· Universitá degli Studi di Verona ·

DIPARTIMENTO DI INFORMATICA
Scuola di Dottorato in SCIENZE INGEGNERIA MEDICINA
Dottorato in INFORMATICA

A SOCIAL SIGNAL PROCESSING PERSPECTIVE ON COMPUTATIONAL AESTHETICS: THEORIES AND APPLICATIONS

Ph.D. Thesis

Advisor:	**Candidate:**
Dott.	**Dott.**
MARCO CRISTANI	**CRISTINA SEGALIN**

XXVIII cycle (January 2013 - December 2015)

A Social Signal Processing Perspective on Computational Aesthetics: Theories and Applications.

ISBN:

Cristina Segalin

A Social Signal Processing Perspective on Computational Aesthetics: Theories and Applications

Ph.D. Thesis

XXVIII cycle (January 2013 - December 2015)

Università degli Studi di Verona
Dipartimento di Informatica

Advisor:
Dr. Marco Cristani

Series N°: **TD-09-16**

Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona
Italy

To my parents Barbara and Roberto

As a student, I studied the Philosophy of Existentialism: it teaches you "fall seven times, stand up eight" (Japanese proverb). Being a researcher is not doing what you "have to" but "love to", spending all the day doing what you are passionate about. Curious, always thinking and craving for new knowledge; focused on aesthetic beauty and simplicity. I love to discover, combine disciplines and go beyond the boundaries of technology, brainstorming with different types of people to solve problems, mixing computer science with several disciplines. I believe that building cross-disciplinary knowledge in various disciplines, for instance, psychology, computer science, behavioral science and neuroscience, is a way for researchers to keep up with the pace of technology development, and this for decades: choosing to be a researcher or a scientist of our time means being a trans-disciplinary thinker. I have witnessed too many times that, "at every crossway on the road that leads to the future, each progressive spirit is opposed by a thousand men appointed to guard the past," (A.G. Bose) but also that, "logic will get you from A to Z; imagination will get you everywhere." (A. Einstein). I believe that cross-disciplinary education opens people's minds to imagine, believe in and do, the impossible.

- Cristina Segalin -

Sommario

Ogni giorno, siamo esposti a varie immagini e video grazie ai social media, come Facebook, Youtube, Flickr, Instagram e altri. In questo scenario, l'esprimere preferenze per un dato contenuto multimediale (per esempio con l'uso del meccanismo di "like") è diventato pervasivo e imponente, diventando un fenomeno di massa sociale. Uno dei principali risultati nelle scienze cognitive è che i processi automatici di cui non siamo a conoscenza, modellano, per la maggior parte, la nostra percezione dell'ambiente. Il fenomeno si applica non solo al mondo reale, ma anche ai dati multimediali che consumiamo giornalmente. Ogni volta che osserviamo una immagine, guardiamo un video o ascoltiamo una registrazione, la nostra attenzione cosciente si concentra sul contenuto osservabile, ma la nostra cognizione percepisce spontaneamente intenzioni, opinioni, valori, attitudini e altri costrutti che, sebbene siano al di fuori della nostra consapevolezza cosciente, modellano le nostre reazioni e comportamenti. Finora, le tecnologie multimediali hanno trascurato questo fenomeno.

Questa tesi discute il fatto che è possibile prendere in considerazione effetti cognitivi per migliore gli approcci multimediali. A questo scopo sono considerati principi di *Computational Aesthetics* e *Social Signal Processing* sotto un punto di vista computazionale. Da un lato la *Computational Aesthetics* ha la funzione di rendere applicabili decisioni estetiche in modo simile a come gli esseri umani sanno fare, permettendo alle tecnologie multimediali di modellare e valutare un senso comune della bellezza. Dall'altro lato il campo del *Social Signal Processing* ha lo scopo di modellare con algoritmi i processi cognitivi che codificano segnali sociali e che ci portano ad interagire in modo particolare con le persone o preferire immagini e video. Questa rappresenta una grande opportunità per la CA perché la risposta estetica umana è formata dalla combinazione di predisposizioni genetiche, assimilazione culturale e esperienze uniche individuali e in questo modo può essere imparata da immagini online usando la saggezza della folla.

La tesi si focalizza sulle immagini come primo tentativo in questa direzione. Le motivazioni del perché concentrarsi sulle immagini sono molte: da un lato, scattare foto è una delle azioni comunemente svolte tramite l'uso di telefoni cellulari, e dell'altro lato, gli utenti postano online immagini originali o video e condividono e redistribuiscono quelli postati da altri.

A questo scopo, la tesi presenta uno studio sull'estetica personale, dove lo scopo è quello di riconoscere le persone e le loro caratteristiche considerando le immagini che piacciono a queste sviluppando diversi approcci ibridi usando modelli generativi e di regressione. L'idea generale assume che, dato un insieme di immagini preferite, è possibile estratte un insieme di attributi che discriminano *pattern* visuali, che possono essere usati per inferire caratteristiche personali del soggetto che le preferisce.

Come primo contributo proponiamo un sistema di *soft biometrics*, che permette di discriminare un individuo rispetto ad altri usando le immagini che le/gli piacciono. Lo studio e sviluppo del sistema biometrico è diventato di primaria importanza sia per l'identificazione di individui che applicazioni di sicurezza è *recommendation system*. Su un dataset di 200 utenti e 40000 immagini, il sistema sviluppato raggiunge il 97% di probabilità di indovinare l'utente corretto usando 5 immagini preferite come modello biometrico; per la capacità di verifica, l'EER è 0.11.

Inoltre, abbiamo sviluppato un sistema capace di inferire la personalità di un soggetto usando le sue immagini preferite. La motivazione è che quando conosciamo una persona per la prima volta, tendiamo ad attribuire tratti di personalità ad esso/essa. Il processo è spontaneo e inconscio. Sebbene non necessariamente accurato, il processo comunque influenza significativamente il nostro comportamento nei confronti degli altri, specialmente quando si tratta di interazioni sociali. Il fenomeno è così diffuso che ha luogo non solo quando conosciamo altri in persona, ma anche quando li osserviamo in registrazioni video, o interagiamo con agenti artificiali che mostrano comportamenti simili agli umani o con materiale multimediale che le persone condividono online.

Come risultato, la tesi mostra che ci sono *pattern* visuali che correlano con i tratti di personalità di utenti Flickr in misura statisticamente significativa, e che i tratti di personalità (sia auto valutati che attribuiti da altri) di questi utenti posso essere inferiti dalle immagini che questi ultimi marcano come preferite. Una della parti più importanti della tesi è stata la collezione del dataset *PyschoFlickr*, composto da 60000 immagini di 300 utenti Flickr annotate in termini di tratti di personalità sia auto attributi che attributi da 22 giudici. La predizione è eseguita usando più approcci (multiple instance regression e deep learning), raggiungendo una correlazione fino a 0.68 e un'accuratezza fino a 0.69 tra tratti reali e predetti.

La predizione dei tratti attribuiti da altri ottiene risultati più alti rispetto a quelli auto attribuiti: la ragione è che le immagini dominano l'impressione della personalità che i giudici percepiscono e il consenso tra loro è statisticamente significativo. Questi due condizioni aiutano la regressione ad ottenere risultati più alti. Quando gli utenti auto giudicano la loro personalità, considerano anche altri informazioni che non sono disponibili nelle immagini che preferiscono, ad esempio, storia personale, la stato interiore, educazione, ecc.. Tuttavia, questo non permette di ottenere alti risultati nella regressione. Questo è un risultato importante che può aiutare a capire meglio il comportamento sociale delle persone a nel progettare agenti artificiali capaci di suscitare la percezione di tratti predefiniti desiderabili e fornire suggerimenti su come gestire le impressioni online usando le immagini preferite.

Abstract

Everyday, we are exposed to various images and videos thanks to the social media, like Facebook, Youtube, Flickr, Instagram and others. In this scenario, the use of expressing preferences for a given multimedia content (for example by the use of liking mechanisms) has become pervasive and massive, becoming a social mass phenomenon. One of the main findings of cognitive sciences is that automatic processes of which we are unaware shape, to a significant extent, our perception of the environment. The phenomenon applies not only to the real world, but also to multimedia data we consume every day. Whenever we look at pictures, watch a video or listen to audio recordings, our conscious attention efforts focus on the observable content, but our cognition spontaneously perceives intentions, beliefs, values, attitudes and other constructs that, while being outside of our conscious awareness, still shape our reactions and behavior. So far, multimedia technologies have neglected such a phenomenon to a large extent.

This thesis argues that taking into account cognitive effects is possible and it can also improve multimedia approaches. For this purpose we take into account Computational Aesthetics and Social Signal Processing principles under a computational point of view. On one side Computational Aesthetics makes applicable aesthetic decision in a similar fashion as human can allowing to multimedia technologies to learn, model and evaluate a common sense of beauty. On the other side,

Social Signal Processing field has the aim of modeling with algorithms cognitive processes that codify social signal and that lead us to interact with a particular way with people or to prefer a particular image or video. This represents an invaluable opportunity for CA because human aesthetic response is formed by a combination of genetic predisposition, cultural assimilation, and unique individual experience and indeed it can be learned from online pictures using the wisdom of crowds.

The thesis focuses on images as a first attempt in this direction. The motivation of why focusing on pictures are many: from one side, taking pictures is the action most commonly performed with mobile phones, on the other side, users either post online original images or videos or share and redistribute those posted by others.

To this aim the thesis presents a study on personal aesthetics, where the goal is to recognize people and their characteristics by considering the images they like by developing several hybrid approaches using generative models and regressors. The general idea assumes that, given a set of preferred images, it is possible to extract a set of features individuating discriminative visual patterns, that can be used to infer personal characteristics of the subject that preferred them.

As first contribution we propose a soft biometric system, that allows to discriminate an individual from another using the images he/she likes. The study and development of biometric system have become of paramount importance for both identification of individual and security applications and recommendation systems. On a dataset of 200 users and 40K images, the developed frameworks gives 97% of probability of guessing the correct user using 5 preferred images as biometric template; as for the verification capability, the equal error rate is 0.11.

Furthermore, we developed a system able to infer the personality of a subject using the images preferred by him/her. The motivation is that whenever we meet a person for the first time, but also when we observe her in video recordings, or we interact with an artifact displaying human-like behavior or with the multimedia material she shares online, we tend to attribute personality traits to her. The process is spontaneous and unconscious. While not necessarily accurate, the process still influences significantly our behavior towards others, especially when in comes to social interactions.

As a supporting proof-of-concept, the thesis shows that there are visual patterns correlated with the personality traits of Flickr users to a statistically significant extent, and that the personality traits (both self-assessed and attributed by others) of those users can be inferred from the images these latter mark as “favorite”. One of the most important part of the thesis has been the collection of the PsychoFlickr corpus, composed of 60K images of 300 Flickr users annotated in terms of personality traits both self and attributed by 22 assessors. The prediction are performed using multiple approaches (multiple instance regression approach and a deep learning framework), reaching a correlation up to 0.68 and an accuracy up to 0.69 between actual and predicted traits.

The prediction of traits attributed from others achieve higher results compared to the self-assessed ones: the reason is that pictures dominate the personality impressions that the judges develop and the consensus across the judges is statistically significant. 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, this does not allow the regression approaches to achieve high performances. This is an important finding as it can help to better understand the social behavior of people, to design artificial agents capable of eliciting the perception of predefined desirable traits and providing suggestions on how to manage online impressions using favorite pictures.

Keywords: Computational Aesthetics, Social Signal Processing, Human Computer Interaction, Pattern Recognition, Social Media Analysis, Nonverbal Behavior, Personality Computing, Soft Biometrics, Feature Extraction, Image Processing.

Preface

This thesis is the result of the work I did during the three years of Ph.D, collaborating with many other colleagues and researchers.

Acknowledgments

This Thesis is the last milestone in the long journey as a student. Many people have contributed in various ways to make this Ph.D study an exciting and memorable journey.

Foremost, I would like to express deep gratitude and respect to my advisor, Dr. Marco Cristani for his enthusiasm and advices along the way.

Beside my advisor, I would like to express a deep gratitude and respect to Dr. Alessandro Vinciarelli for the great opportunity to collaborate with him, and making me discovery the beautifulness of Scotland. His insightful guidance, support, friendship, all I have learned from him, our long conversations about the common research fields made me appreciate and increasingly grow the passion and motivation for this work. He inspired me to continue in the carrier of research.

My sincere thanks goes to my Ph.D tutors, Alessandro Farinelli and Gloria Menegaz for their encouraging and constructive feedbacks.

A grateful thanks also goes to Mirco Musolesi for his advices, friendship during the period spent in Birmingham and Alessandro Perina for all the good advices and help in understanding and applying Machine Learning to my works; he motivated me to explore deeper this field. Thanks to all my colleagues Pietro, Davide, Matteo, Francesco, Giorgio, Michele, Matteo, Francesca, with whom I spent funny moments in and out the office, making it a fun and friendly place to work.

I also thank all my friends, near and far, for their support and care.

I would like to thank Verona, this beautiful city that gave me the opportunity to grow as person during these eight years, to met amazing people, good friends, making my eyes sparkling with its beautiful night lights and places. Overall the places I have been, I think there is not a more beautiful city than it.

My final words go to my family. I would like to thank my parents for their support and love, encouragement, being always on my side, ready to help me packing all my stuff, every time I moved in a new place. Words are not enough for me to express how thankful I am for all the things you have done for me; you are truly my source of inspiration and motivation, thanks for always being there for me!!

Contents

Part II Soft Biometrics for Re-identification

Part III Personality Computing through Multimedia Content Analysis

Part IV Conclusions

List of Figures

List of Tables

1

Introduction

Multimedia data, and in particular pictures, are a basic human artifact, which serve to communicate something.

Until a few years ago, production, diffusion and use of multimedia data required skills and infrastructures that were the privilege of a few individuals and organizations (archives, digital libraries, online repositories, etc.) [25].

Nowadays, technologies as ubiquitous and user-friendly as smartphones and tablets allow one to easily create multimedia material (pictures, videos, soundbites, text and their combinations) and share it with others - typically through social media or other online technologies - by simply pushing a button, becoming an everyday practice of the lay person and spontaneous expression of common individuals involved in everyday social interactions.

Everyday, we are exposed to various images and videos thanks to social media, as Facebook, Youtube, Flickr, Instagram and others [12, 263].

In such a technological landscape, multimedia data is not just a way to transmit knowledge and information - as it used to be traditionally for any type of data [37, 236] - but one of the channels through which we interact with others.

Multimedia data now are associated to well defined agents: the ones that create them (producers) and the ones that consume them (consumers). In this way, the *creation* of an image, for instance, can be traced in social media by means of the first authorized post (eventually supplied with a textual comments); then the *consumption* of the same image can take place in different ways: a simple analysis by looking at it, commenting it, sharing it, rating or marking it as favorite or liking it, the latter two actions eventually supplied with textual comments as well.
The consumption of multimedia data can be captured in several ways:

- in the case of simple analysis, by means like eye tracking
- in the case of sharing, going back to whom have shared it
- in the case of voting, going back to whom have voted it
- in the case of comments, analyzing them

In this thesis we focus on images. The motivation of why focusing on pictures are many: from one side, taking pictures is the action most commonly performed with mobile phones, on the other side, users either post online original images or videos or share and redistribute those posted by others. Following this direction, we go from an anonymous communication to a nominal one. In particular, we are interested in non verbal communication (excluding textual comments), and specifically the communication exploited by voting an image.

Voting an image has two main meanings: it is objectively beautiful and it is liked by someone.

While the first meaning is related to aesthetic issues (an image is beautiful because there are common and shared aesthetics principles) and these have been widely studied, the latter, i.e. the individual preferences, has been taken into account only in the field of paintings [90, 91, 92, 302].

In this thesis we show that the vote of an image given by a person depends on several factors, as the content of the image (which relates to what is represented by the image), or the aesthetics of the image (what is related to color combination, composition, shape, etc., thus how content is represented), to support or give feedback to the person that initially shared the picture, and some others which can be inferred and represented by computational features of the same image and connected to visual image perception and processing of humans, thus being mainly biological, hardwired and therefore universal, making us favor certain images to others: we call this part *Computational Aesthetics*. From human perception, we can immediately agree that what we think is beautiful is connected to our experiences, i.e. what we have learned.

By means of CA we can capture the personal preferences and used them for different purposes (visualizing his/her tastes, identify the subject).

Computational Aesthetics has been somewhat instable over time. For some, the term includes both generative and analytic modes, i.e both the creation and evaluation of art using computers. For others it purely refers to the use of computers in making aesthetic judgments [93].

As a consequence of the recently interest of computer scientists in aesthetics, the term Computational Aesthetics (CA) is defined [116] as a discipline of Computer Science, formulated as follows :

“Computational Aesthetics is the research of computational methods that can make applicable aesthetic decision in a similar fashion as human can”.

The aims of CA is explicitly modeling a common sense of beauty, that is, at reducing the subjectivity that drives the appreciation of a photography, causing the same shot to be appreciated by some viewers but not by others.

CA indeed is an interdisciplinary area at the cross-road between vision, artificial intelligence and pattern recognition, psychology, visual art aesthetics and neuroscience [132]. The interest for the topic is associated with the grow of social media, where “*liking*” multimedia material is one of the most common social activities. This represents an invaluable opportunity for CA because the aesthetic appeal of images can be learned from online pictures using the wisdom of crowds [12]. Such an idea has been used in many studies by using explicit aesthetic scores [67, 193], or relying directly on textual tags and likes [12].

Furthermore, communication through pictures (and in particular by voting) can take place with similar dynamics with respect to face-to-face communication, by means of non-verbal communication.

Social psychologists have studied for long time the non-verbal communication, cognitive processes describing them as a set of temporal changes in neuromuscular and physiological activity that persist for short (millisecond to minute) or long (minute to hour) periods. Studies on the verbal and non-verbal communication, suggest that there is more than words in social interactions, overall the non-verbal social signals seem to be the most predominant sources of information during social interactions [4, 7].

Indeed, the exchange of multimedia data has become a form of human-human communication and, therefore, it should involve and give rise to the cognitive phenomena (e.g. [268, 269]) typically observed in human-human interaction , especially when it comes to expression and mutual attribution of socially relevant characteristics (attractiveness, social status, personality, goals, values, intentions, etc.). This applies in particular to implicit cognitive processes that take place outside our conscious awareness, but still shape to a large extent our perception of the world and our behavior [147], namely the tendency to express and attribute to others goals, values, intentions, traits, beliefs, and any other type of socially relevant characteristics [268].

The problem is that our cognitive processes are the result of a long evolutionary history and cannot change at the pace of technology. Therefore, our cognition keeps following patterns that were shaped during time when technology was far from existing [194]. In particular, a large body of evidence shows that our cognition constantly works to make sense of the world around us and that this happens, to a large extent, “effortlessly, and even unintentionally” [269].

How does social interaction work when none of the cues available in face-to-face settings can be used? This sounds like an artificial and unrealistic situation, but it is exactly what happens on social media where millions of people interact without displaying any of the social signals they typically use in co-located interaction. Most of the interactions involves cues that have no equivalent in ordinary face-to-face social contacts. Pictures provide an example. According to the survey of the Pew Research Center 46% of the American Internet users post original pictures or share online images posted by others [227]. Furthermore, one of the main motivations behind the use of photo-sharing platforms is to maintain contact with others [278]. In this respect, pictures play the role of an online "*social currency*".

In this thesis we take into account also these aspects, and in particular we focus on signals of personality, transferred from who votes a picture and interpreted by who observes what this person is voting specifically in terms of a particular interaction construct, i.e. the Brunswik Lens [36]. The Brunswik Lens, one of the most effective models developed in cognitive psychology, provides a framework suitable for investigating how multimedia data can be adopted as an observable "evidence of attitudes, inferences, goals and theories" as stable and predictable patterns of data producers. Symmetrically, the model helps to explain how data consumers attribute "attitudes, inferences, goals and theories" to data producers.

Psychology and neuroscience have investigated the influence of individual characteristics on aesthetics preferences for 70 years [132]. This applies to style [91] and, to a lesser extent, content of paintings [266]. The main result of these investigations is the identification of correlations between artistic preferences and personality [91], where this is typically described in terms of Big Five traits [228], that are Openness to experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism(N).

These studies have never been taken into account by CA, and the purpose of this thesis is to consider them in a computational manner. To this aim, Social Signal Processing (SSP) methods will be taken into account [288].

"Social Signal Processing aims at modeling with algorithms those cognitive processes that codify social signals and that lead us to interact in a particular way with people, or to prefer a particular image or video, i.e. understand human social signals".

In this thesis we show that individual aesthetical preferences can be captured, and can be also related to personality traits.

1.1 Motivations

There are several reasons to consider CA and SSP important problems for multimedia: the first is that the study of *individual* perception of beauty has been done mostly on paintings (while generic Computational Aesthetics is now focused primarily on photos); our approaches make it possible to extend the investigations to pictures, a type of data particularly important since large amounts of images are shared via online platforms. It is apparent the crucial role that image processing and machine learning have; at the same time, our study delineate new challenges for these areas; for example, discovering visual patterns that correlate with personal traits in a stronger way than ordinary features or fit better cognition processes could be a research mission for the field of deep learning and feature learning [252]. Generative modeling can also be involved, looking for new models that mimic the way diverse visual features should be combined together.

Understanding and modeling of cognitive processes involved in multimedia data consumption are likely to be beneficial for Human Information Interaction (HII), the domain studying the "*relationship between people and information*" [83]. HII researchers are particularly interested in modeling people *proflections*, i.e. individual's conscious and unconscious projections on information objects (*e.g.*, pictures) and the reflections that other people and machines create to those projections (*e.g.*, links and annotations) [179]. This applies in particular to multimedia retrieval technologies that might be enhanced by taking into account not only the data content (like most of current technologies do [157]), but also the interplay between content and perceptual judgments (personality, values, goals, intentions, etc.).

The role of cognitive biases can be of interest for Digital Humanities as well, especially for what concerns the effort towards "*new modes of knowledge formation enabled by networked, digital environments*" and the focus on "*distinctive modes of producing knowledge and distinctive models of knowledge itself*" [37]. In particular, Digital Humanities investigate the impact of media authoring technologies on the transmission of knowledge and information, a phenomenon likely to involve implicit cognitive processes like those described in this work.

According to the European Consumer Commissioner, "*Personal data is the new oil of the Internet and the new currency of the digital world*" [85]. The "states" of data producers often correspond to personal characteristics of potential interest for different bodies (*e.g.*, companies trying to model their customers or governments interested in gathering information about the population). Approaches like those presented in this thesis can help to obtain such information by analyzing publicly available data that people usually post on personal home pages, Youtube, Facebook, etc. [143]. In parallel, the development of technologies capable of going beyond the mere content and infer personal characteristics of data producers require a redefinition of the concept of privacy and a careful analysis of ethical issues [55].

Creating and viewing photographs as a process of self-insight and personal change is the main principle of *phototherapy* and *therapeutic photography* [258, 296], two recent psychology perspectives; for the therapists, "*Images provide an undercurrent of emotion and ideas that enrich interpersonal dynamics, often on a level that is not fully conscious or capable of being verbalized*". Of particular interest for these fields is how the language of composition and visual design intersects with the language of unconscious primary cognitive processes, including emotional/ideational association. Our study suggests that answers to these questions may be found with the help of computers.

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

1.2 Contributions

It is only recently that Computational Aesthetics and Social Signal Processing attracted the interest of the computing community. To the best of our acknowledge, the results presented in this thesis are among the earlies investigation of the CA problem and addressed several issues that, at the beginning of the thesis, were still largely unexplored.

In this thesis, we follow the above direction, with two aims:

i) apply Computational Aesthetics to multimedia data and analyze the relationship between individual characteristics (including both personality traits and information typically available on social media) and aesthetic preferences;

ii) we consider aesthetic preferences as a social signal, that can be made explicit on social media, to show how the individual characteristics of a person can be analyzed and then predicted; furthermore how these characteristics can be perceived by others, through the multimedia data, analyzed and inferred.

The most important contributions are as follows:

- **Feature extraction package**: With the purpose of analyzing individual aesthetics preferences, we implemented Computational Aesthetics features that can be extracted at image level, available as a MATLAB package.

- **Recognition of an individual through his/her aesthetic preferences:** We developed several hybrid approaches using generative models and regressors/classifiers in order to recognize an individual by the images he/she likes as a new *soft biometric trait*. These approaches provide the crucial information towards the development of better soft biometric system, identifying the most discriminative cues.

- **Collection of a large dataset for Personality Computing:** The collection of a dataset was the first crucial step towards the development of our personality computing approaches. We collected a corpus of 60K favorite images from the famous photo-sharing platform Flickr, 200 photos for each of the 300 users together with other demographic information reported in their profile page, including the Big Five personality traits both self and assessed by 22 unique judges (12 European and 10 Asian)[1].

- **Quantitative analysis of the cues influencing aesthetic preferences and personality:** We moved beyond correlational analysis, typically adopted by psychologists, to investigate the effect of behavioral cues on personality perception. We then build hybrid frameworks based on generative models and Multiple Instance Regression and Convolutional Neural Networks to predict both self-assessed and attributed personality traits. In this way we exploit and provide indications of the most influential cues for the personality computing task that psychologists neglected for lack of expertise in signal processing. Furthermore, we provide an online demo able to predict personality traits based on images that a subject can select or upload on the online interface.

It is worth notice that the dataset used for the soft biometrics part is different from the Psychoflickr corpus used in modeling personality traits, both in size, images and selected users. This is because the first has been collected initially solely for re-identification purposes and the personality computing part was conceived later during the first year of PhD and needed more information related to the users, major number of participants as their consensus for participating to the personality questionnaire, answers and demographic information.

1.3 Thesis Outline

The thesis starts with an extensive preliminary part where both the background areas of Computational Aesthetics and Social Signal Processing are illustrated (see Part I). Then it presents a synopsis of the hand crafted aesthetic features proposed in this thesis varying through color statistics, scene descriptors, content and textures. Although it is one of the contributions of the thesis, we present it here as one of the main components of the thesis. Then we introduce the new topic of feature learning using Convolutional Neural Network that has been used to extract automatically the features from the images at a higher level of abstraction. Last we review theoretical foundations of statistical pattern recognition and probabilistic modeling that the thesis exploits.

In Part II original contributions to the pattern recognition community concerning soft biometric for re-identification are presented. We start with a brief introduction to the field of Soft Biometrics and the new personal aesthetic trait proposed in this thesis. Then we show different frameworks developed with this aim, that allows us to increase and improve the performances of the recognition framework, showing the evolution of the approaches. All the experiments of this part are performed on a dataset comprising 200 Flickr users, each one of them associated with 200 preferred images, that is, images that he/she likes, for a total of 40000 images. The results achieved so far in this task have been published at ICIP, ICMI, ACCV conferences, and IEEE Forensics journal.

In Chapter 5 we present the first attempt on distilling and encoding the uniqueness of users' visual preferences. We feed the extracted image features into a LASSO regressor that highlights the most discriminant cues for the individual, allowing authentication and recognition tasks.

[1] The works presented in this thesis focus on the European assessors only; the use of Asian assessors is a work in progress.

In Chapter 6 we present a further step in recognizing people by their aesthetic preferences building a statistical behavioral biometric approach. The approach is called “pump and distill”, since the training set of each user is pumped by bagging, producing a set of image ensembles. In the distill step, each ensemble is reduced into a set of surrogates, that is , aggregates of images sharing a similar visual content. Finally LASSO regression is performed on these surrogates for predicting a user identity.

In Chapter 7 we propose an alternative approach that bypasses the necessity to build an explicit conceptual coding of image preferences, operating directly on the raw properties of the images, extracted with heterogeneous feature descriptors. This is achieved through the *Counting Grid* model, which fuses together content-based and aesthetics themes into a 2D map in an unsupervised way. We show that certain locations in this map correspond to perceptually intuitive image classes, even without relying on tags or other user-defined information. Moreover, we show that users’ individual preferences can be represented as distributions over the map, allowing us to evaluate the affinity between different users’ appreciations.

In Chapter 8 we show a multi-level approach, where each level is intended as a low-dimensional latent space where the images preferred by a user can be projected, and similar images are mapped nearby, through the Counting Grid. Multiple levels of resolution are generated by adopting CG at different resolutions, corresponding to analyze images at different grains. On this space, a set of preferred images of a user produces an ensemble of intensity maps, highlighting in an intuitive way his personal aesthetic preferences. These maps are the used for learning a battery of discriminative classifiers (one for each resolution), which characterizes the user and serves to perform identification. We also show a simple user study to demonstrate its effectiveness in a biometric application.
To overcome the limitations of the object detection features used in the approached proposed so far, we present a further exploitative study. This framework uses the so-called Region-CNN to extract a feature vector that represents the object detection from images but using a deep learning method. This technique first runs a selective search to find for each image a number of region proposals, then extract from them a vector of features using the CNN, and finally for each of them classify the windows over 200 classes of object. We use this state-of-the-art model to extract more robust and reliable object detection feature vector and recompute the identification and recognition task. We show that this framework over-performs the one achieved using Deformable Part Models as object detectors.

Part III focuses on the contribution of modeling personality traits in a Multimedia scenarios, where pictures liked by Flickr users are used to predict their personality, for both self-assessed and attributed traits. The results achieved so far in this task have been published at ACM MM (Brave New Idea) conference, and IEEE Affective Computing journal.

The part starts with an extensive state-of-the-art in personality computing. The survey covers previous works in both psychology and computing. Then we present a overview of the metrics used in psycometrics to measure and evaluate personality.

In Chapter 9 we present our approach aimed at Automatic Personality Recognition and Automatic Personality Perception using as nonverbal behavior, the images the people like, and in particular we try to explore how personal aesthetics of individual are related to the personality traits. We present PsychoFlickr the image dataset collected for this work together with the personal information of the individuals. We investigate several Multiple Instance Learning methods to recognize and predict the personality of the users together with an analysis of the correlation between the pictures, the personality traits and the personal information.

In Chapter 10 we explore a new level of image understanding by means of Convolutional Neural Networks (CNN). The use of CNNs allows us to learn the visual patterns which are predictive of a particular personality trait, without relying on pre-structured visual features designed for a totally different purpose (object detectors, aesthetical features). The results outperform the previous state of the art, both in term of self assessed and attributed traits. We also present a demo in which we used Convolutional Neural Network to build a model that allow us to generalize the most crucial concept about each personality trait, and classify them.

In the last Part IV, conclusive remarks will be reported and future perspectives envisaged.

List of publications

Journals

- **C. Segalin**, D.S. Cheng, M. Cristani, *Social Profiling through Image Understanding: Personality Inference using Convolutional Neural Networks.* Computer Vision and Image Understanding 2016. (submitted)
- **C. Segalin**, A. Perina, M. Cristani, A. Vinciarelli. *The Pictures we Like are our Image: Continuous Mapping of Favorite Pictures into Self-Assessed and Attributed Personality Traits.* IEEE Transactions on Affective Computing 2016.
- P. Lovato, M. Bicego, **C. Segalin**, A. Perina, N. Sebe, 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 2014.

Papers

- **C.Segalin**, F.Celli, L.Polonio, M.Kosinski, D.Stillwell, B.Lepri, M.Cristani. *What your Facebook Profile Picture Reveals about your Personality: A Feature-based Approach.* International Conference on Multimodal Interaction 2016 (submitted)
- **C. Segalin**, M. Musolesi, M. Cristani, A. Vinciarelli. *Visual Contagion: Understanding the Influence of Textual, Visual and Social Cues on Information Propagation in Twitter* ACM International Conference on Information and Knowledge Management 2016. (Submitted)
- **C. Segalin**, A. Perina, M. Cristani. *Recognizing People by Their Personal Aesthetics: a Statistical Multi-level Approach*. Asian Conference on Computer Vision 2014
- **C. Segalin**, A. Perina, M. Cristani. *Personal Aesthetics for Soft Biometrics: a Generative Multi-resolution Approach*. International Conference on Multimodal Interaction 2014
- **C. Segalin**, A. Perina, M. Cristani. *Biometrics on Visual Preferences: a "Pump and Distill" Regression Approach*. IEEE International Conference on Image Processing 2014
- G. Roffo, M. Cristani, F. Pollick, **C. Segalin**, V. Murino. *Statistical analysis of visual attentional patterns for video surveillance*. Pattern Recognition, Image Analysis, Computer Vision, and Applications 2013.
- M. Cristani, A. Vinciarelli, **C. Segalin**, A. Perina. *Unveiling the multimedia unconscious: implicit cognitive processes and multimedia content analysis*. ACM International Conference on Multimedia 2013.
- G. Roffo, **C. Segalin**, V. Murino, M. Cristani. *Reading between the turns: Statistical modeling for identity recognition and verification in chats*. IEEE International Conference on Advanced Video and Signal Based Surveillance 2013.
- **C. Segalin**, A. Pesarin, A. Vinciarelli, M. Cristani. *The expressivity of turn-taking: Understanding children pragmatics by hybrid classifiers*. International Workshop on Image Analysis for Multimedia Interactive Services 2013.
- P. Lovato, A. Perina, D.S. Cheng, **C. Segalin**, N. Sebe, M. Cristani. *We like it! mapping image preferences on the counting grid*. International Conference of Image Processing 2013.
- A. Pesarin, M. Tait, A. Vinciarelli, **C. Segalin**, G. Bilancia, M. Cristani. *Generative modeling of dyadic conversations: characterization of pragmatic skills during development age*. Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction 2013.
- M. Cristani, G. Roffo, **C. Segalin**, L. Bazzani, A. Vinciarelli, V. Murino. *Conversationally-inspired stylometric features for authorship attribution in instant messaging*. ACM International Conference on Multimedia Brave New Idea 2012.
- M. Tait, M. Cristani, A. Pesarin, **C. Segalin**, G. Bilancia. *Lo sviluppo delle competenze pragmatiche tra i 3 e gli 8 anni*. XXI Congresso Nazionale AIRIPA 2012.

Part I

Preliminaries

2

Social Signal Processing for Computational Aesthetics

2.1 Computational Aesthetics

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 [242]). 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 [62, 158, 172, 178, 298]. Unlike Implicit Tagging [209], 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.

Understanding how aesthetic preferences are related to individual differences is one of the fundamental problems of CA. People often get enjoyment from observing images and express preferences for some pictures over others. Being fascinated by a picture is clearly the result of a complex interplay between the undeniable aesthetics of an image (*i.e.*, if a photo is objectively beautiful), its content (some people prefer cars over flowers) and the subjective preferences of a person, which are affected by personal interpretation of visual stimuli, experience, mood, technical or artistic merit, social support, friendship [260, 277] and so on [19]. Honeig [116] in 2005 comprehensively defined Computational Aesthetics as a field of study with many emphasis on the three important factors: computational methods, the human aesthetic point of view and the need to focus on objective approaches.

Up to now, there is no scientifically comprehensive theory that explains what psychologically defines such preferences [154], even if some guidelines which suggest principles of general gratification have been produced [34, 181, 283] – some of them being modeled in a computational sense by the field of Computational Media Aesthetics (CMA) [3]. However, certain guidelines which suggest principles of general gratification have been produced. Some of those guidelines have roots in the cognitive science: facial attractiveness of symmetric faces is one of the most known example. For real-world scenes, there is high agreement in observer's preferences ratings: factors such as naturalness, complexity, coherence, legibility, vista, mystery and the refuge seem to produce shared agreement [135], most probably due to the survival utility of a particular environment or viewpoint. Considering colors, a study reported in [183] showed that human subjects prefer blue and dislike yellow, unveiling intriguing continuity between animal and human color aesthetics. Regarding the shape, the most important principle discussed in the literature is that of the "Golden Ratio": the idea is that a rectangle, whose ratio between height and width is the same as the ratio of their sum to their maximum, is more attractive than other rectangles.

Even if many CMA applications have been developed, only the "aesthetics" side of the problem has been addressed: from aesthetic photo ranking [62, 93, 304] to preference-aware view recommendation systems [257], to picture quality analysis [3, 137, 172, 174]. Nevertheless, these technologies ignore the potential role that factors internal the observer may have on preferences. Recently, researchers have drawn ideas from the aforementioned to address yet more challenging problem such as associating pictures with aesthetics and emotion that they arouse in humans, with low-level image composition beneficiaries of this research include general consumers, media management vendors, photographers, and people who work with art [62, 174]. Researchers in aesthetic quality inference also need to understand and consider human subjectivity and the context in which emotion or aesthetics is perceived. As a result, ties between computational image analysis and psychology, study of beauty and aesthetics in visual art, including photography, are also natural and essential.

The experiments of [62] 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 1664 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 [172] over 6000 images (3000 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 [158], 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 [193] they present AVA, a new dataset for aesthetic visual analysis. It contains a rich variety of meta-data including a large number of aesthetic scores for each image. They took into account three types of annotation: aesthetic, semantic and photographic style. 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 [298] 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 [62]. Like in the other works presented so far, the task is a binary classification and the accuracy is 78.5%. The approach proposed in [178] adopts an alternative approach for what concerns the features. Rather than using features inspired by good practices in photography, 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 12000 pictures and the accuracy in a binary classification task (6000 pictures per class) is close to 90%.

The idea that aesthetic value is connected to features goes back at least to Birkoff in the 1930s [24]. [59] discriminates paintings from photographs taking into account several features derived from the color, edge, and gray-scale-texture information of the image from a collected dataset based on 12000 photos. Datta in 2006 investigated on the features that underpin aesthetic value in artistic and photographic images [62]. Many subsequent studies have applied machine learning to photographic digital images and or art images with the goal of predicting the aesthetic rating of these images from features engineered to measure image properties such as colorfulness, brightness and texture. Examples are [12, 40, 41, 48, 62, 67, 90, 93, 102, 171, 173].

[16] explored how a number of factors relate to human perception of importance. Proposed factors fall into 3 categories: the ones related to composition (size, location), the ones related to semantics (category of object or scene) and the ones related to the likelihood of attribute-object or object-scene to be combined. Most subsequent work focuses on photographic images and studies of art works borrowed features engineered for photographs. For example [48] bases the study on the set of features designed and published by Datta [62]. Additional papers that invent image features that other researchers have used for further works are [63, 137].

[67] developed techniques for estimating high level describable attributes of images that are useful to predict perceived aesthetic qualities and interestingness in images. The term *high level describable attributes* is used to indicate that these are the kind of characteristics that a human might use to describe an image.

The main contribution of [279] is an approach for predicting which photos a user is likely to call favorite based on social, textual and visual signals. Further they provided an in-depth analysis of user behavior on Flickr, and show how this effects their decision to label photos as favorites.

An interactive system, PIXEE [189], was developed to promote greater emotional expression in image-based social media. Images shared on social media were projected onto a large interactive display at public events. A multimodal interface displayed the sentiment analysis of images and invited viewers to express their emotional responses. Viewers could adjust the emotional classification and thereby change the frame color and sound associated with a picture, and experiment with emotion-based composition. An interdisciplinary team deployed this system around the world to explore new ways for technology to catalyze emotional connectedness.

[274] explores the rare combination of speech, electrocardiogram, and a revised self-assessment mannequin to assess people's emotions. Additionally, their personality traits of Neuroticism and Extroversion, demographic information (i.e., gender, nationality, and level of education) were recorded. [171] present a method based on multi-column deep convolutional networks for predicting the aesthetic rating of photographic images.

2.2 Social Signal Processing

There is more than words in social interactions, whether these take place between humans or between humans and computer. This is well known to social psychologists that have studied non-verbal communication for several decades, referring to it as *how people communicate, intentionally or unintentionally without words*.

During social interaction, non-verbal behavior conveys information not only for each involved individuals but it also determines the natures and quality of the social relationships they have with others. This happens through a wide spectrum of non-verbal behavioral cues that are perceived and displayed mostly unconsciously while producing social awareness, i.e. a spontaneous understanding of social situations that require little attention or reasoning [147, 212].

Non-verbal behavior and non-verbal social signal play a major role in shaping the perception of social situations and interactions: [7]. There is a growing research in cognitive science, which argue that our common view of intelligence is too narrow, ignoring a crucial range of abilities that matter immensely for how people do in life. This range of abilities is called *social intelligence* and includes the ability to express and recognize social signals and social behaviors like turn taking, agreement, politeness, and empathy, coupled with the ability to manage them in order to get along with others while winning their cooperation. Social signals and social behaviors are the expression of ones attitudes towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioral cues including facial expressions, body postures and gestures, and vocal outbursts like laughter.

Social Signal Processing (SSP) is the research and technological domain that aims at providing computers with the ability to sense and understand human social signals by modeling, analyzing and synthesizing non-verbal communication. Domains like SSP have shown that non-verbal behavioral cues are reliable evidence for machine understanding of social affective and psychological phenomena.

SSP approach is performed in three steps. First, human science provide suggestions about the behavioral cues related to a social phenomenon of interest. Second, signal processing approaches are applied to the data to extract the behavioral cue of interest. Third, a machine learning approach is used to model the behavioral cue and understand the social phenomenon of interest.

According to SSP paradigm the ultimate goal of this investigation is the development of approaches capable of predicting automatically not only how individuals perceive one other, but also how they react to the non-verbal cues they mutually display.

The psychological point of view of perception

Cognitive psychology is the study of mental processes such as attention, language use, memory, perception, problem solving, and thinking [1]. Social psychology on the other hand, is the scientific study of how people's thoughts, feelings, and behaviors are influenced by the actual, imagined or implied presence of others. The terms mental processes, thoughts, feelings, and behaviors include all psychological variables that are measurable in a human being. Social psychologists typically explain human behavior as a result of the interaction of mental states and immediate social situations. These two areas are becoming even more involved during the perception, whether it concerns person, environment, objects etc..

Person perception and implicit perception are the two major fields of research that are intriguing the social psychologists in the last decades [113, 140]. While person perception is the study of how people form impression of others, implicit perception instead refers to any change in the person's experience thought or action that is attributable to such an event, in the absence of conscious perception of that event.

Social cognition is a growing area of social psychology that studies how people perceive, think and remember about others. Much research rests on the assertion that people think about (other) people differently from non-social targets [190]. Social cognition has shown that people attribute, spontaneously, unconsciously and with automatic cognitive processes of which they are unaware, a wide range of socially relevant characteristics to others [268, 269] and they even occur in zero acquaintance scenarios and early stage of an interaction.

Furthermore, the effect is so pervasive and ubiquitous that it takes place not only when people meet others in person, but also when we see them in pictures[207], we watch them in videos [194] or we listen to them in audio recordings [191, 231]. From a multimedia point of view, the main effect is that the perception of social, cognitive and psychological phenomena taking place in the data influences significantly what we remember about the data we consume.

The key aspect of the phenomenon is the perception-action link, namely the automatic and unmediated activation of behavioral patterns after the very simple perception of appropriate stimuli, whether these correspond to verbal messages, context and environment characteristics or non-verbal behavioral cues [68]. Form a computing point of view the perception-action link is interesting for two main reasons: the first is that it makes human behavior potentially easier to predict. In fact, if a certain stimulus tends to elicit always the same behavioral pattern, the uncertainty about behavior under observation can be reduced. The second reason is that human behavior can possibly be changed by generating or displaying appropriate stimuli. In both cases, the key issue is to understand how people perceive a given stimulus, i.e. what is the social meaning that people tend to attach to it.

Implicit cognitive processes and Multimedia Content Analysis

The exchange of multimedia data has become a form of human-human communication and, therefore, it should give rise to the same cognitive phenomena (e.g., see [268, 269]) typically observed in any human-human interaction.

To the best of our knowledge, this thesis is the first attempt to adopt such a perspective in multimedia technologies. The application of the *socio technical* perspective in studying the use of digital libraries - until a few years ago the most common infrastructure for the exchange of multimedia data - is one of the earliest attempts to take into account social issues in technological applications: "*To understand, use, plan for and evaluate digital libraries, we need to attend to social practice, which we define as people's routine activities that are learned, shaped, and performed individually and together*" [25]. The main difference between the socio technical perspective and the research direction proposed in this thesis is that the former focuses on use and usability issues (especially in professional and institutional settings) while the latter targets the communication between individuals, a step made possible only by recent technologies (social media, mobile devices, etc.).

In parallel, several efforts have been carried out to improve multimedia technologies by automatically detecting and understanding emotional, behavioral and physiological reactions of data consumers (e.g., if a person watching a video laughs, then the video can be tagged as "funny") [5, 157, 210]. The core idea of these trends is that the content of the data produces observable changes in data consumers, then the observation of these latter provides information about the data. The main difference with respect to this thesis is that the accent is on the data content, like in most of the multimedia technologies, and not on the communication process underlying the data exchange between individuals.

More recently, some works investigated the interplay between observable characteristics of multimedia data and cognition [302]. [302] investigates the characteristics of abstract paintings that stimulate certain emotional reactions rather than others. The work shifts the attention from the bare content of images to their potential role in a communication process, namely a painter eliciting emotions. When it comes to personality and images, the literature is scarce as it is a very new research challenge. Most of them take into account other social networks platforms as Facebook and Twitter [99, 103, 117, 143, 222, 223], and use other information publicly available on the user profile. However, unlike the perspective advocated in this thesis, those works take into account only one of the parties involved in the communication process. To the best of our knowledge, few works seem to consider multimedia data as a form of communication [76, 84]. The work in [84] studies the perception of profile pictures on social media and, in particular, the agreement between the actual personality traits of profile holders and traits attributed by others based on the profile picture. The work in [76], does a similar analysis, but it considers all elements that can appear in a profile. Not surprisingly, these works focus on social media, an interaction-oriented technology that allow users to use multimedia material to communicate with others. However early, the approaches in [76, 84] seem to confirm the action of implicit cognitive processes when using multimedia data in a communication scenario, the key-idea advocated in this thesis. Still, both works focus on a specific case and do not try to identify the underlying perspective that can be applied to many different cases.

User content generation and information disclosure

With the proliferation of Internet chat room, e-mail, news-groups and personal website the Internet has become a pervasive medium for social interaction[281]. Social networking websites (SNWs) as well such as Facebook, Twitter, Instagram, Pinterest, Vine are playing an increasingly prominent role in everyday social interactions.

The particular role of SNWs varies across relationships; in some contexts SNWs supplement existing real-world social networks but in other contexts, interactions can be entirely mediated by SNWs. People may even use them to gather information on others (e.g., prospective employers; marketers). The result of all these changes is that SNWs have become a central medium of interpersonal perception [76].

The content created by the users of social networking sites has reached such high levels of quality and variety that it is comparable to that produced by professional agencies. Therefore, understanding what types of content users generate and underlying motivational factors is vital to the success of the sites [307]. In particular, information disclosure is starting to become an important issue in the virtual world. With the use of SNWs other can see not only the visible information that we choose to share in our profile pages, website and posts but still even those information that other can percept in an unconscious way and we freely reveal. For our knowledge the research in the information disclosure faces just one side of the problem, the privacy [105, 185, 215, 248], (e.g. [276] deals with the privacy awareness on Flickr).

3

Image Representation in Computational Aesthetics

This chapter presents the features proposed in this thesis, extracted at image level. Although this is one of the contributions of the thesis, we present it in this preliminary part as it represents as well one of the main components of the thesis.

The adopted features focus on the contributions of *Computational Aesthetics* 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) [161], but the raters cannot access these motivations and can only access the appearance of the pictures.

From a computational perspective, we need to consider steps that are necessary to obtain a prediction from an input image. However, it is important to understand and appreciate certain inherent gaps when any image understanding problem is addressed in a computational way. The aesthetics gap is the lack of coincidence between the information that one can extract from visual data and the aesthetics responses or interpretation of emotion that the visual data may arouse in a particular user in a given situation [132].

We present two different data representations used in this thesis: low level hand crafted and high level deep learning descriptors. We noticed towards this thesis that the first type of descriptors are more interpretable at human level and that for the second there is still a wide need of understanding and interpretation of them.

