International Conference on Machine Learning*, 2001, pp. 425–432.

[10] B. Babenko, "Multiple instance learning: Algorithms and applications."

[11] P. Lovato, M. Bicego, C. Segalin, A. Perina, N. Sebe, and M. Cristani, "Faved! biometrics: Tell me which image you like and i'll tell you who you are," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 364–374, 2014.

[12] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[13] A. Wright, "Current directions in personality science and the potential for advances through computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 292–296, 2014.

[14] J. Uleman, S. Adil Saribay, and C. Gonzalez, "Spontaneous inferences, implicit impressions, and implicit theories," *Annual Reviews of Psychology*, vol. 59, pp. 329–360, 2008.

[15] R. Jenkins, *Social Identity*. Routledge, 2014.

[16] P. Galanter, "Computational aesthetic evaluation: Past and future," in *Computers and Creativity*, J. McCormack and M. d'Inverno, Eds. Springer, 2012, pp. 255–293.

[17] F. Hoenig, "Defining computational aesthetics," in *Proceedings of the Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, 2005, pp. 13–18.

[18] A. Salah, H. Hung, O. Aran, H. Gunes, and M. Turk, "Behavior understanding for arts and entertainment," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 3, pp. 12:1–12:10, 2015.

[19] 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*, 2006, pp. 288–301.

[20] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proceedings of the European Conference on Computer Vision*, 2008, pp. 386–399.

[21] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 236–252, 2009.

[22] L. Wong and K. Low, "Saliency-enhanced image aesthetics class prediction," in *IEEE International Conference on Image Processing*, 2009, pp. 997–1000.

[23] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 1784–1791.

[24] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173–180, 2009.

[25] J. Golbeck, C. Robles, and K. Turner, "Predicting Personality with Social Media," in *Proceedings of the Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 253–262.

[26] S. Bai, T. Zhu, and L. Cheng, "Big-five personality prediction based on user behaviors at social network sites," Cornell University, Tech. Rep., 2012.

[27] F. Celli, "Unsupervised personality recognition for social network sites," in *Proceeedings of the International Conference on Digital Society*, 2012, pp. 59–62.

[28] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Towards discovery of influence and personality traits through social link prediction," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2011, pp. 566–569.

[29] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 213–222.

[30] S. Fitzgerald, D. Evans, and R. 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.

[31] D. 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*, 2008, pp. 45–50.

[32] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on Computational Personality Recognition (shared task)," in *Proceedings of the Workshop on Computational Personality Recognition*, 2013.

[33] C. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima, "Fundamental dimensions of social judgment: Unrdestanding the relations btewen judgments of competence and warmth," *Journal of Personality and Social Psychology*, vol. 89, no. 6, pp. 899–913, 2005.

[34] D. Funder, "Personality," *Annual Reviews of Psychology*, vol. 52, pp. 197–221, 2001.

[35] G. Saucier and L. Goldberg, "The language of personality: Lexical perspectives on the five-factor model," in *The Five-Factor Model of Personality*, J. Wiggins, Ed., 1996.

[36] B. Rammstedt and O. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[37] G. Boyle and E. Helmes, "Methods of personality assessment," in *The Cambridge handbook of personality psychology*, P. Corr and G. Matthews, Eds. Cambridge University Press, 2009, pp. 110–126.

[38] D. Kenny, L. Albright, T. Malloy, and D. Kashy, "Consensus in interpersonal perception: acquaintance and the Big Five." *Psychological Bulletin*, vol. 116, no. 2, pp. 245–258, 1994.

[39] D. Kenny, "PERSON: A general model of interpersonal perception," *Personality and Social Psychology Review*, vol. 8, no. 3, pp. 265–280, 2004.

[40] D. Howell, *Statistical methods for psychology*. Cengage Learning, 2012.

[41] 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*, 2010, pp. 83–92.

[42] K. Mardia and P. Jupp, *Directional Statistics*. Wiley, 2009.

[43] 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*. Springer Verlag, 2006, vol. 3953, pp. 288–301.

[44] P. Valdez and A. Mehrabian, "Effects of color on emotions," *Journal of Experimental Psychology: General*, vol. 123, no. 4, p. 394, 1994.

[45] P. Lovato, A. Perina, N. Sebe, O. Zandonà, 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 *Proceedings of the Asian Conference on Computer Vision*, 2012.

[46] C. Georgescu, "Synergism in low level vision," in *Proceedings of the International Conference on Pattern Recognition*, 2002, pp. 150–155.

[47] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 – 619, 2002.

[48] W. Chu, Y. Chen, and K. Chen, "Size does matter: How image size affects aesthetic perception?" in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 53–62.

[49] H. Tamura, S. Mori, and T. Yamawaki, "Texture features corresponding to visual perception," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.

[50] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[51] M. Johnson, "Subcortical face processing," *Nature Reviews Neuroscience*, vol. 6, no. 10, pp. 766–774, 2005.