3.1 Low Level Representation: Hand Crafted Features

In the last decades, there have been significant contributions to the field of features extraction and image representation for semantics and image understanding. Feature extraction and image representation are prerequisites to any image understanding task, and aesthetics inference makes no exception. There are psychological studies that show that aesthetics responses to a picture may depend upon several dimensions such as composition, colorfulness, spatial organization, emphasis, motion, depth, or presence of humans [62, 174]. Conceiving meaningful visual properties that may have correlation with perceived aesthetics is itself a challenging problem. In the literature, we notice a spectrum from very generic color, texture, shape features to specifically designed features descriptors that are expected to capture the perceptual properties that contribute to the aesthetic value of a picture.

Popular feature extraction techniques like, e.g., SIFT [170] and HOG [60] 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.

Indeed, we are not interested in extracting the classic aesthetic qualities of an image, but the aspects that make an image good for particular users; being this last goal slightly different, many factors and dimensions of analysis can be taken into consideration, each one considered for the purpose of describing different aspects of an image. Therefore, we wanted to span our selection from simple and standard image descriptors up to complex and state-of-the-art ones.

A synopsis of the features adopted is available in Table 3.1. The features cover a wide, though not exhaustive, spectrum of visual characteristics and are grouped into four main categories: *color*, *composition*, *textural properties* and *content*. This follows the taxonomy proposed in [62, 174].

Category	**Name**	**d**	**Short Description**
Color	HSV statistics	5	Average of S channel and standard deviation of S, V channels [174]; *circular variance* in HSV color space [180]; *use of light* as the average pixel intensity of V channel [62]
	Emotion-based	3	Measurement of *valence*, *arousal*, *dominance* [174, 272]
	Color diversity	1	Distance w.r.t a uniform color histogram, by Earth Mover's Distance (EMD) [62, 174]
	Color name	11	Amount of *black*, *blue*, *brown*, *green*, *gray*, *orange*, *pink*, *purple*, *red*, *white*, *yellow* [174]
Composition	Edge pixels	1	Total number of edge points, extracted with Canny detector [167]
	Level of detail	1	Number of regions (after mean shift segmentation) [50, 94]
	Average region size	1	Average *size* of the regions (after mean shift segmentation) [94]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [62, 174]
	Rule of thirds	2	Average of S,V channels over inner rectangle [62, 174]
	Image parameters	2	Size and aspect ratio of the image [62, 167]
Textural Properties	Gray distribution entropy	1	Image entropy [167]
	Wavelet based textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on HSV channels [62]
	Tamura	3	Amount of *coarseness*, *contrast*, *directionality* [261]
	GLCM - features	12	Amount of *contrast*, *correlation*, *energy*, *homogeneousness* for each HSV channel [174]
	GIST descriptors	24	Output of GIST filters for scene recognition [206].
Content	Objects	28	Objects detectors: we kept the number of instances and their average bounding box size [79]
	Faces	2	Number of faces (extracted manually) or number and size of faces after Viola-Jones face detection algorithm [291]

Table 3.1: Synopsis of the features. Every image is represented with 112 features split in four major categories: Color, Composition, Textural Properties, and Content.

3.1.1 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 3.1). The H channel provides information about color diversity through its *circular variance* R [180]:

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

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 [62]. The average of the Hue was not calculated because it cannot be associated to an intensity attribute (low, high), being an angular measure. Figure 3.1 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) [272]:

$$\text{Valence} = 0.69 \cdot \bar{V} + 0.22 \cdot \bar{S} \tag{3.1}$$
$$\text{Arousal} = -0.31 \cdot \bar{V} + 0.60 \cdot \bar{S} \tag{3.2}$$
$$\text{Dominance} = -0.76 \cdot \bar{V} + 0.32 \cdot \bar{S} \tag{3.3}$$

where $\bar{V}$ and $\bar{S}$ are the averages of V and S over an image, respectively. See Figure 3.1 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 [62], 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 3.1 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 [17]: *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 [273] 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.

3.1.2 Composition

The organization of visual elements across an image as distinct from the subject is referred to as *composition*. 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 [167] to identify the edges and calculate the fraction of pixels in an image that lie on an edge (see Figure 3.2 for an example of edge extraction in an image).

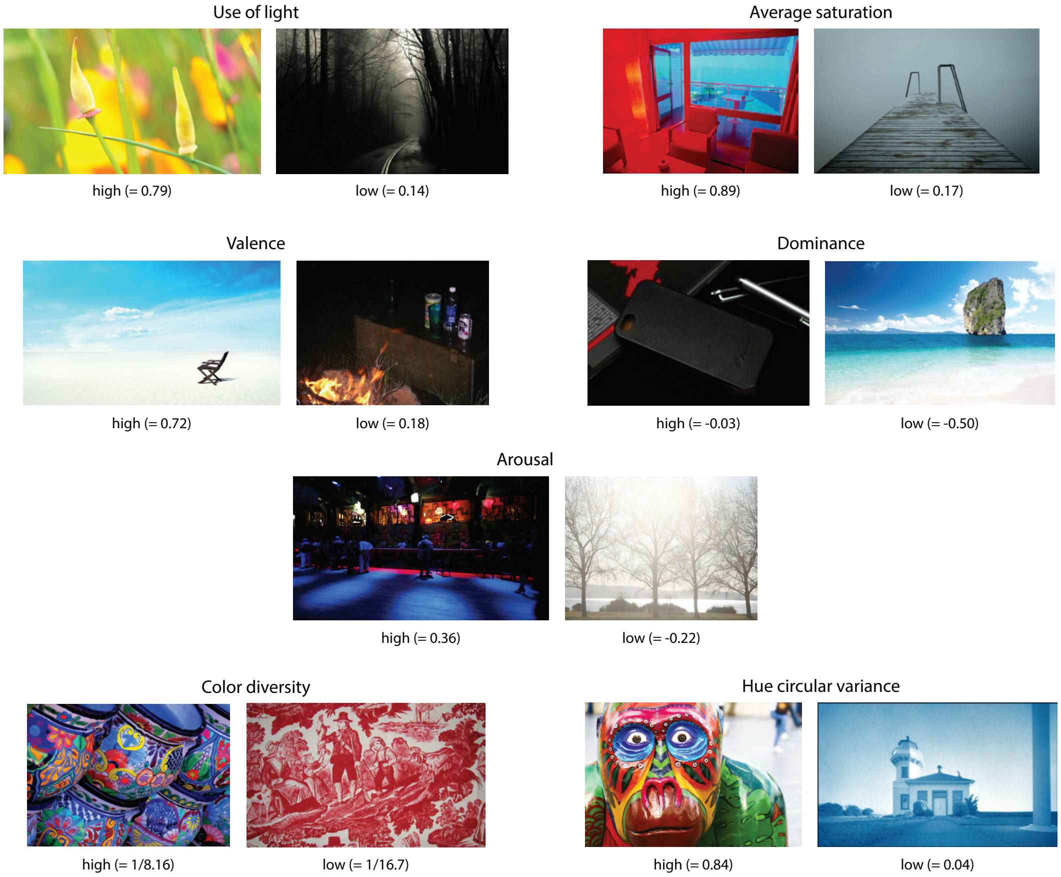


Fig. 3.1: Examples of how the visual properties of a picture change according to several color-related features.

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 [94] of the mean shift algorithm [50], 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 3.2).

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 [62, 174] (see Figure 3.2). 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. (3.7)) 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 $M1, \dots M16$, 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,

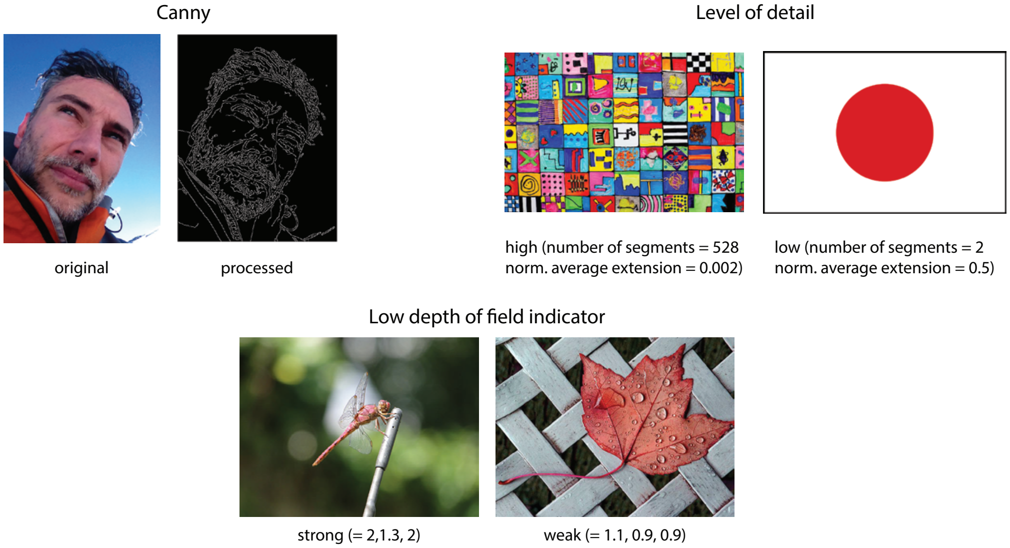

Canny
original
processed
Level of detail
high (number of segments = 528
norm. average extension = 0.002)
low (number of segments = 2
norm. average extension = 0.5)
Low depth of field indicator
strong (= 2,1.3, 2)
weak (= 1.1, 0.9, 0.9)

Fig. 3.2: Effect of the Canny algorithm and an example of the visual properties associated to Level of Detail and Low Depth of Field.

$$\mathrm{DOF}_H = \frac{\sum_{(k,l)\in M_6\cup M_7\cup M_{10}\cup M_{11}} w_3(k,l)}{\sum_{i=1}^{16}\sum_{(k,l)\in M_i} w_3(k,l)} \tag{3.4}$$

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 3.2 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 [62, 174]:

$$f_S = \frac{9}{KL}\sum_{k=K/3}^{2K/3}\sum_{l=L/3}^{2L/3} S_{kl} \tag{3.5}$$

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 parameters

The size and aspect ratio of the image, calculated as the total number of pixels, and ratio between width and length of the image, are used as feature.

3.1.3 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 3.3 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.

$$E = -\sum_i P_i Log_2 P_i \tag{3.6}$$

Wavelet textures

Daubechies wavelet transform can measure the spatial smoothness/graininess in images [62, 174]. 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 3.4). 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.

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:

$$wf_i = \frac{\sum_{k,l} w_i^h(k,l) + \sum_{k,l} w_i^v(k,l) + \sum_{k,l} w_i^d(k,l)}{(|w_i^h| + |w_i^v| + |w_i^d|}, \tag{3.7}$$

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 3.4).

Tamura

In [261], 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

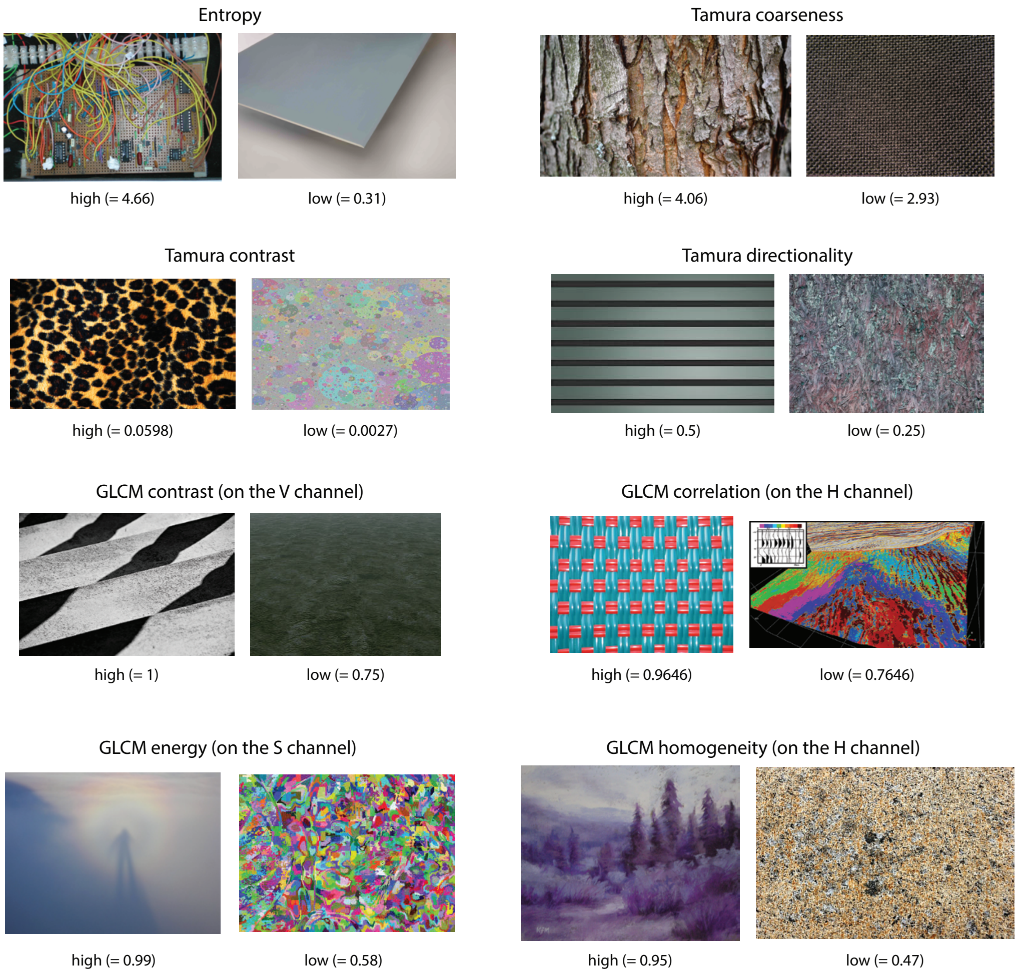


Fig. 3.3: Examples of pictures where the value of the textural features is high and low.

a small number of large texels, while a fine texture contains a large number of small texels. (see Figure 3.3). The coarseness measure is computed as follows. Let X an $I \times J$ matrix of values $X(i,j)$ that can e.g. be interpreted as gray values:

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

$$A_k(i,j) = \frac{1}{2^{2k}} \sum_{n=1}^{2^{2k}} \sum_{m=1}^{2^{2k}} X(i - 2^{k-1} + n, j - 2^{k-1} + m) \tag{3.8}$$

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

$$E_k^h(i,j) = |A_k(i + 2^{k+1}, j) - A_k(i - 2^{k-1}, j)| \tag{3.9}$$

and

$$E_k^v(i,j) = |A_k(i, j + 2^{k+1}) - A_k(i, j - 2^{k-1})| \tag{3.10}$$

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

$$S(i,j) = \arg\max_{k=1\dots 5} \max_{d=h,v} E_k^d(i,j) \tag{3.11}$$

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

$$F_{crs} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} 2^{S(i,j)} \tag{3.12}$$

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

$$F_{con} = \frac{\sigma}{\alpha_4^z} \quad \text{with} \quad \alpha_4 = \frac{\mu_4}{\sigma_4} \tag{3.13}$$

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 3.3.
Directionality: It models how polarized is the distribution of edge orientations. High directionality indicates a texture where the edges are homogeneously oriented, and viceversa. 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). Viceversa, pictures with edges whose orientation is distributed in a uniform manner will have low directionality (~ 0). For an example, see Figure 3.3.

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 [111]:
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 3.3 for examples of pictures with high and low contrast). *Correlation*: It is the coefficient that measures the covariation between neighboring pixels:

$$\sum_{i,j=0}^{L-1} \frac{(i-\mu)(j-\mu)p(i,j)}{\sigma^2}, \tag{3.14}$$

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 3.3).

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 3.3).

Homogeneity is a measure of how frequently neighboring pixels have the same value (see Figure 3.3):

$$H = \sum_{i,j=0}^{L-1} \frac{p(i,j)}{1 + |i-j|} \tag{3.15}$$

The feature tends to be higher when the elements on the diagonal of the GLCM are larger (see Figure 3.3).

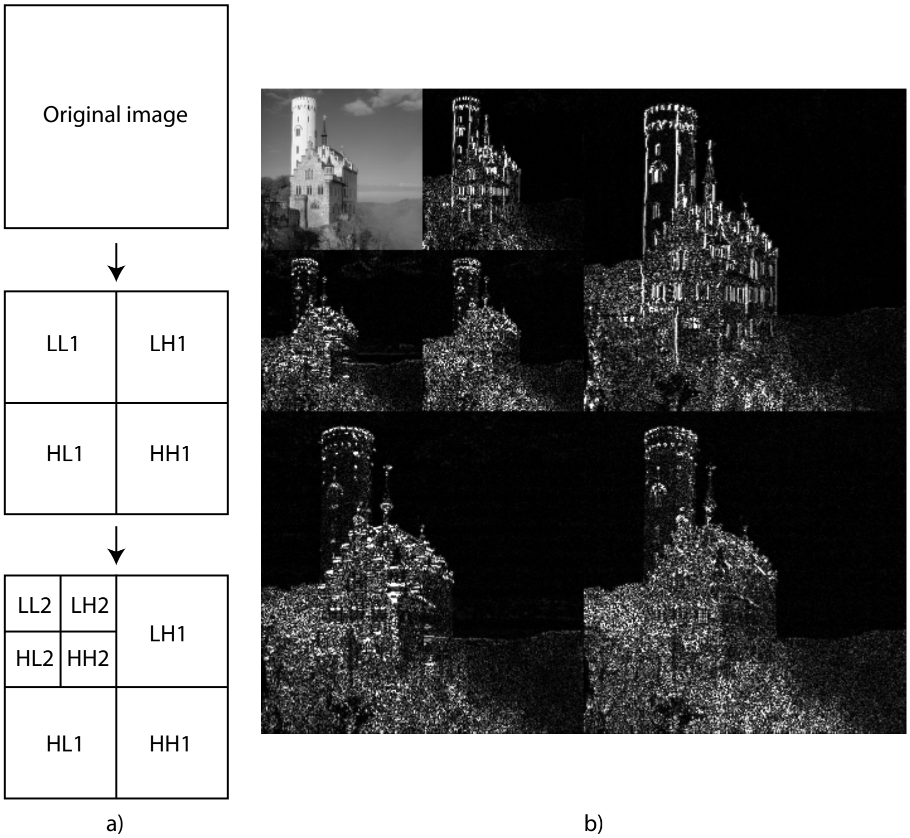


Fig. 3.4: The figure shows how the wavelet decomposition works.

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* [206]. The outputs of the GIST filters are used as features.

3.1.4 Content

The content corresponds to the objects that a picture contains. The semantic variability of the pictures posted online is wide and it is not possible to detect any possible object. However, the literature proposes several approaches aimed at finding objects that appear more frequently than the others. The features described in this section aim at capturing the semantic present inside a picture.

Objects

Once again motivated by [58, 118], we employed the Deformable Part Models [79, 80] system to detect objects. The algorithm works by detecting and localizing a specific object (for example a plane, a cat, a chair or a person), through the use of a model learned from a a set of training examples. The system can detect different objects; in our approach we used as features the number of times every detectable object is present in the image (for a complete list of all the detectable objects see [79]); we also retained the average area (the algorithm gives also the bounding box of the detected objects), to guess if objects are more towards the background of the foreground.

Number of faces

Human faces are frequently portrayed in Flickr pictures and, furthermore, there are neural pathways that make the human brain particularly sensitive to faces [129]. In this thesis, the number of faces is sometimes calculated manually for each of the pictures of the dataset to have a more robust and reliable affability of the feature; in some works instead we extracted the number and size of the faces present int he image employing the standard Viola-Jones face detection algorithm [291] implemented in the OpenCV libraries [1]. Every visible face was counted, irrespectively of its scale, pose, size and occlusion. Facial expressions were not taken into account. For some works 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 [291] identifies only 70% of the faces in the corpus. This introduces noise difficult to model and quantify.

3.2 High Level Representation: Feature Learning

As machine learning extends learning from engineered features set towards a new paradigm of representational learning and features discovery, computational aesthetics has begun to exploit these capabilities also. A number of recent publications related to computational aesthetics more generally, make reference to the developments of deep learning and how these intersect their topics [40, 102, 109, 171, 282]. Ginosar et al. [96] argue for extending the evaluation of computer vision systems for object detection in images to corpora that include art works having characteristic abstractions such as part-reorganization in Cubist paintings and blurring in Impressionist works.

During the first and second year we focused mostly on generative and discriminative approaches to attempt our goal of applying computational aesthetics to multimedia data and to analyze the relationship between individual characteristics (including both personality traits ad information typically available on social media) and aesthetic preferences.

Automatically understanding and modeling a user's liking for an image is a challenging problem. This is because the relationship between images features and user's 'likes' is non-linear, influenced by several factors. Further hand-crafted features are time consuming to extract, they may not disentangle all the explanatory factors of the data. We approached to the field of deep and feature learning trying to understand if there is another possible representation of the data that is human interpretable and representative, easier to use for learning and more efficient than the classical hand crafted features.

Deep learning is a branch of machine learning bases on a set of algorithm that attempt to model high-level abstractions in data by using model architectures, with complex structures or otherwise, composed of multiple non-linear transformations. One of the promises of deep learning is replacing hand crafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchal feature extraction. A key idea of representation learning [14] is the automatic discovery of new high level representations from raw data sources, which express salient features in a form that is more powerful than any features set of human design, and using this representation to perform typical machine learning task such as classification, regression, clustering. In the case of probabilistic model, for instance, a good representation is often the one that captures the posterior distribution of the underlying factors for the observed input. In the case of deep learning data representation can be seen as a hierarchical representation to develop statistical model that can discover more abstract and useful aspects of the data, composing multiple non-linear transformations.

Deep learning algorithms are based on distributed representations. The underlying assumption behind distributed representations is that observed data is generated by the interaction of many factors on different levels, corresponding to different levels of abstraction or composition. Deep learning helps to

[1] http://opencv.willowgarage.com/wiki

disentangle these abstractions and pick out which features are useful for learning.

There are several state of the art approaches of deep learning: Convolutional Neural Network [153], Deep Belief Network [115], Auto Encoders [15, 234, 235], Deep Boltzmann Machine [196, 243, 254] and Hierarchical Deep model [244, 306]. In this thesis we use the Convolutional Neural Network framework for a first exploratory analysis and classification of our data in this wide field of research.

3.2.1 Convolutional Neural Network

Computational models of neural networks have been around for more than half a century. Beginning from the Perceptron model and going to the Feed-Forward Neural Network model, the back-propagation algorithm [239] has been defined over a multilayer Feed-Forward Neural Network, or FFNN. A FFNN can be thought in terms of neural activation and the strength of the connections between each pair of neurons. Because we are only concerned with feed forward networks, the pools of neurons are connected together in some directed, acyclic way so that the networks activation has a clear starting and stopping place (i.e. an input pool and an output pool). The pools in between these two extremes are known as hidden pools. The flow of activation in these networks is specified through a weighted summation process. Each neuron sends its current activation any unit is connected to, which is then multiplied by the weight of the connection to the receiving neuron and passed through some squashing function, typical a sigmoid, to introduce nonlinearities (if this were a purely linear process, then additional layers would not matter, since adding two linear combinations together produces another linear combination). Since we typically assume each layer to be fully connected to the next layer, these calculations can be done via multiplying the vector of activations by the weight matrix and then passing all of the results through the squashing function.

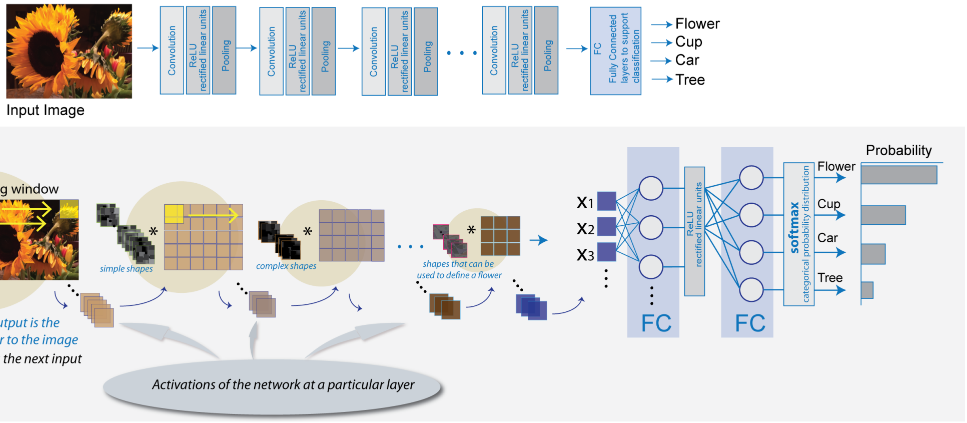

Input Image
Convolution
ReLU rectified linear units
Pooling
Convolution
ReLU rectified linear units
Pooling
Convolution
ReLU rectified linear units
Pooling
Convolution
ReLU rectified linear units
Pooling
FC Fully Connected layers to support classification
Flower
Cup
Car
Tree
Sliding window
Filters
light and dark
simple shapes
complex shapes
shapes that can be used to define a flower
X1
X2
X3
FC
ReLU rectified linear units
FC
softmax
categorical probability distribution
Probability
Flower
Cup
Car
Tree
Every feature map output is the result of applying a filter to the image
The new feature map is the next input
Activations of the network at a particular layer

Fig. 3.5: Convolutional Neural Network architecture and layers.

Learning in these networks occurs through changing the weights so as to minimize some error function, typically specified as the difference between the output pool's activation vector and the desired activation vector. Normally this is accomplished incrementally via the previously mentioned back-propagation algorithm, in which the partial derivative of the error with respect to last layer of weights is calculated (and generally scaled down) and used to update the weights. Then the partial derivatives can be calculated for the second-to-last weight layers and so on, with the process repeating recursively until the weight layer connected to the input pool is updated.

Despite being a universal function approximation in theory, FFNNs were not good at dealing with many sorts of problems in practice. The example relevant to the current discussion is the FFNN's poor ability to recognize objects presented visually. Since every unit in a pool was connected to every unit in the next pool, the number of weights grew very rapidly with the dimensionality of the input, which led to slow learning for the typically high dimensional domain of vision. Even more disconcerting was the spatial ignorance of the FFNN. Since every pair of neurons between two pools had their own weight, learning to recognize a object in one location would not transfer to the same object presented in a different part of the visual field; separate weights would be involved in that calculation. What was needed was an architecture that exploited the two dimensional spacial constraints imposed by its input modality whilst reducing the amount of parameters involved in training. Further NNs and back-propagation limitation are that they need many labeled examples, the initialization strongly affects the performances and it get stuck in local optima.

The first breakthrough idea was to add a layer-wise unsupervised pre-training [75, 115] that could help for a better initialization of the net and to escape poor local optima, learning the features by either reconstructing the input with deterministic auto-encoder [148, 234, 284] or predicting the input with stochastic RMBs [114]. But it was necessary to fine-tuning the parameters. Then, Convolutional neural networks are the architecture (see Figure 3.5).

Convolutional Neural Network (CNN) is a particularly interesting and special class of feed forward networks that are very well-suited to image classification [151, 153]. The solution to FFNNs' problems with image processing took inspiration from neurobiology, Yann LeCun and Yoshua Bengio tried to capture the organization of neurons in the visual cortex of the cat, which at that time was known to consist of maps of local receptive fields that decreased in granularity as the cortex moved anteriorly [149]. There are several different theory about how to precisely define such a model, but all of the various implementations can be loosely described as involving the following process:

1. Convolve several small filters on the input image.
2. Subsample this space of filter activations.
3. Repeat steps 1 and 2 until your left with sufficiently high level features.
4. Use a standard a standard FFNN to solve a particular task, using the results features as input.

Indeed the input and output of each stage are sets of arrays called features maps. For example, if the input is a color image, each feature map would be a 2D array containing a color channel of the input image. At the output, the feature map represents a particular features extracted at all locations of the input. Each stage is composed of three layers: a filter bank layer or convolutional layer, a non-linearity layer and a feature pooling layer or subsampling.

Convolution

Convolution is a mathematical term, defined as applying a function repeatedly across the output of another function. In this context it means to apply a "filter" over an image at all possible offsets. A filter consists of a layer of connection weights, with the input being the size of a small 2D image patch, and the output being a single unit. Since this filter is applied repeatedly, the resulting connectivity looks like a series of overlapping receptive fields, which map to a matrix of the filter outputs (or several such matrices in the common case of using a bank of several filters). The result of one filter applied across the image is called feature map (FM) and the number of feature maps is equal to the number of filters. If the previous layer is also convolutional, the filters are applied across all of it's FMs with different weights, so each FM is connected to each output FM. The intuition behind the shared weights across the images is that the features will be detected regardless of their location, while the multiplicity of filters allows each of them to detect different set of features. The module indeed computes $y_j = b_j + \sum_i k_{ij} * x_i$ where $*$ is the 2D discrete convolution operator, k_{ij} is a trainable filter (kernel) in the filter bank and connects input feature map x_i to the output feature maps y_j and b_j is a trainable bias parameter.

Non-Linearity layer

In traditional CNN this simply consists in a point-wise $tanh()$ sigmoid function applied to each site of the input. An important subtlety here is that why there are still a good deal of connections between the input layer and the filter output layer, the weights are tied together. This means that during back-propagation, you only have to adjust a number of parameters equal to a single instance of the filter, a drastic reduction from the typical FFNN architecture. Another nuance is that we could sensibly apply such filter to any input that's spatially organized, not just a picture. This means that we could add another bank of filters directly on top of our first filter bank's output. However, since the dimensionality of applying a filter is equal to the input dimensionality, we would not be gaining any translation invariance with these additional filters, we'd be stuck doing pixel-wise analysis on increasingly abstract features. In order to solve this problem, we must introduce a new sort of layer: a subsampling layer.

Subsampling

Subsampling, or down-sampling, refers to reducing the overall size of a signal. In many cases, such as audio compression for music files, subsampling is done simply for the size reduction. But in the domain of 2D filter outputs, subsampling can also be thought of as increasing the position invariance of the filters. The specific subsampling method used in LeNets [151] for instance, is known as "max pooling". This involves splitting up the matrix of filter outputs into small non-overlapping grids (the larger the grid, the greater the signal reduction), and taking the maximum value in each grid as the value in the reduced matrix. Semantically, this corresponds to changing the question answered by the convolution layer from "how well does this filter apply right here" to "how well does this filter apply to this area". Now, by applying such a max pooling layer in between convolutional layers, we can increase spatial abstractness as we increase feature abstractness. This results in a reduced-resolution output feature map which is robust to small variation in the location of features in the previous layer. There are also other pooling ways like average pooling or stochastic pooling. The last subsampling layer is usually connected to one or more fully connected layers, the last of which represents the target data.

Put it all together

Even with the convolution and subsampling layers specified, there are still many free hyper-parameters. Namely, how many filters per convolutional layer? How big should the filters and subsamples be? And how many overall layers should there be? None of these questions can be answered definitively, as the effectiveness of each hyper-parameter setting depends on the task setting, but LeCun attempted to provide some plausible values for roughing simulating human performance on natural image classification tasks. In LeNet5 [152], for instance, there are five functional layers in the total architecture, corresponding to 2 sets of convolution-max-pooling pairs and a FFNN for solving the actually classification problem. While training takes quite some time, this network learns much faster than a standard FFNN and performances quite well as a pure piece of computer vision software. Training is performed using back-propagation that takes the subsampling layers into account and updates the convolutional filters weights based on all values to which that filters is applied. Supervised training is performed using a form of stochastic gradient descent to minimize the discrepancy between the desired output and the actual output of the network. The gradients are computed with the back-propagation method. Recently it has been proved that putting on the top of the network a Softmax layer to perform classification reach better performances.

In [146] for instance, they trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet into 1000 different classes. The neural network consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax.

4

Probabilistic Modeling

In this chapter we introduce a brief review of the concept of *statistical pattern recognition*, an approach to machine intelligence which is based on statistical modeling of data and *probabilistic modeling*.

With a statistical model in hand, one applies probability theory and decision theory to get an algorithm. We exploit the basic model on a typical pattern recognition scenario. Then we review the differences between supervised and unsupervised learning, generative and discriminative paradigms. Many common pattern recognition algorithms are probabilistic in nature, in that they use statistical inference to find the best label for a given instance. Unlike other algorithms, which simply output a "best" label, often probabilistic algorithms also output a probability of the instance being described by the given label.

In this direction, we present and overview about *probabilistic modeling*, the area that focus on how to use probability to model and analyze data and in particular on generative probabilistic models. Data in real world almost always involves uncertainty. This uncertainty may come from noise in the measurements, missing information, or from the fact that we only have a randomly sampled subset from a larger population. Probabilistic models are an effective approach for understanding such data, by incorporating our assumptions and prior knowledge of the world. These ideas are important in many areas of computer science, including machine learning, data mining, natural language processing, computer vision, and image analysis.

4.1 Introduction: Statistical Pattern Recognition

Pattern recognition is a field of study developed significantly in the 1960s. It was very much an interdisciplinary subject, covering developments in the areas of statistics, engineering, artificial intelligence, computer science, psychology and physiology, among others. Some people entered the field with a real problem to solve. The large number of applications, ranging from the classical ones such as automatic character recognition and medical diagnosis to the mere recent ones in data mining, have attracted considerable research efforts, with many methods developed and advances made. Other researchers were motivated by the development of machine with "brain-like" performance, that in some way could emulate human performance.

Interest in the area of pattern recognition has been renewed recently due to emerging applications which are not only challenging but also computationally demanding. These applications include data mining (identifying pattern, correlation or an outlier in millions of multidimensional patterns), document classification (efficiently searching text documents), financial forecasting, organization and retrieval of multimedia databases, and biometrics (personal identification based on various physical or behavioral attributes). Picard [217] has identified a novel application of pattern recognition, called *affective computing* which will give computer the ability to recognize and express emotions, to respond intelligently to human emotion, and to employ mechanisms of emotions that contribute to rational decision making.

To some extent there exist strong parallels with the growth of research on knowledge-based system in the 1970s and neural networks in the 1980s. The topic could be easily described under the term machine learning that describes the study of machines that can adapt to their environment and learn from example. The emphasis in machine learning is perhaps on computationally intensive models and less on statistical approach, but there is strong overlap between the research areas of statistical pattern recognition and machine learning.

Statistical pattern recognition (or learning) is a term used to cover all stages of an investigation from problem formulation and data collection to discrimination and classification, assessment of results and interpretation.

4.1.1 The Basic Model

In a typical pattern recognition scenario, the main goal is to analyses processes or sets of data using statistical models, that assign the data or the processes to classes of membership w_c with $c = 1, 2, \ldots, C$.

Statistical models are built by using training observations, i.e. a set of delegates that represent the data or the processes being modeled [122].

We shall use the term *pattern* to denote the d-dimensional data vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ of measurements, whose components x_i are measurements of the features - they could be continue or discrete random variable representing the probability of an event or just quantitative measurements - that represents the statistical model.

The recognition system is operated in two modes: training (or learning - the step with which a statistical model is built) and classification (testing). The role of preprocessing module is to segment the pattern of interest from the background, remove noise, normalize the pattern, and any other operation which will contribute in defining a compact representation of the pattern.

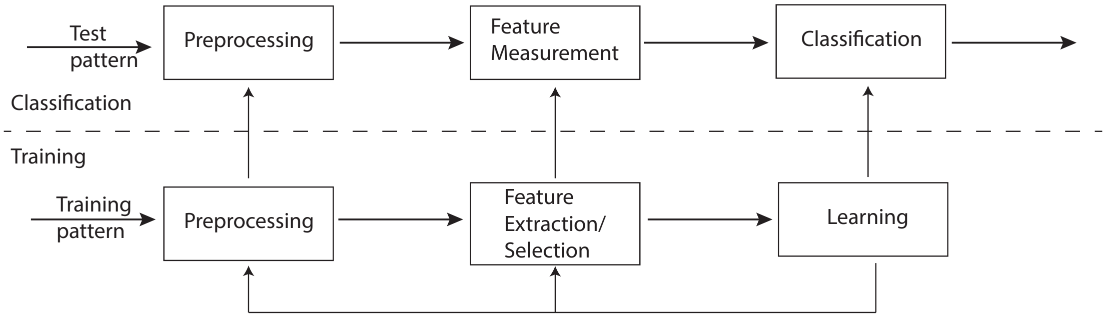


Fig. 4.1: The typical classification model.

4.1.2 Supervised and Unsupervised Learning

The learning mode is defined *supervised* when the membership class label of each sample is known a priori. If the training set is a collection of data where we do not know the membership class of the samples, i.e. the data are not labeled, and we do not have information about the classes, then we call the learning mode *unsupervised*. In this case, the learning step permits also to discover automatically the natural clusters of the observations. Note that the recognition problem here is being posed as a classification or categorization task, where the classes are either defined by the system designer (in supervised classification) or are learned based on the similarity of patterns (in unsupervised classification).

In the supervised training mode the feature extraction/selection module finds the appropriate features for representing the input pattern and a classifier is trained to partition the feature space. The feedback path allows a designer to optimize the preprocessing and feature extraction/selection strategies.

In the classification mode, the trained classifier assign the input pattern to one of the pattern classes under consideration based on the measured features. Regression is just like classification expect the response variable is continuous.

The decision making process in statistical pattern recognition can be summarized as follows [122, 192]: a given pattern is assigned to one of the c categories $w_1, w_2, \ldots, w_c$ based on a vector of d features values $x = (x_1, x_2, \ldots, x_d)$. The features are assumed to have a probability density or mass (depending on whether the features are continuous or discrete) function conditioned on the pattern class.

If we have to chose the class we can follow the simple *decision rule*: Decide w_1 if $p(w_1) > p(w_2)$ for instance; otherwise we consider x to be a continuous random variable whose distribution depends on the state of the nature expressed as $p(x, w_1)$, that is a pattern vector x belonging to a class w_i is viewed as an observation drawn randomly from the class-conditional probability function $p(x, w_i)$. The joint probability density of finding a pattern of category w_i and has feature value x then can be written as: $p(w_i, x) = p(w_i, x) \cdot p(x) = p(x|w_i) \cdot p(w_i)$. So the Bayes' decision rule can be derived as

$$p(w_i|x) = \frac{p(x|w_i) \cdot p(w_i)}{p(x)} \rightarrow \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

where the evidence can be seen as a scaling factor that guaranties that the posterior probability sum to 1. The *posterior* probability indeed is the probability of assigning observations to groups given the data, the *prior* is the probability that an observation will fall into a group before you collect the data and the *likelihood* is a function of the parameters of the model, the hypothetical probability that an event that has already occurred would yield a specific outcome (how probable the observed data is for different settings). In the general case of more classes $p(x) = \sum_{i=1}^{C} p(x|w_i) \cdot p(w_i)$, called also marginal probability, because it is obtained by marginalizing or summing out the other possible variables.

A number of well known decision rules, including the Bayes' decision rule, the maximum likelihood rule (which can be viewed as a particular case of the Bayes' rule) etc., are available to define the decision boundary.

Various strategies are utilized to design a classifier in statistical pattern recognition, depending on the kind of information available about the class-conditional densities. If all the class-conditional densities are completely specified, the Bayes' rule can be used to design a classifier. However, class-conditional densities are usually not known in practice and must be learned from the available training patterns. If the form of the class-conditional densities is known (e.g multivariate Gaussian), but some of the parameters of the densities (e.g mean vectors and covariance matrices) are unknown, then we have a parametric decision problem. A common practice for this kind of problem is to replace the unknown parameters in the density function with their estimated values. If the form of the class-conditional densities is not known then we operate in a nonparametric mode. In this case we must estimate the density function or directly construct the decision boundary based on the training data.

Bayesian methods view the parameters as random variables having some known a priori distribution. Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters. In the Bayesian case, we shall see that a typical effect of observing additional samples is to sharpen the a posteriori density function, causing it to peak near the true values of the parameters, phenomena known as Bayesian learning.

In the case of nonparametric mode, we can take advantages of the *maximum likelihood estimation* that views the parameters as quantities whose values are fixed but unknown. The best estimate of their values is defined to be the ones that maximizes the probability of obtaining the samples actually observed.

No matter which classification or decision rule is used, the classifier must be trained using the available training samples. The goal of designing a recognition system is to classify future test samples which are likely to be different from the training samples. Therefore, optimizing a classifier to maximize its performance on the training set may not always results in the desired performances on a test set. The generalization ability of a classifier refers to its performance in classifying test patterns which were not used in the training stage. A poor generalization ability of a classifier can be attributed to any one of the following factors:

1. the number of features is too large relative to the number of training samples (curse of dimensionality)
2. the number of unknown parameters associated with the classifier is large (e.g. polynomial classifiers are a large neural network)
3. a classifier is too intensively optimized on the training set (overtrained); this is analogous to the phenomenon of overfitting in regression when there are too many free parameters.

In many application of pattern recognition, it is extremely difficult or expensive, or even impossible, to reliably label a training sample with its true category. Unsupervised classification refers to situations where the objective is to construct decision boundaries based on unlabeled training data.

Unsupervised classification is also known as data clustering which is a generic label for a variety of procedures designed to find natural groupings, or cluster, in multidimensional data, based on measured or perceived similarities among the patterns.

4.1.3 Generative and Discriminative Paradigms

In the machine learning literature two kinds of approaches to classification have been developed: the generative and the discriminative paradigms [30, 131, 192].

Discriminative models are concerned with defining boundaries between categories, they are not interested in the data per-se, but in how classes differentiates. To do this, they predict the distribution of the class variable given the input: $P(class|input)$. Discriminative approaches work best when the data is extensively preprocessed so that the amount of data relative to the complexity of the task is increased. Such preprocessing involves analyzing the unprocessed inputs that will be encountered. This task is performed by a user who may or may not use automatic data analysis tools, and involves building a model of the input, $P(input)$.

Once this model is available, the combination of $P(input)$ and the discriminative model $P(class|input)$ corresponds to a particular decomposition of a generative model: $P(class, input) = P(class|input) \cdot P(input)$ [97]. Discriminative models have performed well in many scientific areas; for example object or scene recognition [28, 160], economics [119], bio-informatics [163] and text recognition [27, 211] have known a huge improvement with the use of support vector machines (SVMs) over generative approaches. Depending on the features, it has been proved that discriminative methods provide performances higher than those achieved by the generative models, especially when large training sets are available [200]. Despite of this, they have no modeling power, it is very difficult to inject prior knowledge, and they need to be re-trained if a new class is inserted [82].

Generative models provide a more general way to combine the preprocessing task and the discriminative task. In particular, by jointly modeling the input and the output (i.e., the class variable), they do not require to build $P(input)$ and they can discover useful, compact representations and use them to better model the data. When limited training data are available, generative models work betters and their ability to model uncertainty while still absorbing prior knowledge, is what provoked the transfer from traditional techniques mostly to generative probabilistic modeling [200].

Generative models are built to understand how samples were generated. Without any notion of discrimination, they simply explain the data providing a rigorous platform to combine prior knowledge with observed data and exploiting useful cues/interaction between the data. In the generative models, other than the observed quantities, we model the generative process that produced the observations, which is not directly visible from the data, it is hidden, eventually encoded in an a-priori knowledge. To perform classification with such model one has to learn a model per-class and assign a new point to the category whose model fits the point best, i.e. the model with highest likelihood.

More formally, a generative model is specified by a joint distribution $P(input, class)$ and it can be used for discrimination by computing $P(class|input)$ using marginalization and Bayes' rule.

Summarizing, in standard regimes, discriminative classifiers achieve better performances [130]. This is rather intuitive since they are concerned with defining boundaries between categories; they are not interested in the data per-se, but in how classes differentiates [280]. On the other hand generative models are built to understand how samples were generated; without any notion of discrimination, they simply explain the data. Generative models are powerful tools to model natural processes upon which one can build powerful classification algorithms.

4.2 Probabilistic Models

In order to build robust classifiers, one has to capture and model various aspects of the data at the same time, often dealing with uncertainties. Uncertainties are implicitly in the data: uncertainties about which features are most useful for processing or classifying the data, uncertainties in the relationships between variables, and uncertainties in the value of the action that is taken as a consequence of inference. Probability theory offers a mathematically consistent way to model problems and formulate inference algorithms when reasoning under uncertainty. In particular, probabilistic modeling focuses on how to use probability to model and analyze data. Depending on the task at hand, it can be convenient to employ model that further exploits the information contained in the data, or feature vectors. This is done in order to increase the performances, to explain the process of data generation (if the goal is classification or clustering), to highlight particular facets of the data providing a more interpretable description (if the goal is visualization/interpretation) or to validate the observed representation.

Beginning form the definition of a key distributions, probabilistic manipulation can be expressed in terms of two simple equations, known as the sum rule and the product rule [26]. In general, given two random variables a and b we can write the following equations:

$$\textbf{(Sum rule)} \quad p(a) = \sum_b p(a,b) \tag{4.1}$$

$$\textbf{(Product rule)} \quad p(a,b) = p(b|a)p(a) \tag{4.2}$$

The sum rule is sometimes called marginalization, and the sum is over all possible values b can take. Note also that the summation must be replaced by an integral if b is continue rather than discrete. All of the probabilistic inference and learning manipulations discussed in this thesis, no matter how complex, amount to repeated application of these two equations. For example, by applying two times the product rule, one can easily derive the Bayes' rule, which states the following:

$$\textbf{(Bayes' rule)} \quad p(a|b) = \frac{p(b|a)p(a)}{p(b)} \tag{4.3}$$

This basic equation serves as the main ingredients of the probabilistic *generative models* [86].

4.2.1 Generative Models

The goal of generative modeling, as stated in the previous section, is to formally develop statistical models that can explain the input data, or visible variable $\mathbf{v}$, as tangible *effects* which are generated from a combination of hidden variables $\mathbf{h}$, representing the *causes*, in case coupled with conditional interdependencies. As a result, a generative model jointly models the input and the causes via the joint distribution

$$\text{Generative model} = P(\mathbf{v}, \mathbf{h}) = p(\text{effects, causes}) \tag{4.4}$$

A generative model is hence a statistical model for which the observed data is an event in the sample space. So, sampling from the model generates a sample of possible observed data. If the training data has high probability, the model is a good fit. However, the goal is not to find the model that is best fit, but to find a model that fits the data well and is consistent with prior knowledge.

We turn now to a general problem of building a generative model. Given a set of observations considered as a set of instances of i.i.d random variable $\mathcal{D}$, the steps that lead to the related generative model are:

1. **Intuitive definition**: definition of the intuitive hidden causes that generated the observed variables. The hidden causes can be related by conditional dependencies.

2. **Statistical definition**: this step is needed to manage the uncertainty, both in the definition of the hidden variables and in the observations themselves, eventually corrupted by noise. It consists on coding each hidden cause as hidden random variable and describing the whole generation process with a joint distribution. This distribution should be factorized, taking into account the relations among the causes defined in the previous step. Conditional dependencies lead to conditional densities, independences bring to independent densities, and so on. In this step, the definition of conditional independence needs to be included, in order to simplify the factorization of the joint distribution, by canceling some useless conditioning.

3. **Parametrization**: this step consists on the parametrization of all the densities involved in the factorization. The parameters should be thought as hidden quantities either and, together with the values assumed by the hidden variables, forming the set $\mathbf{h}$. Therefore, the joint distribution can be indicated as $P(\mathbf{v}, \mathbf{h})$.

4. **Inference**: at this point, the most important step in the learning of the hidden quantities using the observations, i.e., choosing a possible instance of the value for $\mathbf{h}$ that maximize the posteriori distribution $P(\mathbf{h}|\mathbf{v})$.

An effective method aimed at representing the generative process of the data is given by graphical models. Graphical models use graphs to represent and manipulate joint probability distributions, and are very useful instruments aimed to the construction of efficient generative models.

4.2.2 Graphical Models and Bayesian Networks

Graphical models are an important tool for representing the dependencies between random variables in a generative model. They are important for two reasons. First, graphs are an intuitive way of visualizing dependences. They are used to represent graphical depictions of dependency, for example, in circuit diagrams and in phylogenetic trees. Second, by exploiting the structure of the graph it is possible to advise efficient algorithms for computing marginal and conditional probabilities in a complex model.

The main statistical properties represented explicitly by the graph are conditional and marginal independence between variables.
Consider three variables a, b and c, and suppose that the conditional distribution of a, given b and c, is such that it does not depend on the value of b, so that

$$P(a|b, c) = P(a|c).$$

We say that a is conditionally independent of b given c. This property can be expressed in a slightly different way if we consider the joint distribution of a and b conditioned on c, which we can write in the form

$$p(a, b|c) = P(a|b, c)P(b|c) = P(a|c)P(b|c)$$

Thus we see that, conditioned on c, the joint distribution of a and b factorizes into the product of the marginal distribution of a and the marginal distribution of b (again both conditioned on c). This says that the variables a and b are statistically independent, given c. Conditional independence is really different from marginal independence which states that $P(a, b) = P(a)P(b)$, being the latter the usual notion of independence between a and b. Conditional independence and the concept of d-separation [214] play and important role in using probabilistic models for pattern recognition by simplifying both the structure of a model and the computations needed to perform inference and learning under that model. We will make an example in the following.

A graphical model is a probabilistic model for which a graph denotes the conditional independence structure between random variables. There are several different graphical formalisms for depicting conditional independence relationships, namely undirected graphs like Markov Random Field, directed graphs like Bayesian networks and Factor graphs.

A Bayesian network for random variables (RVs) $\mathbf{x} = \{x_1 \ldots, x_N\}$ is a directed acyclic graph (no directed cycles) on the set of RVs.

The nodes x_i of the graphs represent the variables, while the arcs represent the probabilistic dependencies between them. There are two kind of nodes, the hidden nodes h_i which model the hidden variables, and the observables nodes v_i which model the visible variables.

The joint distribution (i.e. the generative model) over the N variables is given by

$$P(\mathbf{x}) = P(x_1, \ldots, x_N) = P(\{v_i\}, \{h_i\}) \tag{4.5}$$

By repeated application of the product rule of the probability, this joint distribution can be written as a product of conditional distributions, one for each of the variables

$$P(x_i, \ldots, x_N) = \prod_i P(x_i | x_1, \ldots, x_{N-1}) \ldots P(x_2|x_1)P(x_1) \tag{4.6}$$

This factorization can be applied to any kind of graph, however the absence of links in the graphs conveys interesting information about the properties of the class of distributions that the graph represents. The conditional independence property specified with the topology of a Bayesian networks states that a node is necessarily conditionally independent of its non-descendants given its parents. What this mean is simply that given a node x_i and its parents x_{A_i} (there may not be any parents), and given another node y not reachable from x_i, we have that

$$P(x_i, y|x_{A_i}) = P(x_i, x_{A_i})P(y|x_{A_i}) \tag{4.7}$$

This properties can be used to simplify the factorization of a generative model which turn to be given by the product of all the conditional probability functions associated to the models, i.e.

$$P(\mathbf{x}|\theta) = P(x_1 \ldots x_N|\theta) = \prod_{i=1}^{N} P(x_i|X_{A_i}, \theta_i) \tag{4.8}$$

where N is the number of RVs, and θ is the set of all the parameters, i.e. $\theta = \bigcup_{i=1}^{N} \theta_i$. This key equation express the factorization properties of the joint distribution for a directed graphical model. For example the generative models represented by the Bayesian network in Figureure 4.2 factorize as

$$P(a, b, c) = P(a)P(b)P(c|a, b)$$
$$P(a, b, c) = P(a|c)P(b|c)P(c)$$

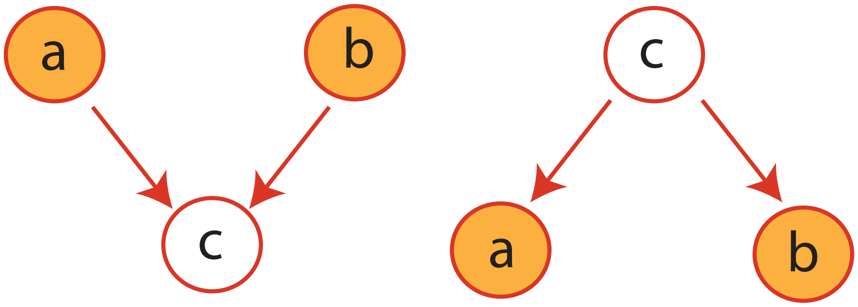


Fig. 4.2: The Bayesian network with three variables (a, b, c). a, b are hidden variables, c is visible.

Summarizing a Bayesian network is characterized by nodes which represent the quantities in hands, arcs that express conditional independences properties and by one conditional probability function for each RV given its parents, $P(x_i|x_{A_i})$ (see Eq. 4.8). Each conditional probability function $P(x_i|x_{A_i})$ can be governed by a set of (hidden) parameters θ_i which give a known parametric form to that distribution. For example we can model a conditional distribution (i.e. the probability of x_i given its parents) as a Gaussian probability density function, with mean μ and variance ψ; in this case we have that

$$p(x_i|x_{A_i}, \theta_i) = \mathcal{N}(x_i; \mu, \psi), \quad \theta_i = \{\mu, \psi\} \tag{4.9}$$

Although we have considered each node to correspond to a single variable, we can equally well associate sets of variables and vector-valued variables with the nodes of a graph.

Example: Occlusion model

To better understand the design of a Bayesian network we present a vision example.

The occlusion model explains an input image with pixel intensities $z_1, \ldots, z_N$ as a composition of a foreground image and a background image, and each of these images is selected respectively from a library of F and B possible images. Foreground, background and the mask represent the causes of the image generation; what we have just done is complete the first step, the *intuitive definition*, toward the creation of a generative model. Figureure 4.3 summarized this concept; on the left some images sampled from the occlusion model, on the right the causes of the image generation.

The *statistical definition* requires the specification of the generative process which can be described by a graphical model. The generative process is the following and it is illustrated in Figureure 4.4A.

1. A foreground image is randomly selected from the library by choosing the class index f from the distribution, $P(f)$.

2. Depending on the class of the foreground, a binary mask $m = (m_1, \ldots, m_K), m_i \in \{0, 1\}$ is randomly chosen; $m_i = 1$ indicates that pixel z_i is a foreground pixel, whereas $m_i = 0$ indicates that pixel z_i is a background pixel. The distribution over the mask depends on the foreground class, since the mask must "cut out" the foreground object. However, given the foreground class, the mask RVs are chosen independently: $P(m|f) = \prod_{i=1}^{K} P(m_i|f)$.

3. A background image index is randomly selected from the library by choosing the class index b.

4. At the end the image pixels are generate, given al the other causes: $P(z|m, f, b) = \prod_{i=1}^{K} P(z_|m_i, f, b)$.

This generative process is encoded by the Bayesian network depicted in Figure 4.4B. The joint distribution can be obtained by applying Eq. 4.8 to the graphical model:

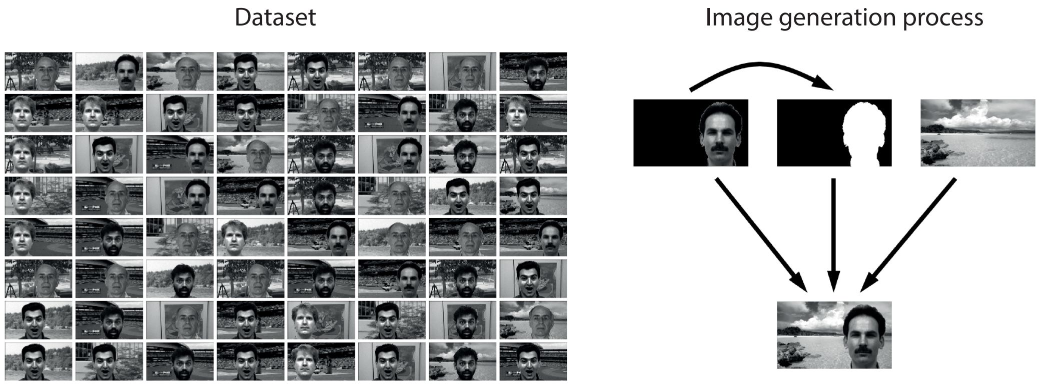


Fig. 4.3: On the left some images generated by the causes indicated on the right.

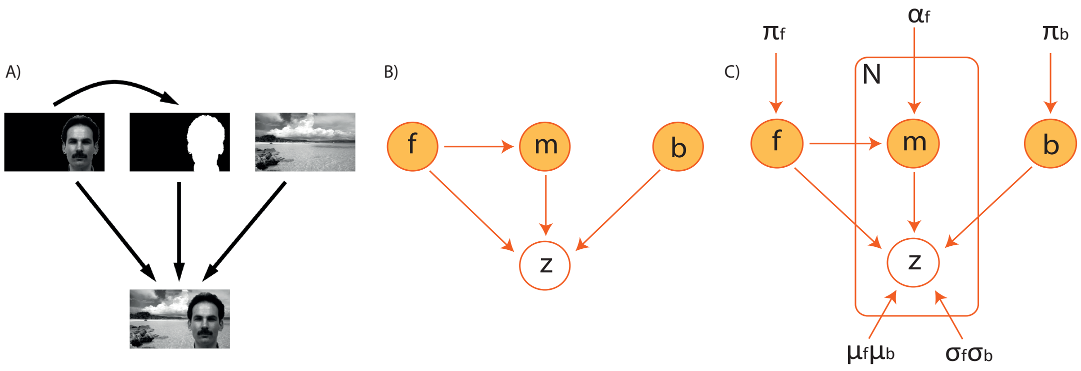


Fig. 4.4: Bayesian network that encodes the generative process explained in the text.

$$P(z, m, f, b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i|f)\right)\left(\prod_{i=1}^{K} P(z_i|m_i, f, b)\right) \tag{4.10}$$

This equation $P(z_i|m_i, f, b)$ can be further factorized by noticing that if $m_i = 0$ the pixel value z_i is given by b, otherwise from f. So, we can write

$$P(z_i|m_i, f, b) = P(z_i|f)^{m_i} P(z_i|b)^{1-m_i} \tag{4.11}$$

where $P(z_i|f)$ and $P(z_i|b)$ are distributions over the i-th pixel intensity given by the foreground and background respectively. The joint distribution becomes:

$$P(z, m, f, b) = P(b)P(f)\left(\prod_{i=1}^{K} P(m_i|f)\right)\left(\prod_{i=1}^{K} P(z_i|f)^{m_i}\right))\left(\prod_{i=1}^{K} p(z_i|b)^{1-m_i}\right) \tag{4.12}$$

The third step in defining a generative model is the *parametrization* where one is asked to specify a model using parametric distributions for each conditional distribution. Given a foreground (background) class index $f(b)$, we assume z_i is equal to μ_{f_i} plus zero-mean Gaussian noise with variance ψ_{f_i}:

$$p(z_i|m, b, f) = \mathcal{N}(z_i; \mu_{f_i}, \psi_{f_i})^{m_i} \mathcal{N}(z_i; \mu_{b_i}, \psi_{b_i})^{1-m_i} \tag{4.13}$$

where $\mathcal{N}$ stands for the Gaussian probability function. We denote the probability of (either a background of foreground) class k by π_k

$$P(f) = Mult(f|\pi_f) \quad P(b) = Mult(b|\pi_b)$$

We let the probability that $m_i = 1$ given that the foreground class is d, be α_{f_i} Since the probability that $m_i = 0$ is $1 - \alpha_{f_i}$, we have

$$P(m_i|f) = \alpha_{f_i}^{m_i}(1 - \alpha_{f_i})^{1-m_i}$$

Once performed the parametrization it is usual including the parameters in the Bayesian network (see Figureure 4.4C). We used the plate notation for representing variables that repeat in a graphical model. In practice, instead of drawing each repeated variable individually (m_i, $z - i$ for each pixel in our case), a plate or rectangle is used to group variables into a subgraph that repeat together, and a number is drawn on the plate to represent the number of repetitions of the subgraph in the plate.

Combining the parametric forms for the conditional distributions and Eq. 4.8, the joint distribution becomes

$$P(z, m, f, b) = \pi_b \pi_f \left(\prod_{i=1}^{K} \alpha_{f_i}^{m_i}(1 - \alpha_{f_i})^{1-m_i} \mathcal{N}(z_i; \mu_{f_i}, \psi_{f_i})^{m_i} \mathcal{N}(z_i; \mu_{b_i}, \psi_{b_i})^{1-m_i} \right) \tag{4.14}$$

4.2.3 Learning Graph Parameters

Training data $\mathcal{D}$ can be used to infer plausible configurations of the model parameters. We imagine that there is a setting of the parameters that produced the training data. However, since we only see the training data, there will be many settings of the parameters that are good matches to the training data, so the best we can do is compute a distribution over the parameters. Since the parameters θ are still hidden quantities, we can include them in the set of the hidden RV and treat them like an hidden variable. So, in the following, using $\boldsymbol{h}$ are either described hidden RVs and parameters, unless differently specified. The set $\boldsymbol{h}$ can be divided into the parameters, denoted by θ, and one set of hidden RVs $\mathbf{h}^{(t)}$, for each of the training cases, $t = 1 \dots T$. Similarly, there is one set of visible RVs for each training cases, so $\mathbf{v}^{(1)} \dots \mathbf{v}^{(\mathbf{T})}$.

Assuming the training cases are independent and identically distributed (i.i.d), the distribution over all visible RVs and hidden RVs (including parameters) is

$$P(\mathbf{h}, \mathbf{v}) = P(\theta) \prod_{t=1}^{T} P(\mathbf{h}^{(t)}, \mathbf{v}^{(t)}|\theta) \tag{4.15}$$

where $P(\theta)$ is the parameter prior, $P(\mathbf{h}^{(t)}, \mathbf{v}^{(t)}|\theta)$ is the generative model which factorizes following the Bayesian network rule (Eq. 4.8) for each sample at hand. Using marginalization and/or Bayes' rule we can calculate two other interesting quantities: the *marginal likelihood* and the *posterior distribution*.

The probability of the visible data given the parameters is called marginal likelihood and is obtained summing out all the contributes of the hidden variables, thus

$$P(\mathbf{v}, \theta) = \prod_{t=1}^{T} \sum_{h} P(\mathbf{h}^{(t)} = h, \mathbf{v}^{(t)}|\theta)) \tag{4.16}$$

The sum is replaced with an integral in the case of continuous hidden variables. The other important quantity is the posterior distribution, needed for computing estimates or making decisions. From Bayes' rule:

$$P(\mathbf{h}|\mathbf{v}) = \frac{P(\mathbf{h}, \mathbf{v})}{\sum_h P(\mathbf{h} = h, \mathbf{v})} \tag{4.17}$$

The denominator normalizes the distribution, but if only a proportional function is needed, $P(\mathbf{h}, \mathbf{v})$ suffices since with regard to $\mathbf{h}$, $P(\mathbf{h}|\mathbf{v}) \propto P(\mathbf{h}, \mathbf{v})$.

The major task in Bayesian networks, as in generative modeling, is *learning* or (estimating) the parameters θ. The learning problem can be divided into learning the graph parameters for a known structure, and learning of the topology (i.e., which edges should be present or absent in the graph). Note that we have made no assumption on the hidden variables $\mathbf{h}$, which can be either present or absent.

Once the model is learned, that is, we have an estimate of the parameters θ ot he model structure $\mathbf{m}$, another important task is the *hidden variable inference*.

Probabilistic inference in a graph usually refers to the problem of computing the conditional probability of some variable $\tilde{h}_i$ given the observed values $v = \hat{v}$, while marginalizing out all other variables; in formula

$$P(\{\tilde{h}_i\}|v)$$

Starting from the joint distribution $P(x_i, \ldots, x_N)$, we can divide the set of all variables into three exhaustive and mutually exclusive sets

$$\{x_1, \ldots, x_N\} = \{\tilde{\mathbf{h}} \cup \mathbf{v} \cup \mathbf{h}_{other}\}$$

The variables belonging to the first set are the ones for which we want to calculate the conditional probability $P(\tilde{\mathbf{h}}|v)$, the second set is formed by the visible variables and the third set includes all the other hidden variables $\mathbf{h}_{other} = \mathbf{h} \backslash \tilde{\mathbf{h}}$. We want to compute

$$P(\tilde{\mathbf{h}}|\mathbf{v} = c) = \frac{\sum_o P(\tilde{\mathbf{h}}, \mathbf{h}_{other} = o, \mathbf{v} = c)}{\sum_o \sum_{o'} P(\tilde{\mathbf{h}} = o', \mathbf{h}_{other} = o, \mathbf{v} = c)}$$

but the problem is that the sumer over o is exponential in the number of the variables in $\mathbf{h}_{other}$; for example, if there are M variables in $\mathbf{h}_{other}$ and each is binary, there are 2^M values for o. If the variable are continuous, then the desired conditional probability is the ratio of two high-dimensional integrals, which could be intractable to compute.

There are several algorithms for computing these sums and integrals which exploit the structure of the graph to get the solution efficiently for certain graph structures. For example the belief propagation algorithm is a message passing algorithm for computing conditional probabilities of any variable given the values of some set of other variables in a singly-connected directed acyclic graph. For multi-connected graphs, the standard exact inference algorithms are based on the notion of junction tree [78]. The basic idea of the junction tree algorithm is to group variables so as to convert the multiply connected graph into a singly-connected undirected graph over sets of variables, and so inferences in this tree. We only consider the problem of learning graph parameters when the model structure is known. The absence of a known topology complicates much the situation.

The complete data case

Assume that the parameter θ_i controlling each conditional distribution $P(x_i|x_{A_i}, \theta_i)$ are distinct and that we observe T i.i.d. instances of all N variables in our graph. The data set is therefore $\mathcal{D} = \{\mathbf{x}^{(1)} \ldots \mathbf{x}^{(T)}\}$. With $x_j^{(t)}$ we indicate the j-th variable of the t-th sample. We are in the complete data case and so each variable is visible ($\mathbf{x} = \boldsymbol{v}$). At this point we can write the likelihood in natural way ad the product over the samples, of the factorizations over the visible variables

$$p(\boldsymbol{v}|\theta) = \prod_{t=1}^{T}\prod_{i=1}^{N} p(v_i^{(t)}|v_{A_i}^{(t)}, \theta_i) \tag{4.18}$$

and applying the "log" to the last equation we can write the log likelihood, whose manipulation is often preferred to avoid numerical underflow and to simplify the calculations

$$\text{log}P(\boldsymbol{v}|\theta) = \sum_{t=1}^{T}\sum_{i=1}^{N} \text{log}P(v_i^{(t)}|v_{A_i}^{(t)}, \theta_i) \tag{4.19}$$

Maximizing the log likelihood with respect to the parameters, results in N decoupled optimization problems, one for each family, since the log likelihood can be written as a sum of N independent (logarithmic) terms.

The incomplete data case

When the model is characterized by missing information (and this is the usual case), the likelihood no longer factors over the variables. As previously seen the variables $\{x_i\}_{i=1}^{N}$ can be divided into observed and hidden variables, $\mathbf{v}$ and $\mathbf{h}$. The observed data is now $\mathcal{D} = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(T)}\}$ and the marginal likelihood is obtained through marginalization of the hidden variables:

$$P(\boldsymbol{v}|\theta) = \prod_{t=1}^{T}\sum_{[h_j]} P(\boldsymbol{v}^{(t)}, h_j^{(t)} = h|\theta) \tag{4.20}$$

where with $[h_j]$ we mean all the values that the hidden variable h_j can assume. This sum required by the marginalization yields to a cost function which can no longer be written as a sum of N independent terms since the parameters are all coupled and special algorithms are required to solve this problem.

4.2.4 Maximum Likelihood, Maximum a Posteriori and the EM Algorithm

In all the learning problems, the key quantities we want to optimized are the marginal likelihood or the posterior distributions. A natural choice is to pick the most probable parameters value given the data, this is known as the *maximum a posteriori* or MAP parameter estimate

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max_{\mathbf{h}} P(\mathbf{h}|\mathbf{v}, \theta) \\ &= \arg\max_{\mathbf{h}} \text{log}P(\mathbf{h}|\mathbf{v}, \theta) \end{aligned} \tag{4.21}$$

Another natural choice is the maximum likelihood or ML parameter estimate

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\max_{\mathbf{h}} P(\mathbf{v}|\theta) \\ &= \arg\max_{\mathbf{h}} \sum_{\mathbf{h}} P(\mathbf{h}|\theta) \cdot P(\mathbf{v}|\mathbf{h}, \theta) \\ &= \arg\max_{\mathbf{h}} \text{log}P(\mathbf{v}|\theta) = \mathcal{L}(\theta) \end{aligned} \tag{4.22}$$

where in the last term of both equations we passed in the log-domain. The natural choice for optimizing the function is to use the so-called Expectation Maximization algorithm [65, 74]. The expectation-maximization (EM) algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, when the model depends on unobservable latent variable $\mathbf{h}$. EM is an iterative procedure that alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed:

$$E_{\mathbf{h}|\mathbf{v},\theta_n}[\log P(\boldsymbol{h},\boldsymbol{v}|\theta_n)] \tag{4.23}$$

and a maximization (M) step, which computes the maximum likelihood estimates ot the parameters by maximizing the expected likelihood found on the E step.

$$\theta_{n+1} = \arg\max_{\theta} E_{\boldsymbol{h}|\boldsymbol{v},\theta_n} \tag{4.24}$$

The parameters found on the M step are then used to begin another E step, and the process is repeated.

An EM algorithm can also find MAP estimates [198], be performing MAP estimation in the M step, rather than maximum likelihood.

There are other methods for finding maximum likelihood estimates, such as gradient descent, conjugate gradient or variations of the Gauss-Newton method. Unlike EM, such methods typically require the evaluation of first and/or second derivatives of the likelihood function.

Standard view of the Expectation Maximization algorithm

Part of the reason for the popularity of EM algorithms is that an EM iteration does not decrease the observed data likelihood function $P(\mathcal{D}|\theta) = \prod_t P(\boldsymbol{v}^{(t)}|\theta)$, so EM is guaranteed to find a local optimum of data (log) likelihood. In fact we have that

$$\begin{aligned}\mathcal{L}(\theta) &= \log P(\mathbf{v}|\theta)\\ &= \sum_t \log P(v^{(t)}|\theta)\\ &= \sum_t \log \sum_{\mathbf{h}} P(v^{(t)},\mathbf{h}|\theta)\\ &= \sum_t \log \sum_{\mathbf{h}} P(\mathbf{h}|\theta)P(v^{(t)}|\mathbf{h},\theta)\end{aligned} \tag{4.25}$$

where we note there is a summation inside the log. This couples the parameters θ_i and if we try to maximize the marginal log likelihood by setting the gradient to zero, we will find that there is no longer a nice closed from solution, unlike the complete joint log likelihood with complete data.

The Expectation Maximization algorithm (EM) is an iterative procedure that maximizes the marginal log likelihood $\mathcal{L}(\theta)$ by constructing a concave, easy to optimize lower bound $B(\theta,\theta^k)$, where θ is the variable, while θ^k is the (fixed) set of parameters at the previous iteration. This lower bound has this interesting property

$$B(\theta,\theta^k) = \mathcal{L}(\theta) \tag{4.26}$$

This mean that each choice of θ^{k+1} that maximizes B, is guaranteed to have $B \geq \mathcal{L}(\theta)$ and since B is a lower bound on the marginal log likelihood, we have that $\mathcal{L}(\theta^k) \leq \mathcal{L}(\theta^{k+1})$.

We can obtain the lower bound of B using the Jensen inequality which states that a concave function of a convex combination of points in a vector space is greater than or equal to the convex combination of the concave function applied to the points; in formula

$$\log \sum_i p_i \cdot f_i \geq \sum_i p_i \log f_i \tag{4.27}$$

where p_i's are a proper probability distribution. Given this and recalling that the logarithmic function is concave, we can derive the bound B as follows

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_t \log \sum_{\mathbf{h}} P(v^{(t)}, \mathbf{h}|\theta) \\ &= \sum_t \log \sum_{\mathbf{h}} P(\mathbf{h}|v^{(t)}, \theta) \cdot \frac{P(v^{(t)}, \mathbf{h}|\theta)}{P(\mathbf{h}|v^{(t)}, \theta)} \\ &= \sum_t \sum_{\mathbf{h}} P(\mathbf{h}|v^{(t)}, \theta) \cdot \log \frac{P(v^{(t)}, \mathbf{h}|\theta)}{P(\mathbf{h}|v^{(t)}, \theta)} \\ &\equiv B(\theta, \theta^k)\end{aligned} \tag{4.28}$$

Note that in Eq. 4.28 we introduces the posterior $P(\mathbf{h}|v^{(t)}, \theta)$, separately from each sample $x^{(t)}$, and this is what the E-step of EM step computes, in fact we have

$$E_{\boldsymbol{h}|v^{(t)},\theta}\left[\log P(\boldsymbol{h}, v^{(t)}|\theta)\right] = p(\mathbf{h}|v^{(t)}, \theta) \tag{4.29}$$

The M-step maximizes the lower bound. Most important is that using B we can now set its gradient to zero to obtain a closed form solution.

4.2.5 Simple Generative Models

In order to build robust machine learning algorithms, is necessary that generative models are capable of capturing various aspects of the data at the same time. These models should be fairly simple, but capable of adapting to the data. Flexible models, as defined in the machine learning community, are minimally structured probability models with a large number of parameters that can adapt so as to explain the input data. In this chapter we introduce many flexible models which are later used as blocks for building complex generative models.

Mixture Models

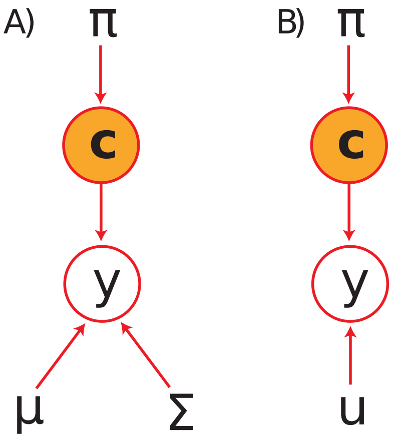


Fig. 4.5: The Bayesian network depicting A) Mixture of Gaussian, B) Mixture of Bernoulli.

To model data with complex structure such as clusters, it is very useful to consider mixture models. We will present here the most straightforward cases: the mixture of Gaussians and the mixture of Bernoulli. The density of each point in a mixture model can be written as

$$P(y|\theta) = \sum_{c=1}^{C} \pi_c \cdot P(h|\theta_c) \tag{4.30}$$

where we have C components, and the parameters θ_c represents the parameters. For example in case of Gaussian distribution we have $\theta_c = \{\mu_c, \psi_c\}$ and π_c is the mixing proportion for the component c, such that $\sum_{c=1}^{K} \pi_c = 1$ and $\pi_c \geq 0 \; \forall c$.

A different way to think about mixture models is to see the mixture variable as latent, and to associate each point to a C-ary discrete latent variable c which has the interpretation that $c = k$ if the data point was generated by component k. This can be formally written as

$$p(y, c|\theta) = \sum_{k=1}^{K} P(c = k|\pi) \cdot P(y|c = k, \theta) \tag{4.31}$$

where $P(c = k|\pi) = \pi_k$ is the prior of the k-th component, and $P(y|c = k, \theta) = P(y|\theta_k)$ is the density under component k; for a mixture of Gaussian this is equal to

$$P(y|\theta_k) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) \tag{4.32}$$

where D is the dimensionality of y. For a mixture of Bernoulli, the probability function is

$$p(y|\theta_k) = \prod_{d=1}^{D} u_{d,k}^{y_d} \cdot (1 - u_{d,k})^{(1-y_d)} \tag{4.33}$$

where y are now binary variables of dimension D.

Latent Dirichlet Allocation (LDA)

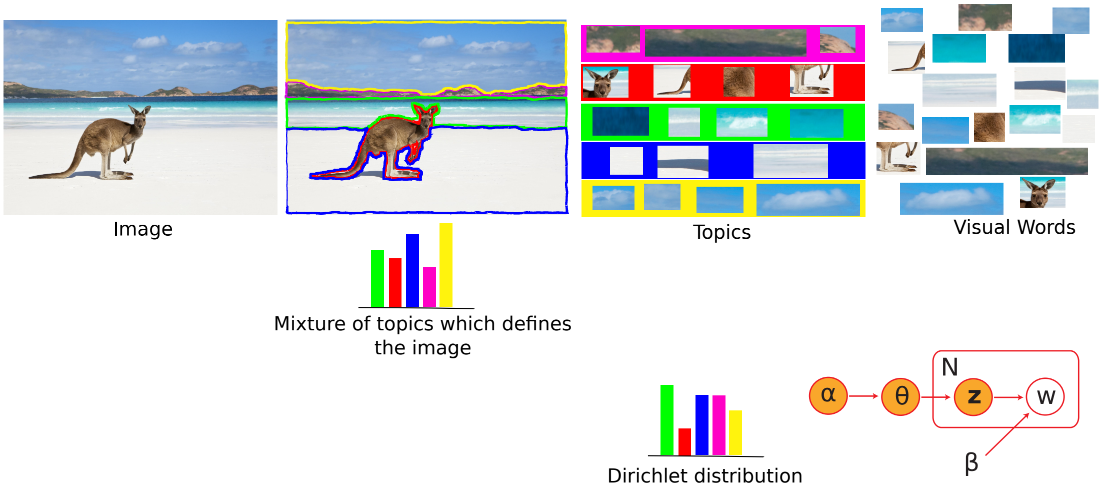


Fig. 4.6: The Bayesian network depicting the latent Dirichlet allocation.

Latent Dirichlet Allocation (LDA), first introduced by Blei [28], is a generative model that can be used to explain how documents are generated given a set of Z topics and a vocabulary of words. This is done looking for co-occurring words clustering them into topics. In the LDA model, words w_n are the only observable variables and they implicitly reflect a latent structure, i.e., the set of Z topics used to generate the document. Generally speaking, given a set of documents $\mathcal{D} = \{\mathbf{w}^{(1)}, \ldots \mathbf{w}^{(T)}\}$ where $\mathbf{w}^{(t)} = \{w_1^{(t)}, \ldots, w_N^{(t)}\}$ the latent topic structure lies in the set of words itself. Figureure 4.6 shows the graphical model for LDA; in generating the document for each word-position a topic is sampled and, conditioned from the topic, a word is selected. Each topic z is chosen on the basis of the random variable θ that is sampled for convenience from a Dirichlet distribution $p(\theta|\alpha)$ where α is the hyperparameter. The topic z conditioned on θ and the words w conditioned on the topic and on β are sampled from multinomial distributions $p(z|\theta)$ and $p(w|z,\beta)$ respectively. β represents the word distribution over the topics. The joint probability of the model can be written as

$$P(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) = \prod_{t=1}^{T} \left(\prod_{k=1}^{Z} P(\theta_k|\alpha) \prod_{n=1}^{M} P(z_n^{(t)} = k|\theta_k) P(w_n^{(t)}|z_n = k, \beta) \right)$$

Learning the various distributions (the set of topics, their associated word probabilities, the topic of each word, the particular topic mixture of each document) is an intractable problem and one must turn to variational approximations.

LDA model was originally developed for topic discovery in a text corpus, where each document is represented by its word frequency, but recently it has successfully applied to the vision domain [29, 162]. In this field the documents are images while the visual words are particular interest points detected; to assign a visual word index to each interest points a visual vocabulary is created. This visual vocabulary is obtained by vector quantizing descriptors computed from the training images using k-means. Beside scene classification, LDA has been applied with success in a number of computer vision scenarios. For instance, in [176] LDA is used in object segmentation and labeling for a large dataset of images. For each image in the dataset multiple segmentations are obtained via different methods these are treated as documents for LDA. An histogram of visual words (SIFT descriptors) is then computed for each segment-document. They then train an LDA model in order to discover topics in the set of documents. Segments corresponding to an object are those well explained by the discovered topics.

Part II

Soft Biometrics for Re-identification

Soft Biometrics

In the last two decades, the study and development of biometric systems have become of paramount importance, from both a scientific and a practical point of view [123].

Classical biometrics offers a natural and reliable solution for establishing the identity of an individual. The use of human physical characteristics has been increasingly adopted in security applications due to various advantages such as universality, robustness, permanence and accessibility. Currently state-of-the-art intrusion detection and security mechanism systems include meanwhile by default at least one biometric trait. The latest addition in this field, is soft biometrics, inheriting a main part of the advantages of classical biometry and furthermore endorsing by its own assets.

The beginnings of soft biometric science were laid by Alphonse Bertillon in the nineteenth century, who firstly introduced the idea for a personal identification system based on biometric, morphological and anthropometric determinations [112]. He used traits like color eye, hair, beard and skin; shape and size of the head; general discriminators like height or weight and also description of indelible marks such as birth marks, scars or tattoos. A great majority of those descriptors fall at the present time into the category of soft biometrics.

Jain et al. first introduced the term soft biometrics to be *a set of characteristics that provide some information about the individual, though these are not able to individually authenticate the subject because they lack distinctiveness and permanence* [121]. Further research has shown that a larger set of soft biometric traits can be used to identify individuals.

A redefinition of Soft Biometrics was proposed by Reid and Nixon as any characteristic which can be naturally described by humans [232]. Later on, the work in [120] additionally noted that soft biometrics are not expensive to compute, can be sensed at a distance, do not require the cooperation of the surveillance subjects and have the aim to narrow down the search from a group of candidate individuals; even more notably, the identification and verification operation can be conducted without letting the user know what is going on [61]. Moreover we here note that the lack of human compliance requirement of soft biometrics is a main factor, which differentiates soft biometrics from classical biometric offering new application fields.

Soft biometric traits are physical, behavioral or adhered human characteristics, classifiable in predefined human compliant categories. These categories are, unlike in the classical biometric case, established and time-proven by human with the aim of differentiating individuals. In other words the soft biometric traits instances are created in a natural way, used by humans to distinguish their peers. The plethora of soft biometrics related benefits motivates the application examination of employing solely soft biometrics traits with the purpose of human identification. This approach is new and has several advantages over classical biometry human identification, as non obtrusiveness, computational and time efficiency to name a few.

Several biometrical traits have been designed, everyone analyzed under different perspective like accuracy, efficiency, usability, acceptability and others; from a very general point of view, they can be divided in three main classes:

- **Physical /physiological** biometric traits, encoding physical characteristic of a person: the face, the fingerprint, the iris [38, 123, 124, 177] - just to cite the most striking examples;

- **Behavioral** biometric traits: more than a physical feature, such traits encode a characteristics linked to the behavior of a person [301], like the gait of the signature [293, 309]. This class can be further partitioned into *authorship-based* (linked to style peculiarities of the individual - how she/he write a book), *motor skill-based* (how a person performs a particular physical task), *purely behavioral* (how a person solves a mentally demanding task) and *HCI-based biometrics*.

- **Adhered** human characteristics, like clothes colour, tattoos, accessories.

The so called **HCI-based behavioral biometrics** [301], are based on the idea that every person has a possibly unique way to interact with a personal computer: for example some methods successfully investigated the possibility of characterizing a person on the base of keystrokes or mouse dynamics [220, 241]. In the same context, very recently some other approaches investigated the exploration of Internet-based biometrical traits, like browsing histories [72, 202].

A novel Soft Biometric Trait: Personal Aesthetics

A very recently brand-new HCI-based biometric trait emerged, exploiting the "*personal aesthetic*" of people, that is, those image preferences that distinguish people from each other [167]. The approach assumes that, given a set of preferred images, it is possible to extract a set of features individuating discriminative visual patterns; these patterns can be used as biometric template, and employed for identification. A further step along this direction has been reached, starting from [167] which focuses on characterizing each individual by his personal aesthetic taste, defined by the most discriminative image features that distinguish him from the rest of the community.

In particular, we take a crowdsearch approach [32] and we focus on Flickr, a popular website where every user can select his/her preferred photos, by tagging them as "*favorite*". This creates, for every user, a set of favorite photos, which is often very heterogeneous and whose modeling/recognition goes beyond standard computer vision tasks such as object/scene recognition. To this purpose we adopted a wide, though not exhaustive, spectrum of features (see Section 3.1) mostly used in the literature. On one side we have the cues that focus on aesthetic aspects [62, 174]: the reason is that Flickr corpus is composed by pictures posted as "*favorite*", i.e. likely to represent the aesthetic and visual preferences of the user under examination. On the other side, we focus on the content of the images.

Having this capability transferred into a machine is without doubts a great benefit for many application: from recommender system that suggest images of interest of a particular user, to social aggregators which foster connection among individual sharing similar aesthetics preferences.

5

“Faved” Biometrics: A New Biometric Trait based on Aesthetic Preferences

5.1 Introduction

This work is the first attempt aimed at investigating how *identifiable* aesthetics traits are, namely if it is possible to model the visual preferences of an individual in a unique way. To do that, a biometric recognition/authentication system is built: in the enrollment phase, the “preference model” of a user is learned from a set of *preferred images*; in the verification/recognition phase, such model is tested with an unseen set of favorites preferred by a probe subject. More in detail, we take a crowdsearch approach [32] and we focus on Flickr[1], a popular website where every user can select his preferred photos, by tagging them as “favorites”. This creates, for every user, a set of favorite photos, which is often very heterogeneous and whose modeling/recognition goes beyond standard computer vision tasks such as object/scene recognition (see Figure 5.1 for an example). In order to infer the *personal aesthetics* trait of a given subject, we analyze his “favorites set”: we characterize each image with different features (the ones proposed in Section 3.1), ranging from low-level color/edge statistics up to more high-level and semantic descriptors such as object detectors and overall scene statistics. LASSO regression is then exploited to learn the most discriminative aesthetic attributes, i.e., the aspects a user likes that distinguish her/him from the rest of the community: such aspects represent the template. In the experiments, involving both

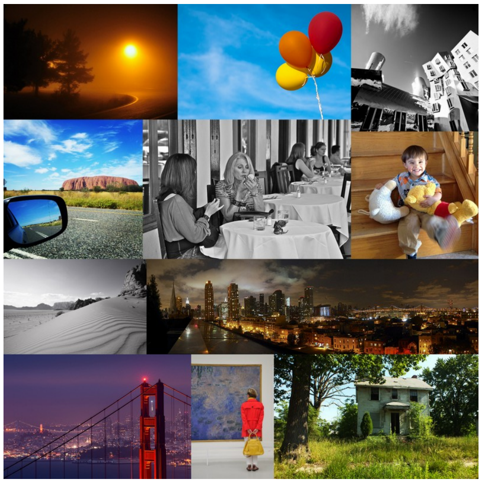


Fig. 5.1: Some samples of favorite images taken at random from a Flickr user.

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

verification and identification, we show that personal tastes act like a blueprint for a user, allowing to recognize him against a set of 200 users with high accuracy; in particular, given just one image from an unknown user, his identity is recognized better than with a random classifier, and this dramatically raises when considering a higher number of images.

5.2 The Proposed Approach

This section describes the main ingredients of our approach. We first introduce the features extracted form the images; then, the learning of the user specific preference model is detailed. Finally, the matching score computation is determined.

5.2.1 Feature extraction

We adopted a wide, though not exhaustive spectrum of features reported in Section 3.1 . In this work we grouped the features into two families, here reported in Table 5.1. On one side, we considered the cues that focus on aesthetic aspects [62, 174], which we will refer to in the remainder as *perceptual* the reason is that the Flickr corpus is composed by pictures posted as “*favorite*”, i.e. likely to represent the aesthetic and visual preferences of the users under examination. On the other side, we focus on the content of the images.

The concatenation of all the descriptors, a vector $\mathbf{x}_m$, represents the proposed signature for the image m. It is worth noting that for the sake of reproducibility, every parameter of he different off-the-shelf computer vision libraries has been left as the default setting.

Category	**Name**	**d**	**Short Description**
Perceptual	HSV statistics	5	Use of light, mean of S channel and standard deviation of S, V channels, circular variance [62, 174, 180]
	Emotion-based	3	Amount of *Pleasure*, *Arousal*, *Dominance* [174, 272]
	Color diversity	1	Colorfulness measure based on Earth Mover’s Distance (EMD) [62, 174]
	Color name	11	Amount of *Black*, *Blue*, *Brown*, *Green*, *Gray*, *Orange*, *Pink*, *Purple*, *Red*, *White*, *Yellow* [174]
	Gray distribution entropy	1	Image entropy [167]
	Wavelet based textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on the HSV channels [62]
	Tamura	3	Amount of *Coarseness*, *Contrast*, *Directionality* [261]
	GLCM-features	12	Amount of *Contrast*, *Correlation*, *Energy*, *Homogeneity* for each HSV channel [174]
	Edges pixels	1	Total number of edge points, extracted with Canny [167]
	Level of detail	1	Number of regions (after mean shift segmentation) [50, 94]
	Average region size	1	Average *size* of the regions (after mean shift segmentation) [50, 94]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [62, 174]
	Rule of thirds	2	Mean of S,V channels in the inner rectangle of the image [62, 174]
	Image parameters	2	Size and aspect ratio of the image [62, 167]
Content	Objects	28	Objects detectors [79]: we kept the number of instances and their average bounding box *size*
	Faces	2	Number and *size* of faces after Viola-Jones face detection algorithm [291]
	GIST descriptors	24	Level of openness, ruggedness, roughness and expansion for scene recognition [206].

Table 5.1: Summary of all features. The column ‘d’ indicates the feature vector length for each type of feature.

5.2.2 Learning the preference model

The preference model for the user u, is built starting from a set of N favorite images (that is, their D-dimensional feature vectors) $\mathbf{x}_1^{(u)}, \ldots, \mathbf{x}_N^{(u)}$, representing the biometrical trait $X^{(u)} = \{\mathbf{x}_1^{(u)}, \ldots, \mathbf{x}_N^{(u)}\}$ In our case, $D = 112$.

Given the biometrical trait $X^{(u)}$, we can build the template as follows. First, we partitioned his favorite images in two sets, one for the training, $X_{\text{tr}}^{(u)}$, composed by N_{tr} images, and one for the testing, $X_{\text{te}}^{(u)}$, composed by N_{te} images. We will see in the next section how crucial is the choice of the value of N_{tr} and N_{te}. In the following, we will also say that the images $X_{\text{tr}}^{(u)}$ will define the *gallery* biometrical trait for the user u, while the images $X_{\text{te}}^{(u)}$ will define the *probe* biometrical trait for the user u.

Since we are interested in characterizing which are the particular personal tastes of the given user, we decided to train a binary classifier, using as positive examples $X_{\text{tr}}^{(u)}$ and as negative the favorites of other Flickr users $\{X_{\text{tr}}^{(v)}\}_{v \neq u}$: this will permit to extract what really makes the subject different from the others. In particular, we represent the discriminative aesthetical aspects of each user as a subset of all the features considered, opportunely weighted. To do that, we perform a sparse regression analysis using LASSO [264].

LASSO is a general form of regularization in a regression problem. In the simple linear regression problem, every n-th training image, described by the proposed feature vector $\mathbf{x}_n$, is associated with a target variable y_n (a positive label is given to all training images coming from user u, that is, the one we want to characterize, whereas the favorites of other users v, $v \neq u$, have a negative label). Then, we can express the target variable as a linear combination of the image features:

$$y_n = \mathbf{w}^{(u)T} \mathbf{x}_n \tag{5.1}$$

The standard least square estimate calculates the D-dimensional weight vector $\mathbf{w}^{(u)}$ by minimizing the error function

$$E(\mathbf{w}^{(u)}) = \sum_{n=1}^{N_{\text{TR}}} \left(y_n - \mathbf{w}^{(u)T} \mathbf{x}_n \right)^2 \tag{5.2}$$

where in our case N_{TR} corresponds to the total number of images of all the users we have in the training set. The regularizer in the LASSO estimate is simply expressed as a threshold on the L1-norm of the entries $\{w_d\}_{d=1,\ldots,D}$ of the weight $\mathbf{w}$:

$$\sum_{d=1}^{D} |w_d| \leq t \tag{5.3}$$

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 w_d become exactly zero [264]. Since each component w_d of the weight vector weighs a different feature, it is possible to understand which features are the most important for a given user, and which ones are neglected. By looking at the values in the "user-specific" weight vector $\mathbf{w}^{(u)}$ for user u, we have that only the most important image features that characterize the preferences of that user are retained. Therefore, we can call $\mathbf{w}^{(u)}$ the *template* for user u.

More in detail, a positive weight for a feature indicates that in the pool of preferred images of a user that feature is present, and is discriminative for the user. Vice versa, the presence of a negative weight for a feature indicates that a presence of a particular feature for a user is unlikely, and this could well characterize him.

5.2.3 The matching score

At this point, we may want to match the probe biometrical trait of the user v, represented by his positive testing images $X_{\text{te}}^{(v)}$ with the gallery biometrical trait of the user u, represented by his positive training images $X_{\text{tr}}^{(u)}$.

Intuitively, a single image does not contain every facet of the visual aesthetics sense of a person; the idea is to consider a *set* of testing images, and guess if the set contains enough information to catch the preferences of the user, allowing to identify him among all the others. Given a template $\mathbf{w}^{(u)}$ of the user u, the matching score is aimed at measuring how likely the set $X_{\text{te}}^{(v)}$ of the user v contains images which are in accord with those favorites by the user u. In order to determine it, we compute for every image $\mathbf{x}_n^{(v)} \in X_{\text{te}}^{(v)}$ the regression score $\beta_n^{(u,v)}$, as described by eq. 5.1:

$$\beta_n^{(u,v)} = \mathbf{w}^{(u)T}\,\mathbf{x}_n^{(v)} \tag{5.4}$$

Then, the final matching score for the whole set (the biometrical trait) is determined as the averaged regression scores of the images belonging to it, i.e.:

$$\beta^{(u,v)} = \frac{1}{N_{\text{te}}} \sum_{n=1}^{N_{\text{te}}} \beta_n^{(u,v)} \tag{5.5}$$

5.3 Experiments

In this section the experimental evaluation is proposed. In particular, we first introduce the dataset, followed by authentication and recognition results. Finally some interpretability issues are reported.

5.3.1 Data collection

To test our approach, we consider a real dataset of 40000 images, belonging to 200 users chosen at random from the Flickr website. For each user, we retained the first 200 favorites.

Please note that the process of adding favorites is a continuous time process, which can last for months. In particular, in our dataset, the minimum amount of time elapsed from the oldest and the newest favorite is 23 weeks (the maximum is 441 weeks) – this ensuring reliable multisession acquisitions. For all the images of the dataset we computed the image signature. In order to guarantee robust testing, we randomly split the images of each user into two parts, one used to build the gallery biometrical trait X_{tr}, and, consequently, its template and one used to build the probe trait X_{te}, needed for testing the algorithm. Since the value ranges are very heterogeneous, each feature is normalized across all training images to have zero mean and unit standard deviation (note that the testing set is normalized with the constants calculated on the training set). In all experiments, the parameter t of the LASSO has been determined by crossvalidation.

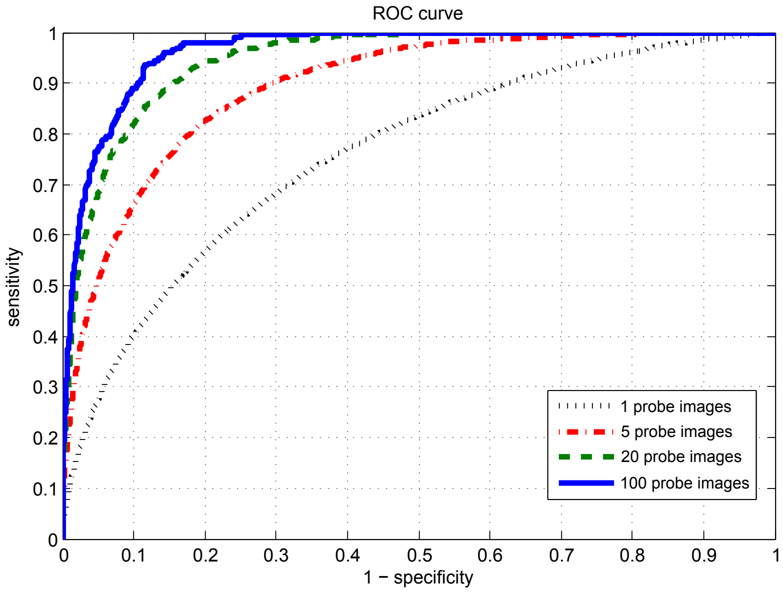
5.3.2 Authentication results

In this section the system is tested in an authentication scenario: a ROC curve is computed for every user v, where:

- client images are taken from the probe set of the user v
- impostor images are taken from all the other probe sets

In particular, different kinds of client/impostor signatures may be built, depending on the number of images we take into account: in detail, the smallest signature is formed by a single image; pooling together more pictures gives rise to composite signatures, intuitively carrying more information. Matching a signature composed by more than one image occurs by following what is described in Sec. 5.2.3, i.e., roughly speaking, by averaging the matching scores derived from the set of probe images. Given an “authentication threshold”, i.e. a value over which the subject is authenticated, sensitivity (true positive rate) and specificity (true negative rate) can be computed. By varying this threshold the ROC curve is finally obtained.