[52] B. Berlin, *Basic color terms: Their universality and evolution*. University of California Press, 1991.

[53] J. van de Weijer, C. Schmid, and J. Verbeek, "Learning color names from real-world images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[54] R. Haralick and L. Shapiro, *Computer and Robot Vision*. Addison-Wesley, 1992.

[55] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, 2001.

[56] P. Costa, R. MacCrae, and I. Psychological Assessment Resources, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual*. PAR, 1992.

[57] S. Ray, "Learning from data with complex interactions and ambiguous labels," Ph.D. dissertation, 2005.

[58] D. R. Dooly, Q. Zhang, S. A. Goldman, and R. A. Amar, "Multiple instance learning of real valued data," *Journal of Machine Learning Research*, vol. 3, pp. 651–678, 2003.

[59] J. Wang, et Jean-Daniel Zucker, and J. daniel Zucker, "Solving the multiple-instance problem: A lazy learning approach," in *Proceedings of the International Conference on Machine Learning*, 2000, pp. 1119–1125.

[60] K. L. Wagstaff, T. Lane, and A. Roper, "Multiple-instance regression with structured data." in *Workshops of International Conference on Data Mining*, 2008, pp. 291–300.

[61] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Jornal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[62] A. Perina and N. Jojic, "Image analysis by counting on a grid," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1985–1992.

[63] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1994.

[64] S. Cloninger, "Conceptual issues in personality theory," in *The Cambridge handbook of personality psychology*, P. Corr and G. Matthews, Eds. Cambridge University Press, 2009, pp. 3–26.

[65] A. Eagly, R. Ashmore, M. Makhijani, and L. Longo, "What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype." *Psychological bulletin*, vol. 110, no. 1, p. 109, 1991.

[66] J. Biesanz and S. West, "Personality coherence: Moderating self–other profile agreement and profile consensus." *Journal of Personality and Social Psychology*, vol. 79, no. 3, pp. 425–437, 2000.

[67] S. Vazire and S. Gosling, "e-Perceptions: personality impressions based on personal websites." *Journal of Personality and Social Psychology*, vol. 87, no. 1, p. 123, 2004.

[68] D. Coutu, "We googled you," *Harvard Business Review*, vol. 85, no. 6, pp. 1–8, 2007.

[69] 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 ACM International Conference on Multimedia*, 2012, pp. 349–358.

[70] I. Poggi and F. D'Errico, "Social Signals: a framework in terms of goals and beliefs," *Cognitive Processing*, vol. 13, no. 2, pp. 427–445, 2012.

[71] A. Vinciarelli and A. Pentland, "New social signals in a new interaction world: The next frontier for social signal processing," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 1, no. 2, pp. 10–17, 2015.

Cristina Segalin received the B.Sc. degree in multimedia computer science, and the M.Sc. degree in engineering and computer science from the University of Verona in 2010 and 2012, respectively, where she is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science. Her research interests mainly focus on machine learning, pattern recognition, social signal processing, image processing, social media analysis, and computational aesthetics techniques for multimedia applications. Her doctoral work investigates the effects of nonverbal behaviors on personality perception.

Alessandro Perina received the Ph.D. degree in computer science from the University of Verona with a thesis on classification with generative models. From 2006 to 2010, he was a member of the Vision, Image Processing, and Sound Group, University of Verona. He is currently a data scientist at Micrsoft Corporation, Redmond, WA, USA. His area of expertise lies in machine learning, statistics and applications, particularly in probabilistic graphical models, unsupervised learning, and optimization algorithms.

Marco Cristani is an Associate Professor with University of Verona, Department of Computer Science, since 2015, where he teaches and does research within the Vision, Processing and Sound Laboratory. He is a Research Affiliate with Istituto Italiano di Tecnologia, Genova, Italy, where he was a Team Leader from 2009 to 2012. His interests are focused on generative modeling, and in particular, on generative embeddings with applications on social signal processing and multimedia. Dr. Cristani is the co-author of more than 120 papers in important international journals and conferences. He is in the technical program committee of social signaling/pattern recognition conferences, and organizer of social signaling/video surveillance.

Alessandro Vinciarelli (www.dcs.gla.ac.uk/vincia) is with University of Glasgow where he is Senior Lecturer (Associate Professor) of the School of Computing Science and Associate Academic of the Institute of Neuroscience and Psychology. Overall, he has published more than 100 works, including one authored book and more than 30 journal papers. He has participated in the organization of the IEEE International Conference on social computing as a program chair in 2011 and as a general chair in 2012, he has initiated and chaired a large number of international workshops. Furthermore, he is or has been Principal Investigator of several national and international projects, including a European Network of Excellence (the SSPNet, www.sspnet.eu). He is the cofounder of Klewel (www.klewel.com), a knowledge management company recognized with several awards. He is a member of the IEEE.