In Figure 5.2 the authentication ROC curves are portrayed; in addition, we reported also the area under the curve (AUC) and the averaged equal error rate (EER), namely, the error when sensitivity and 1-specificity have an equal value. Typically, these values represent a compact and meaningful way to summarize the ROC curve.



	┆┆┆┆	-·-	- -	━
AUC	0.76	0.89	0.95	0.96
EER	0.69	0.81	0.87	0.90

Fig. 5.2: ROC curves for the user authentication, varying the number of images per signature. Each ROC curve has been obtained by averaging over all the ROC curves for each user.

As expected, augmenting the images per signatures increments the performance. This confirms the suitability of using this trait as a biometrical trait, even if, as all behavioral biometrics, with a not so outstanding performance.

5.3.3 Recognition results

In this section the recognition capability of the proposed biometrical trait is investigated. In particular, given a probe image or a set of probe images, we want to guess the gallery user who tagged them. To do that, we compute the matching score of the probe image (or set) using all the templates $\{\mathbf{w}^{(u)}\}$. Hopefully, the gallery user with highest score is the one who originally faved the photo (or group of photos).

In order to evaluate the recognition rate, we built a CMC curve [188], a common performance measure in the field of person recognition/re-identification [13]: given a probe set of images coming from a single user and the matching score previously defined, the curve tells the rate at which the correct user is found within the first k matches, with all possible k spanned on the x-axis (they are also called *ranks*). Figure 5.3 shows various CMC curves for our dataset, where the curves have been obtained by averaging the CMC curves of 20 different experiments with different gallery/probe splits.

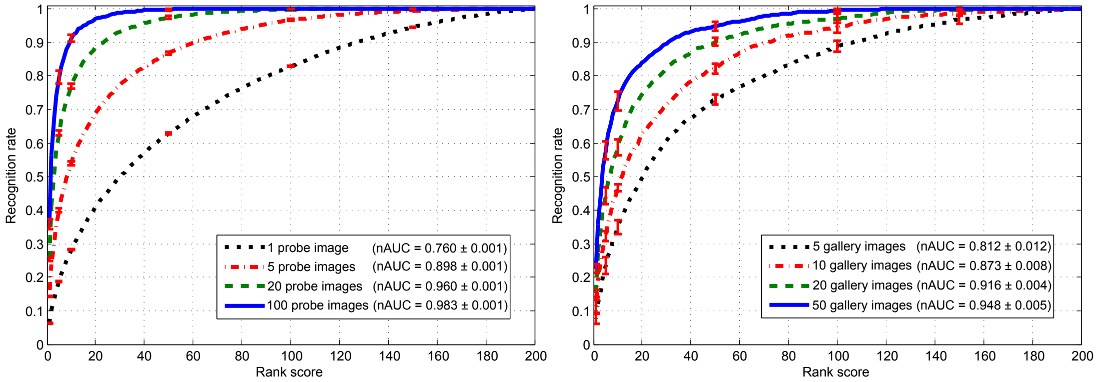


Fig. 5.3: CMC curves for our dataset: the curves have been obtained by averaging the CMC curves of 20 different experiments with different gallery/probe splits. On the left: for each curve, we varied the number of probe images to be considered as a single “set”, while keeping the number of gallery images fixed to 100. On the right: for each curve, we varied the number of gallery images used to train Lasso, while keeping the number of probe images to 20. Since we performed 20 random splits of gallery/probe, we report also the standard deviation of results. Table 5.2 reports more in detail the values of both curves at rank 1-5-20-100, in order to provide a better quantitative idea of the probability of having the correct match within the first 1-5-20-100 signatures.

Protocol	**rank 1**	**rank 5**	**rank 20**	**rank 100**
1 probe image	0.063	0.188	0.408	0.829
5 probe images	0.143	0.399	0.688	0.966
20 probe images	0.254	0.629	0.883	0.998
100 probe images	0.359	0.796	0.970	0.999
5 gallery images	0.076	0.236	0.496	0.889
10 gallery images	0.113	0.322	0.628	0.948
20 gallery images	0.152	0.443	0.743	0.971
50 gallery images	0.225	0.578	0.838	0.995

Table 5.2: CMC values for different ranks. Values represent the probability of having the correct match within the first 1-5-20-100 signatures, considering different numbers of probe (first 4 rows of the table) and gallery images (last 4 rows of the table).

On the left, we reported four different CMCs, varying the parameter N_{te}, which tells how many images are aggregated to form a single probe object, while keeping the number of gallery images fixed to 100.

From the figure it is evident that performing the task of identifying correctly a user with a single image (black dotted line) is very difficult. However, as soon as the number of probe images grouped together increase a little, a consistent improvement can be noted. This is in line with our hypothesis: we are aggregating information from heterogeneous images, each one characterizing only a small portion of the user subjective tastes.

On the right, we assessed the importance of the gallery set size N_{tr} by keeping the probe parameter N_{te} fixed to 20. As expected, by lowering the number of gallery elements, it is more difficult to learn the users’ preferences and their aesthetic sense uniqueness. For both figures, the normalized Area Under the Curve (nAUC) has been reported in the legend.

As a further comment, it is also worth noting that, even if at CMC rank 1 we achieve in the worst case a 6.3% rate of correct identification, this is higher than the probability of recognizing the user by mere chance (which amounts to 0.5%).

A final interesting question can be made: how is our signature when compared to other biometrical cues? Having clear in mind that realizing a proper and exhaustive comparison is not so trivial, we would like to provide here some intuitions. We focus on the field of people re-identification, where the signature of a user is composed by set of his full-body images (i.e., the appearance). In particular, we take into account the experiment done on the CAVIAR4REID dataset in [13], where multiple methods of re-identification have been tested on small images of people (averagely, around 50×120 pixel). In order to create a fair comparison with the CAVIAR4REID re-identification experiment, we randomly select the same number of users (72) and we used 5 images for the gallery, 5 for the probe, repeating the experiments 10 times. In our case, we obtain an nAUC of 74.8%, whereas with the classical re-identification approaches the n-AUC (reported in the paper) is 78.5%. This result is quite intriguing, as it states that having images chosen and marked by an user as his favorites is not too far from actually looking at that subject directly. This witnesses the potentiality of our biometrical strategy.

5.3.4 Feature analysis

This section is aimed at providing a qualitative evaluation of the proposed approach, showing that the regression score β provides a valid measure of the preferences of a user, while the weight coefficients in the vector $\mathbf{w}$ provide an interpretable description for his visual aesthetic sense.

In the first experiment, given the gallery user u, we considered all the probe images of all the users $\{v\}$, and we sorted them according to their regression score $\beta^{(u,v)}$. The higher the score of an image, the higher the probability that the user may have actually faved that image. Figure 5.4 gives an excerpt of the results; each column corresponds to a different Flickr user u; given the template of that user, the first 10 rows are the favorite gallery photos which exhibit high regression scores, ranked in descending order from the highest one. In other words, these 10 images are the ones which better represent the user u, as modeled by the related template. The second 10 images are taken from the probe images of all the users (not only from user u), which exhibit the 10 highest regression scores (again, w.r.t the template of the user u), with a blue frame indicating actually those images which belong to that user.

The figure reveals some interesting information: although the highest test image for the template of the user is not on his favorites set, it can have some visual appeals reflected on some of the images on his gallery set (see for example the black and white faces and scenes in the first column, or the airplanes in the second). It seems that a sort of "internal coherence" starts to show up.

We then looked into the weight coefficients for some users after learning the sparse regression model. For two random users, we reported the vector $\mathbf{w}$ in Figure 5.5, on the right of their gallery and predicted preferred images. For visualization purposes, we labeled the most prominent features (i.e. the ones with highest – in absolute value – weight value).

For user 41, the rule of thirds (i.e., its computational aesthetics version) plays an important role, and actually most of his images report an object in the central rectangle of the image. This is visible also in the probe images, selected by regression among the probe images of all the users; all the images appear with high luminosity, and the same happens in the probe images of that user. Note also that there are few regions in the gallery/probe images, this being reflected by the corresponding negative weight. For user 182, faces, the white color, hue homogeneity and edge/textural properties are important, and this is visible since many black/white images of many people (many edges/textures) characterize his favorite set of shots. Similarly, probe images which conform with the classifier of that user report faces and edgy pictures, some of them with few colors (high hue homogeneity), with a strong presence of white. The negative weight on aspect ratio indicates that images composing the favorite set of user 182 are more "rectangular" than the preferred images of others.

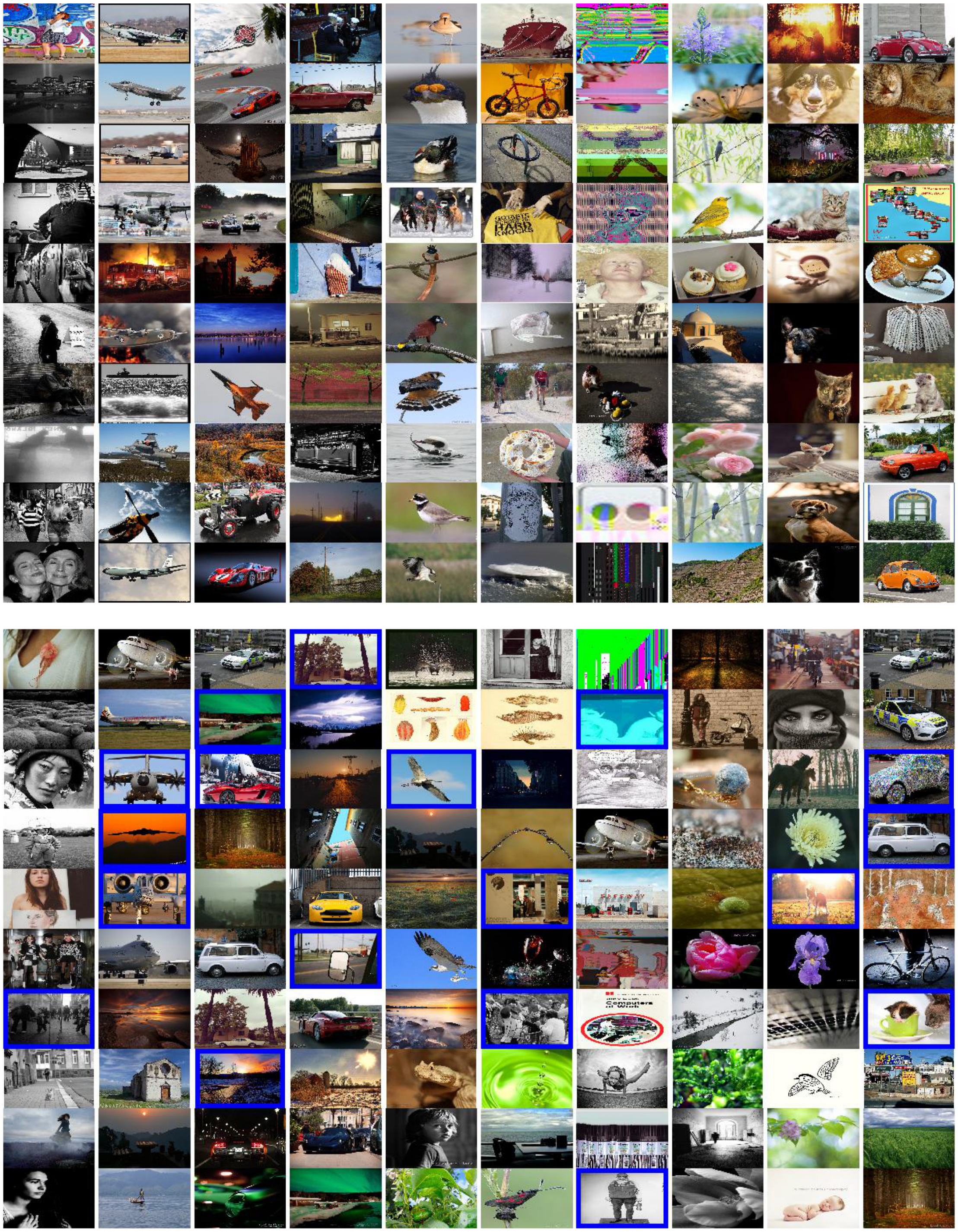


Fig. 5.4: Gallery and recognized probe images for different users. Each column is a user, and the first 10 images come from his gallery set. In the half-bottom part, we show the first 10 probe images for that user, ranked on the basis of their regression score (the first being the one with highest score). In blue, correct matches are highlighted. A “coherence” between gallery and probe images can be seen.

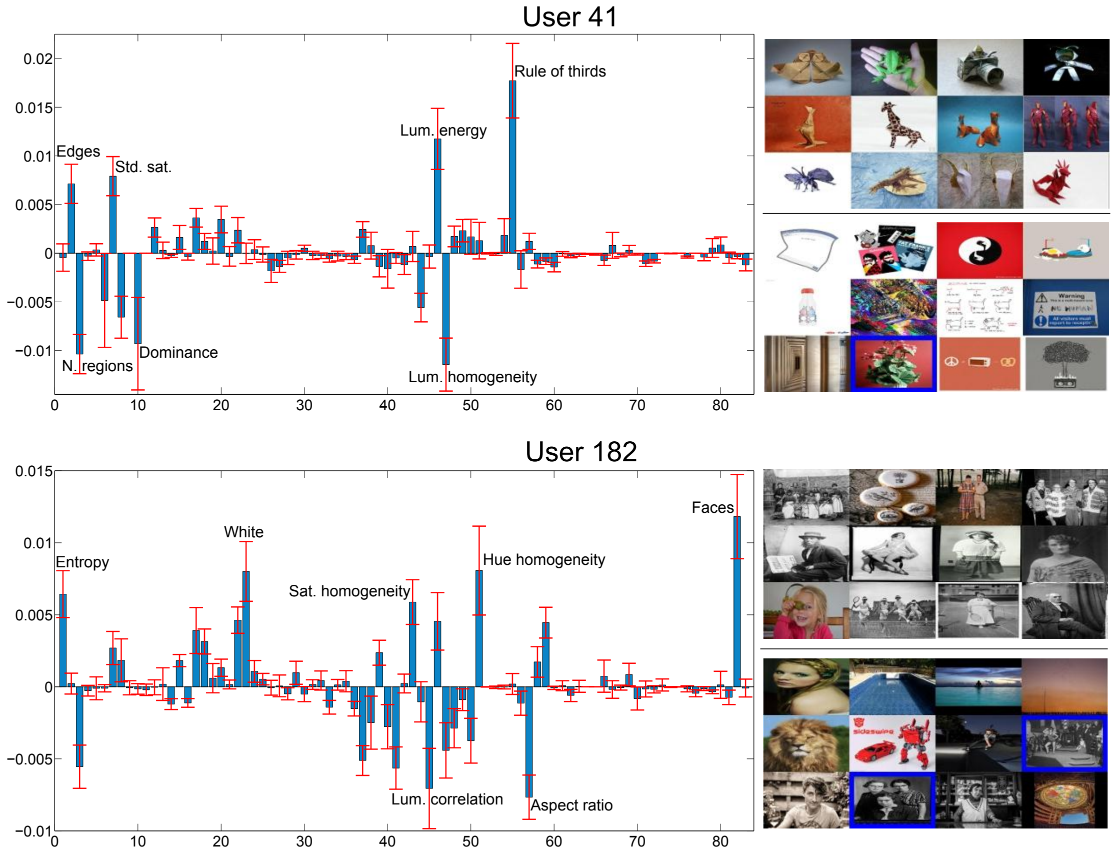


Fig. 5.5: Most prominent features for 2 users taken from the dataset. On the left, for each user, a bar plot of each feature's importance is shown. The height of each bar represents the value of the corresponding weight, and the standard deviation along the 20 different experiments on different partitions is also visualized. On the right, gallery and probe elements are shown, in the same fashion of Figure 5.4.

6

A “Pump and Distill” Regression Approach for Soft Biometrics

6.1 Introduction

In this chapter we present a novel framework for personal aesthetics biometrics, based on a statistical classification, where the training stage is characterized by a “pump and distill” strategy. In the “pump” step, the training set of each user (a set of liked images) is augmented by bagging, generating a set of ensembles of preferred images. In the “distill” step, each image of an ensemble is associated to a single *thematic exemplar*, chosen in a set of exemplars, learned beforehand by clustering. For example, we could have in principle clusters of cars, humans, etc., depending on the nature of the clustering procedure, and a thematic exemplar is the centroid of a cluster. All the images of an ensemble linked to the same exemplar are fused together, by averaging them, thus obtaining a surrogate. In practice, a surrogate encodes a customized version of a cluster of a user, capturing what kind of cars, humans, etc. a user likes. Finally, LASSO regression is performed on these surrogates predicting a user identity.

Experiments have been performed on a set of 200 Flickr users, considering 200 preferred images per user, the same as Chapter 5. Our approach overcomes definitely the previous method [168] presented in Chapter 5 on recognition and authentication tasks in average, promoting our idea as an effective strategy for learning the personal aesthetics.

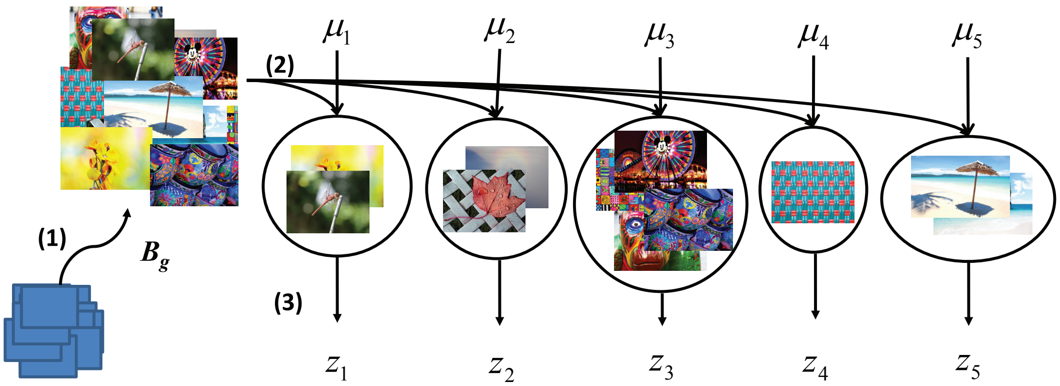


Fig. 6.1: “Pump and distill” approach: (1) “pumping” a bag B_g from the original image training set; (2) image assignation to a thematic exemplar μ_i; (3) “distilling” the assigned images to the same exemplar to a surrogate $\mathbf{z}_i$.

6.2 Proposed Approach

Each image $\mathbf{x}$ is composed by the same feature vector used in Chapter 5. In the following, with the term “image” we imply its feature description. For more details on the features, see Section 3.1.

The dataset is divided in two partitions, a *gallery* set used for training and *probe* set for testing, both formed by 100 preferred images per user.

We assume that our approach uses the results of a clustering algorithm operating on the image representations, and in particular we suppose to have K cluster centers, here called “thematic exemplars” μ_k, $k = 1, ..., K$. These exemplars represent different image typologies, depending on the clustering approach employed. For simplicity, here we adopt a simple K-means on the raw feature space fed with the training samples, but other, more advanced, partitioning techniques can be applied.

Our approach can be dubbed “pump and distill” to highlight two important steps which characterize the training of the classifiers, one for each user.

In the “pump” step, we augment the training set by employing bagging [33], a strategy designed to improve the stability and accuracy of machine learning algorithms. More in the detail, for each user u, $u = 1, ..., U$, bagging generates G new training sets, the “bags” ${}^{(u)}B_g$, $g = 1, ..., G$, each obtained by sampling uniformly with repetition M times from his training images. In practice, a bag represents a small exert of what is liked by the user.

In the “distill” step, for each bag ${}^{(u)}B_g$ we want to generate a set of *surrogates* $\{\mathbf{z}\}$. In practice, we assign each image $\mathbf{x} \in {}^{(u)}B_g$ to the nearest thematic exemplar μ_k, adopting whatever plausible distance. In our case, we use the simple Euclidean distance. After the assignation, all the images $\mathbf{x} \in {}^{(u)}B_g$ that have been associated to a given cluster are fused together by averaging their feature vectors, creating a surrogate $\mathbf{z}$. For each bag, the number of surrogates may vary from 1 (a single cluster is associated to all the images in the bag) to K (all the clusters have at least one image in the bag associated with them), the last case obviously holding only if M, the number of images per bag, is $>= K$. In this last case, with G bags we end up with $G \times K = N^+$ surrogates per user. In practice, a surrogate encodes a user-customized representation of a cluster, a sort of epitomic representation of a bunch of similar images liked by a user.

At this point we use the pool of surrogates of all the users as distilled training set, to learn a per-user classifier. For this sake, we perform a sparse regression analysis using LASSO [264], assigning to all the training (positive) surrogates $\{\mathbf{z}_n\}$, $n = 1, ..., N^+$ of a user the $y_n = +1$ label, and -1 to all the other N^- surrogates; that is, the negative samples are the positive surrogates of all the other users.

In the testing step, we want to match the probe images of the user v (that is, their real-valued feature vectors) with the gallery biometrical traits of the user u, represented by his positive surrogates $\{{}^{(u)}\mathbf{z}_n\}$, $n = 1, ..., N^+$. Experimentally, we observed that applying the pump and distill processing to the testing images leads to inferior performance. Clearly, a single image scarcely represent the visual aesthetics sense of a person; therefore, the idea is to consider a *pool* of testing images TE as test biometrical trait, and guess if the pool contains enough information to catch the preferences of the user, allowing to identify him among all the others. In particular we can perform two different operations, that is, user recognition and user authentication.

User recognition

In the user recognition, given the classifier template ${}^{(u)}\mathbf{w}$ of the user u, the matching score is aimed at measuring how likely the set $\{{}^{(v)}\mathbf{x}\}$ of the user v contains images which are in accord with the surrogates $\{{}^{(u)}\mathbf{z}\}$ by the user u. In order to determine it, we compute for every image ${}^{(v)}\mathbf{x}_m$ in the testing pool TE of user u the regression score $\beta_m^{(u,v)}$, as described by Eq. 5.1: Then, the final matching score for the whole pool is determined as the averaged regression scores of the images belonging to it, i.e.:

$$\beta^{(u,v)} = \frac{1}{M_{\mathrm{TE}}} \sum_{m=1}^{M_{\mathrm{TE}}} \beta_m^{(u,v)} \tag{6.1}$$

where M_{TE} is the cardinality of the testing pool. For the recognition, we compute the matching score of the probe image (or pool) using all the classifiers $\{^{(u)}\mathbf{w}\}$. Hopefully, the gallery user with the highest score is the one who originally faved the photo (or pool of photos). In order to evaluate the recognition rate, we built a CMC curve [188].

User authentication

In this case the system is tested in a authentication scenario: a ROC curve is computed for every user u, where *client* images are taken from the probe set of the user u, and *impostor* images are taken from all the other probe sets. In particular, different kinds of client/impostor signatures may be built, depending on the number of images we take into account as testing pool. Matching a signature composed by more than one image occurs by following what is described previously, i.e., roughly speaking, by averaging the matching scores derived from the set of probe images.

6.3 Experiments

The experiments focus on evaluating how effective is our surrogate representation in capturing the personal aesthetics. For each user, we have 200 images: we partition them into a training and a testing set (100 images each), crossvalidating using a 2-fold scheme as in Chapter 5, and repeating each experiment 5 times, shuffling the the partitions. To explore different ways of creating the surrogates, for processing the training set we fix $G = 50$ bags and we vary the number of images used to fill them. In particular, we use $M = 5, 10, 20, 50, 100$ images per bag. As number of clusters, we fix $K = 6$. Later in this section, we discuss about varying G and K. Given the surrogates, to learn a LASSO classifier, we decide the best α by a 10-fold cross-validation on a subset of the training set (Eq. 5.1). As comparison, we consider the approach of Chapter 5, which essentially can be thought as having 100 surrogates for training, each formed by a single image. For a fair comparison with the present approach, we run the code of Chapter 5 following the same protocol, that is, repeating each experiment 5 times, shuffling train and testing sets. Actually, in Chapter 5, experiments were run for a single 2-fold cross-validation run, and authentication results were run for a single test image.

6.3.1 Identification results

Table 6.1 shows the recognition results in terms of normalized area under the CMC curve (nAUC CMC). Other than changing the number M of images per bag, we also vary M_{TE}, i.e., the cardinality of the testing set. Please remember, the approach in Chapter 5 can be thought as having bags formed by one image.

Many observations can be made:

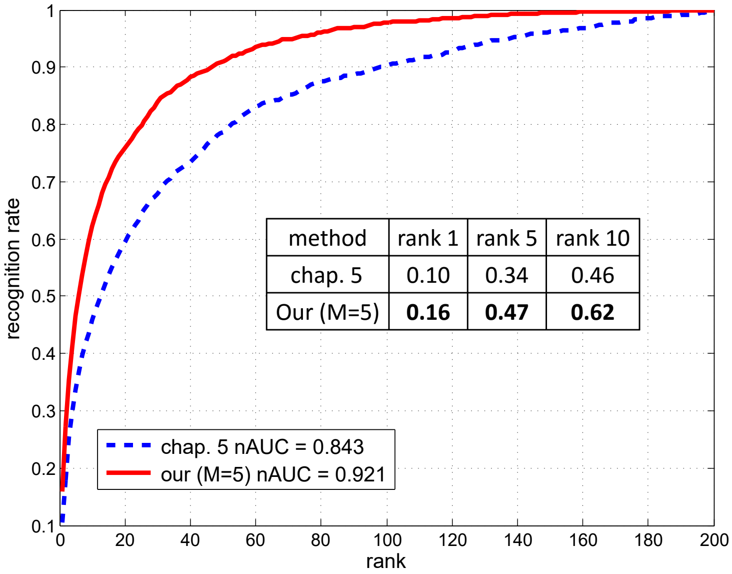
1. the “pump and distill” approach overcomes the approach based on simple images of Chapter 5: this suggests that pooling together images into surrogates produces more discriminative information, which is successfully exploited by LASSO;
2. in general, the “pump and distill’ gives its best with testing pools composed by multiple images: the more the images, the higher the nAUC. Considering the increment in performance w.r.t. Chapter 5, the highest difference holds for $M_{\text{TE}} = 20$, increasing the nAUC of 15%;
3. changing the number of images per bag is not very important, especially when having a high number of test images.

Anyway, there is a tendency of having higher results with less images per bag. More into detail, having bags formed by 5 images creates in average 166 surrogates that are formed by more that 1 image (otherwise, we have simple images) in the 38% of the cases, and in particular, 73% of these *proper* surrogates are formed by 2 images, 22% by 3 images, and 5% by 4 images, respectively. In practice, coupling just 2 images into a surrogate substantially ameliorates the classifier.

M (=$\left\|{}^{(v)}B_g\right\|$)	nAUC CMC			
	M_{TE}=1	M_{TE}=5	M_{TE}=20	M_{TE}=100
5	**0.69**	**0.83**	**0.92**	**0.96**
10	0.69	**0.83**	**0.92**	**0.96**
20	0.68	0.82	0.91	**0.96**
50	0.68	0.80	0.90	0.95
100	0.66	0.79	0.89	0.95
chap. 5	0.66	0.77	0.84	0.88

Table 6.1: nAUC values of the CMC curves varying the num. of images per bag (M), and the num. of test images (M_{TE}). The last row shows the performance of Chapter 5.

In Figure 6.2 we show the CMC curve of our approach using $M = 5$ images per bag, and $\mathrm{TE} = 20$ test images, against the Chapter 5 approach. In this way, we can detail how or approach overcomes the competitor. Actually, the highest difference in terms of recognition rate (the probability of having the correct match in the first r ranked positions) is localized in the first ranks (see the table in the figure), which is very beneficial in terms of a real biometric system, where is expected to find the correct match in the top ranked positions. For example, the probability of having the correct match in the first 10 position, in our approach, is 62%, against the 46% of the competitor.



method	rank 1	rank 5	rank 10
chap. 5	0.10	0.34	0.46
Our (M=5)	**0.16**	**0.47**	**0.62**

Fig. 6.2: Recognition results: CMC curves and recognition rates at rank 1,5,10.

6.3.2 Verification results

Following what described in Sec. 6.2, we show in Table 6.2 the nAUC values related to the ROC curves. In a similar way of the previous section, here we vary the number M of images per bag, and the number M_{TE} of test images. Here, similar considerations to those of the recognition case can be carried out, so having few images per bag (5,10) gives the best performance.

As for changing the values of the number of clusters K, we observe on most of the results that the performance (on both recognition and authentication) is similar for $K = 5, 6, 7, 8$, while start decreasing

M (=$\left\|{}^{(u)}B_g\right\|$)	nAUC ROC			
	M_{TE}=1	M_{TE}=5	M_{TE}=20	M_{TE}=100
5	**0.70**	**0.83**	**0.91**	**0.95**
10	0.67	**0.83**	**0.91**	**0.95**
20	0.67	0.82	0.90	0.94
50	0.67	0.80	0.90	0.94
100	0.66	0.79	0.88	0.94
chap. 5	0.65	0.76	0.82	0.87

Table 6.2: nAUC values of the ROC curves varying the number of images per bag (M), and the number of test images (M_{TE}). The last row shows the nAUC scores of the ROC curves related to the Chapter 5 approach.

for smaller and bigger values of K. We monitored also the performance while changing the number of bags G. In this case, in average, the results exhibit a slight increase in the interval $[20, 50]$ bags, decreasing for a lower number of bags, and keeping constant after 50 bags until 200, where overfitting starts to reduce the performance.

7

Mapping Image Preferences on the Counting Grid

7.1 Introduction

Modeling preferences in photographic images is often reduced to analyze intermediate explicit representations (e.g. textual tags) as means of capturing the objective and subjective properties of image perception, trying to distill the essence of what gives pleasure.

In this chapter, we bypass the problem of finding a comprehensive intermediate coding for conceptual images: instead of processing images to extract explicit aesthetics codes or content-based tags in an independent way, we directly exploit the information within the images to map heterogeneous image features together in a seamless way. In particular, we arrange the set of photos taken from Flickr in a 2-dimensional map through a *Counting Grid* [213]: this generative model considers each image as a specific distribution of generic features (color, SIFT features, etc.) and places it alongside similar images.

In our case, we adopt this model in an unconventional and novel way, feeding it with aesthetical and content-based features: the trained Counting Grid fuses the two worlds in an unsupervised way, thus defining a manifold where local regions mark semantic areas that smoothly transit between each other, and thus allowing fine thematic shifts. Besides revealing image classes, we can use this model to retrieve a Flickr user aesthetic profile, by locating his preferred images in the Counting Grid and building a map of his specific tastes. We used these profiles to calculate a novel "subjective aesthetics" metric between people, and we tested its usefulness by checking on users that subscribe to the same Flickr photo groups, under the assumption that they indicate shared tastes. In fact, this technique can be used to tell users about other groups they may like.

Summarizing, in this chapter we present:

- how to fuse in a principled way aesthetics traits and content-based features in a low-dimensional "spatial" latent manifold – the Counting Grid;
- how to highlight the image preferences of users directly on the Counting Grid, without the need of intermediate cross-modal representations (*e.g.*, text);
- a way to suggest and encourage groupings, by exploiting a novel distance defined over Counting Grids.

7.2 The Counting Grid Model: Organizing and Making Sense of Bags of Words

In machine learning, data samples are often represented as disordered "bags of words" of image features. This choice is typically motivated by the difficulty or computational efficiency of modeling the feature structure [2, 28, 57, 78]. However BoW method has some disadvantages since in many situation it looses a lot of important information. For instance, BoW approach does not take into account words relation or co-occurrences. In fact, these features should be highly discriminative so that most categories of images of interest are uniquely identifiable by the presence of a handful of features. In practice, however, individual features are not sufficiently discriminative, and modeling joint variation in feature counts becomes an interesting machine learning problem.

It is tempting to use here the existing discrete models, such as histograms, multinomial mixtures, or LDA [229], PLSA [28, 78, 82], already extensively validated on text data showing inter-relations among words. Topic models [27, 28, 64], were introduced by text analysis community and have been particularly successful in representing text documents. These simplified model of text assume that a text document has been generated simply by mixing words from a subset of possible topics. In typical applications, the number of possible topics is large, and these topics are inferred from the data by analyzing word co-occurrence patterns, and so the topic scope can vary from very narrow to quite broad, e.g., from near homonyms, to words found in most stories on US politics.

An individual document is assumed to use only a fraction of all possible topics, and so the resulting bag of words will exhibit strong co-occurrence patterns: when the president is mentioned, so is the congress, as both appear in the same topic. These models can be used in other domains by simply replacing word with some other set of feature of interest. In image processing, for example, images are represented by the frequency of visual words, i.e. document are replaced by images, words by features extracted from them and their counts by the frequency of these features. Visual descriptors are extracted from pictures and clustered into "visual words" replacing traditional bag of words.

Among topic models one of the best known is the Latent Dirichlet Allocation [27] (see Section 4.2.5 for a general overview). To formally define this model, we will index possible words (features) by z and denote the set of observed word (or feature) counts in the t-th bag of words by $\{c_z^t\}$. The latent (hidden) variable describe the choice of topics indexed by k. The choice of topics follows a distribution $p(k|\theta) = \theta_k$, and each topic has its own distribution over all the words $p(z|k, \beta) = \beta_{z|k}$. The vector that describe the topic distribution for one document θ is sampled from Dirichlet distribution with parameters α. The following probability of generating a particular document is induced by this simple generative process (after picking the topic distribution θ, pick a topic, then pick a word from the topic, then pick a topic and a word from it again and again till all the words in the document are generated):

$$p(\{c_z^t\}|\alpha, \beta) = \int p(\theta|\alpha) \cdot \prod_z (\sum_k (p(z|k, \beta) \cdot p(k|\theta))^{c_z^t}) d\theta \tag{7.1}$$

The model parameters are estimated based on a training set so as to maximize the product of probabilities of all training documents. The topic proportion θ for individual document can be used as a compact representation of the bag of words that discards the superfluous aspects of the data.
The counts c_z are not independent. This means that we need a model that takes into account and can capture these correlations and this is precisely what the probability model of bags of words were meant to do for text documents. However LDA, such as other discriminative techniques retain just the counts, and the spatial distribution is typically forgotten with the justification that established correspondence for individual image locations across different images of the same thing would be prohibitively expensive, and that in practice only the presence or absence of feature is informative, not their spatial distribution.

The bags of features extracted form images have an imprint of the images' spatial structure, which is evident when the bags from related images are considered together, and ignoring the constraints imposed on the feature counts may have negative consequences in classification tasks. However, if we consider set of such bags of words from related images we can see that the features counts in these disorganized bags of features may still indirectly follow the rules of spatial organization. Further LDA mixes small number of topics in bag of words, while for instance, a document can evolve in one to another in a smooth way, with some feature dropping and new one being introduced. For this reason, to enhance the representation of an image and its structure we use in our approach the CG model.

The CG exploits not only word co-occurrence but also topological relations among words, modeling smooth changes as we mention the example of the stories in times. In particular, with CG an ordering procedure between BoWs is introduced by allowing BoWs to lie in n-dimensional grid structure. For our application, we can think that user aesthetics preference can be model in the same way. Two subjects usually not present the same aesthetic preferences, but still very similar. Thus, we should take into account these effects, i.e. in a particular position of the grid, where is mapped a preference, let's say cars, we can have a limited number of preferences around it, and we will have as neighbors, user that share these preferences. In the LDA model, vice versa we can mix more topics, making difficult to map the aesthetics preference of a user.

7.2.1 The Model

The Counting Grid model [213], is a generative model which cluster together similar observations, highlighting the compactness of a class and its underlying structure.

Each t-th observation is characterized by a vector - often called count vector $\{c_z^t\}$ - containing the number of occurrences of each feature z, the higher the feature level, the "more present" the feature is in such image. Each sample can be seen as independent observation. Roughly speaking, a CG is 2D finite discrete grid that starts assuming that images are represented as histograms $\{c_z\}$ over unordered bags of features (see Figure 7.1), where each location $\mathbf{i} = (x, y)$ contains a normalized count of feature $\pi_{\mathbf{i},z}$; each c_z counts the occurrences of feature z. Each of these distributions is relatively tight, with only a few features having significant probability. The underlying generative process draws an image (i.e. its bag of features $\{c_z\}$) by locating a small windows in the grid, averaging the feature counts within it to obtain a local probability mass function over the features, and then generating from it an appropriate number of features in the bag (see Figure 7.3). The key idea of topic models is still present: a document is abstracted in an intermediate representation of "topics", which are probability distribution over words that picks out a coherent cluster of correlated terms. However, here topics are arranged on a discrete grid, learned in a way that "similar" topics are closely arranged. Figure 7.2 pictures the idea and compare it with the PLSA model. Given that the size $E_1 \times E_2$ of a Counting Grid is usually small compared to the number of images, this also forces windows linked to different images to overlap, and co-exist by finding a shared compromise in the feature counts located in their intersection. The overall effect of these constraints is to produce locally smooth transitions between strongly different feature counts by gradually phasing feature in/out (i.e. dropping certain words and adds new ones) in the intermediate locations where the window is shifting in the grid.

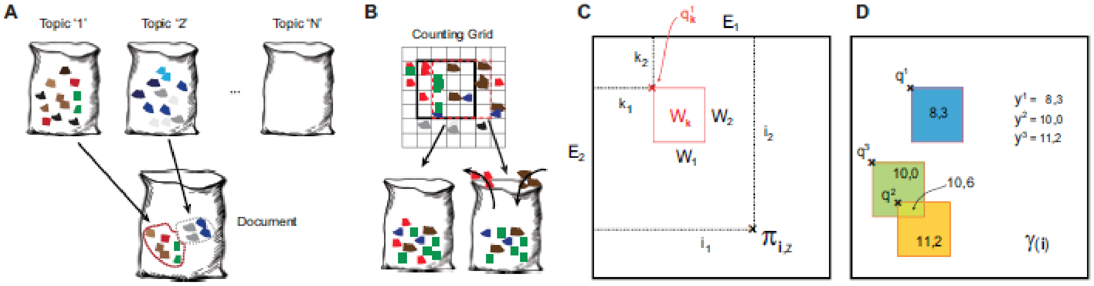


Fig. 7.1: Capturing dependences in Bag of Words. A) Bag of Words example; B) Counting Grid as Bag of Words paradigm; C) An example of a Counting Grid geometry; D) Label Embedding γ_i

Formally, the basic Counting Grid $\pi_{\mathbf{i},z}$ is a set of normalized counts of features indexed by z on the 2-dimensional [1] discrete grid indexed by $\mathbf{i} = (x, y)$ where $x \in [1 \dots E_1], y \in [1 \dots E_2]$ and $\mathbf{E} = [E_1, E_2]$ describes the extent of the Counting Grid. Since π is a grid of distributions, $\sum_z \pi_{\mathbf{i},z} = 1$ everywhere on the grid. A given bag of features, represented by counts $\{c_z\}$ is assumed to follow a count distribution

[1] N-dimensional in general, here we focus on 2 dimensions.

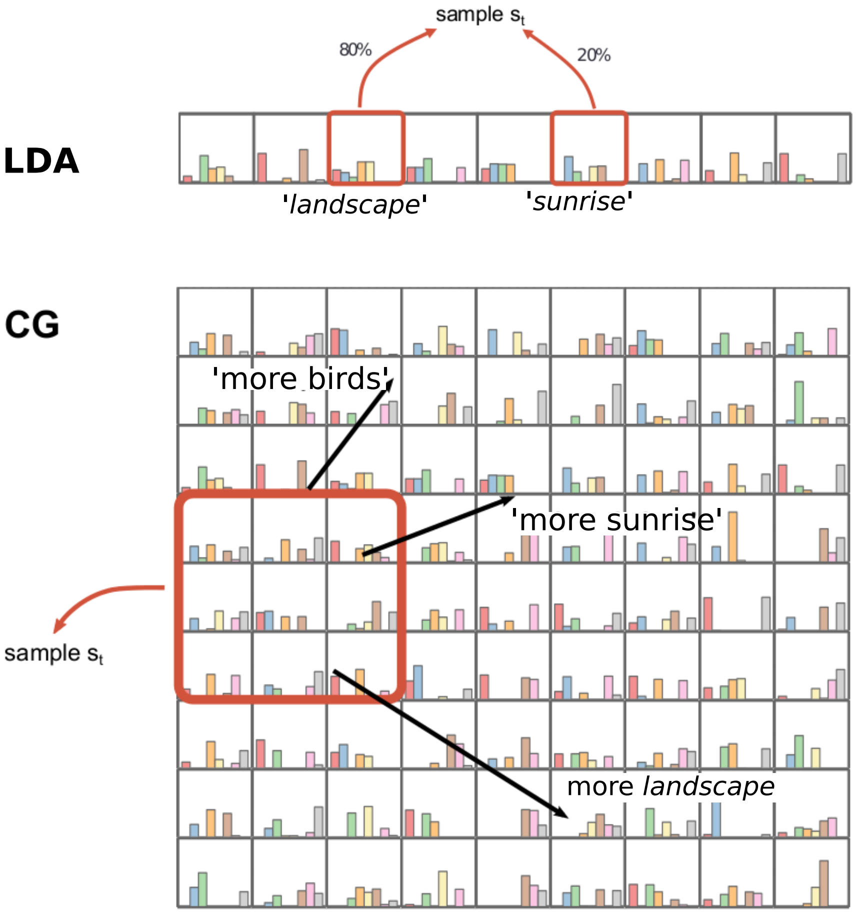


Fig. 7.2: In the LDA model an image is an admixture of independent topics. In the examples document s^t is composed by the topics 'landscapes' and 'sunrise' with a 0.8:0.2 proportion. In the CG model, neighboring topics are similar, and a document is generated from one window in the grid (see Figure 7.3). Traveling in any direction on the grid lead to a smooth topic transition.

found in a patch of the Counting Grid. In particular, using a window of dimension $\mathbf{W} = [W_1, W_2]$, each bag can be generated by first selecting a position $\mathbf{k}$ on the grid and then by placing the windows in the grid such that $\mathbf{k}$ is its upper left corner. Then, all counts in this patch are averaged to form the histogram $h_{\mathbf{k},z} = \frac{1}{\prod_d W_d} \sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z}$, and finally a set of features in the bag is generated. In other words, the position of the windows $\mathbf{k}$ in the grid is a latent variable given the probability of the bag of features $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{k}) = \prod_z (h_{\mathbf{k}}, z)^{c_z} = \frac{1}{\prod_d W_d} \prod_z (\sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z})^{c_z} \tag{7.2}$$

where with $W_{\mathbf{k}}$ we indicate the particular window placed at location $\mathbf{k}$ (see Figure 7.1C).

We will also often refer to the ratio of the CG area and the windows are $\kappa = \frac{\prod E_d}{\prod W_d}$, as the capacity of the model, as term of an equivalent number of topics, as this is how non-overlapping windows can be fit onto the grid. Fine variation achievable by moving the windows in between any two close by but non-overlapping windows is useful if we expect such smooth thematic shifts to occur in the data. An example of a 2D grid is depicted in Figure7.4 on the left; on the right, the Bayesian network for the model is depicted.

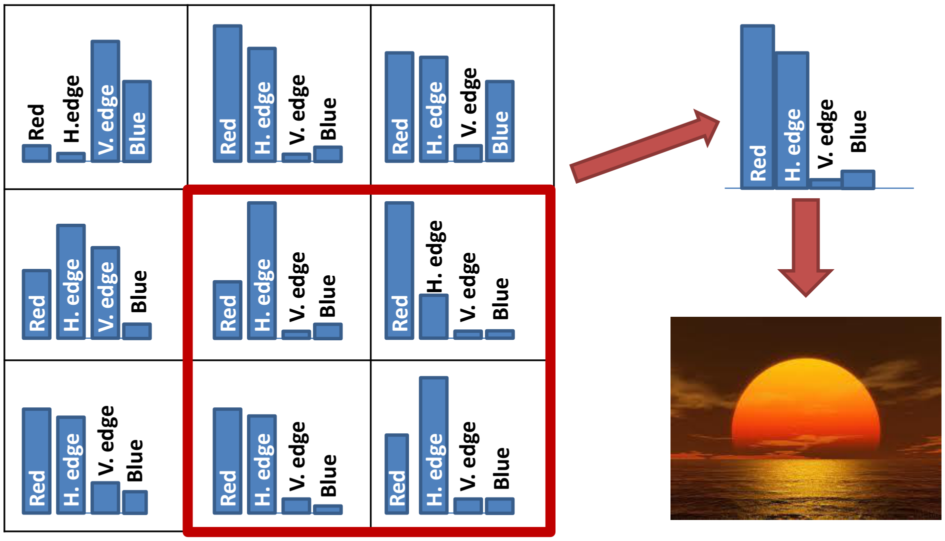


Fig. 7.3: The Counting Grid and image generation.

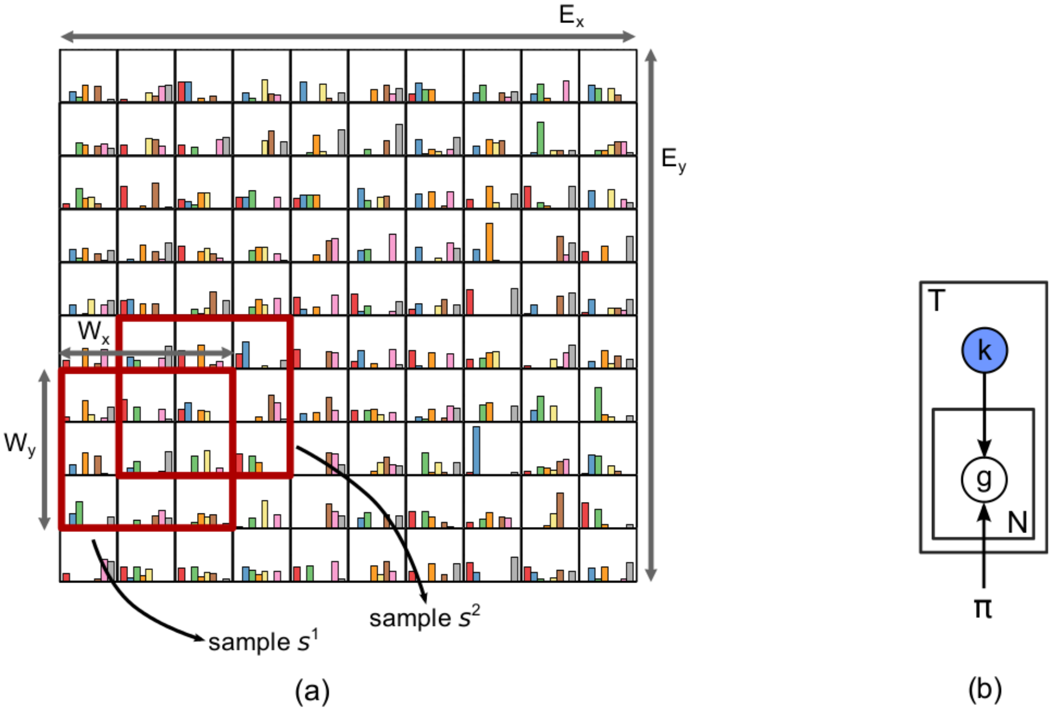


Fig. 7.4: (a) The Counting Grid model. Closely mapped samples s^1 and s^2 share some topics. (b) Bayesian network representation for the Counting Grid model.

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

$$p(\{\{c_z^t\}, \mathbf{k}^t\}_{t=1}^T) \propto \prod_t \prod_z (h_{\mathbf{k},z}^{c_z^t}) \tag{7.3}$$

$$\log P = \sum_t \log \left(\sum_{\mathbf{k}} \cdot \prod_z (h_{\mathbf{k},z}^{c_z^t}) \right) \tag{7.4}$$

The sum over the latent variable $\mathbf{k}$ makes it difficult to perform assignment to the latent variables while also estimating the model parameters; therefore it is necessary to employ an iterative EM algorithm. The E-step aligns all bags of features to grid windows, to match the bags' histograms, inferring the posterior distribution $q_{\mathbf{k}}^t$ over all windows $\mathbf{k}$ so that a better match between $\{c_z^t\}$ and $h_{\mathbf{k},z}$ across all features z yields a higher value for the match, that is infer $q_{\mathbf{k}}^t = p(\mathbf{k}^t|\{c_z^t\}) \propto \exp \sum_z c_z^t \cdot \log h_{\mathbf{i},z}$. In other words, $q_{\mathbf{k}}^t$ is a probabilistic mapping of the t-th bag to the grid windows $\mathbf{k}$. This mapping is usually peaky, i.e. each bags tend to map to a few nearby location in the grid. In the M-step the model parameter, i.e. the Counting Grid π, is re-estimated so that these same histogram matches are even better. For details on the learning algorithm and on its efficiency see [213]. For our purposes, the most interesting outputs are the posterior probabilities $p(\mathbf{k}^t|\{c_z^t\})$, which localize each image in the grid. By construction, similar images will be placed in the same location or nearby.

Once the CG is learned, we show here how one can may embed continuous values y^t on the grid. This is achieved using the posterior probabilities $q_{\mathbf{k}^t}$ for each bag already inferred and embedding the corresponding user inside the entire mapped window(s), and then averaging all the overlapping windows (Figure7.1D), which is similar how M step re-estimates the distributions π:

$$\gamma(\mathbf{i}) = \frac{\sum_t \sum_{\mathbf{k}|\mathbf{i}\in W_{\mathbf{k}}} q_{\mathbf{k}^t} \cdot y^t}{\sum_t \sum_{\mathbf{k}|\mathbf{i}\in W_{\mathbf{k}}} q_{\mathbf{k}^t}} \tag{7.5}$$

The function γ can be then used for regression, in what is essentially a nearest-neighbor strategy: when a new data point is embedded based on its bag of words, the target is simply read out from γ, which is dominated by the training points were mapped in the same region.

Summarizing CG seems to be very suitable in the modeling aesthetic preferences for the following reasons:

- The CG provides a powerful representation which permits to capture evolution of patterns in the experiments, that can be clearly visualized.
- The CG is well suited for data that exhibits smooth variation between samples.

7.3 The Proposed Approach

7.3.1 Counting grid training

In this chapter each image is described by a count vector of features as we used in Chapter 5. After each image have been processed into a feature vector, a Counting Grid is learned. This way, we obtain for each image the 2D coordinates $\mathbf{k}$, described by the posterior probability $p(\mathbf{k}|\{c_z\})$ which tells where the image is most likely to be placed. We cannot visualize the resulting 2D map directly (each location contains a bag of features), but we can create an image mosaic using images with the highest posteriors $p(\mathbf{k}^t|\{c_z^t\})$ at each location $\mathbf{k}$ in the map. To give an idea of the representation returned by the model, in figure 7.5 we show some regions in a 70-by-70 discrete space.

Looking around the center of the figure, where images cannot be clearly distinguished, a dark area can be noted slightly to the left. In the enlargement shown, a cluster of faces and another where the transition between sharp textures, sunsets and red flowers seems very natural and interesting. Indeed, we want to stress that smooth transitions and grouping between images is done both on an aesthetical and content-based level, taking into account jointly all the feature we extract in the previous step.

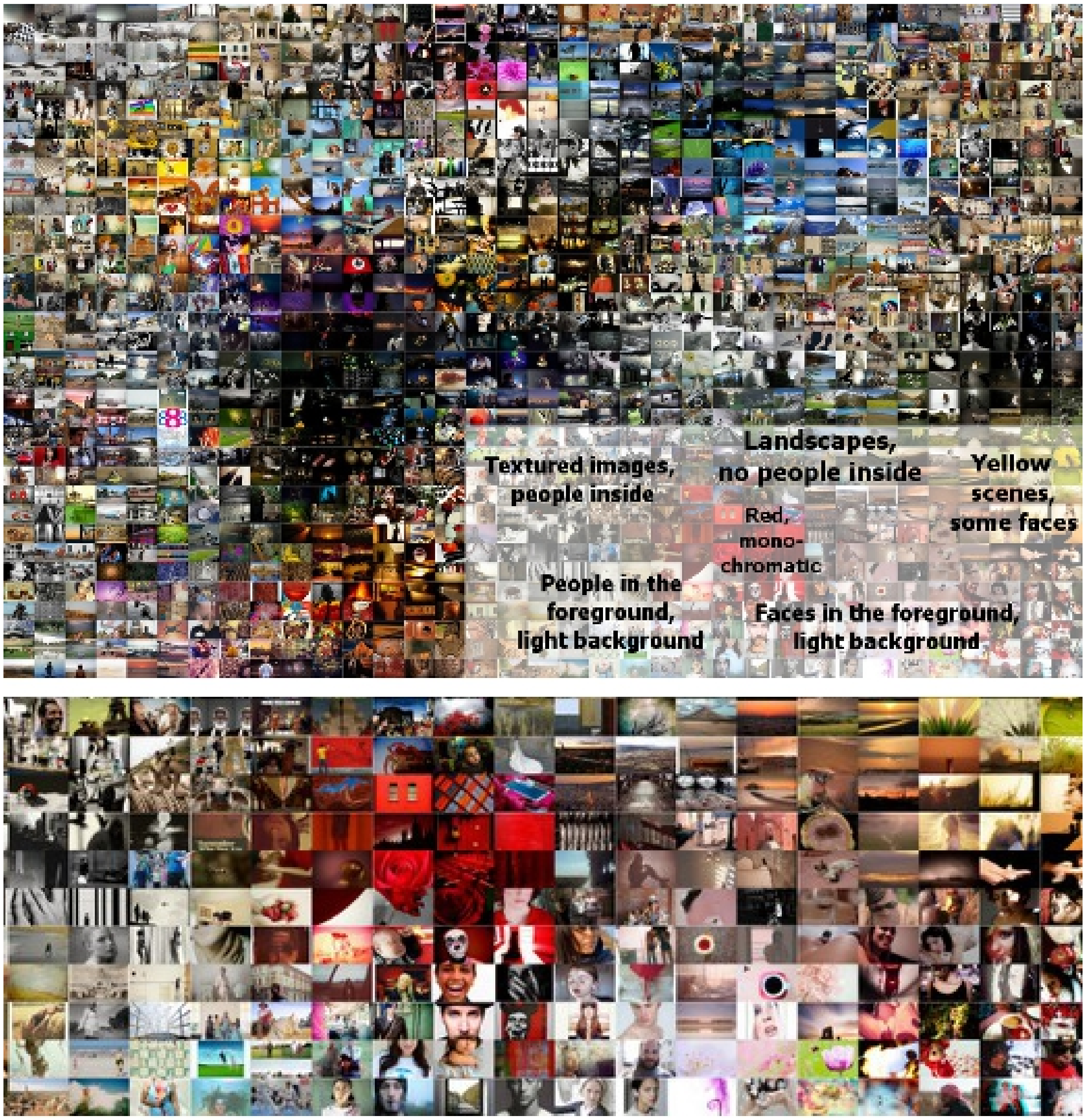

Textured images, people inside
Landscapes, no people inside
Yellow scenes, some faces
Red, mono-chromatic
People in the foreground, light background
Faces in the foreground, light background

Fig. 7.5: On the top, part of a trained 70x70 CG: at each location we show the image with the highest posterior. Thematic areas naturally emerge everywhere: on the bottom, we zoom in on a region indicating perceptively coherent groups.

7.3.2 User analysis as inference in the CG

Given the trained CG, as a novel contribution for the CG model, we discover the preferences of a subject by aggregating the posteriors of his preferred images. Technically, we calculate the following *user map*

$$\gamma_{\mathbf{i},u} = \sum_{t\in T_u} \sum_{\mathbf{k}|\mathbf{i}\in W_{\mathbf{k}}} p(\mathbf{k}^t|\{c_z^t\}), \tag{7.6}$$

where u is a user and T_u identifies his set of images. In the case of a user from our training set, we already possess the posterior probabilities, while in the case of a new user, we can calculate them using the E-step formula. We can think of $\gamma_{\mathbf{i},u}$ as a summary of the personal preferences of a subject: it is an accumulation of the positions in the grid where the liked images for a unique individual u are placed. In practice, different locations in this user map correspond to different aesthetic characteristics of images: black-and-white photos of faces, sharp textured city landscapes, and so on. Finally, given the compact representation of the users map, we can estimate the affinity between users by calculating the Euclidean distance of their user maps, and exploit such measures to cluster them together: the goal is to discover groups of people who share visual interests.

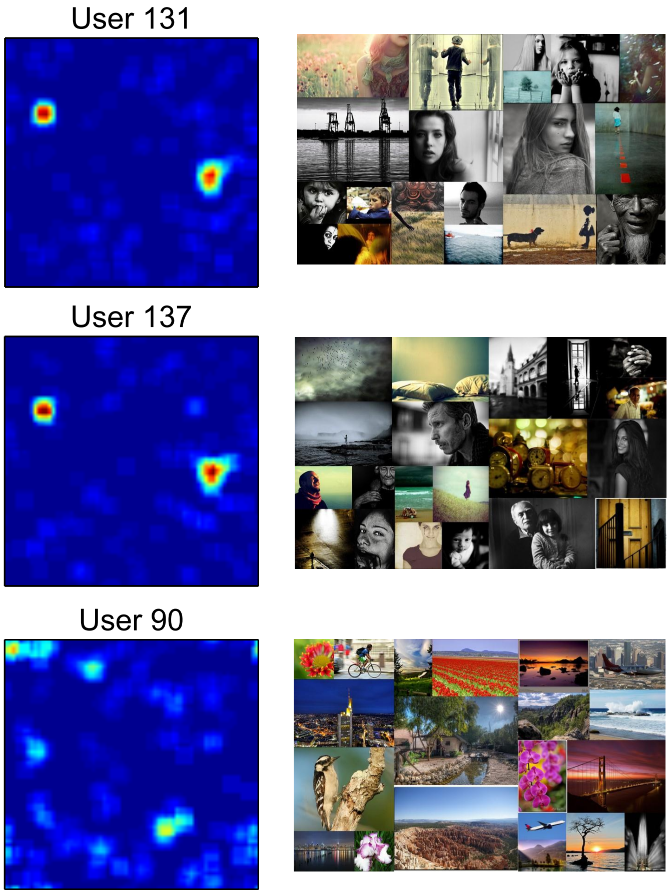


Fig. 7.6: User maps for 3 users from the dataset: red peaks correspond to clusters of many favorites. The first two users are very similar: indeed, their set of favorites (shown partially on the right) is visually similar. On the contrary, the third one shows a more sparse map, with his preferred images looking very different from the first two.

7.4 Experimental Evaluation

To assess our proposal, we consider the 200 users Flickr dataset as in previous chapters, and for each one of them we extracted the features of Table. 5.1. For the CG learning algorithm, the parameters of the model are the size of the grid and the size of the window. Different parameterizations offer the same conceptual analysis at different resolutions: in our case, we choose a 70×70 grid and 5×5 windows. Figure 7.5 shows a quarter of the trained CG: we can immediately observe that rich semantic groups pop out, highlighted in the lower part of the figure; we created an overlay, tentatively explaining the most evident themes in words, but observation alone gives a better picture of the thematic clusters. Then, using Eq. 7.6, we can infer the personal preferences of the Flickr users. Figure 7.6 shows three user maps with highly multimodal distributions that reveal personal tastes diversified over several themes. The first two users have similar maps, and a random sample of images shows very similar characteristics. If we

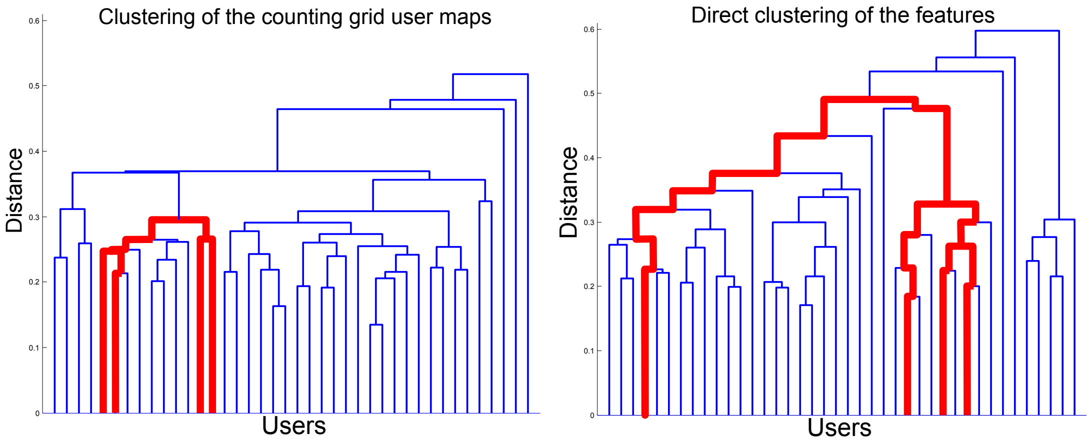


Fig. 7.7: Clustering of users: on the left, dendrogram based on distances between CG user maps; on the right, based directly on the average feature vectors of his favorites. The highlighted path shows a cluster of 4 users who share lots of groups related to urban street photography.

were to describe in words their tastes, we could say that they like faces and yellow-tinted photos, but this would be undeniably very reductive. The user maps are a much more holistic representation of their tastes.

To quantitatively evaluate the expressiveness of the user maps, we calculate pairwise Euclidean distances between all users – called *CG distance* from now on. Then, we cluster them hierarchically (by average link), obtaining the dendrogram shown (partially) in Figure 7.7 (left). On the right, we show the alternative dendrogram obtained by the Euclidean distance between each user's average feature vector (shortly, *mean feature distance*). In red, we highlight users that share various street and urban Flickr photo groups (mostly in black and white): the CG distance puts them close in the dendrogram, while the mean feature distance is not so informative.

As a further test, given the distance matrix between users (employing the CG distance and the mean feature distance), we computed the mean between every pair of users sharing at least 3 groups, assuming this as indication of shared common interests. In addition, we do the same by discarding the content-based features. Of the original 200 users, 183 have a link with each other, and 141 groups have been considered in total. Of course, normalization of distances has been done in order to have a fair comparison. The results are shown in Table 7.1, revealing three aspects: 1) our approach clusters users that share groups in a more compact way; 2) using the mean feature distance, the inclusion of content-based features do not compact the clusters, while 3) CG distance benefits from the addition of content-based information, compacting more the groups. An experiment considering only the content-based feature cannot be directly performed, as images with detectable objects are very few with respect to the size of the dataset.

Features retained	User maps distance	Feat distance
Non content-based	**0.2400**	0.3500
All features	**0.2125**	0.3552

Table 7.1: Mean intra-cluster distances.

8

A Statistical Generative Multi-resolution Approach

8.1 Introduction

In this chapter a novel generative embedding approach for managing the personal aesthetics for soft biometrics is proposed. The approach is based on the projection of the images into different latent spaces, each one of them representing a particular level with which to consider the preferred images. These spaces are 2D Counting Grids (CGs) [213] where each CG is characterized by a particular resolution, that in rough words models how much visually similar should be the images in order to be close on the grid: the higher the resolution, the stronger the visual similarity of close images. The presence of multiple resolution brings to evaluate differently grained similarity relations among images. The approach assumes the same dataset used in Chapter 5 and it consists in a serial pipeline of initialization, enrollment and identification/verification stage.

In the initialization stage, multiple levels corresponds to CGs of different resolutions, which are learned with the gallery images of all the users without using ID labels. In the enrollment, the training data of a single user is projected on the CGs at different resolutions, resulting in different *embedding maps*. These maps are then fed into Support Vector Machines (SVMs), one for each CG. In particular the SVMs are trained as exemplars SVM, that is, using a single map as positive sample, and as negative samples all the maps of the other users at that CG resolution. In the identification/verification stage, probe images are projected into the CGs, forming another set of maps which are then classified by each of the SVMs, and producing a joint prediction; this last is used to provide or verify the identity of the user. It is worth noting that our method works with a varying number of images, both for the enrollment and the identification/verification stage, providing a versatile approach.

Through some explicative experiments, it is easy to capture the advantages of our method. The use of 2D CGs allows to see the kind of images liked by some user and disliked by the others; projecting on low-dimensional spaces permits to use any kind and number of counting features for encoding images, contrarily to our previous approaches presented in Chapter 5, 6 and in [167], which are based on an explicit cues weighting; having CGs at multiple resolutions avoids to deal with model selection issues (deciding the "correct" resolution for a CG is a problem [213]). We also performed an extensive on the kind of features which can be used to describe the images showing that using color and composition cues gives the best results.

Finally, in order to demonstrate the effectiveness of our approach, we set up user study which analyzes a random subset of users; each user is asked to select a number of preferred images from a finite pool of available test images, as it could happen in a standard biometric application where the user has to select a signature from a finite number of alternatives. in this case, the available aesthetic variability diminishes dramatically, and as a consequence, identification performances drop.

8.2 The proposed approach

The proposed three-step approach is sketched in Figure 8.1. The initialization step is applied on the training image set: it consists on creating a bag of features for each image, and learning a set of Counting Grids, each at a different window size (i.e., the resolution of the CG). In the enrollment stage, the preferred images of each user x_u, $u = 1, \ldots, U$ of the gallery set are mapped on the CG latent spaces, and the resulting maps (one for each CG space) are fed into a discriminative classifier. In the identification/verification stage, the test images of a probe subject are transformed into bags of features, and projected into the CGs; in particular, in the identification scenario, the resulting maps are given as input to all the U gallery classifiers, producing U identification scores. These scores are used to decide the best gallery user. In the case of the verification task, the maps are given to a single gallery classifier (the one which is supposed to match the identity of the probe), which accepts or rejects the signature considering a given threshold.

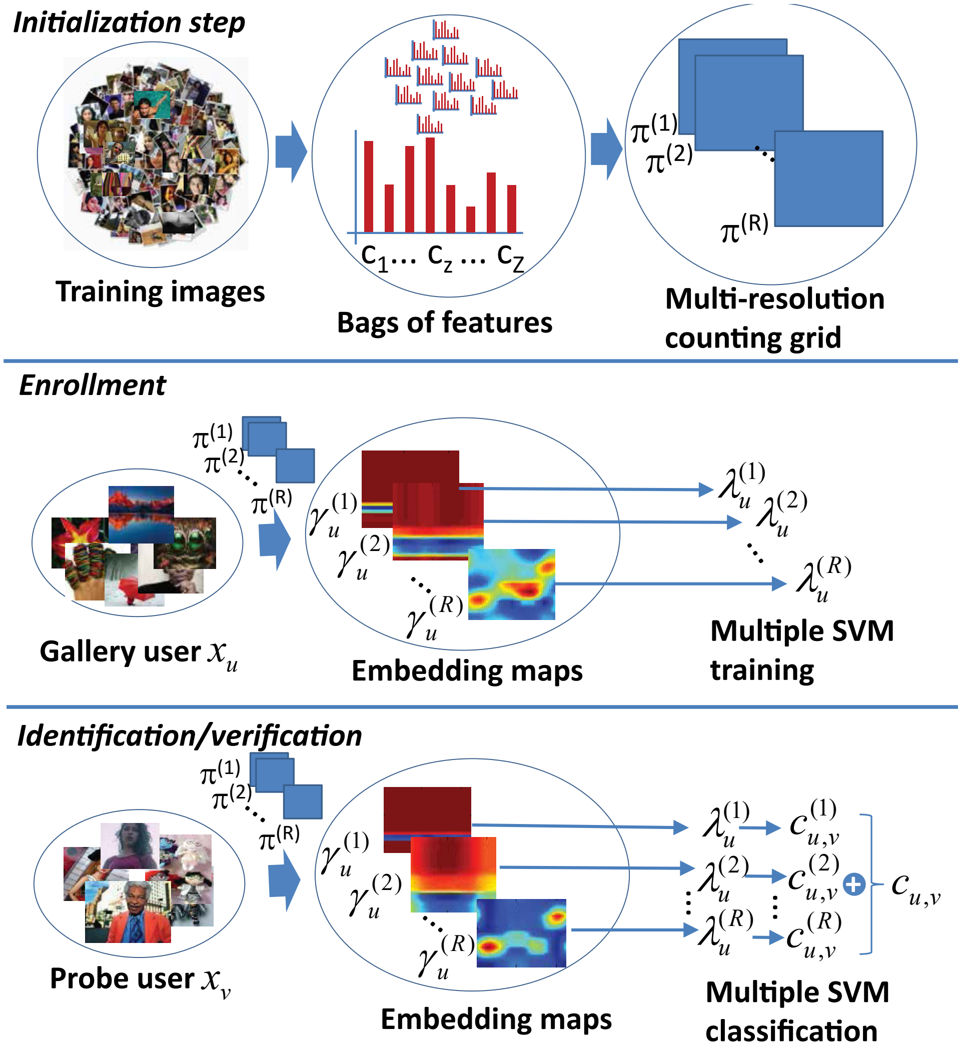


Fig. 8.1: The proposed approach, composed by three stages: *initialization*, where the multi-resolution Counting Grid is learnt; *enrollment*, where the classifiers for each user are trained, and *identification/verification* stages, where unknown personal aesthetics are matched with the gallery.

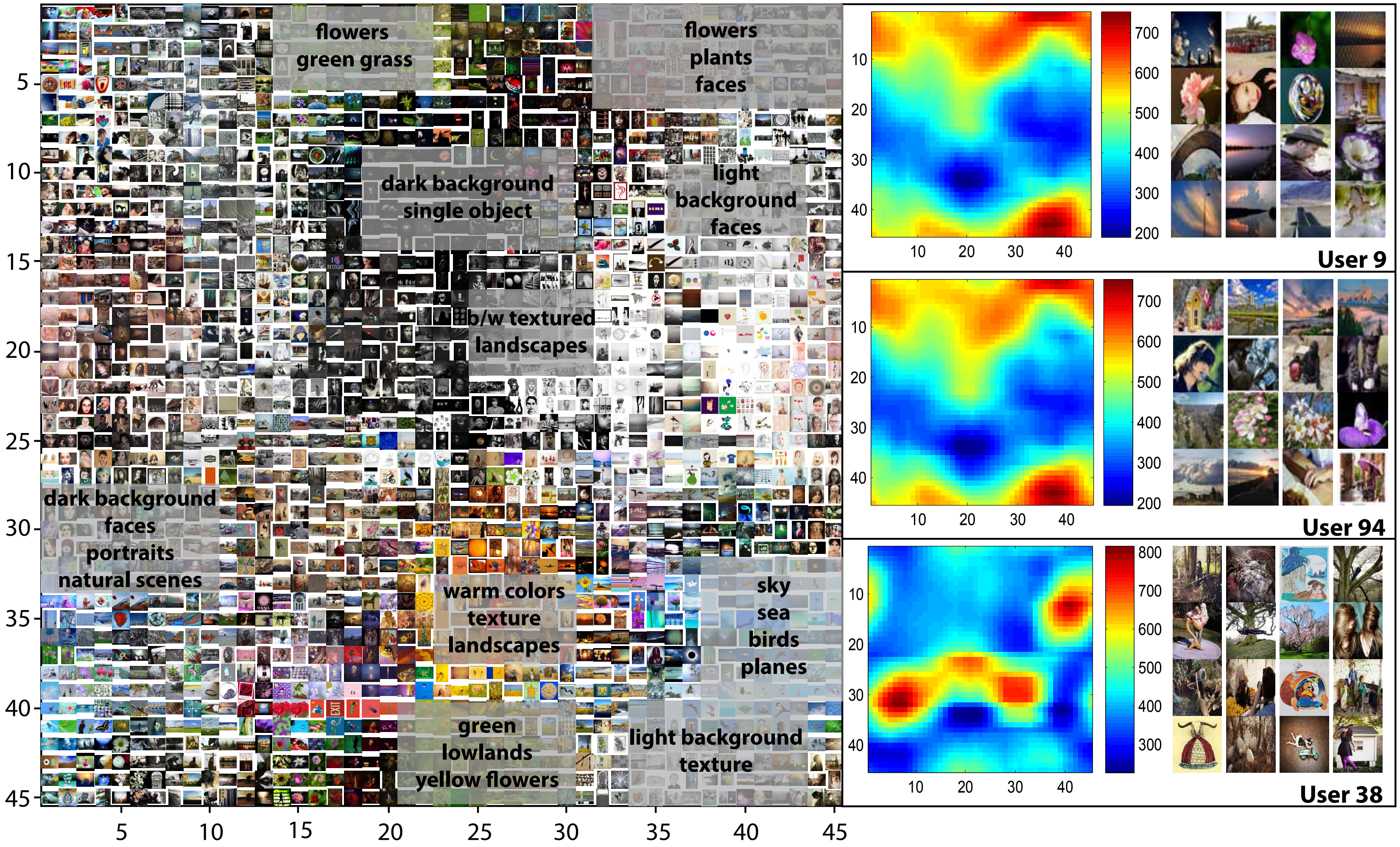
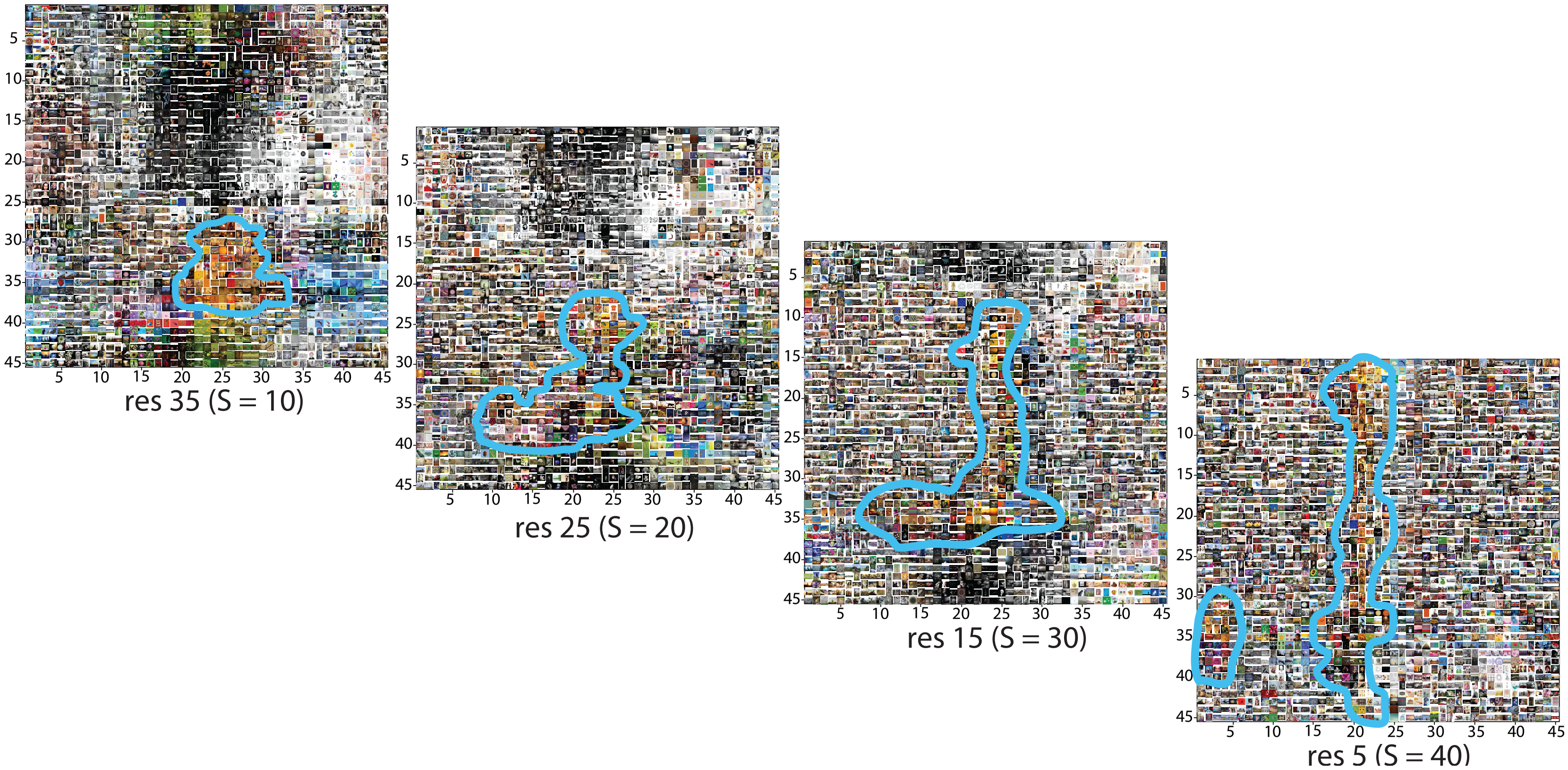


Fig. 8.2: Visualization of Counting Grids: on the left, CGs with $E = 45$ at resolutions r=5 (S=40, top) and 35 (S=10, bottom). On the left, the $S = 10$ grid is visualized as a collage of images (see the text for the details on how the collage is created). On the right, the embedding maps of a single resolution level (r=R) are reported for three subjects, together with some random images preferred by them (better viewed in colors).



res 35 (S = 10)

res 25 (S = 20)

res 15 (S = 30)

res 5 (S = 40)

Fig. 8.3: Visualization of Counting Grids: from the left, CGs with $E = 45$ at resolutions r=5 , 15, 25, 35.

8.2.1 Initialization Stage: Creating the Bags of Features

For the sake of comparison, in this work the dataset used in Chapter 5, 6 and 7 has been considered.

From each image $\mathbf{x}_t$, the same set of cues of Section 3.1 has been extracted, composed by 19 types of features resulting in a 112-dimensional real vector (see Table 5.1); the goal is to manage highly heterogeneous image features, letting them smoothly interact in the CG.

It is worth noting that each feature extracted in the proposed approach indicates the level of presence of a particular cue, i.e. an intensity count. This is needed for the modeling with the Counting Grid. For this reason, features working on angular measures (as those modeling the Hue channel in the HSV color space) have been discarded. Since the range values are very heterogeneous, each feature is normalized across all training images to have zero mean and unit standard deviation. The same normalization is then applied to the features extracted from test data.

8.2.2 Initialization Stage: Multi-view Counting Grid Training

Given the bags of features of the training images, the extent E of the Counting Grid and its window size S, a multi-resolution CG is learned. This amounts to learn $R = E - S$ Counting Grids, starting at resolution $r = 1$ (the lowest resolution level) with the window of size $E - 1$, decreasing the window size of one pixel at each time, until the minimum size S (the highest resolution level $r = R$) is reached. At each resolution level r (except the first one), we used the CG learned at the previous step, i.e., $\pi^{(r-1)}$ as initialization for $\pi^{(r)}$. At the first resolution level, the initialization is random.

Using different windows sizes corresponds to vary the topology of the CG latent spaces: a large window size leads to an embedding map where loosely similar images are near-uniformly distributed over a large area and only the very different images are strongly separated. Conversely, a small window size will create a peaked map, where only highly similar images are projected nearby, and weakly similar pictures are separated. Initializing a training model using the CG of the previous level allows to mitigate local minima problems (as in the case of a too sparse CG, with many images mapped very close) ensuring to use all the CG extent for the mapping. In addition, this initialization strategy permits to show how the mapping evolves at the different resolutions, refining spread and unfocused projections into defined and intuitive thematic regions.

As stated in Chapter 7, the Counting Grids can not be directly visualized (each location contains a distribution of features), but it is possible to create an image mosaic using those images $\{c_z^t\}$ which give the highest posterior at each location $\mathbf{k}$, i.e., $p(\mathbf{k}^t|\{c_z^t\})$, at a given resolution level r. Adopting this visualization strategy, Figure 8.3 shows CGs at resolutions r=5 (S=40), 15 (S=30), 25 (S=20), 35 (S=10). The aim of the figure is to show how, while going from coarse (bottom-right) to finer (top-left) resolution, the semantics of the CG emerge as peaked regions, where each region carries out a different type of images. For example, in this case a set of images where the orange is predominant are highlighted. As visible, at the coarser resolution the orange images lie in two regions, where other tonalities are also present. Going to finer resolutions has the effect of compacting those images, until they form a compact area at resolution 35 (S=10).

Such a representation is shown in Figure 8.2 (center) for a Counting Grid with $E = 45$ at resolution $r = R$ ($S = 10$, maximum resolution). As visible, close images are visually similar, and semantic topics do emerge.

8.2.3 Enrollment Stage

Once the different Counting Grids are learned, the images of each gallery user can be projected within it, obtaining R maps per user, one map for each resolution. The projection corresponds to a generative embedding, calculating a posterior probability at each location $\mathbf{k}$; once we have fixed a user u and a resolution r the posterior is

$$\gamma_u^{(r)} = \sum_{t \in T_u} p(\mathbf{k}^t | \{c_z^t\}, \pi^{(r)}) \tag{8.1}$$

where T_u identifies the set of images of the user u: T_u can be different, depending on how many gallery images are available for user u. Roughly speaking, the main idea is to sum all the mappings of the images belonging to a given user, thus highlighting the zones of the latent space where the images have been located. The presence of Counting Grids at multiple resolutions allows to map the preferences of the user from a very rough resolution (on the Counting Grids obtained with large windows) until the finest resolution (the Counting Grid being learned with a small sized window), where the map is usually peaked.

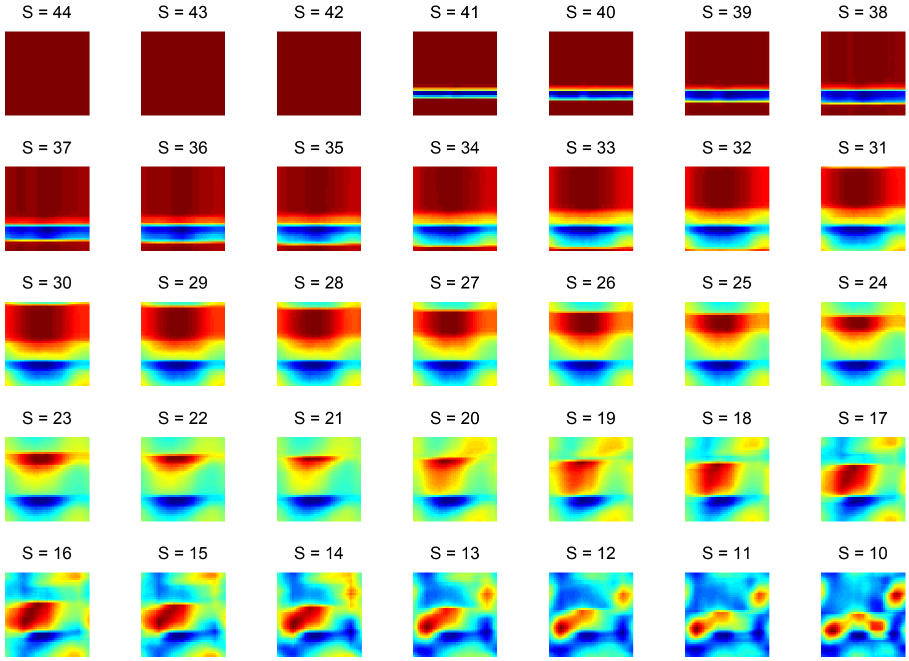


Fig. 8.4: Embedding maps for user 38 of Figure 8.2. Starting from the lowest resolution (r=1, S=44) and going towards higher resolutions, the maps show refined blobs and areas, identifying more precisely semantic areas, easily interpretable, on the grid.

A graphical explanation of the mapping process is shown in Figure 8.2 and Figure 8.4; in Figure 8.2, together with the collage of the CG, on the right are reported the embedding maps of a single resolution level (the maximum, i.e., r=R) for three subjects, together with some random images preferred by them. One can notice two facts: 1) given a user, looking at his map and at the CG collage as reference, does allow to easily understand which kind of images are his preferred; 2) comparing the maps of different users, one can understand possible similarities: first two users from the top appear to share much the same preferences, while the third one has radically diverse preferences. This fact is confirmed by checking the random pictures of the users, on the right.

In Figure 8.4 are reported the R mappings for the user 38 of Figure 8.2. Starting from very blurred and unstructured maps corresponding to the lower resolutions, going toward higher resolution maps, blobs and distinct areas start to emerge, refining the "semantic" knowledge of the preferences a user exhibits.

After the mapping step, the maps $\{\gamma_u^{(r)}\}_{r=1,\ldots,R}$ can be used as ID template for user u; to this sake, a battery of exemplar SVMs $\{\lambda_u^{(r)}\}_{r=1,\ldots,R}$ are learned (one for each resolution), using as positive samples the maps $\gamma_u^{(r)}$ at the different resolutions r (one map for each SVM) and as negative samples the maps of the other users. In this study, Support Vector Machines with radial basis functions have been employed. This step concludes the enrollment stage.

8.2.4 Identification and Verification Stage

In the identification/verification stage, all the probe images of a user v are first encoded as bags of features. Subsequently, they are mapped on the multi-resolution CG, and the resulting maps $\{\gamma_v^{(r)}\}_{r=1,...,R}$ are used as input of the SVMs related to the gallery user u; they classify the maps producing R scores $\{c_{u,v}^{(r)}\}_{r=1,...,R}$ that, once mediated, provide a single classification score $c_{u,v}$. In other words, each user produces R probe maps; each of them is given as test input to the correspondent SVM of the gallery user, providing a confidence score (the distance from the separating hyperplane). Averaging these scores over all the resolutions gives the final confidence score. In the identification case, a confidence score is associated to each gallery user; this allows to rank the scores, keeping the highest ranked user as the best match with the probe. In the verification of the probe user, assumed to be the $v-$th, the confidence score given by the $v-$th classifier is simply evaluated, accepting or rejecting the signature depending on a threshold opportunely decided.

8.3 Experimental evaluation

Several experiments are carried out to understand the potentialities of our approach. First of all, we investigate the ability of the features in capturing what is liked by an user, ensuring the highest identification and verification performance. Then, we compare our approach against a set of competitors, including our previous work Chapter 5: to this sake, the same experiments carried out in Chapter5 have been taken into account. Finally, we analyze how beneficial is to exploit CGs at different resolutions, capturing also how informative is each single resolution.

Identification and verification applications are considered. In both the cases, the parametrization of the Counting Grids is the same: the size is fixed at $E = 45$ pixels for all of them, while the (smallest) window size is set to $S = 10$; this generates a set of 35 maps per user. The extraction of the image features takes 60 minutes per user (100 images), on a not optimized MATLAB code run on a 3.4 GHz processor with 16 Giga of RAM. The learning of the Counting Grid at a single resolution takes in total 2 minutes, while the mapping + SVM training operation requires 3 seconds for $N = 100$ images of the same user, on the same computer. Regarding the variability of the results in relation to the E and S values, the proposed approach maintains similar performance when the ratio between E and S (also dubbed "capacity" in [213]) is bounded in the interval [3,5]. Even if E and S respect the capacity ratio, performances seem to decrease when $E < 10$ and $E > 70$.

8.3.1 Feature Analysis

Following the Table 5.1, for each feature category, we instantiate a identification task: given a probe signature built from an image or a set of images, the goal is to guess the gallery user who tagged them; to do that, fixing a gallery user, the average of the confidence scores produced by the exemplar SVMs (one score for each resolution) is calculated. Hopefully, the gallery user with highest averaged score corresponds to the probe user. As identification figure of merits, we use the Cumulative Matching Characteristic (CMC) curve [188]. In all the following experiments CMC plots are obtained averaging the CMC curves of 5 different experiments with different gallery/probe splits. In this experiment, we use 100 images as forming the gallery signatures, and 5 images for the probe signatures.

In Table 8.1 are reported the CMC values at different ranks, together with the normalized Area Under the Curve (nAUC). As visible, the color category is the most significative, followed by composition, texture and content. The poor performance of the content features, that is, object detectors, is due to

category	rank 1	rank 5	rank 20	rank 50	nAUC
color	0.38±0.21	0.65±0.01	0.86±0.01	0.97±0.01	0.96± <0.01
composition	0.11±0.01	0.25±0.02	0.45±0.02	0.69±0.12	0.81±0.01
texture	0.10±0.01	0.21±0.01	0.39±0.03	0.64±0.02	0.79±0.01
content	0.10±0.01	0.20±0.01	0.38±0.03	0.61±0.03	0.78±0.01
all	0.36±0.02	0.64±0.02	0.86±0.01	0.97± <0.01	0.96± <0.01
color + composition + textures	0.37±0.02	0.64±0.02	0.86±0.01	0.98±0.01	0.96± <0.01
color + composition	**0.42±0.02**	**0.71±0.02**	**0.91±0.01**	**0.99±0.01**	**0.97± <0.01**

Table 8.1: CMC scores for the identification task, 100 images fo reach gallery user and 5 images for the probe user

the fact that object detectors produce many errors, both in precision and recall: this is due to the nature of the Flickr photos, which are artistic and not reminiscent those of the object recognition benchmark (PASCAL, CALTECH and the like). We evaluate all the possible combinations of group of features (some of them are reported in the table), with the best one formed by color and composition, which will be used in the following. Interesting, the textures seem to slightly degrade the performances.

8.3.2 Identification Results

The results of the identification task are carried out following the protocol of Chapter 5. We cross-validate the parameters of the SVM classifier with Gaussian kernel obtaining the best configuration with $C = 1000$ and $g = 0.001$. As competitors, we report the performance of Chapter 5 (with the acronym *LASSO*) and Chapter 6 (*PaD*). In addition, we set up some baselines, which may help in motivating some technical choices we have made with our framework. The *Ensemble* method is the same as our proposal, with the only difference that the CGs are learned independently, without sharing their parameters; the *PCA* approach, which actually uses Principal Component Analysis to create a low dimensional space projection space where all the images can be projected. Once the projection of a probe signature is performed, the resulting map containing the projected images (opportunely quantized in order to be of the same dimension irrespective of the nature and cardinalities of the signatures) is fed into the exemplar SVMs. In Figure 8.5 the various CMC curves are reported by fixing the number of gallery images to 100, and the number of probe images to 5. As visible, our approach, here named *MRCG*, overcomes all the competitors.

T_{te}	*rank 1*	*rank 5*	*rank 20*	*rank 50*	*nAUC*
1	0.19	0.42	0.66	0.86	0.90
5	0.42	0.71	0.71	0.99	0.97
20	0.63	0.87	0.7	1.00	0.99
100	0.73	0.92	0.98	1.00	0.99

T_{tr}	*rank 1*	*rank 5*	*rank 20*	*rank 50*	*nAUC*
5	0.29	0.59	0.83	0.94	0.95
10	0.46	0.80	0.95	0.99	0.98
20	0.63	0.89	0.98	1.00	0.99
50	0.71	0.93	0.99	1.00	0.99
100	0.71	0.92	0.98	1.00	0.99

Table 8.2: Identification results, varying the number T_{te}/T_{tr} of images of gallery/probe signatures (and fixing the other cardinality to 100 for each user). All the results are with a variance of less than the 1%.

In Table. 8.2 we report (in the upper part) the performance of our approach while varying the number of test images used to compose the probe signature of a user, while keeping the number of images used

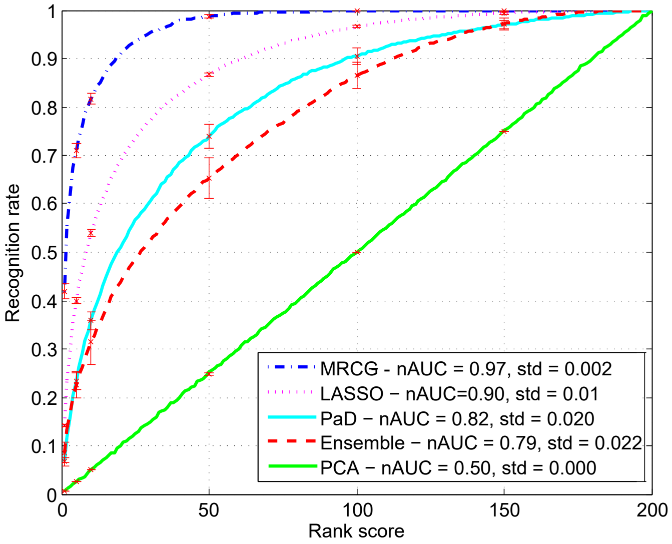


Fig. 8.5: Comparative results for the identification task, with 5 images for the probe signatures and 100 images for the gallery signatures.

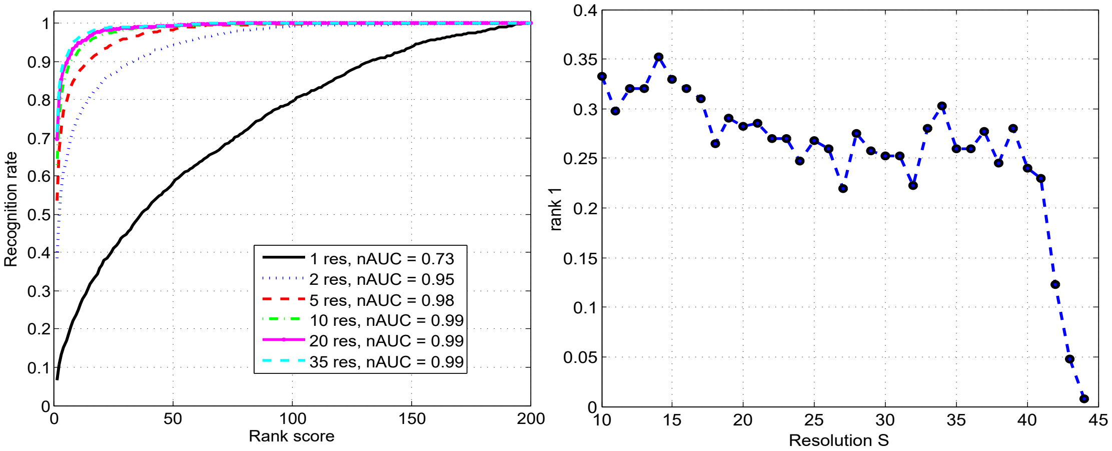


Fig. 8.6: Identification scores while varying the number of resolution employed, and analysis at rank 1.

to build the gallery signature fixed to 100; in the lower part we report the analogue figure while varying the cardinality of the gallery signatures and keeping fixed to 100 the cardinality of the probe signature. Intuitively, augmenting the cardinality of the gallery/probe signature does ameliorate the identification performance.

To test the importance of having different CG resolutions, we perform a set of identification trials while using 100 images of gallery and 100 of probe, with 1, 2, 5, 10, 20 and 35 different resolutions (35 is the total number of resolutions employed). In the case of a single resolution, all the S windows size between 10 and $E-1=44$ have been independently evaluated, averaging their recognition performance. For evaluating higher numbers of resolutions, different windows size have been sampled without replacement (depending on the cardinality being evaluated) and ranked in descending order. After that, the window with the largest size has been learned with random initialization; the obtained CG has been used as prior for the second ranked one and so on. Results (averaged over 35 gallery/probe splits) are portrayed in Figure 8.6.

As expected, increasing the number of resolution levels does augment the identification capabilities. To better understand the role of each resolution, each one of them has been evaluated independently (under the same experimental protocol, $T_{tr} = T_{te} = 100$, 35 repetitions), reporting in Figure 8.6 the rank-1 identification score (standard deviation $< 1\%$ in all the cases). It emerges that performance is better while going toward higher resolutions, even if no one of them can reach the same score one can get when using the joint framework (that is, 0.71, see Table. 8.2). This means that every resolution level carries out a different complementary analysis of the images.

8.3.3 Verification Results

In the verification scenario, the capability of the system to verify if a signature matches a given identity is evaluated. For this purpose, a ROC curve is computed for every user u, where client images are taken from the probe set of the user u and impostor images are taken from all the other probe sets. In Figure 8.7 the authentication ROC curves are portrayed; other than AUC, the equal error rate (EER) is also reported.

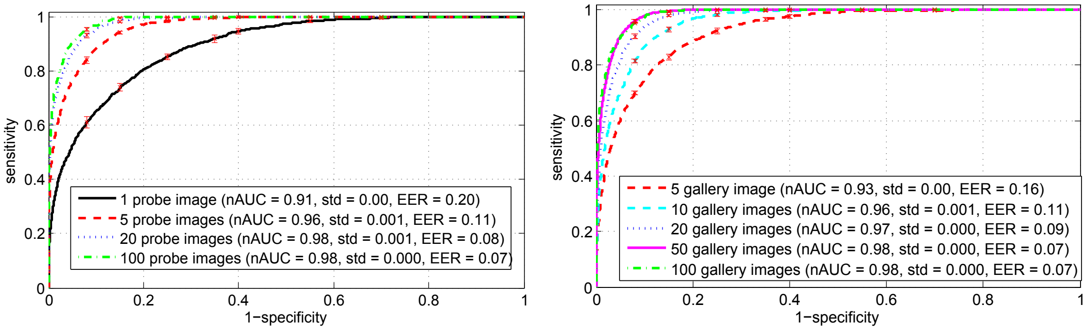


Fig. 8.7: Verification scores: the ROC curves (together with AUC and EER score) are reported while varying the number of probe (left) and gallery (right) images employed.

Even in this case, augmenting the number of test images per signatures increments the performance; as for varying the number of images used for the gallery signatures, and the number of resolutions for producing the multi-scale CG, analogue results than those obtained for the recognition task can be observed, so the results have been omitted.

8.3.4 Limitations of our approach

So far, all the works on personal aesthetics did the general assumption that all the images selected from the users, both of training and testing, were not overlapping, that is, no common preferred images are shared among users. In the dataset used so far this hypothesis holds, being the number of repeated images less than the 0.2%. But this is not always the case, especially when a much larger number of users is occurring, or when the images to select come from a restricted number of available pictures. In this last experiment we take into account this situation with a user study: as first operation we build a "reduced" test dataset, by sampling one image from each pool of the 200 originally liked images of all the 200 users, clearly avoiding repeated images. These 200 images have been organized on a web interface, where the users can select them. Then, 16 users of the original dataset have been asked to select from this interface 5, 10, 20 images. After that, the selected images have been used as a test signature for our approach, and compared with the gallery signatures (which for simplicity have been kept equal to the experiments of the previous section), generating three different CMC curves. As comparison, we use the test signatures coming from the original test images of the dataset, and not from the reduced set. The results are shown in Figure 8.8, which show that the images coming from the reduced dataset have obviously less discriminative power than the ones coming from the original one: this is because of the less aesthetical variability contained within, and to the possible number of images which have been selected by more than one user. In this sense, it is interesting to note that, while increasing the number of images used for building the signature from 5 to 20, the performance of the reduced dataset slightly diminish, while in the original case they obviously augment.

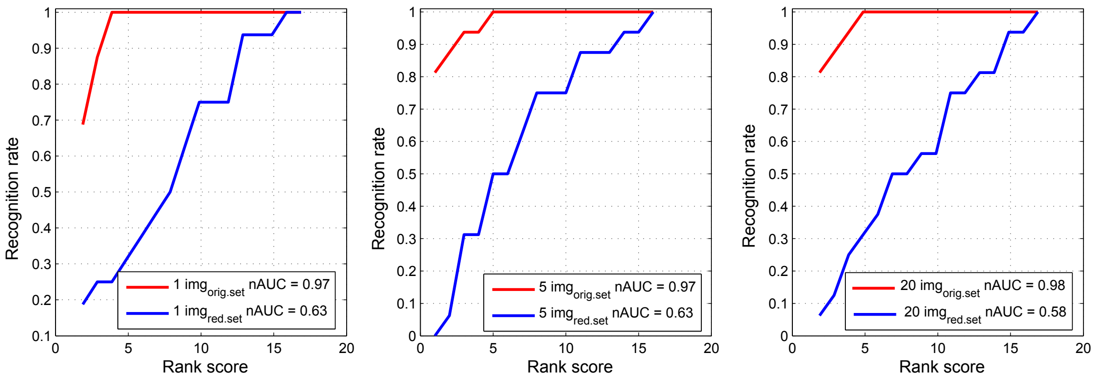


Fig. 8.8: Identification performance while using the original test dataset (img$_{\text{orig. set}}$) and the reduced dataset (img$_{\text{red. set}}$).

8.4 Overcome Content Features Limitation through R-CNN

As shown in Section 8.3.1, our method found a limitation for what concerns the performances when dealing with the objects detectors features. To this purpose we tried to overcome this lack using a novel approach based on deep learning.

Deep learning proposed several frameworks with the aim of objects detection, and reach the state of the are performances so far [46, 146, 250] compared to the old fashion hand crafted methods of [77, 79]. Taking advantages from these encouraging results we propose to replace the object detection features used so far in our experiments with these novel and more robust techniques.

The last decade of progress on various visual recognition tasks has been based considerably on the use of SIFT [170] and HOG [60]. But if we look at performances on the canonical visual recognition task, PASCAL VOC object detection [77], it is generally acknowledged that progress has been slow during 2010-2012, with small gains obtained by building ensembles systems and employing minor variants of successful methods. SIFT and HOG are blockwise orientation histograms, a representation we could associate roughly with complex cell in V1, the first cortical area in the primate visual path-way. But we also know that recognition occurs several stages downstream, which suggests that there might be hierarchical, multi-stage processes for computing features that are even more informative for visual recognition.

Fukushima's "neocognitron" [87] , a biologically inspired hierarchical and shift-invariant model for pattern recognition, was an early attempt at just such a process. The neocognitron, however, lacked a supervised training algorithm. Building on Rumelhart et al. [238], LeCun at al. [150] showed that stochastic gradient descent via back-propagation was effective for training convolutional neural networks (CNNs) (see Section 3.2.1), a class of models that extend the neocognitron.

CNNs saw heavy use in the 1990s but then fell out of fashion with the rise of SVMs. In 2012, [146] rekindled interest in CNNs by showing substantially higher image classification accuracy on the

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [66, 240]. Their success resulted from training a large CNN, together with a few twists on LeCun's CNN. The central issue on the results can be distilled to the following: to what extent do the CNN classification results on ImageNet generalize object detection results on the PASCAL VOC Challenge? [98] answer this question by bridging the gap between image classification and object detection. Here we follow the same direction and inspire our solution to this approach.

8.4.1 Breakthrough idea and results

[98] shows that CNN can lead to dramatically higher object detection performance on PASCAL VOC as compared to system based on simpler HOG-like features.

To achieve these results they focused on two problems: localizing objects with a deep network and training a high capacity model with only a small quantity of annotated detection data. They solve the CNN localization problem by operating within the "recognition using regions" paradigm [106], which has been successful for both object detection and semantic segmentation [42]. At test time, their method generates around 2000 category-independent region proposals for the input image, extracts a fixed-length feature vector from each proposals using a CNN, and then classifies each region with category specific linear SVMs. They use a simple technique, affine image wrapping, to compute the fixed-size CNN input from each region proposal, regardless of the region's shape. Figure 8.9 presents an overview of the method. Since the method combines region proposals and CNNs, they dub it R-CNN.

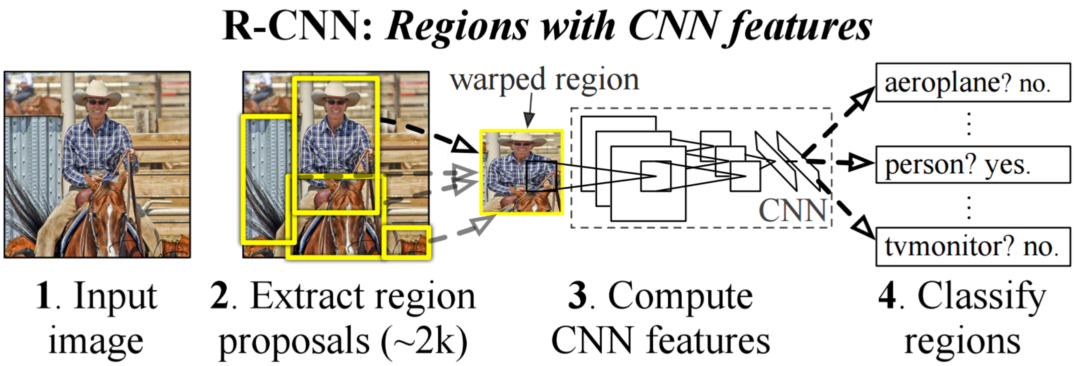


Fig. 8.9: Object detection system overview. The system (1) takes an input image, (2) extracts around 2000 bottom-up region proposal, (3) computes features for each proposal using a large CNN, and then (4) classifies each region using class specific linear SVMs.

Here, using the Caffe Toolbox [126] - a fully open source deep learning framework that affords clear access to deep architectures - we run detection of the R-CNN model for ImageNet on our Flickr dataset. Inspired by [98], we exploit selective search algorithm to retrieve the region proposals and the Caffe R-CNN ImageNet model: selective search is the region proposer used by R-CNN. Selective search combines the strength of both an exhaustive search and segmentation. Like segmentation it uses the image structure to guide the sampling process; like exhaustive search, the aim is to capture all possible object location. It is subject to the three design considerations: it needs to take into account all possible object scales, need to consider diverse set of strategies to group regions together as there is no optimal strategy and finally it has to compute the set of possible object location reasonably fast.

The first consideration is addressed by hierarchical grouping, used as a basis of the selective search. In particular [267] used bottom-up grouping for segmentation [50, 81] Because the process of grouping itself is hierarchical, we can naturally generate locations at all scales by continuing the grouping process until the whole image becomes a single region. This satisfies the condition of capturing all scales. As regions can yield richer information than pixels, [267] wants to use region-based features whenever possible. To get a set of small starting regions which ideally do not span multiple objects, they used the fast method of [81]. So the grouping procedure, first uses [81] to create initial regions, then uses a greedy algorithm to iteratively group regions together. First the similarities between all neighboring regions are calculated. The two most similar regions are grouped together, and new similarities are calculated between the resulting region and its neighbors. The process of grouping the most similar regions is repeated until the whole image becomes a single region.

The second design criterion is to diversify the sampling and create a set of complementary strategies whose location are combined afterwards. [267] used (1) a variety of color spaces with different invariance properties, (2) different similarity measures and (3) varied the starting regions.

Then the object hypotheses of several variations of our hierarchical grouping algorithm are combined. Ideally, the wish is to order the object hypotheses in such a way that the locations which are most likely to be an object come first. This enables one to find a good trade-off between the quality and quantity of the resulting object hypothesis set, depending on the computational efficiency of the subsequent feature extraction and classification method. The order to combine object hypotheses set is based on the order in which the hypotheses were generated in each individual grouping strategy.

The selective search outputs a DataFrame with the filenames, selected windows and their detection scores to an HDF5 file; the R-CNN detector outputs class scores for the 200 detection classes of ILSVRC. The number of proposals will vary from image to image based on its contents and size. From each of this HDF5 file related to each image, given the classes labels, we retained for each class, the window with the highest posterior probability from the classification.

As new feature vector then we use the highest posterior probability of each class and the area of the window. In this way, we retain both the probability that an object is present in the image and that is not on it, so the higher the posterior, the higher the probability that an object is present on the image. We use this new feature vector $\mathbf{x}_{R-CNN}$ to repeat the experiments performed in Section 8.3, for what concern the category related to the object.

The new feature vector based on the R-CNN object detection outperform the results achieved in Section 8.3, here named as *MRCG*. Figure 8.10 reports the result for the identification task. We can notice that just with only one image the AUC reach 0.94 compared to the 0.72 achieved with the object detector of [79, 80]. Table 8.3 shows the comparison of the identification results of the two approaches, while varying the number of probe images, highlighting that R-CNN framework reaches higher identification scores. The same applies for the recognition task as shown in Figure 8.11. With only one image we reach an AUC of 0.90 compared to 0.72 achieved using [79, 80].

MRCG T_{te}	*rank 1*	*rank 5*	*rank 20*	*rank 50*	*nAUC*
1	0.19	0.42	0.66	0.86	0.90
5	0.42	0.71	0.71	0.99	0.97
20	0.63	0.87	0.7	1.00	0.99
100	0.73	0.92	0.98	1.00	0.99

R-CNN T_{te}	*rank 1*	*rank 5*	*rank 20*	*rank 50*	*nAUC*
1	0.50	0.82	0.83	0.97	0.99
5	0.79	0.95	1.00	1.00	1.00
20	0.83	0.98	1.00	1.00	1.00
100	0.85	1.00	1.00	1.00	1.00

Table 8.3: Identification result varying the number of T_{te} of images of probe signatures. On the top the results form Section 8.3, on the bottom the results achieved with the proposed method.

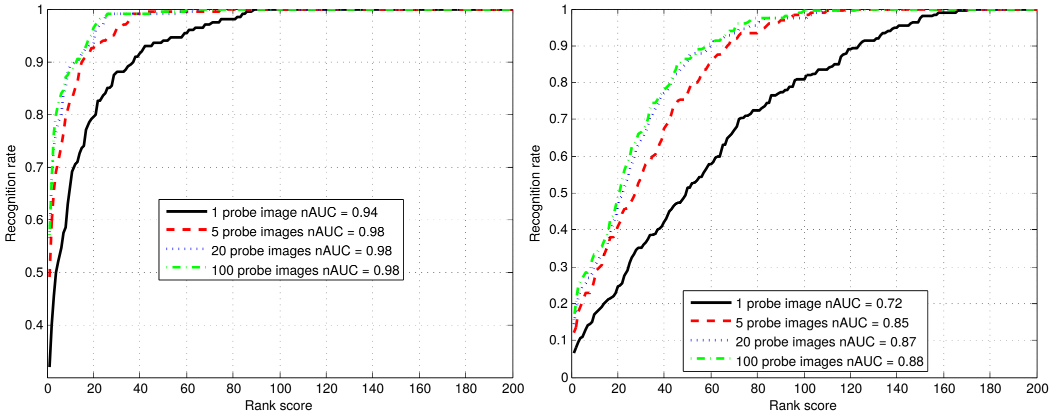


Fig. 8.10: Comparative results for the identification task, varying the number of images for the probe signatures and 100 images for the gallery signatures. On the left the CMC from the use of R-CNN detection object, on the right the CMC from the standard Deformable Part Models [79, 80] system to detect objects.

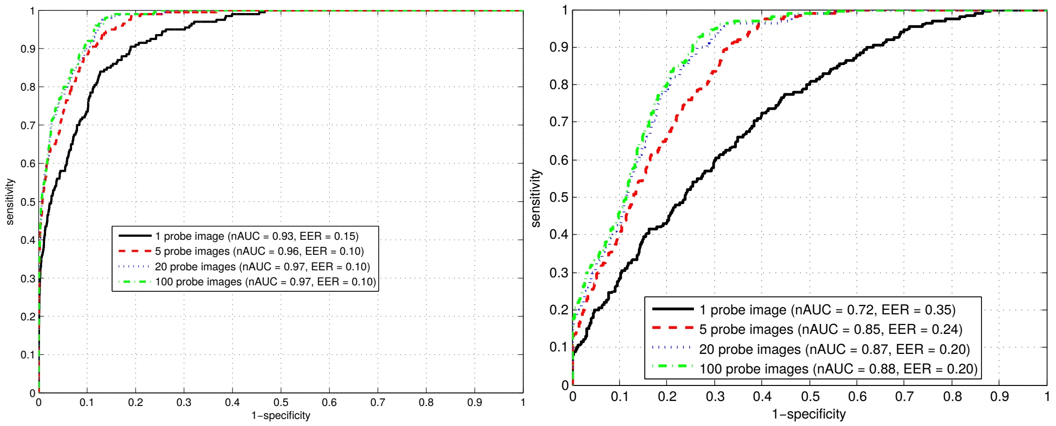


Fig. 8.11: Comparative results for the recognition task, varying the number of images for the probe signatures and 100 images for the gallery signatures. On the left the CMC from the use of R-CNN detection object, on the right the CMC from the standard Deformable Part Models [79, 80] system to detect objects.

Part III

Personality Computing through Multimedia Content Analysis

Personality Computing

Personality, one of the most important factor influencing our life, 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*" [88, 89]. Human sciences show not only that personality plays a major role in every aspect of our life, but also that we unconsciously and spontaneously tend to attribute personality traits to others [268].

As a construct, personality aims at capturing stable individual characteristics, typically measurable in quantitative terms, that explain and predict observable behavioral differences [182]. Current personality models successfully predict "*patterns thought, emotion and behavior*" [89] as well as important life aspects, including "*happiness, physical and psychological health, [...] quality of relationships with peers, family, and romantic others [...] occupational choice, satisfaction, and performance, [...] community involvement, criminal activity, and political ideology*" [208]. Furthermore, attitude and social behavior towards a given individual depend, to a significant extent, on the personality impression others develop about her [268].

Such an effectiveness in capturing the crucial aspects of an individual is probably the main reason behind the interest of the computing community for personality. After the earliest, pioneering approaches aimed at integrating personality psychology in Human-Computer-Interaction [195], the interest of the topic was fueled by three main phenomena in the technological landscape. The first is the increasing amount of personal information, often self-disclosing beyond intention [134], available on social networking platforms [226]. The second is the possibility of collecting everyday spontaneous, fine-grained behavioral evidence through mobile technologies and, in particular, smartphones [225]. The third is the attempt of endowing machines with social and effective intelligence, the ability of interacting with humans like humans do [289]. The three phenomena are probably the reason of the sudden rise of interest for the topic in the mid 2000s.

Overall personality is relevant to any computing area involving understanding, prediction or synthesis of human behavior. Still, while being different and diverse in terms of data, technologies and methodologies, all computing domains concerned with personality consider the same three main problems, namely the recognition of the true personality of an individual (Automatic Personality Recognition), the prediction of the personality others attribute to a given individual (Automatic Personality Perception), and the generation of artificial personalities through embodied agents (Automatic Personality Synthesis). It is only recently that APP and APR attracted the interest of the computing community. To the best of our knowledge, the results presented in this thesis are among the earliest investigation of the APP and APR problem and addressed several issues that, at the beginning of the thesis, were still unexplored.

According to [216] the three areas stem from different aspects of the Brunswik Lens [36] (see Section BLM), the cognitive model depicted in Figure 8.12. Originally proposed to explain how living beings gather information in the environment, the Brunswik Lens was later adopted to describe externalization and attribution of socially relevant characteristics during human-human [246] and, more recently, human-machine [216] interactions. The following section describe how personality is measured and the Lens Model in detail and shows the correspondence between the phenomena the Lens accounts for and the three Personality Computing problems mentioned above.

The Big Five Model: Personality and its measurement

The key-assumption of personality psychology is that stable individual characteristics result into stable behavioral patterns that people tend to display, at least to a certain extent, independently of the situation. Therefore, the main goals of personality psychology are "*to distinguish internal properties of the person from overt behaviors, and to investigate the causal relationships between them*" [182]. In other words, personality psychology aims at predicting observable individual differences based on stable possibly measurable, individual characteristics.

Different theories adopt different "*internal properties*" as a personality bias, including psychology (the biological perspective), unconscious (the psychoanalytic perspective), environment (the behaviorist perspective), inner states (the humanistic perspective), etc. [89]. However the model that most effectively predict measurable aspects in the life of people as those based on traits, a construct widely recognized as "*on of psychology's major achievements*" [51]. Trait model build upon human judgment about semantic similarity and relationship between adjectives that people use to describe themselves and the others [101].

While numerous and widely different, the terms used to describe people typically account only for a few, major dimensions. These latter, if sufficiently stable, are the adopted as *personality traits*, i.e. as factors capable of capturing stable individual characteristics underlying overt behavior.

Several decades of research and experiments have shown that the same traits appear with surprising regularity across a wide spectrum of situations and cultures, suggesting that they actually correspond to psychologically salient phenomena [51]. These traits, known as *Big-Five (BF)* or *Five-Factor Model (FFM)* are today the "*dominant paradigm in personality research, and one of the most influential model in all of psychology*" [166]. Trait models represent personality in terms of values, a form particularly suitable for computer processing.

The personality model most commonly applied in the literature, known as the Big-Five Model (BF), relies on five broad dimensions that not only capture most of the observable differences between people but also are stable across cultures and situations [245]. The BF personality factors "*appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences*" [297].

"*Highly replicable dimensions*" means that it can apply any possible situation, individual; "*parsimoniously and comprehensively*" means that if you add a dimension the model becomes redundant while, if you crop one it can lose information; "*phenotypic individual differences*" means that is not directly observable but explain what we can observe.

Over the last 50 years, the Big-Five Model has become a standard in psychology, and experiments using the Big Five have repeatedly confirmed the influence of personality traits on many aspects of individual behaviour including leadership, general job performance, attitude toward machines, sales ability, and teacher effectiveness. Big Five traits have been shown to influence the human/technology relationship, affecting attitudes toward computers in general as well as toward specific technologies such as adaptive systems, conversational agents, and tutoring systems. For all these reasons, most of the works concerned with the automatic prediction of personality have addressed the Big Five. The BF have been identified by applying Factor Analysis to the large number of words describing personality in everyday life (around 18000 in English) [245].

The Big-Five traits are as follows:

- **Extraversion**: Active, Assertive, Energetic, Outgoing, Talkative, etc.
- **Agreeableness**: Appreciative, Kind, Generous, Forgiving, Sympathetic, Trusting, etc.
- **Conscientiousness**: Efficient, Organized, Planful, Reliable, Responsible, Thorough, etc.
- **Neuroticism**: Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying, etc.
- **Openness**: Artistic, Curious, imaginative, insightful, Original, Wide interests, etc.

In this perspective, the clusters are interpreted as the trace that salient psychological phenomena leave in language (the lexical hypothesis [245]), one of the main evidences supporting the actual existence of the BF [182]. In light of the above, the BF model represents a personality with five scores (one per trait) that can be thought of as the position on an ideal personality map. Thus, in the BF perspective, personality assessment means essentially to obtain those scores. As the BF account for "phenotypic individual differences" (see quote from [245] above), the main instruments for score assignment are questionnaires where a person is assessed in terms of observable behaviors and characteristics, i.e. in terms of what a person does or how a person appears to be. Intuitively, assessing the personality of an individual means to measure how well the adjectives above describe her. Questionnaires where people rate their own behavior with Likert scales are the instrument most commonly adopted for such a purpose [31]. Table 8.4 shows the Big Five Inventory 10 (BFI-10), the questionnaire used in these works [228].

ID	Question	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1	This person is reserved	(1)	(2)	(3)	(4)	(5)
2	This person is generally trusting	(1)	(2)	(3)	(4)	(5)
3	This person tends to be lazy	(1)	(2)	(3)	(4)	(5)
4	This person is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
5	This person has few artistic interests	(1)	(2)	(3)	(4)	(5)
6	This person is outgoing sociable	(1)	(2)	(3)	(4)	(5)
7	This person tends to find fault with others	(1)	(2)	(3)	(4)	(5)
8	This person does a thorough job	(1)	(2)	(3)	(4)	(5)
9	This person gets nervous easily	(1)	(2)	(3)	(4)	(5)
10	This person has an active imagination	(1)	(2)	(3)	(4)	(5)

Table 8.4: The BFI-10 questionnaire used in this thesis. The version reported here is the one that has been proposed in [228]

BFI-10 is a short version of full BFI including only 10 items of original questionnaire. Each question is associated to a 5 points Likert scale (from "Strongly Disagree" to "Strongly Agree") mapped into the interval [-2, 2]. The BFI-10 includes the ten items that better correlate with the assessments obtained using the full BFI (44 items). The personality scores can be obtained using the answers provided by the assessors as follow (Q_i is the answer to item i) and mapped into the interval [-4,4]:

- Extraversion: $Q_6 - Q_1$
- Agreeableness: $Q_2 - Q_7$
- Conscientiousness: $Q_8 - Q_3$
- Neuroticism: $Q_9 - Q_4$
- Openness: $Q_{10} - Q_5$

The main advantage of the BFI-10 is that it takes less than one minute to be completed. Furthermore, it can be transformed into a self-assessment questionnaire by simply changing the way questions are formulated (e.g., item 1 becomes "I am reserved") and result into the personality others attribute to a given individual. In the latter case, every subject must be rated by several assessors and each of these must rate all subjects involved in an experiment. Statistical criteria (e.g., reliability proposed in [237]) allow one to set the number of assessors based on their mutual agreement.

The main limitation of self-assessment is that the subjects might tend to bias the ratings towards socially desirable characteristics, especially when the assessment can have negative consequences like, e.g. filling a job interview. However, extensive experiments have shown that the correlation tends to be high between self-assessments and assessments provided by acquainted observers (spouses, family members, etc.) [31].

The Multimedia Lens Model

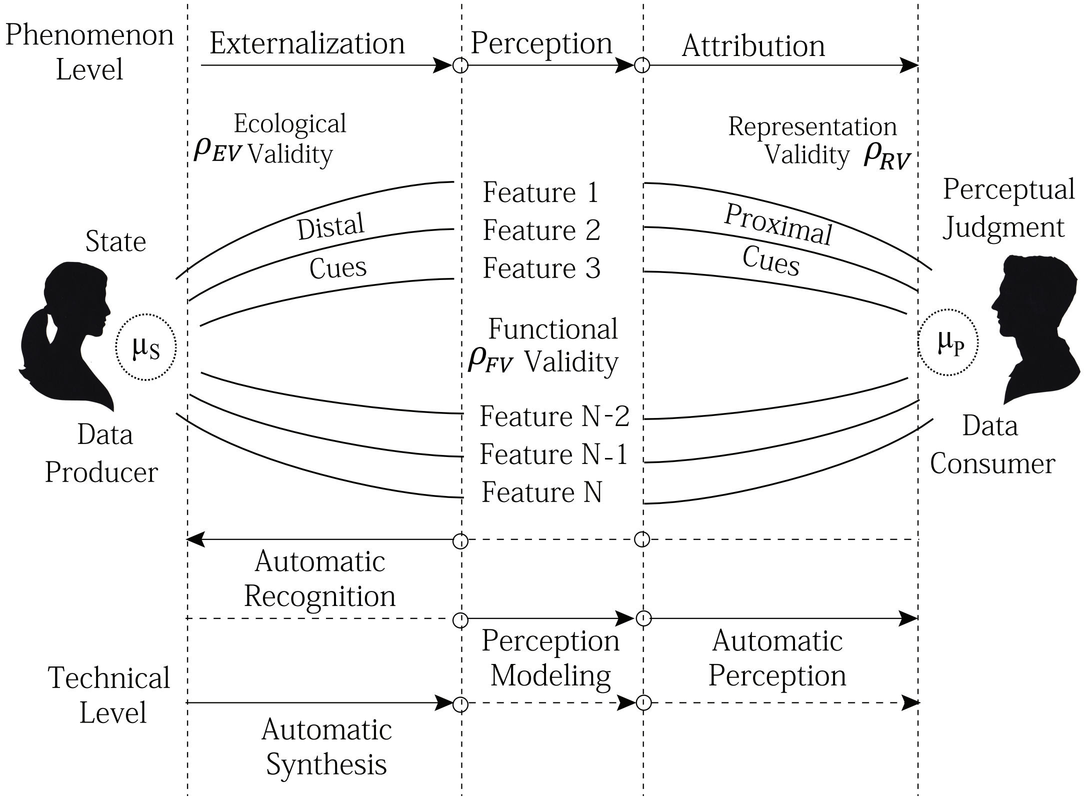


Fig. 8.12: The picture shows a simplified version of the Brunswik's Lens Model adapted to the exchange of multimedia data between a Data Producer and a Data Consumer.

Humans deploy knowledge about personality to attribute traits to other people, even those they never met before and even on the basis of very short sequences (down to a few seconds) of expressive behavior (so-called *thin slices* i.e. ability to find patterns in events based only on narrow windows, of experience). The human attribution process can be described by means of Brunswik's Lens model [36], to investigate, among other interaction phenomena, the influence of non-verbal behavior in face-to-face interactions [247] or the judgment of rapport [18].

In the model of Figure 8.12, the multimedia data is considered a form of communication between "*Data Producers*" (DP) and "*Data Consumers*" (DC). The key-idea is that the process includes not only the exchange of content, a problem that the multimedia indexing and retrieval community has extensively investigated, but also implicit cognitive processes typical of any human-human interaction like, *e.g.*, the spontaneous attribution of socially relevant characteristics (attractiveness, trustworthiness, etc.) or the development of impressions.

The DP is always assumed to be in a certain *state* that can be either *transient* (*e.g.*, emotions, attitudes, goals, physiological conditions etc.) or *stable* (*e.g.*, personality traits, values, social status, etc.). In operational terms, the states are defined as quantitative measures (identified as μ_S in Figure 8.12) to be obtained via objective processes depending on the particular case under observation. For example, in the case of the social status, the measure can be the yearly income of the DP, while for the physiological condition it can be the heart rate or the galvanic skin conductance. In many cases, the states correspond to psychological constructs (*e.g.*, personality traits or interpersonal attractiveness) and the measures are the outcome of psychometric questionnaires. These latter are typically administered to the DPs and include questions associated to *Likert* scales.

According to the model, *distal cues* are an *externalization* of the DP state, i.e. any form of observable behavior that can be perceived by others [216] (see left hand side of Figure 8.12). In other words, while being an abstract psychological construct non accessible to direct observation, personality leaves physical traces - or markers - in virtually everything observable individuals do [247]. In our case, individuals *externalize* their personality through multimedia data. Furthermore, the data is all the DCs know about a DP. From an operational point of view, the data correspond not only to the actual multimedia material (*e.g.*, pictures, video, soundbites, etc.), but also to any *feature* that can be extracted, manually or automatically, from the material itself. The empirical covariation of state measures and features quantifies the *ecological validity* of these latter indicated with ρ_{EV}, i.e. their effectiveness in accounting for the DP state.

When the DCs consume the data, they attribute to the DP a state of measure μ_P. Distal cues that reach an observer undergo a perception process that results into a percept, i.e. the "mental representation of something that is perceived" (central part of Figure 8.12) [108]. The Lens Model distinguishes between distal and *proximal* cues, these latter being the ones the observer actually perceives. Proximal cues activate the *attribution* process (see right hand side of Figure 8.12), i.e. the development of *perceptual judgment*, referred ad μ_P, that accounts for the personality traits an observer attributes to a person being observed. For example, the DCs can attribute a certain yearly income to the DP based on the pictures and videos this latter shows. In the case of the psychological constructs, the attribution process is typically unconscious and it takes place spontaneously, whether the DC needs it (wants it) or not [268, 269]. In principle, μ_S and μ_P should have the same value (or at least similar values), but communication processes are always noisy, especially when the communication takes place through ambiguous channels like multimedia data are. The empirical covariation between features and perceptual judgments accounts for the *representation validity* of the features (identified as ρ_{RV}), i.e. for the influence these latter have on the attribution process. Like in the case of ρ_{EV}, the most common measurements of ρ_{RV} are correlation and Spearman coefficient.

The cognitive processes my thesis focuses on are active in particular at the perceptual judgment stage, when DCs unconsciously develop an impression about the DP even if all they know about this latter is the multimedia data they are consuming. However, the processes are important for the DP as well because, in a communication scenario, there is no data production without an attempt to convey an impression, i.e. to ensure that μ_S and μ_P are close to each other. The empirical covariation of μ_S and μ_P (identified as ρ_{FV} in Figure 8.12) accounts for the latter aspect and it is called *functional validity*.

Furthermore, the full version of the Brunswik Lens [36] allows one to include contextual information. This makes it possible the integration of the technologies into context aware approaches [10].

Humans have limited access to network-type information and are not specifically attuned to it. Computer systems, in turn, can access and exploit the huge amounts of information about the people's networks that are contained in the digital traces of individuals' and groups' behaviors provided by wearable sensors, smartphones, e-mails, and the like.

One way to endow computers with the ability to predict people's personality is to adopt some variants of the (modified) Brunswick's model introduced above and apply it to zero-acquaintance cases exploiting thin slices of expressive behavior. The task can be modeled either as a classification or a regression task.

The ground truth for the target variable - personality assessments - can be provided by means of standard questionnaires that are either compiled by the subjects themselves (self-assessment) or by other people (other-assessment) who are either well acquainted with the target subjects (e.g., relatives) or, more frequently, strangers who see them for the first time. Each thin slice was then summarized by a feature vector consisting of the means and standard deviations of the behavioral cues.

Personality Externalization and APR

Automatic Personality Recognition targets the externalization process and it is the task of inferring self-assessed personalities from machine detectable distal cues. The task is called "*Recognition*" because it aims at inferring traits resulting from self-assessments, traditionally considered to be the true traits of an individual [228]. In most cases, APR approaches adopt methodologies typical of Affective Computing, Social Signal Processing, sociolinguistics and the other domains aimed at inferring emotional and social phenomena from machine detectable behavioral evidence. In computing research, covariation studies are often aimed at performing features selection, i.e. at identifying the distal cues most likely to lead to high APR performance.

APR approaches presented so far in the literature consider a wide spectrum of distal cues, including texts, nonverbal behavior, data collected via mobile or wearable devices and online games. Several works investigate the interplay between personality and computing by measuring the link between traits and use of technology [107, 221, 224, 305]. The core principle behind this line of research is that users externalize their personality through the way they use technology. Therefore, personality traits should be predictive of users' behavior. Predicting individual traits based on various cues such as samples of written text, answers to psychometric test, or the appearance of space people inhabit [104], has a long history.

APR and Text

Language psychology shows that the choice of words is driven not only by meaning, but also by psychological phenomena such as emotions, relational attitudes, power status and personality traits [262]. Therefore integrating sociolinguistics in techniques for automatic text analysis makes it possible, among other task, to infer personality traits from written text. One of the earliest efforts in this direction was proposed in [6]. The experiments were performed over 2263 essays written by roughly 1200 students that filled the NEO-FFI [53]. The goal of the experiments was to discriminate between subjects at the opposite extreme of Extraversion and Neuroticism. They grouped the words into four physiologically meaningful category: function, cohesion, assessment, appraisal; the texts were represented with the relative frequencies of the words appearing in each category. The same data and approach were used in [175] but adopting 88 categories from LIWC [262]. More recent efforts aimed at inferring personality traits from texts tend to focus on blogs, being a way to focus on personal issues and experiences, showing traces of their author's personality [99]. In [95], the experiments aimed at predicting whether a blogger was, "low", "medium", or "high" with respect to each of the Big Five traits. The analysis allowed the authors to identify that for example, neurotic authors tend to blog to release their tensions while extroverts tend to talk about their life. The works presented so far in this section adopts lexical approaches, i.e. they are based on statistics over the use of individual words. Other approaches consider word N-grams [201].

APR and Nonverbal Communication

Psychology suggests that nonverbal communication is, at the same time, an externalization of personality [73] and a cue that influences the traits that others attribute to an individual [268]. From APR point of view, this means that people's personality can be inferred, at least in principle, from automatically detected nonverbal behavioral cues. An example is [11] where the experiments were performed over 89 video of self-presentations delivered by Skype. Prosodic features, eye-gaze direction, frown, posture, hand movements, head shake/nods, fidgeting and duration of videos were fed to SVM and Naive Bayes classifiers to predict whether one individual was above or below median with respect to the Big Five traits. In [88] he builds an approach called Realistic Accuracy Model (RAM), beginning form the assumption that personality traits are the real attributes of an individuals.

APR on Social Media

Social media are one of the main channels through which people interact with others, an ideal means for self-disclosure and, therefore, an excellent ground for research on personality computing. The approach proposed in [100] 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 training items (roughly 150) and the high number of features might have led to overfitting. In a similar way, the experiments presented in [9] 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.

[22] addresses the task of predicting personality impressions from vloggers based on what they say in their YouTube videos. First, they use manual transcripts of vlogs and verbal content analysis techniques to understand the ability of verbal content for the prediction of crowdsourced Big-Five personality impressions. Second, they explore the feasibility of a fully-automatic framework in which transcripts are obtained using automatic speech recognition (ASR). The results show that the analysis of error-free verbal content is useful to predict four of the Big-Five traits, three of them better than using nonverbal cues, and that the errors caused by the ASR system decrease the performance significantly. [21] addresses the problems of crowdsourcing the annotation of first impressions of video bloggers (vloggers) personal and social traits in conversational YouTube videos, and mining the impressions with the goal of modeling the interplay of different vlogger facets. First, they designed a human annotation task to crowdsource impressions of vloggers that extends a tradition of studies of personality impressions with the addition of attractiveness and mood impressions. Second, they proposed a probabilistic framework using Topic Models to discover prototypical impressions that are data driven, and that combine multiple facets of vloggers. Finally, they addressed the task of automatically predicting topic impressions using nonverbal and verbal content extracted from videos and comments. The study of 442 YouTube vlogs and 2210 annotations collected in Mechanical Turk supports the literature showing the feasibility to crowdsource interpersonal human impression with comparable quality to what is reported in social psychology research, and provides insights on the interplay among human first impressions. They also showed that topic models are useful to discover meaningful prototypical impressions that can be validated by humans, and that different topics can be predicted using different sources of information from vloggers' nonverbal and verbal content, as well as comments from the audience.

The study in [107] shows that personality traits, in particular Neuroticism and Openness, predict to a significant extent whether a person activates or not a blog. Other works aim at predicting the effect of personality on observable social media behavior [221, 224, 305]. The study in [221] applies Linguistic Inquiry Word Count to analyze the Tweets produced over one month by 142 Twitter users that have filled the Big Five Inventory. The results show that there is a significant correlation between the frequency of several LIWC categories and Big five traits. The only trait that does not show significant correlation with any of the LIWC categories is Conscientiousness. A similar approach was used by the same authors to predict the personality of Twitter users [99]. The privacy on Facebook is the focus of [224], where personality traits of 1323 users - collected with *myPersonality* [142] - are used as features to predict their respective Item Response Theory (IRT) scores, i.e. the psychometric measurements of their attitude towards the privacy problem. The investigation shows that personality traits explain in part the tendency to self-disclosure. In the case of [305], the personality traits of 652 subjects are used to predict the motivations behind the use of Youtube. All traits are correlated, to a statistically significant extent, which at least one of the dimensions adopted to represent the motivations behind the use of Internet. In [161] they analyze the relationship between recent photos uploaded by user's connection and the favorite actions. In [271] they study the friendship graph i.e. they study users interaction in Flickr, showing what fraction of its users interact with other, how they interact and how these interaction evolve in time. It has been shown also that age, gender, occupation, education level, and even personality can be predicted form people's Web site browsing logs [281].

APR on Facebook profiles was the subject of an international benchmarking campaign[1] the results of which appear in [45]. 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 [44, 197]. In the case of [44], 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 [197], 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. [143] they predict not only the personality traits form the *Likes* of Facebook but also psychodemographic information. They also dealt with a very important problem of the information we disclose through the Internet and social media unintentionally.

APR via Mobile and Wearable Devices

Mobile phone, as social media, have penetrated our everyday life as quickly and deeply as only a few other technologies did. While being conceived to exchange phone calls and SMS, standard mobile phones carry an increasing number of sensors and can be used as wearable devises to "measure" the life of individuals in naturalistic settings [225]. Such a phenomenon has attracted the attention of computing researchers trying to infer personality traits from data collected via mobile phones and wearable sensors. The experiments [204] use wearable devices to collect behavioral evidence such as speaking activity,

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

movement, proximity, face-to-face interactions, ad position in the social network resulting from mutual proximity to predict personality traits. The approach proposed in [256] focuses on the possibility of using mobile phones to extract social networks based, e.g. on who calls whom over certain period of time. The experiments were aimed at predicting whether each individual was above or below median with respect to the Big five traits. The effect of personality on the tendency to use or not mobile phone in certain contexts is the focus of[169]. The work considered 42 individuals and measures their attitude towards incoming class, when other are more or less close. [47] investigates the relationship between behavioral characteristics derived from rich smartphone data and self-reported personality traits. Their data stems from smartphones of a set of 83 individuals. From the analysis, they showed that aggregated features obtained from smartphone usage data can be indicators of the Big-Five personality traits. Additionally, they developed an automatic method to infer the personality type of a user based on cellphone usage using supervised learning. They show that their method performs significantly above chance and up to 75.9% accuracy.

APR and Computer Games

The literature proposed approaches aimed at inferring personality traits from strategies and options players adopt. The work in [303] analyzes the behavior of 1040 players in *World of Warcraft*, where the players are represented in terms of options, actions, strategies, number of days in activity, role played, number of competitors killed etc. Finally the experiments of [128] measure the correlation between the Big-Five Traits of 214 game players, Genre Preference and Player Experience of Need Satisfaction (PENS), a construct that includes five dimensions accounting for presence/immersion, relatedness, intuitive controls, competence and autonomy.

Personality Attribution and APP

Automatic Personality Perception is the task of inferring the personality observers attribute to a given individual from proximal cues.

The target of APP is not the true personality of individuals, but the personality these are attributed by others. Therefore APP adopts assessments made by others about the subjects under examination. The methodologies that work for APR are effective for APP as well, but current approaches are unable to use proximal cues and use distal cues as approximation instead. Like in the case of APR, APP works often include covariation studies aimed at identifying the cues most likely to result into high APP performance. APP approaches typically aim at predicting the average of the traits attributed by multiple raters. The individual assessments are the results of an actual attribution process. The average assessment does not result from an attribution process and, hence, cannot be considered a real personality. However, the prediction of the average remains an important task because it captures what is common across individually assigned traits and, furthermore, it can provide indications on the factors driving the attribution process.

APP approaches focus mainly on nonverbal behavior and social media. The number of subject tend to be lower in the case of APR because the collection of multiple assessments per subject limits the number of individuals that can be involved in the experiments. The first attempts of person perception appears in the fields of social psychology, cognitive psychology, social and cognitive science [123].

APP from Paralanguage

Psychologist have been observing that, at least in certain experiments conditions, “judgment made from speech alone rather consistently [have] the highest correlation with whole person judgments”, where speech means here not only what people say, but also paralanguage, i.e. everything accompanies words (prosody, vocalization, fillers) etc. Most of the works that deal with personality perception of the last decades, come from the domain of Social Signal Processing, specially for what concern the personality perception form speech [73, 186, 187, 219, 290]. Speech based on APP was the focus of a recent benchmarking campaign [249] that has lead to the first, rigorous comparison of different approaches over the same data and using the same experimental protocol.

APP and Nonverbal Behavior

The key idea is that nonverbal behavior can be considered as the physical, machine detectable evidence of social and psychological phenomena [287]. Nonverbal behavior includes a large number of cues that, on one hand, are likely to influence the attribution of personality traits and, on the other hand, have already been shown to allow the inference of socially and psychologically relevant information [289]. The experiments of [20] were performed over a corpus of 442 video blogs lasting 50 to 70 seconds. The nonverbal cues included speaking activity, prosody, gaze behavior, framing, motion and combination visual and audio cues. The goal was to predict the personality traits as per assessed by 5 observers. The experiments of [255] adopt 3907 clips extracted from movies where the characters are assessed in terms of the Big-Five traits. The features include not only nonverbal behavior, but also lexical features such as the popularity of the words used. In [199] they use videotaped interviews to show that people that are judged as attractive are perceived and described with positive qualities and attributes. In [295] he applies the Brunswick Lens (see Section III) to show that strangers, i.e. zero acquaintance people can, without efforts, perceive the real personality of people.

APP and Social Media

Unlike APR the problem of personality perception over social media has received only limited attention. The work in [76, 84] 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 [84], the focus is on 440 profiles pictures on the traits assigned by 736 unique observers. The pictures were represented in terms of content, body portion, facial expression, appearance and gaze and the experiments show that the agreement is higher when the picture subjects smile and do not wear hats. The other work [76] 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 whether they can account for the average impression a subject conveys. One of the most cited work is the one of Gosling et al. [103] where they predict personality traits from Facebook profiles and used other psychometric measures to find the *observer accuracy* and *inter-observer consensus* (the extent to which independent judges agree in a personality impression). In [76] they study the personality impression form Facebook, computing the *inter-observer agreement* and the accuracy (functional validity if we use the term used by [36, 246]).

Overall, the best performances tend to be observed for Extraversion and Conscientiousness. This is not surprising because personality psychologists have observed that these are the traits that human observers tend to perceive more clearly as well [133]. 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 [299]. This is the case, e.g., of the higher performances obtained for Openness in [76].

Psychometrics for Personality Computing

Psychometrics is the field of study concerned with the theory and technique of psychological measurement, which includes the measurement of knowledge, abilities, attitudes, personality traits, and educational measurement. The field is primarily concerned with the construction and validation of measurement instruments such as questionnaires, tests, and personality assessments. In the following we present the common metrics used in personality computing.

Pearson r correlation coefficient:

This is the number most people mean when they use the word “correlation” in conversations about the strength of associations. It is very sensitive to outliers, skewing, and nonlinearity (but not in any usefully systematic way). Therefore, Pearson’s *r* should only be used with well-behaved (preferably “jointly normal”) data that exhibit strictly linear associations. If *r* is -1 or 1, then the linear association is perfect (and either negative or positive, respectively). If *r* is (close to) 0 then any linear association that may be present is very weak, indeed.

Coefficient of determination r^2:

This coefficient is literally the percent of variation in responses explained by the variation in explanatory values, via the linear association between the two variables. Because r^2 is a percent, it is always between 0 and 1 (inclusive). It can be calculated by squaring Pearson’s *r*, hence the name “r^2”.

Spearman ρ correlation coefficient:

Spearman’s ρ [253] is Pearson’s *r* applied not to the data themselves, but to their ranks. In this way, Spearman’s ρ can be applied to nonlinear associations. Unfortunately, in such a case, it is not valid to interpret ρ^2 as a percentage of anything, in contrast to the interpretation of r^2. Spearman’s ρ tends to be a bit lower than Pearson’s r, which is a good thing in my opinion, especially when any amount of nonlinearity is present in the data.

Point-biserial correlation coefficient r_{pb} :

The point-biserial correlation coefficient [35, 156], is a statistic used to estimate the degree of relationship between a naturally occurring dichotomous nominal scale and an interval (or ratio) scale. In brief like the Pearson r, the r_{pb} can range from 0 to 1.00 if the two scales are related positively (that is, in the same direction) and from 0 to -1.00 if the two scales are related negatively (that is, in opposite directions).

Measure of association η and η^2:

A measure of association is a statistical quantity used to indicate the strength of the relationship between two variables. Measures of association take on values ranging from -1.0 to 1.0, with the positive and negative signs indicating the direction of the relationship, not the strength of the relationship. η [203] was invented specifically for the situation in which you have a nominal explanatory (column) variable and a numeric response (row) variable. η has the same kind of interpretation as Pearson’s r, but η does not assume the association is linear. (It can’t be, when one of the variables is not numeric). η is always between 0 and 1. η requires the counts to be “large enough” for its interpretation as the strength of an association to be reliable. It also performs better when the number of categories of the nomial variable is “large”. η close to 0 means no association; η close to 1 means any association there may be is strong. η^2 can be thought in the same way as the coefficient of determination.

Inter-rater Reliability:

In Statistics, inter-rater reliability, inter-observer reliability, inter-judge agreement are terms that refer to the extent of agreement among raters, judges, observers [110]. Inter-rater reliability data are usually presented as a table where the first column represents the subjects, with subsequent columns representing the raters and the different scores they assign to these subjects. With these data we can compute two types of analysis: we can compute an agreement coefficient in the form of a single number that expresses the extent of agreement among the raters; study the different factors that affect the model that describes several aspects pertaining to the rating process, like log-linear model.

When the assessment of the characteristics of subjects is not the result of their self-assessment but the assessment of other two or more observers, we need to check that the agreement among observers in determining the score or category membership of the subjects is as high as possible. In this case the issue is not the reliability assessment between observers (inter-rater reliability), i.e the their internal consistency as group of assessors measured by the Cronbach's α coefficient [56], but rather if they all say the same thing, i.e their agreement (inter-rater agreement). Thus, inter-rater reliability is defined as the "*degree to which the ratings of different judges are proportional when expressed as deviations from their means*" [265]. While inter-rater agreement is defined as the "*extent to which the different judges tend to make exactly the same judgments about the rated subject*"[265].

When the absolute value of the ratings matters we can use Krippendorff's α [144]. Krippendorff's α is a reliability coefficient developed to measure the agreement among observers, coders, judges, raters, or measuring instruments drawing distinctions among typically unstructured phenomena or assign computable values to them. It is applicable to any number of coders, each assigning one value to one unit of analysis, to incomplete (missing) data, to any number of values available for coding a variable, to binary, nominal, ordinal, interval, ratio, polar, and circular metrics, and it adjusts itself to small sample sizes of the reliability data. Krippendorff's α defines a large family of reliability coefficients and embraces several known ones (as Scott's pi, Cohen kappa, Fleiss' k). α also allows for reliability data with missing category or scale points. α' s general form is: $\alpha = 1 - \frac{D_0}{D_e}$ where D_o is the observed disagreement among values assigned to units of analysis and D_e is the disagreement one would expect when coding of units is attributable to chance rather than to the properties of these units. When raters agree perfectly, observed disagreement $D_o = 0$ and $\alpha = 1$ which indicate perfect reliability. When raters agree as if chance had produced the results $D_o = D_e$ and $\alpha = 0$, which indicates the absence of reliability, units and the values assigned to them are statistically unrelated. $\alpha < 0$ when disagreements are systematic and exceed what can be expected by chance.

Social scientists commonly rely on data with reliabilities $\alpha \geq 0.800$, consider data with $0.800 > \alpha \geq 0.667$ only to draw tentative conclusions, and discard data whose agreement measures $\alpha < 0.667$. Inasmuch as mathematical statements of the statistical distribution of α are always only approximations, it is preferable to obtain α's distribution by bootstrapping [145]. α's distribution gives rise to two indices: the confidence intervals of a computed alpha at various levels of statistical significance and the probability that α could be below a chosen minimum, required for data to be considered sufficiently reliable (one-tailed test).

9

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

9.1 Introduction

Is a picture worth a thousand words? It seems to be so when it comes to mobile technologies and social networking platforms.

Pictures streams are "*often seen as a substitute for more direct forms of interaction like email*" [276] and interacting with connected individuals appears to be one of the main motivations behind the use of online photo-sharing platforms [278]. The pervasive use of *liking* mechanisms, further confirms the adoption of pictures as a social glue [259]: on Flickr, *likes* between connected users are roughly 10^5 times more frequent than those between non-connected ones [161].

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 [143]. Furthermore, online expressions of aesthetic preferences convey an impression in terms of characteristics like prestige, differentiation or authenticity [164]. For this reason, this chapter 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 chapter proposes an approach aimed at mapping the pictures people like into personality traits adressing APR and APP problems.

The reason is that self-assessed and attributed traits tend to relate differently to different aspects of an individual [299] and, therefore, both 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 while they have a major impact on attributed ones. 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 [285] for an extensive survey). This is an important advantage because it allows one to assess the effectiveness of a given type of behavioral evidence (the favorite pictures in this case) in conveying information about personality.

Regression analysis appears to be the most suitable computational framework for the problem. This applies in particular to Multiple Instance Regression (MIR) [8, 230] 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 (see part II). This work extends such a principle to the inference of personality traits and addresses the problem with a multiple instance strategy. In particular, we 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 results 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 [299]. 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 [268]. 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 recognize that identification is often most consequential as the categorization of others, rather than as self-identification*" [125]. For this reason, the literature proposes approaches aimed at predicting both self-assessed and attributed traits [285] and this work addresses both problems.

9.2 PsychoFlickr: Pictures and Personality

The experiments have been performed over *PsychoFlickr*, a corpus of 60000 pictures tagged as favorite by 300 *Pro* Flickr users[1] (200 randomly selected favorite pictures per user).

9.2.1 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)[2]. 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).

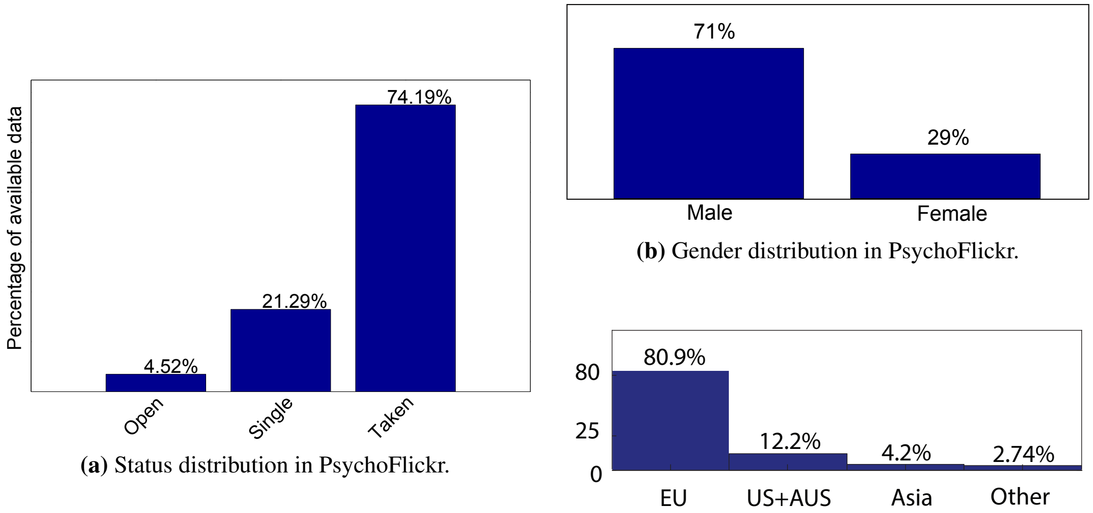
The personal data of each subject has been downloaded from the Flickr profile and, in some cases, there is missing information (see Table 9.1 for the details). The histograms of all kinds of personal information contained in PsychoFlickr, are listed in 9.1. We have *Status* (*open, single, taken*) Figure9.1a, *Gender* Figure9.1b, *Country* (Nationality) Figure9.1c, *#photos* Figure9.2a, *#faved* Figure9.2b, *#galleries* Figure9.2c, *#contacts* Figure9.2d, *#groups* Figure9.2e, *Age* (i.e., the birth date) Figure9.2f.

Info	#
Gender	300
Status	155
Age	44
Country	288
#Photos	300
#Faved	300
#Galleries	300
#Contacts	300
#Groups	300

Table 9.1: Personal information acquired for the 300 users. On the right, the number of users whose corresponding information (on the left) was available.

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

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



(a) Status distribution in PsychoFlickr.

(b) Gender distribution in PsychoFlickr.

(c) Country distribution in PsychoFlickr.

Fig. 9.1: Dataset distribution of a) status, b) gender and c) country

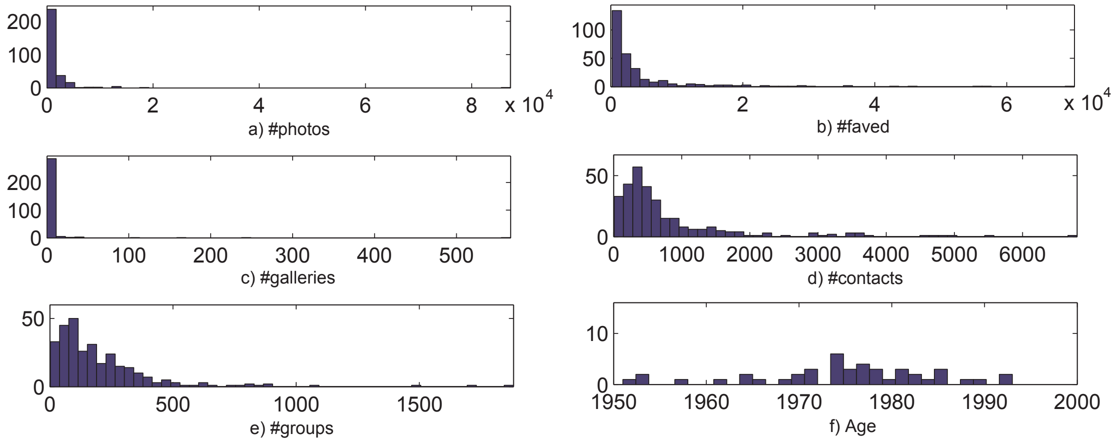


Fig. 9.2: Personal information distributions in PsychoFlickr.

9.2.2 Personality and its Measurement

As personality models of PsychoFlickr we adopt the Big-Five (BF) Traits [245]. The reason behind this choice is twofold: on the one hand the BF is the model most commonly applied in both personality computing [285] and personality science [299]. On the other hand, the BF model represents personality in terms of five numerical scores, a form particularly suitable for computer processing.

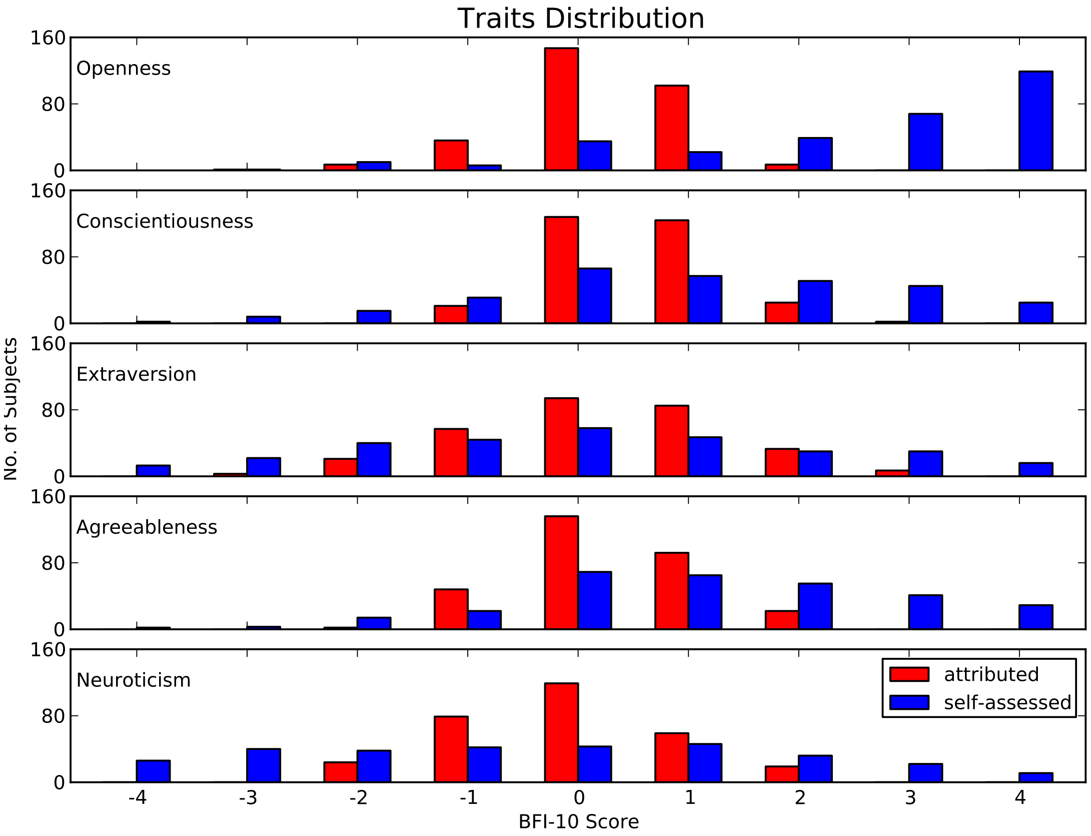


Fig. 9.3: Distributions of the Big Five Inventory test results: self-assessed (top); assessed (bottom); in this last case, the distributions have more bins, since their values are result of the averaging on the 12 assessors' tests (becoming real value, and not integers).

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. In Figure 9.3 are reported the distributions of the Big Five Inventory test results where the left chart of figure 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) [31]. In the case of PsychoFlickr, this applies in particular 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. 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 right chart of Figure 9.3 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*".

The agreement among the assessors was measured with the Krippendorff's α [144], a reliability coefficient suitable for a wide variety of assessments (binary, nominal, ordinal, interval etc.), and robust to small sample sizes. The value of α is 0.06 for Openness, 0.12 for Conscientiousness, 0.26 for Extraversion, 0.17 for Agreeableness and 0.22 for Neuroticism. The values are statistically significant and comparable to those observed in the literature for zero acquaintance scenarios [23], i.e. situations where assessors and subjects being rated do not have any personal contact (like in the *PsychoFlickr* corpus). Furthermore, the results of the APP experiments confirm that the judgments are sufficiently reliable to allow statistically significant performances.

Having both self-assessed and attributed assessements, according to the terminology introduced in [285], it is possible to perform both Automatic Personality Recognition (APR) and Automatic Personality Perception (APP).

Here, we report 400 random images of people whose personality trait scores (assessed and self-assessed) are the highest and the lowest, for all the five traits, that is: Openness Figure 9.4 (assessed) and Figure 9.5 (self-assessed); Consciousness Figure 9.6 (assessed) and Figure 9.7 (self-assessed); Extraversion Figure 9.8 (assessed) and Figure 9.9 (self-assessed); Agreeableness Figure 9.10 (assessed) and Figure 9.11 (self-assessed); Neuroticism Figure 9.12 (assessed) and Figure 9.13 (self-assessed). The purpose of these pictures is twofold: on one side, one can check *qualitatively* the relationship among the features and the personality traits. On the other side, one can individuate particular characteristics which are not captured by our features, and that should be modeled for a better understanding and modeling of the trait.

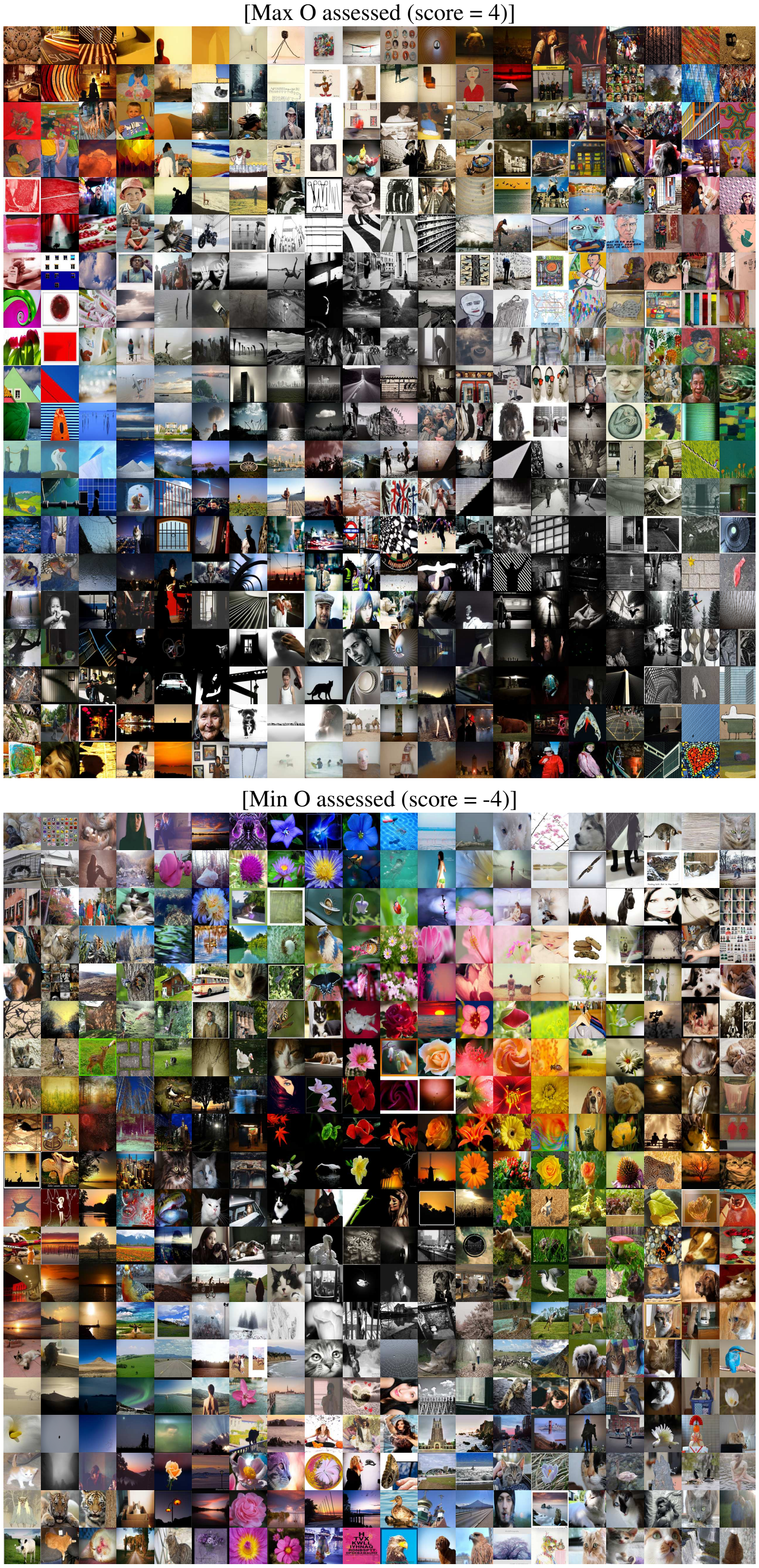
9.3 Feature Extraction

The goal of this work is to map the aesthetic preferences of Flickr users into personality traits. Therefore, the features adopted in this work focus on the contributions of *Computational Aesthetics* (see Section 3.1) and do not take into account feature extraction techniques like SIFT or HOG that, while being widely applied in the computer vision community, do not account for aesthetic preferences. A synopsis of the features adopted in this work is available in Table 3.1, where we do not take into account features that cannot be converted to counts; with this assumption 82 features are retained over all the 112 of Table 3.1, here reported in Table 9.2 for convenience. The number of faces are ubiquitous in the images and, furthermore, the human brain is tuned to their detection in the environment [129]. For this reason this feature has been annotated manually.

9.4 Inference of Personality Traits

This section presents the regression approaches adopted to map the features into personality traits. The regression is performed separately for the Big-Five traits because these result from the application of Factor Analysis to behavioral data [52] and, therefore, they are independent.

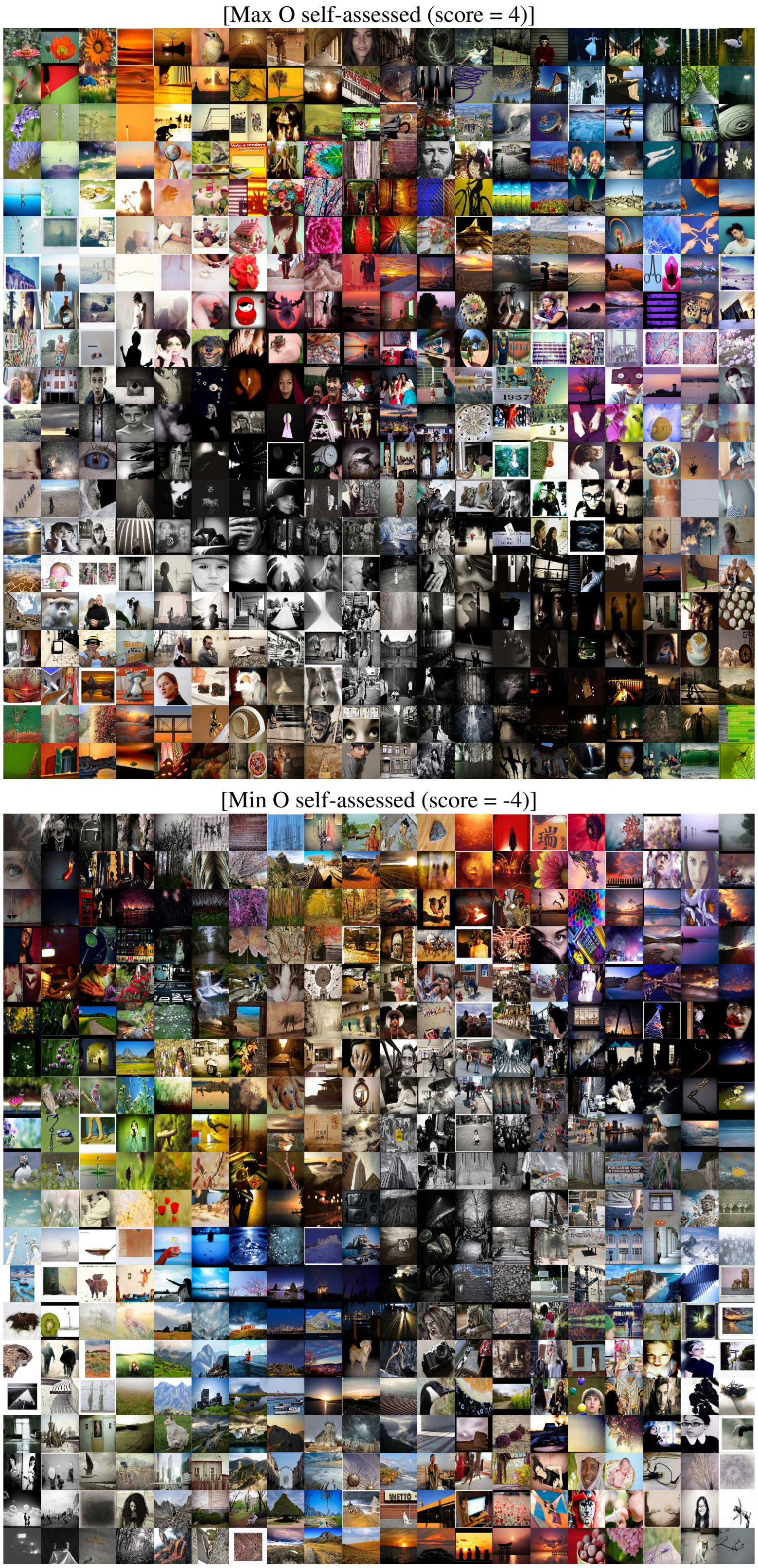
Being the goal of the experiments the inference of the personality traits - both self-assessed and attributed - from the multiple pictures that a user tags as favorite, Multiple Instance Regression (MIR) [8, 230] appears to be the most suitable computational framework because it addresses problems where 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. This can be useful in case only part of the pictures tagged as favorite are actually predictive of the traits of an individual.



[Max O assessed (score = 4)]

[Min O assessed (score = -4)]

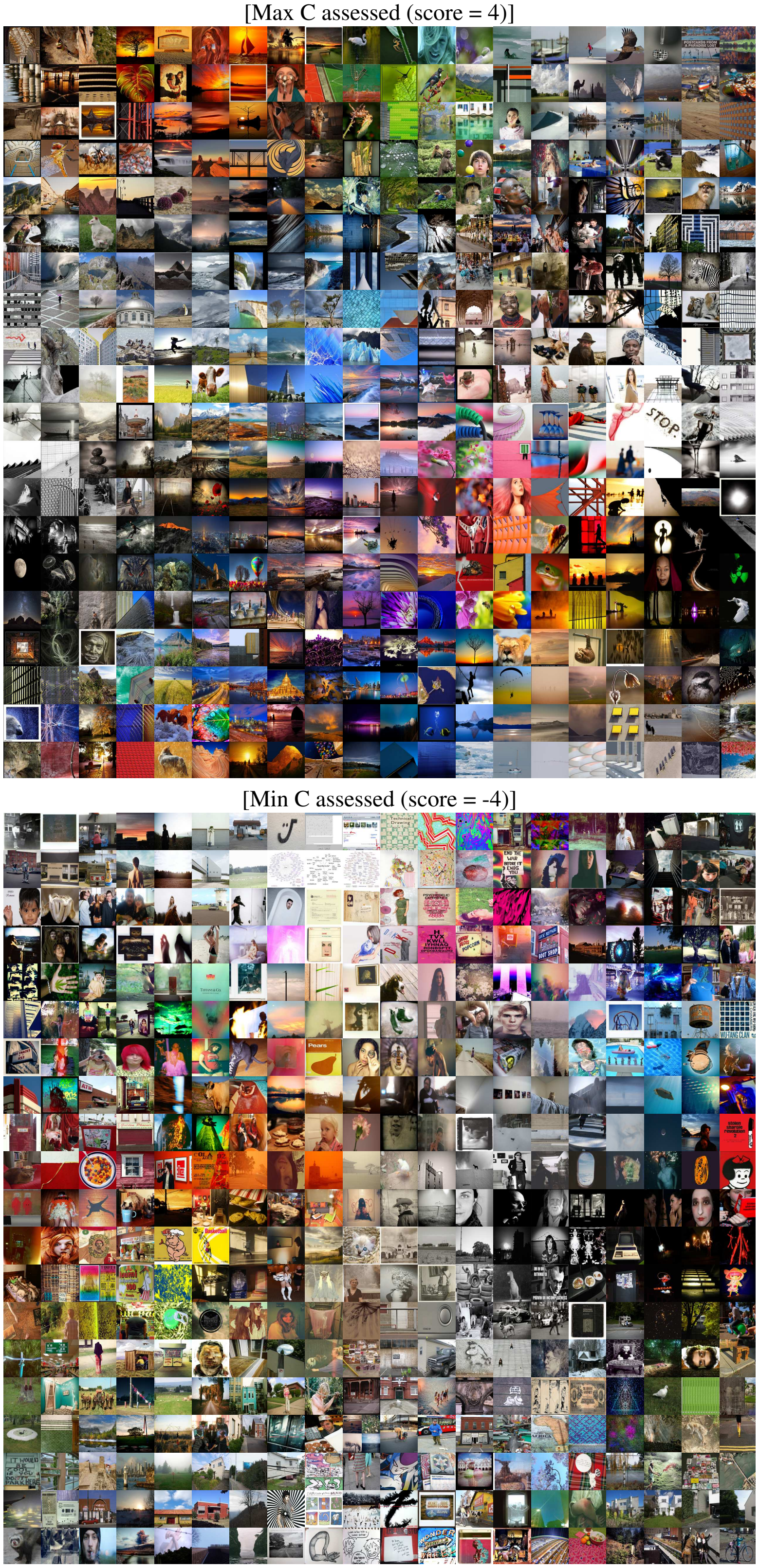
Fig. 9.4: A random sampling of images of users related to the assessed Openness trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max O self-assessed (score = 4)]

[Min O self-assessed (score = -4)]

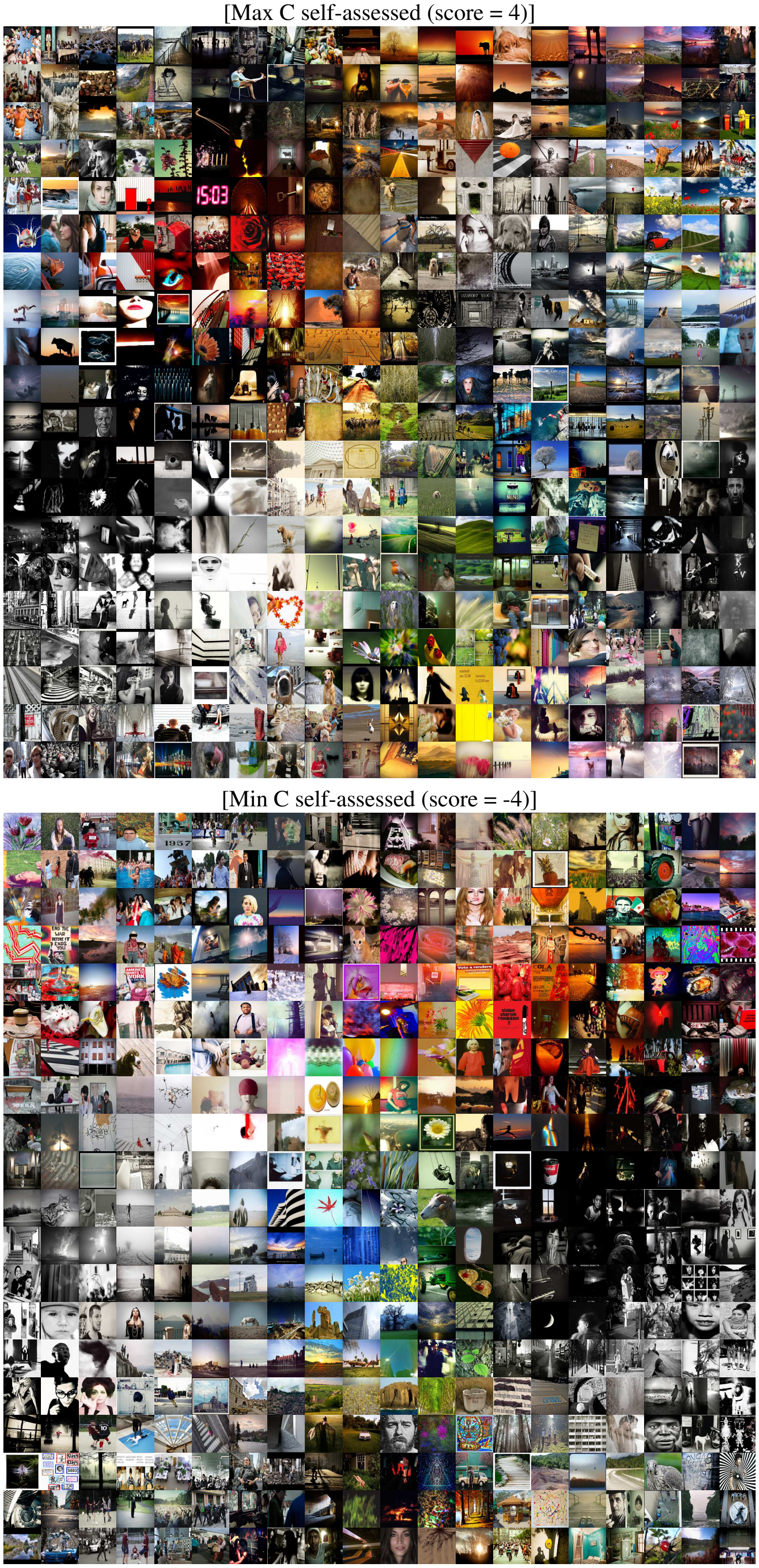
Fig. 9.5: A random sampling of images of users related to the self-assessed Openness trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max C assessed (score = 4)]

[Min C assessed (score = -4)]

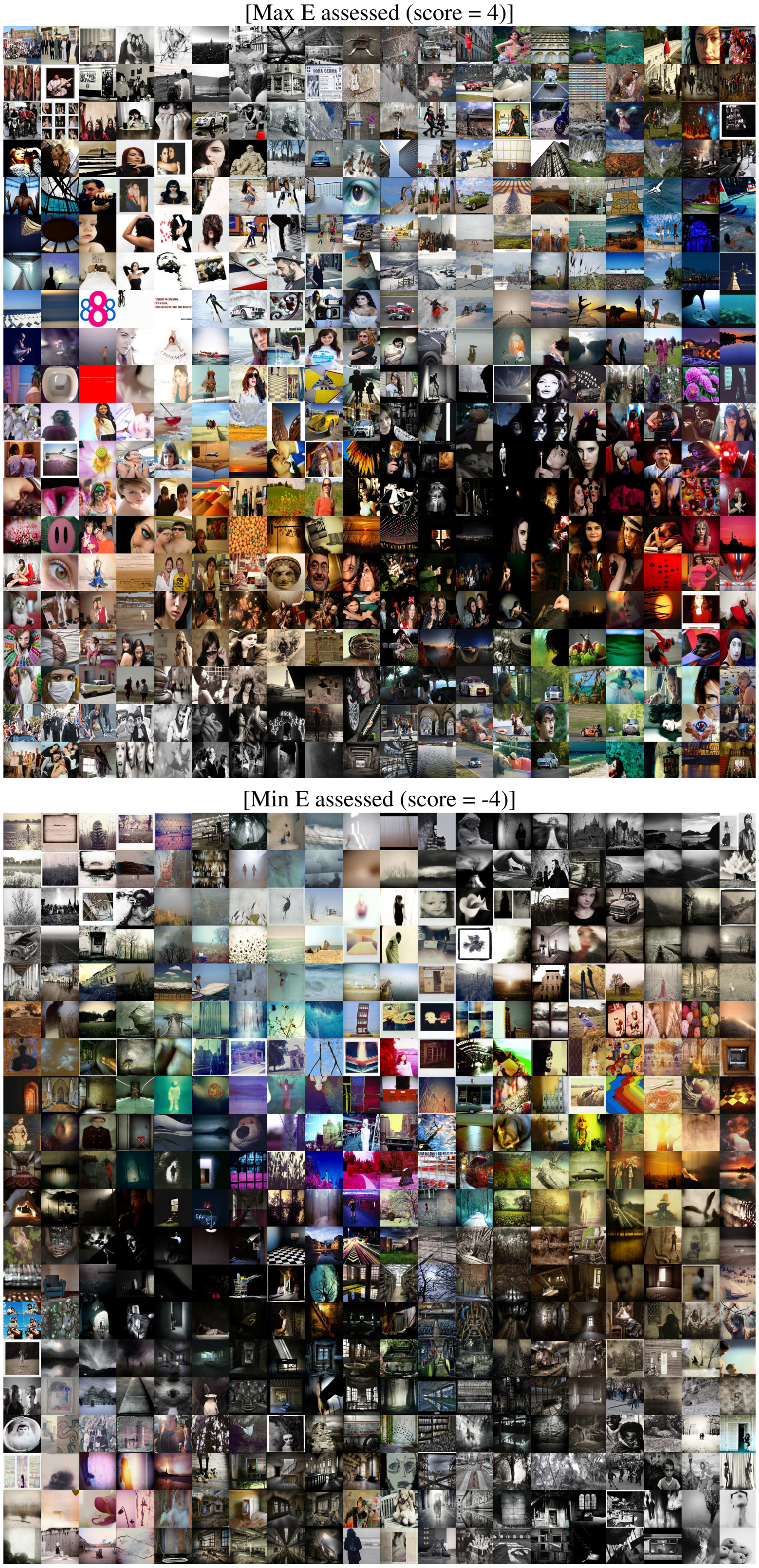
Fig. 9.6: A random sampling of images of users related to the assessed Conscientiousness trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max C self-assessed (score = 4)]

[Min C self-assessed (score = -4)]

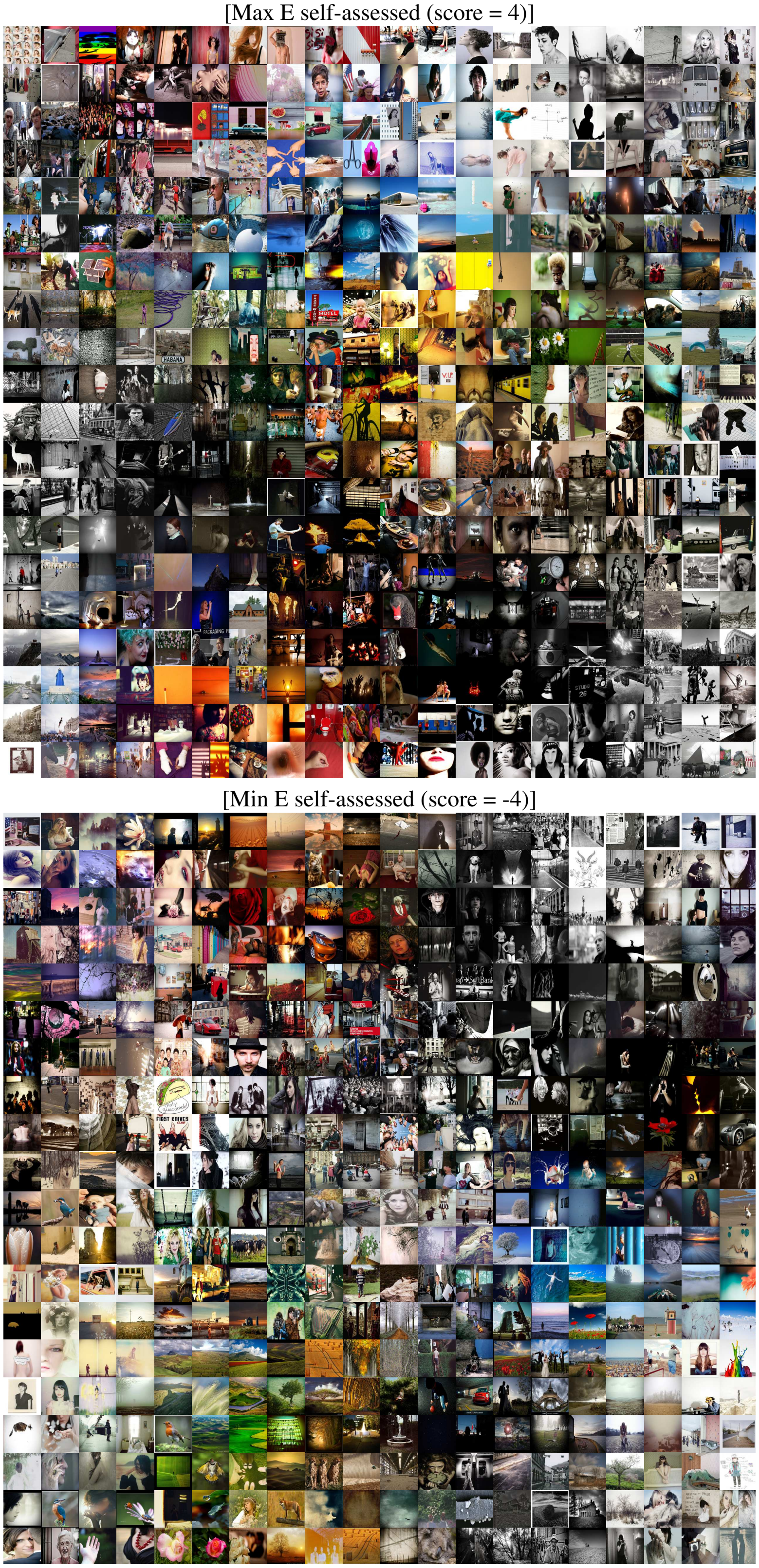
Fig. 9.7: A random sampling of images of users related to the self-assessed Conscientiousness trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max E assessed (score = 4)]

[Min E assessed (score = -4)]

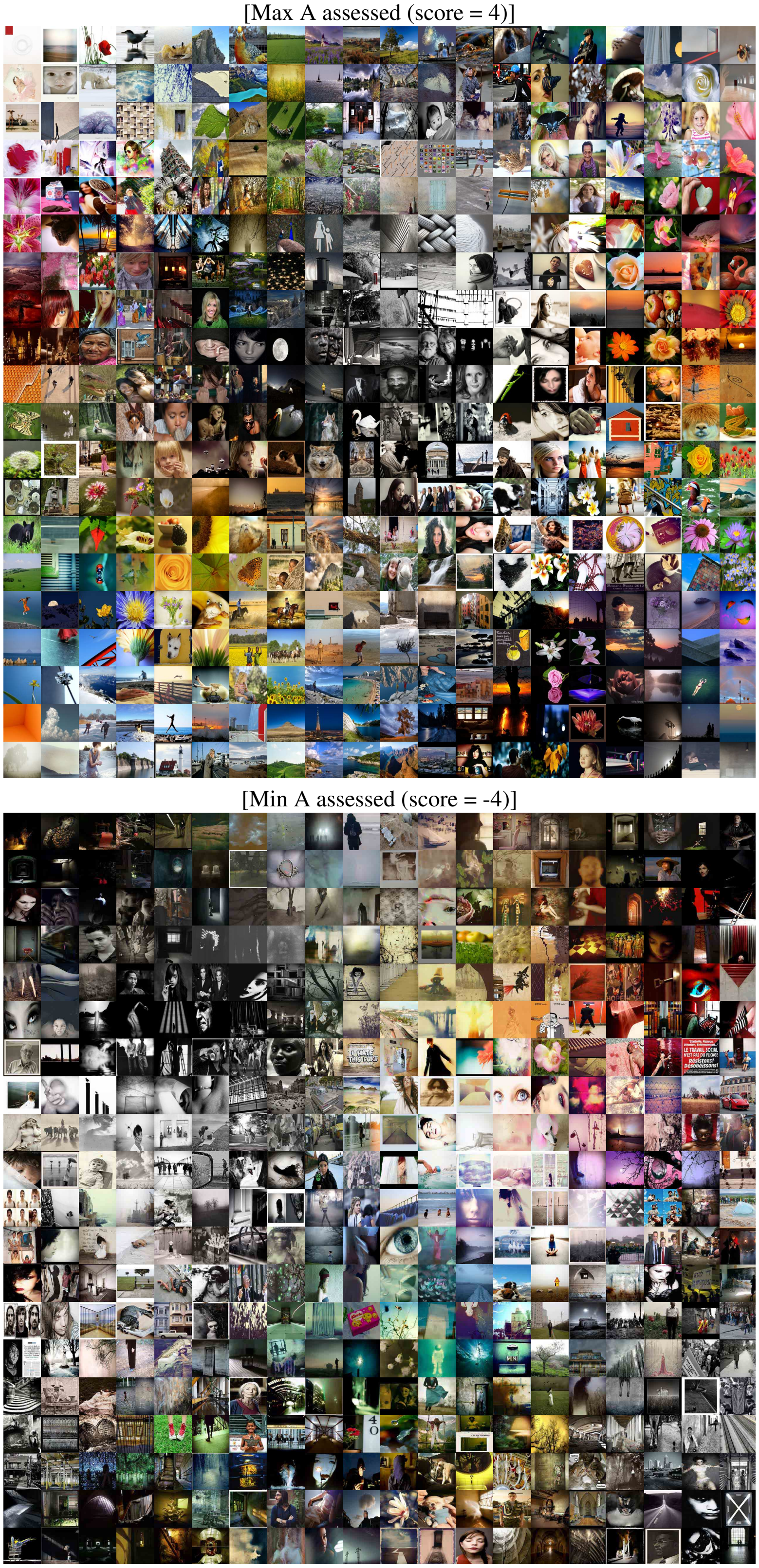
Fig. 9.8: A random sampling of images of users related to the assessed Extraversion trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max E self-assessed (score = 4)]

[Min E self-assessed (score = -4)]

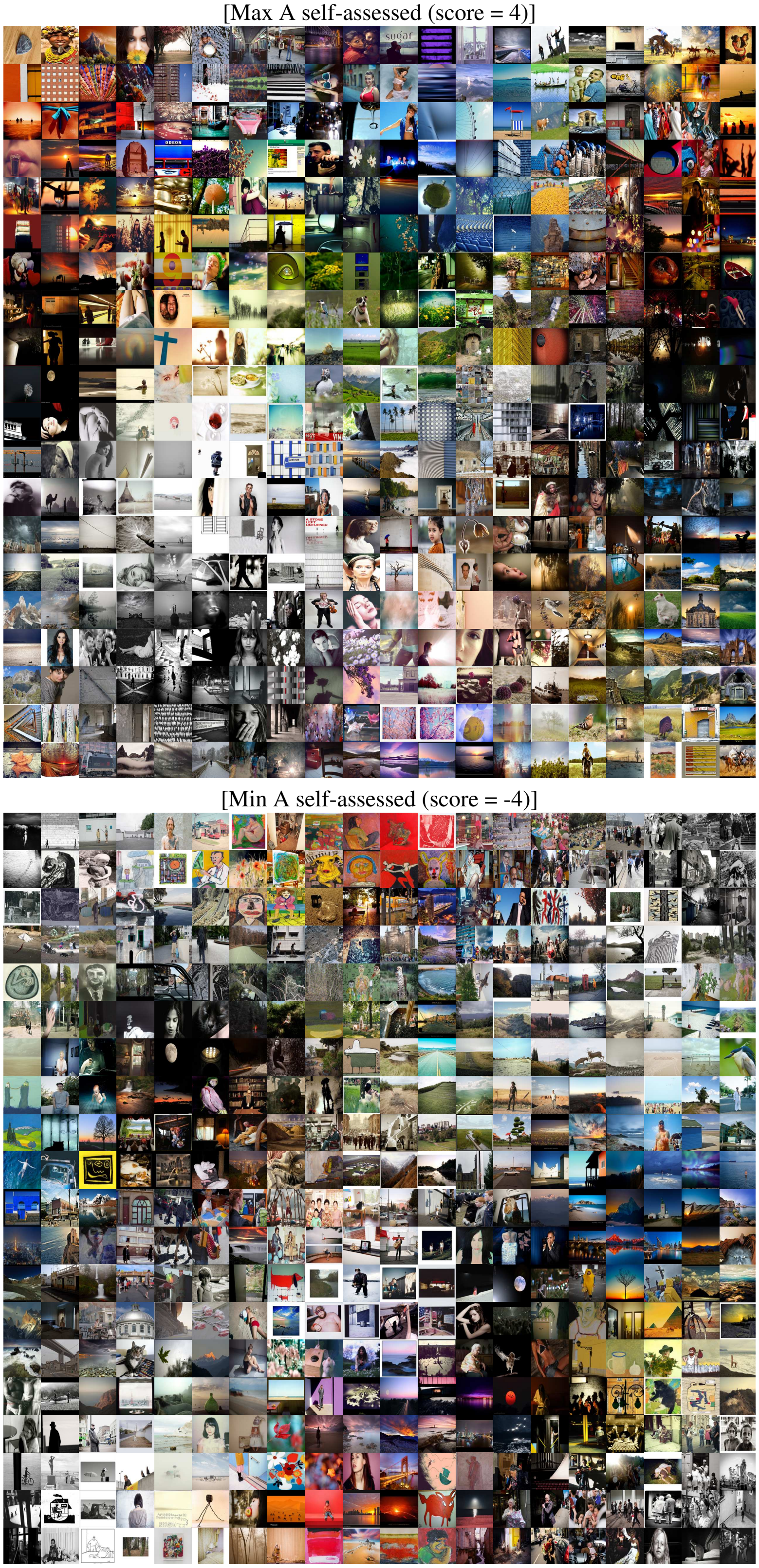
Fig. 9.9: A random sampling of images of users related to the self-assessed Extraversion trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max A assessed (score = 4)]

[Min A assessed (score = -4)]

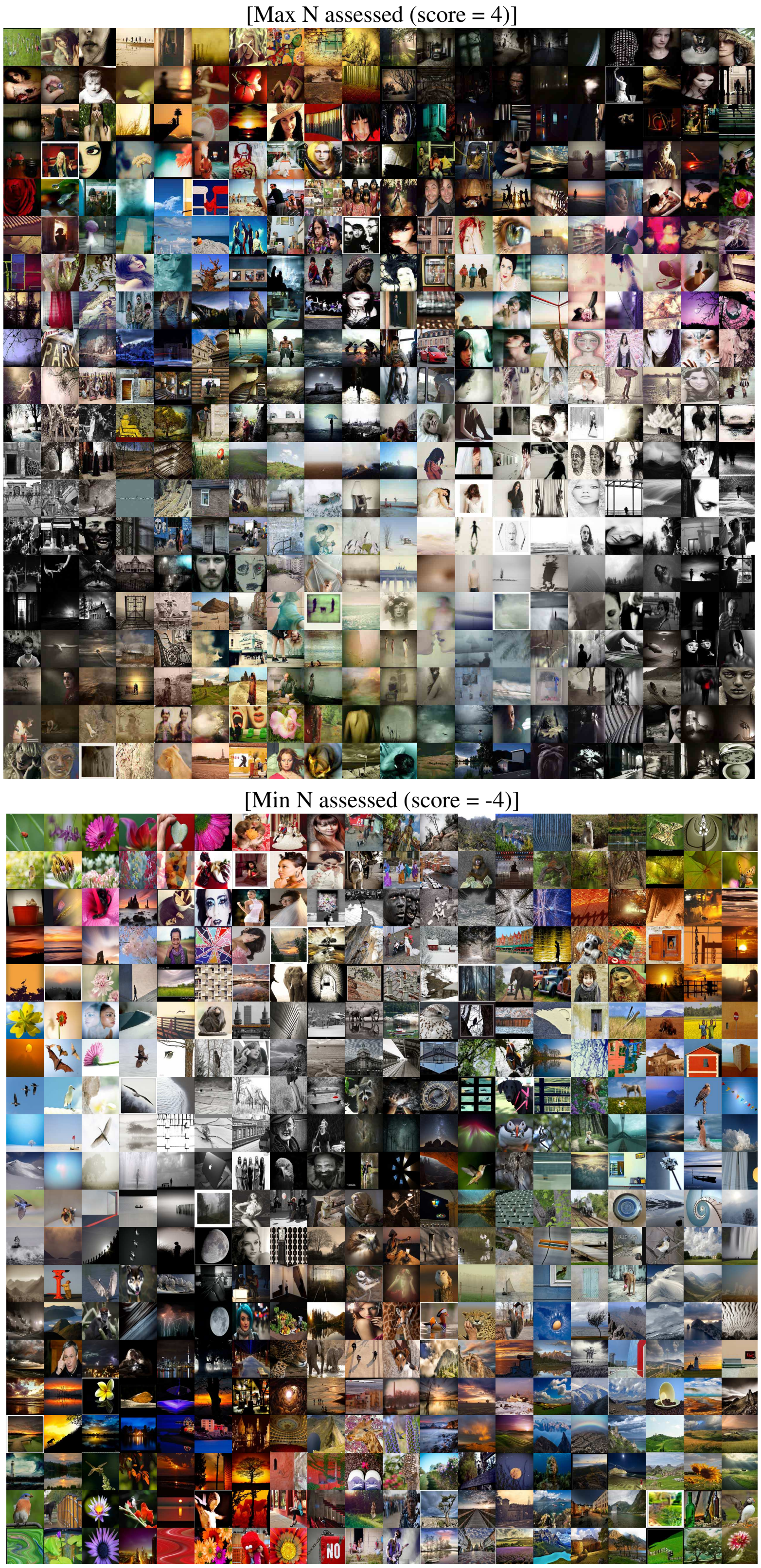
Fig. 9.10: A random sampling of images of users related to the assessed Agreeableness trait. The collage has been produced ensuring a pleasant smoothing between images.



[Max A self-assessed (score = 4)]

[Min A self-assessed (score = -4)]

Fig. 9.11: A random sampling of images of users related to the self-assessed Agreeableness trait. The collage has been produced ensuring a pleasant smoothing between images.


[Max N assessed (score = 4)]

[Min N assessed (score = -4)]

Fig. 9.12: A random sampling of images of users related to the assessed Neuroticism trait. The collage has been produced ensuring a pleasant smoothing between images.

[Max N self-assessed (score = 4)]

[Min N self-assessed (score = -4)]

Fig. 9.13: A random sampling of images of users related to the self-assessed Neuroticism trait. The collage has been produced ensuring a pleasant smoothing between images.

Category	**Name**	**d**	**Short Description**
Color	HSV statistics	5	Average of S channel and standard deviation of S, V channels [174]; *circular variance* in HSV color space [180]; *use of light* as the average pixel intensity of V channel [62]
	Emotion-based	3	Measurement of *valence*, *arousal*, *dominance* [174, 272]
	Color diversity	1	Distance w.r.t a uniform color histogram, by Earth Mover's Distance (EMD) [62, 174]
	Color name	11	Amount of *black*, *blue*, *brown*, *green*, *gray*, *orange*, *pink*, *purple*, *red*, *white*, *yellow* [174]
Composition	Edge pixels	1	Total number of edge points, extracted with Canny detector [167]
	Level of detail	1	Number of regions (after mean shift segmentation) [50, 94]
	Average region size	1	Average *size* of the regions (after mean shift segmentation) [94]
	Low depth of field (DOF)	3	Amount of focus sharpness in the inner part of the image w.r.t. the overall focus [62, 174]
	Rule of thirds	2	Average of S,V channels over inner rectangle [62, 174]
	Image size	1	Size of the image [62, 167]
Textural Properties	Gray distribution entropy	1	Image entropy [167]
	Wavelet based textures	12	Level of spatial graininess measured with a three-level (L1,L2,L3) Daubechies wavelet transform on HSV channels [62]
	Tamura	3	Amount of *coarseness*, *contrast*, *directionality* [261]
	GLCM - features	12	Amount of *contrast*, *correlation*, *energy*, *homogeneousness* for each HSV channel [174]
	GIST descriptors	24	Output of GIST filters for scene recognition [206].
Faces	Faces	1	Number of faces (extracted manually)

Table 9.2: 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 pictures content.

Before applying the MIR approaches developed for this work, the features are discretized using $Q = 6$ quantization levels (Q values between 3 and 9 were tested, but no significant performance differences were observed). 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 can be interpreted as counts.

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, $\upsilon(t)$ is a function that takes as input a picture index t and returns the corresponding user-index u.

9.4.1 Baseline Approaches

The general MIR formulation is NP-hard [230] 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 $\upsilon(t) = u$ - 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.

Naive-MIR [8]

The simplest baseline 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 $\upsilon(t) = u$. 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 [69]

Given a test bag B^u, this methodology adopts the minimal Hausdorff distance [294] 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 [292]

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 ($i = 1, \dots, C$) - each obtained by training the model in Section 9.4.3 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 $\hat{y}_p^u$. 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.

9.4.2 Latent representation-based methods

The baseline approaches presented in Section 9.4.1 operate in the feature space where the pictures are represented. Such an approach 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. For this reason, this section proposes new MIR approaches that map the pictures of the training bags onto an intermediate latent space $\mathcal{Z}$ expected to capture most of the information necessary to perform the prediction of the trait scores. 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 learned 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 9.4). The topic model adopted in this work is the Latent Dirichlet Annotation (LDA) [27]. The LDA expresses the topics as probability distributions of features $p(\mathcal{C}_z^u|k)$, with $k = 1, \dots, K$ and $K << D$ (D is the dimension of the feature vectors). Once the topics are learned from the training bags, a test bag B^u can be expressed as a mixture of topics:

$$p(B^u) = \sum_{k=1}^{K} p(\mathcal{C}_z^u|k)p(k|u), \tag{9.1}$$

where $\mathcal{C}_z^u$ 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 9.4.3 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 $\hat{y}_p^u$.

Gen-LDA

This approach learns a LDA model [27] 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 9.4.3. 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 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 $c = 1, \dots, C$:

$$Z^u(c) = \sum_{t:\upsilon(t)=u} p(c|t) \tag{9.2}$$

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* [292], 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 $p(c|t)$.

Counting Grid

This approach is based on the *Counting Grid* (CG) [213]. After the 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 9.4.3 to predict the trait scores.

9.4.3 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:

$$\hat{y}_p^u = \sum_{k=1}^{K} \beta_k x_k^u, \tag{9.3}$$

where the $\beta = (\beta_1, \dots, \beta_K)$ are the regressor parameters and the values x_k^u 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). 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.

9.5 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.

9.5.1 Correlational Analysis between Pictures and Personality Traits

Table 9.3 shows the absolute covariation, measured with the Spearman Coefficient ρ, of features and traits, both self-assessed and attributed (color of the cell account for the absolute value of ρ). The covariation with the features 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 influence 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.

Category	z	Feature	Ope(S)	Ope(A)	Con(S)	Con(A)	Ext(S)	Ext(A)	Agr(S)	Agr(A)	Neu(S)	Neu(A)
Color	1	use of light	-	-0.13	-	-	-	-	-	0.17	-	-0.26
	2	avg saturation	-	-0.20	-	0.21	-	-	-	0.40	-	-0.55
	3	std saturation	-	-0.22	-	0.18	-	0.15	-	0.38	-	-0.52
	4	std brightness	-	-	-	-	-	0.35	-	-	-	0.15
	5	valence	-	-0.18	-	-	-	-	-	0.27	-	-0.40
	6	dominance	-	-	0.15	0.29	-	-	-	-	-	-
	7	arousal	-	-0.18	0.12	0.28	-	-	-	0.38	-	-0.52
	8	hue circular variance	-	-0.19	-	-	-	-	-	0.20	-	-0.36
	9	colorfulness	0.15	0.28	-	-	-	-0.13	-	-0.21	-	0.36
	10	black	-	0.26	-	-	-	-	-	-0.22	-	0.34
	11	blue	-0.12	-0.17	-	0.17	-	-	-	0.36	-	-0.52
	12	brown	-	-0.24	-	-	-	-	-	0.24	-	-0.34
	13	gray	-	-	-	-0.14	-	-	-	-0.29	-	0.34
	14	green	-	-0.19	-	-	-	-0.18	-	0.16	-	-0.28
	15	orange	-	-0.18	-	0.31	-	-	-	0.45	-	-0.56
	16	pink	-	-0.13	-	-	-	-	-	0.19	-	-0.26
	17	purple	-	-0.15	-	-	-	-	0.12	0.22	-0.14	-0.27
	18	red	-	-	-	0.18	-	0.18	-	0.30	-	-0.40
	19	white	-	0.27	-0.16	-0.31	-	-	-	-0.24	-	0.36
	20	yellow	-	-	-	0.12	-	-0.12	0.14	0.29	-	-0.38
Composition	21	pers. perc. of edges pixels	-	-	-	-	-0.12	-0.32	-	-	-	-
	22	level of detail	-	-0.30	-	0.19	-	0.12	-	0.33	-	-0.44
	23	avg region size	0.14	0.32	-	-	-	-0.27	-	-0.14	-	0.14
	24	low DOF - hue	-	-0.23	-	-	-	-	-	0.30	-	-0.43
	25	low DOF - saturation	-	-0.21	-	-	-	-	-	0.27	-	-0.42
	26	low DOF - brightness	-	-	-	0.16	-	-	-	-	-	-
	27	rule of thirds - saturation	-	-0.22	-	0.22	-	-	-	0.41	-	-0.56
	28	rule of thirds - brightness	-	-	-	-	-	-	-	0.21	-	-0.29
	29	image size	-	-	-	-	-	-	-	-	-	-
Textural Properties	30	gray distribution entropy	-	-0.27	-	-0.17	-	0.15	-	-	-	-
	31	hue wavelet - lev 1	-	-0.21	-	-	-	0.23	-	0.12	-	-0.26
	32	hue wavelet - lev 2	-	-0.22	-	-0.12	-	0.23	-	0.12	-	-0.25
	33	hue wavelet - lev 3	-	-0.20	-	-0.12	-	0.26	-	0.15	-	-0.25
	34	saturation wavelet - lev 1	-	-0.19	-	-	-	0.15	-	0.19	-	-0.35
	35	saturation wavelet - lev 2	-	-0.21	-	-	-	0.19	-	0.23	-	-0.37
	36	saturation wavelet - lev 3	-	-0.21	-	-	-	0.23	-	0.29	-	-0.41
	37	brightness wavelet - lev 1	-	-0.11	-	-	-	-	-	-	-	-
	38	brightness wavelet - lev 2	-	-0.13	-	-	-	0.17	-	-	-	-
	39	brightness wavelet - lev 3	-	-0.11	-	-	-	0.30	-	-	-	-
	40	hue wavelet avg	-	-0.21	-	-0.12	-	0.25	-	0.13	-	-0.25
	41	saturation wavelet avg	-	-0.21	-	-	-	0.21	-	0.26	-	-0.39
	42	brightness wavelet avg	-	-0.12	-	-	-	0.23	-	-	-	-
	43	Tamura coarseness	-	-	-	-0.17	-	-	-	-	-	-
	44	Tamura contrast	-	0.20	-	-	-	0.25	-	-0.12	-	0.34
	45	Tamura directionality	-	-	-	0.23	-	-0.33	-	-	-	-0.23
	46	GLCM contrast - hue	-	-0.18	-	-	-	0.26	-	-	-0.12	-0.22
	47	GLCM correlation - hue	-	-0.21	-	-	-	-	-	0.26	-	-0.40
	48	GLCM energy - hue	-	0.25	-	0.13	-	-	-	-0.14	-	0.29
	49	GLCM homogeneity - hue	-	0.21	-	0.13	-	-0.15	-	-	-	0.22
	50	GLCM contrast - saturation	-	-0.12	-	0.15	-	0.20	-	0.16	-	-0.29
	51	GLCM correlation - saturation	-	-0.22	-	-	-	-	-	0.29	-	-0.44
	52	GLCM energy - saturation	-	0.28	-	-	-	-	-	-0.28	-	0.45
	53	GLCM homogeneity - saturation	-	0.26	-	-	-	-0.13	-	-0.18	-	0.33
	54	GLCM contrast - brightness	-	-	-	0.15	-0.13	-	-	-	-	-
	55	GLCM correlation - brightness	-	-	-	-0.14	0.14	-	-	-	-	0.19
	56	GLCM energy - brightness	-	0.35	-	-	-	-0.18	-	-0.15	-	0.21
	57	GLCM homogeneity - brightness	-	0.17	-	-	-	-	-	-	-	-
	58	GIST - channel 1	-	-	-	-0.13	-	0.14	-0.12	-0.29	-	0.27
	59	GIST - channel 2	-	-	-	-	-	-	-0.12	-0.17	-	-
	60	GIST - channel 3	-	-	-	-0.12	-	0.15	-0.14	-0.26	-	0.26
	61	GIST - channel 4	-	-	-	-	-	-	-0.14	-0.15	-	-
	62	GIST - channel 5	-	-	-	-	-	-	-	-0.23	-	0.19
	63	GIST - channel 6	-	-	-	-	-	-	-	-0.13	-	-
	64	GIST - channel 7	-	-	-	-	-	0.12	-0.12	-0.22	-	0.20
	65	GIST - channel 8	-	-	-	-	-	-	-0.12	-	-	-
	66	GIST - channel 9	-	-	-	-	-	-	-	-0.22	-	0.14
	67	GIST - channel 10	-	-	-	0.14	-0.12	-	-	-	-	-
	68	GIST - channel 11	-	-	-	-	-	-	-	-0.19	-	0.13
	69	GIST - channel 12	-	-	-	0.16	-0.12	-	-	-	-	-
	70	GIST - channel 13	-	-	-	-	-	-	-	-0.25	-	0.22
	71	GIST - channel 14	-	-	-	-	-	-	-	-0.12	-	-
	72	GIST - channel 15	-	-	-	-	-	-	-0.13	-0.22	-	0.19
	73	GIST - channel 16	-	-	-	-	-	-	-	-0.11	-	-
	74	GIST - channel 17	-	-	-	-	-	0.25	-0.12	-0.23	-	0.24
	75	GIST - channel 18	-	-	-	-	-	0.14	-	-	-	-
	76	GIST - channel 19	-	-	-	-	-	0.26	-0.13	-0.21	-	0.24
	77	GIST - channel 20	-	-	-	-	-	0.14	-0.12	-	-	-
	78	GIST - channel 21	-	-	-	-	-	0.22	-	-0.20	-	0.17
	79	GIST - channel 22	-	-0.13	-	-	-	-	-	-	-	-
	80	GIST - channel 23	-	-	-	-	-	0.23	-0.11	-0.17	-	0.17
	81	GIST - channel 24	-	-0.13	-	0.14	-	-	-	-	-	-
	82	number of faces	-	-	-	-0.20	0.13	0.53	-	-0.17	-	0.28

Table 9.3: Spearman ρ correlation coefficients scores between features and self-assessed(S)/assessed(A) personality traits.

Color properties 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 viceversa. 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. Viceversa, 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. The features of this category that covariate most with the two traits 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 appear to covariate with the perception of most traits, especially when it comes to the properties of the *Gray-Level Co-occurrence Matrix*.

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 [133]. However, different data can make different traits more or less *available*, i.e. more or less accessible to human observers [299]. The correlational analysis shows that favorite pictures convey impressions more effectively for Agreableness 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.

9.5.2 Correlational Analysis between Pictures and Personal Information

In Table 9.4 we report the correlation scores (Spearman ρ correlation coefficients) between the image features and personal data (see Table 9.1). In Table 9.5 we report the correlation scores (point-biserial correlation coefficients r_{pb} [156]) between the image features and dichotomous data and in Table 9.6 the measure of association η^2 between image features and categorical data [203]. Correlation values in italic are significative at 5%, underlined at 1%.

The features show to have significant correlations also with the age ($\hat{\rho} = 0.0.39$), #photos ($\hat{\rho} = 0.16$), #fave ($\hat{\rho} = -0.18$), #galleries ($\hat{\rho} = 0.06$), #contacts ($\hat{\rho} = -0.07$), #groups ($\hat{\rho} = 0.08$), gender ($\hat{r_{pb}} = 0.38$), status ($\hat{\eta}^2 = 0.04$), country ($\hat{\eta}^2 = 0.46$), where the operator "^" indicates the average, ρ is the Spearman correlation coefficient, r_{pb} the point-biserial correlation coefficient and η^2 is a strength of association measure.

9.5.3 Regression Analysis: 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 user included in the corpus . 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 9.4.1 and Section 9.4.2 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 $\{20\times20, 25\times25, \ldots, 65\times65\}$. 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 [27, 213], and the original papers of cit-kNN [69] for what concerns the number of citers and neighbors.

9.5.4 Prediction Results

Figure 9.14 reports the results obtained with the regression methods 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, 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 [285], 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. Furthermore, the consensus across the judges is statistically significant. 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

Category	z	Features	Age	#Photos	#Fave	#Galleries	#Contacts	#Groups
Color	1	use of light	-	0.1754	-	0.0769	-	-
	2	avg saturation	-	0.2314	-	-	-0.1489	0.1511
	3	std saturation	-	0.2335	-0.1136	-	-0.1187	0.1562
	4	std brightness	-	-0.1159	-0.2280	-	-	-
	5	valence	-	0.2183	-	0.0139	-	-
	6	dominance	-	-	-0.1333	-	-	-
	7	arousal	-	0.2037	-	-	-0.1514	0.1550
	8	hue circular variance	0.3254	0.1906	-	-	-	-
	9	colorfulness	-	-0.1741	-	-	-	-
	10	black	-	-0.1811	0.1164	-	0.1228	-
	11	blue	-	0.1992	-	-	-0.1155	-
	12	brown	-	0.2018	0.1455	-	-	-
	13	gray	-	-	0.2125	-	-	-
	14	green	0.3958	0.1948	0.1460	0.0432	-	-
	15	orange	-	0.2077	-	-	-0.1555	0.1737
	16	pink	0.3135	0.1326	-	-0.0014	-	-
	17	purple	0.4423	0.1167	-	-	-	-
	18	red	-	0.1876	-	-	-	0.1143
	19	white	-	-0.1258	0.2048	-	0.2065	-
	20	yellow	-	0.2285	-	-0.0297	-0.1247	0.1906
Composition	21	pers. perc. of edges pixels	-	-	0.2193	-	-	-
	22	level of detail	-	0.1487	-	-	-0.1595	-
	23	avg region size	-	-0.1429	-	-	-	-
	24	low DOF - hue	-	0.2201	-	-	-	-
	25	low DOF - saturation	-	0.2159	-	-	-	-
	26	low DOF - brightness	-	-	-	-	-	-
	27	rule of thirds - saturation	-	0.2377	-	-	-0.1449	0.1639
	28	rule of thirds - brightness	-	0.1652	-	0.0471	-	-
	29	image size	-	-	0.3479	-	-	-
Textural Properties	30	gray distribution entropy	-	-	-	-	-	-
	31	hue wavelet - lev 1	-	0.1320	-	-	-	-
	32	hue wavelet - lev 2	0.3086	0.1355	-	-	-	-
	33	hue wavelet - lev 3	0.3184	0.1386	-	-	-	-
	34	saturation wavelet - lev 1	-	0.1278	-0.1721	-	-0.1258	-
	35	saturation wavelet - lev 2	-	0.1488	-0.1578	-	-0.1305	-
	36	saturation wavelet - lev 3	-	0.1730	-0.1618	-	-0.1367	-
	37	brightness wavelet - lev 1	-	-	-0.1804	-	-0.1228	-
	38	brightness wavelet - lev 2	-0.3066	-	-0.2066	-	-	-
	39	brightness wavelet - lev 3	-	-	-0.2237	-	-	-
	40	hue wavelet avg	0.3119	0.1360	-	-	-	-
	41	saturation wavelet avg	-	0.1582	-0.1643	-	-0.1341	-
	42	brightness wavelet avg	-	-	-0.2169	-	-	-
	43	Tamura coarseness	0.3729	-	-	-	0.1221	-
	44	Tamura contrast	-	-0.1657	-	-	0.1410	-
	45	Tamura directionality	-0.3021	-	-	-	-0.1320	-
	46	GLCM contrast - hue	-	-	-	-	-	-
	47	GLCM correlation - hue	-	0.2302	-	-	-	0.1161
	48	GLCM energy - hue	-0.3390	-0.1972	-	-	-	-
	49	GLCM homogeneity - hue	-0.3112	-0.1579	-	-	-	-
	50	GLCM contrast - saturation	-	-	-0.2329	-	-0.1302	-
	51	GLCM correlation - saturation	-	0.2544	-	-	-	0.1169
	52	GLCM energy - saturation	-	-0.2210	-	-	-	-
	53	GLCM homogeneity - saturation	-	-0.1593	-	-	-	-
	54	GLCM contrast - brightness	-0.3339	-	-0.2141	-0.0142	-0.1229	-
	55	GLCM correlation - brightness	-	-	0.1251	-	0.1765	-
	56	GLCM energy - brightness	-	-0.1429	-	-	-	-
	57	GLCM homogeneity - brightness	-	-	0.1370	-	-	-
	58	GIST - channel 1	-	-0.1429	-0.1207	0.0657	-	-0.1325
	59	GIST - channel 2	-	-	-0.1701	-0.0091	-	-
	60	GIST - channel 3	-	-0.1505	-0.1532	-	-	-0.1407
	61	GIST - channel 4	-	-	-0.2006	-0.0233	-	-
	62	GIST - channel 5	-	-	-0.1416	0.0513	-	-
	63	GIST - channel 6	-	-	-0.1791	-0.0341	-	-
	64	GIST - channel 7	-	-0.1307	-0.1659	0.0187	-	-
	65	GIST - channel 8	-	-	-0.1967	-0.0420	-	-
	66	GIST - channel 9	-0.3475	-	-0.1364	0.0387	-	-
	67	GIST - channel 10	-0.3287	-	-0.1785	-0.0433	-0.1425	-
	68	GIST - channel 11	-0.3500	-	-0.1509	0.0076	-	-
	69	GIST - channel 12	-0.3615	-	-0.1754	-0.0565	-0.1367	-
	70	GIST - channel 13	-	-0.1247	-0.1446	0.0416	-	-
	71	GIST - channel 14	-	-	-0.1858	-0.0218	-	-
	72	GIST - channel 15	-	-0.1228	-0.1686	0.0251	-	-
	73	GIST - channel 16	-	-	-0.2053	-0.0516	-0.1205	-
	74	GIST - channel 17	-	-	-0.1572	-	-	-0.1306
	75	GIST - channel 18	-	-	-0.2076	-0.0296	-	-
	76	GIST - channel 19	-	-	-0.1739	0.0174	-	-0.1361
	77	GIST - channel 20	-	-	-0.2362	-0.0442	-	-
	78	GIST - channel 21	-	-	-0.1749	0.0432	-	-
	79	GIST - channel 22	-	-	-0.2213	-0.0496	-0.1234	-
	80	GIST - channel 23	-	-	-0.1827	0.0131	-	-
	81	GIST - channel 24	-	-	-0.2321	-0.0654	-0.1275	-
	82	number of faces	-	-0.1452	-	-	0.2381	-

Table 9.4: Spearman ρ correlation coefficients scores between features and personal information (see Table 9.1).

Category	z	Features	Gender
Color	1	use of light	**-0.2668**
	2	avg saturation	-
	3	std saturation	**-0.1097**
	4	std brightness	**-0.3277**
	5	valence	**-0.2037**
	6	dominance	**-0.2901**
	7	arousal	**-0.1425**
	8	hue circular variance	-
	9	colorfulness	**-0.0975**
	10	black	**0.3280**
	11	blue	**0.3385**
	12	brown	**0.3179**
	13	gray	**0.2926**
	14	green	**0.3503**
	15	orange	**0.3855**
	16	pink	**0.3892**
	17	purple	**0.3938**
	18	red	**0.3889**
	19	white	**0.3462**
	20	yellow	**0.3752**
Composition	21	pers. perc. of edges pixels	-
	22	level of detail	**0.2471**
	23	avg region size	**0.4070**
	24	low DOF - hue	-
	25	low DOF - saturation	-
	26	low DOF - brightness	**-0.1102**
	27	rule of thirds - saturation	-
	28	rule of thirds - brightness	**-0.3115**
	29	image size	-
Textural Properties	30	gray distribution entropy	**-0.5209**
	31	hue wavelet - lev 1	**0.2818**
	32	hue wavelet - lev 2	**0.2314**
	33	hue wavelet - lev 3	**0.1932**
	34	saturation wavelet - lev 1	**0.2478**
	35	saturation wavelet - lev 2	**0.2142**
	36	saturation wavelet - lev 3	**0.2394**
	37	brightness wavelet - lev 1	**0.2109**
	38	brightness wavelet - lev 2	**0.2231**
	39	brightness wavelet - lev 3	**0.2568**
	40	hue wavelet avg	**0.2013**
	41	saturation wavelet avg	**0.1955**
	42	brightness wavelet avg	**0.2231**
	43	Tamura coarseness	**-0.6561**
	44	Tamura contrast	**0.1060**
	45	Tamura directionality	-
	46	GLCM contrast - hue	**0.3338**
	47	GLCM correlation - hue	**-0.4703**
	48	GLCM energy - hue	**-0.3064**
	49	GLCM homogeneity - hue	**-0.6831**
	50	GLCM contrast - saturation	**0.3309**
	51	GLCM correlation - saturation	**-0.4958**
	52	GLCM energy - saturation	**-0.1533**
	53	GLCM homogeneity - saturation	**-0.6187**
	54	GLCM contrast - brightness	**0.3176**
	55	GLCM correlation - brightness	**-0.7102**
	56	GLCM energy - brightness	**0.1656**
	57	GLCM homogeneity - brightness	**-0.6189**
	58	GIST - channel 1	**0.2413**
	59	GIST - channel 2	**0.1655**
	60	GIST - channel 3	**0.2420**
	61	GIST - channel 4	**0.1071**
	62	GIST - channel 5	**0.1928**
	63	GIST - channel 6	**0.1086**
	64	GIST - channel 7	**0.1980**
	65	GIST - channel 8	**0.1176**
	66	GIST - channel 9	**0.2950**
	67	GIST - channel 10	**0.2623**
	68	GIST - channel 11	**0.2881**
	69	GIST - channel 12	**0.2485**
	70	GIST - channel 13	**0.2736**
	71	GIST - channel 14	**0.2316**
	72	GIST - channel 15	**0.2341**
	73	GIST - channel 16	**0.1517**
	74	GIST - channel 17	**0.1672**
	75	GIST - channel 18	**0.1099**
	76	GIST - channel 19	**0.1487**
	77	GIST - channel 20	**0.0829**
	78	GIST - channel 21	**0.2260**
	79	GIST - channel 22	**0.1446**
	80	GIST - channel 23	**0.2517**
	81	GIST - channel 24	**0.1446**
	82	number of faces	**0.4019**

Table 9.5: Point-biserial r_{pb} correlation coefficients scores between features and personal information (see Table 9.1).

Category	z	Features	Status	Country
Color	1	use of light	-	-
	2	avg saturation	-	**0.2075**
	3	std saturation	-	**0.1833**
	4	std brightness	-	-
	5	valence	-	**0.1815**
	6	dominance	-	-
	7	arousal	-	**0.1924**
	8	hue circular variance	-	**0.2551**
	9	colorfulness	-	**0.2729**
	10	black	-	**0.1818**
	11	blue	-	**0.2091**
	12	brown	-	**0.2246**
	13	gray	-	-
	14	green	-	**0.2361**
	15	orange	-	-
	16	pink	-	**0.1773**
	17	purple	-	**0.2189**
	18	red	-	-
	19	white	-	-
	20	yellow	-	**0.2111**
Composition	21	pers. perc. of edges pixels	-	-
	22	level of detail	-	**0.2101**
	23	avg region size	-	-
	24	low DOF - hue	-	**0.2135**
	25	low DOF - saturation	-	**0.2042**
	26	low DOF - brightness	-	-
	27	rule of thirds - saturation	-	**0.1902**
	28	rule of thirds - brightness	-	-
	29	image size	-	-
Textural Properties	30	gray distribution entropy	-	**0.2810**
	31	hue wavelet - lev 1	-	**0.2436**
	32	hue wavelet - lev 2	**0.0448**	**0.2343**
	33	hue wavelet - lev 3	**0.0488**	**0.2210**
	34	saturation wavelet - lev 1	-	**0.1740**
	35	saturation wavelet - lev 2	-	-
	36	saturation wavelet - lev 3	-	-
	37	brightness wavelet - lev 1	-	**0.1806**
	38	brightness wavelet - lev 2	-	**0.1793**
	39	brightness wavelet - lev 3	-	-
	40	hue wavelet avg	**0.0461**	**0.2295**
	41	saturation wavelet avg	-	-
	42	brightness wavelet avg	-	**0.1774**
	43	Tamura coarseness	-	-
	44	Tamura contrast	-	**0.1859**
	45	Tamura directionality	**0.0450**	-
	46	GLCM contrast - hue	**0.0452**	**0.2407**
	47	GLCM correlation - hue	-	**0.2299**
	48	GLCM energy - hue	-	**0.2758**
	49	GLCM homogeneity - hue	**0.0404**	**0.2958**
	50	GLCM contrast - saturation	-	-
	51	GLCM correlation - saturation	-	**0.2195**
	52	GLCM energy - saturation	-	**0.2429**
	53	GLCM homogeneity - saturation	-	**0.2434**
	54	GLCM contrast - brightness	-	-
	55	GLCM correlation - brightness	-	-
	56	GLCM energy - brightness	-	**0.2690**
	57	GLCM homogeneity - brightness	-	**0.2288**
	58	GIST - channel 1	-	**0.2209**
	59	GIST - channel 2	-	**0.2037**
	60	GIST - channel 3	-	**0.2164**
	61	GIST - channel 4	-	**0.2040**
	62	GIST - channel 5	-	**0.1925**
	63	GIST - channel 6	-	**0.1825**
	64	GIST - channel 7	-	**0.1867**
	65	GIST - channel 8	-	-
	66	GIST - channel 9	-	-
	67	GIST - channel 10	-	-
	68	GIST - channel 11	-	-
	69	GIST - channel 12	-	-
	70	GIST - channel 13	-	**0.1869**
	71	GIST - channel 14	-	**0.1816**
	72	GIST - channel 15	-	**0.1810**
	73	GIST - channel 16	-	**0.1753**
	74	GIST - channel 17	-	**0.1982**
	75	GIST - channel 18	-	-
	76	GIST - channel 19	**0.0427**	**0.1827**
	77	GIST - channel 20	-	-
	78	GIST - channel 21	-	-
	79	GIST - channel 22	-	-
	80	GIST - channel 23	-	-
	81	GIST - channel 24	-	-
	82	number of faces	**0.0460**	**0.1798**

Table 9.6: Correlation scores (η^2 measure of association) between features and personal information (see Table 9.1).

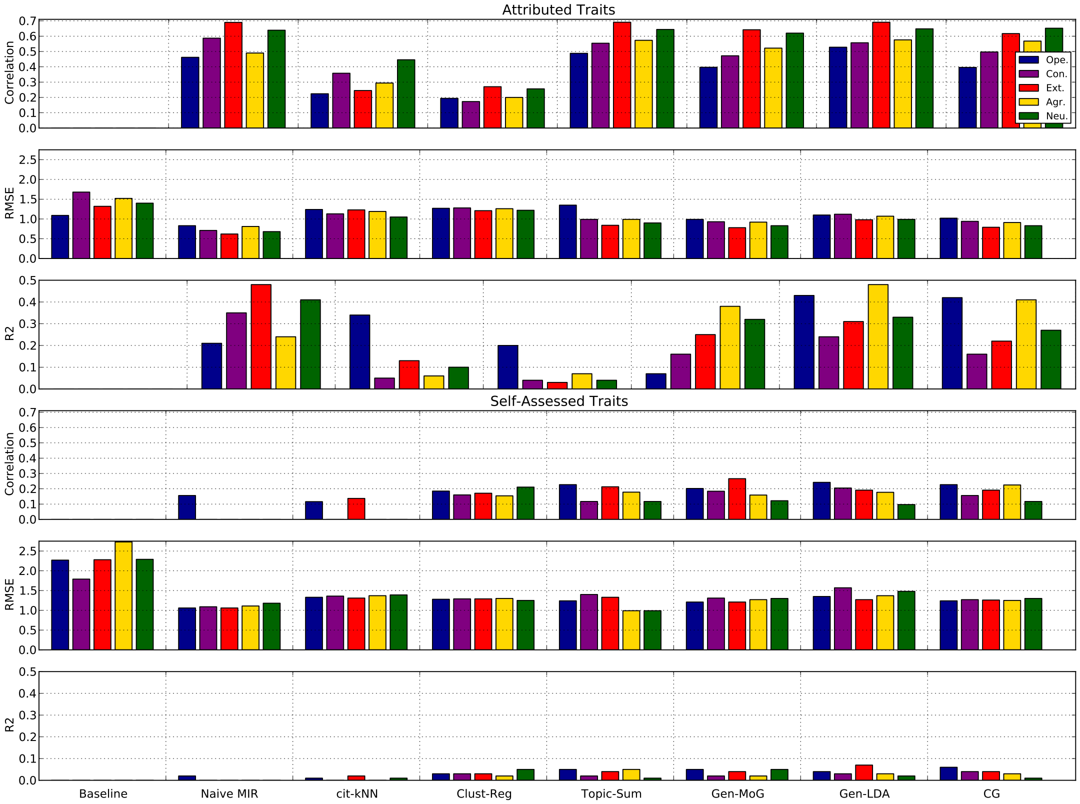


Fig. 9.14: The figure shows APP (upper chart) and APR (lower chart) performance. The missing bars correspond to correlations that are not significant at 0.05 level.

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*" [49]. 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 9.14). 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. 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 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 [299]. 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 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 [285] and Personality Psychology [133]. The reason is that this trait is the most socially oriented and, therefore, it leaves more traces in observable behavior [299]. 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. 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 9.3, 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.

9.5.5 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 9.15 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. 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 (Table 9.3 shows that Openness, Agreeableness and Conscientiousness are more weakly correlated with the features than the other two traits).

The central and rightmost plots of Figure 9.15 show the relationship between performance and size of test and training set bags, respectively. 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.

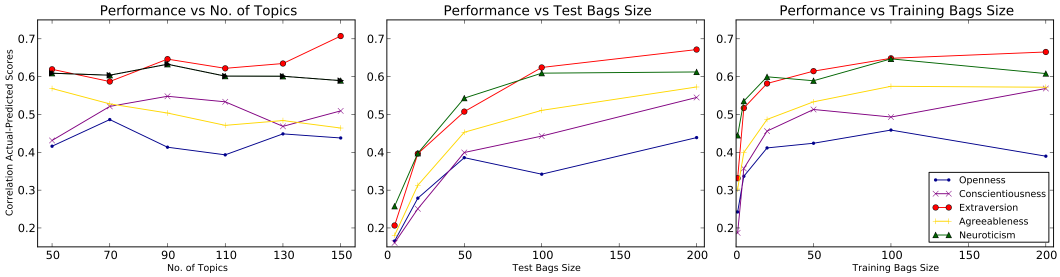


Fig. 9.15: From left to right, the first plot shows the performance as a function of the number of topics, the second and the third report the performance as a function of test and training bags size, respectively for APP.

9.5.6 Reading between the lines of MI-Gen-LDA

To get further insight on how the MI-Gen-LDA works, we analyzed the subset of features which is more related to a trait: this can be done by looking at the regression coefficients β_k found by LASSO. We consider the most important topic $k_{\text{best}} = \text{argmax}_k \beta_k$ in the latent space. As each topic is a tight distribution over the features, k_{best} represents the pool of features which, *when taken together*, are more related with a high value of the trait. This differs substantially from what we did in Section 9.5.2, where features have been analyzed *independently*. We list the most important features for that topic (i.e., that with higher $p(z|k_{\text{best}})$), together with those images where the presence of these features is the highest in Fig 9.16-Figure 9.20. In practice, as each topic is a tight distribution over the features, k_{best} represents the pool of features which, *when taken together*, are more related with a *high* value of the trait.

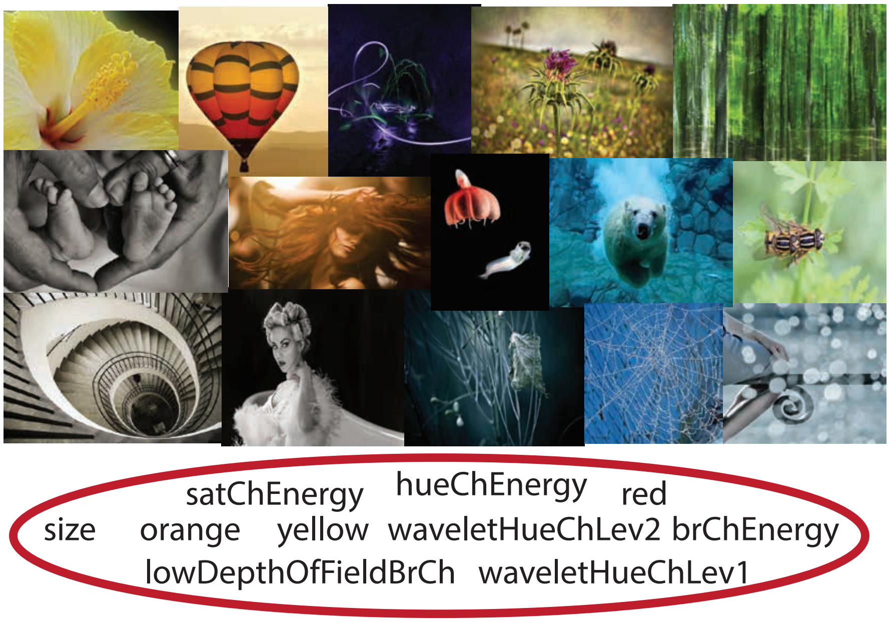


Fig. 9.16: Images related to the most representative features for the Openness trait. Here we can see a topic which models colorful images with big and few regions, and characterized by a blurred background surrounding the subject of the image.

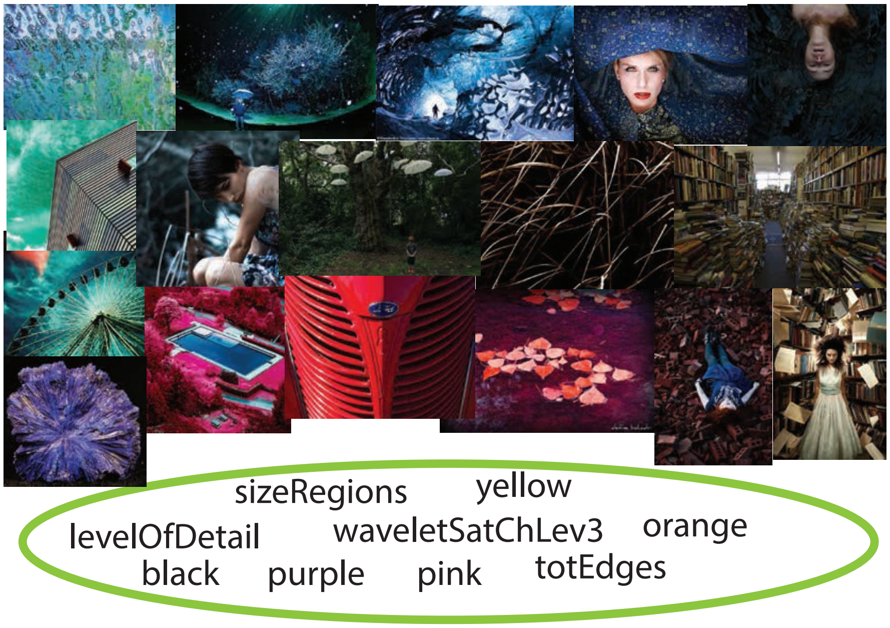


Fig. 9.17: Images related to the most representative features for the Conscientiousness trait. Here we can see a topic which models the joint presence of hot and cold colors, with many sharp regions (high number of edges and level of detail).

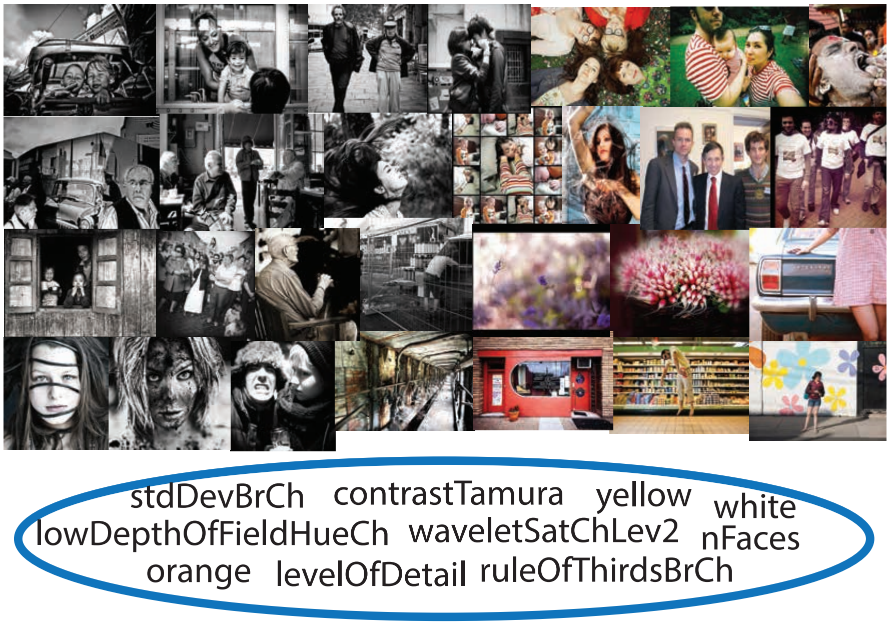


Fig. 9.18: Images related to the most representative features for the Extraversion trait. Here we can see a topic which models people in bright images, with a high quantity of white.

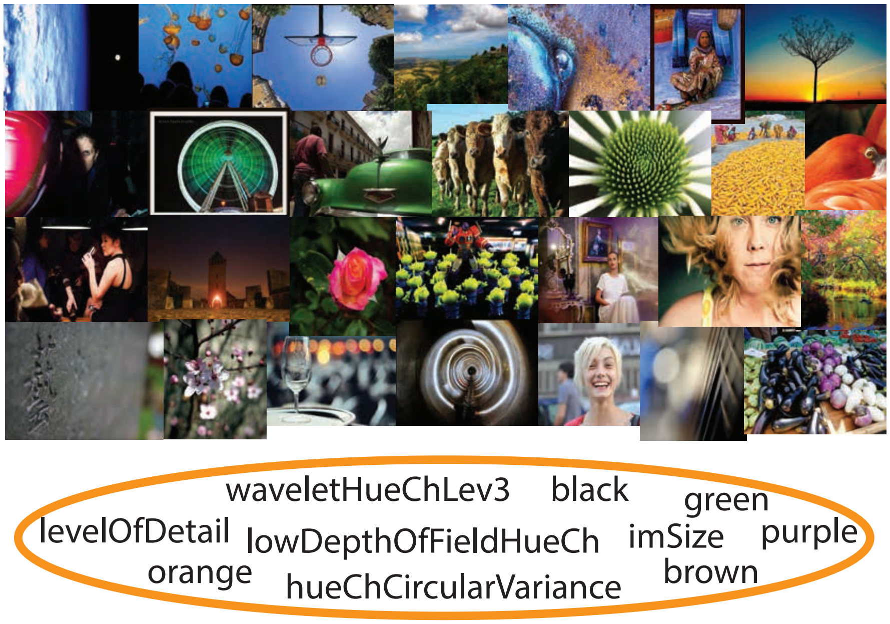


Fig. 9.19: Images related to the most representative features for the Agreeableness trait. Here we can see a topic which models colorful images.

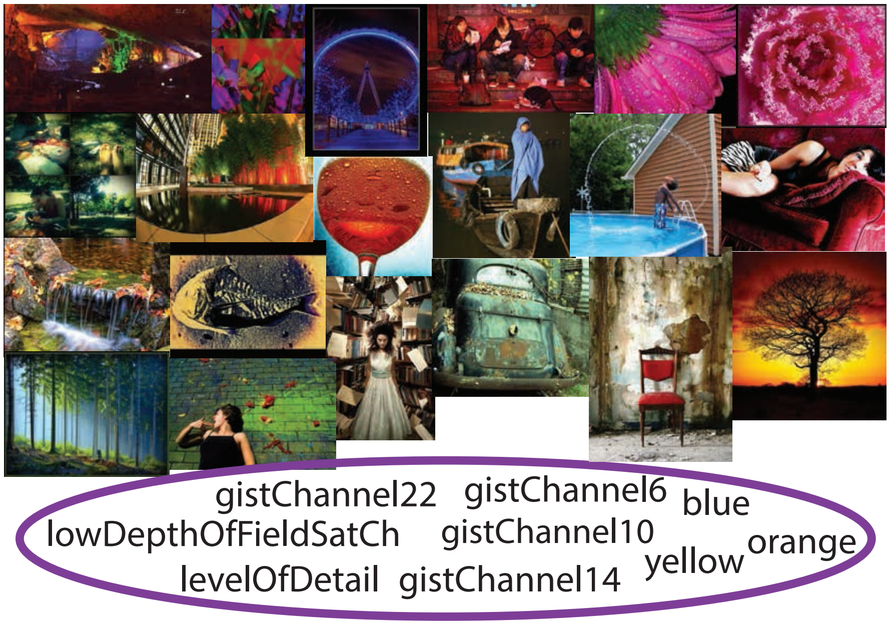


Fig. 9.20: Images related to the most representative features for the Neuroticism trait. Here we can see that the features that belong to the most "important" topic focus on the characterization of the scene (the GIST features) with a prominent role of some colors.

9.6 Discussion

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. 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 [161] 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*" [71].

The performance of the approach proposed in the experiments tends to be higher when attributed traits are closer to the extremes of the scales, 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 [138, 139]. In particular, Section 9.2.2 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 [138, 139] 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 [23]. Figure 9.3 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 [285]). 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)*" [299]. 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 co-variate 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 [285]. 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) [299]. 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 9.5 shows that a large number of features covariate with the attributed traits to a statistically significant extent. The covariation is particularly high in the case of two traits - Neuroticism and Agreeableness - and the features related to the colors . 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*" [281] 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 [54].

10

Social Profiling through Image Understanding: Personality Inference using Convolutional Neural Networks

10.1 Introduction

A limitation of the approach in Chapter 9 is that the features used to describe the images are taken from the Computational Aesthetics (CA) literature; in practice, CA often focuses on designing features that explain how a particular image has been captured, discarding the content of the images. In addition, given the wide spectrum of subjects appearing in database images, standard object recognition and feature extraction techniques might not be sufficient to capture significant dependencies between the pictures and the personality traits of their owner. This leads to the development of more advanced techniques such as feature learning, carried out in this paper by Convolutional Neural Networks.

Differently from many Computational Aesthetics studies that describe pictures through an array of feature extractors and scene analyzers [62, 174], our characterization of each image is locked within the layers of a Convolutional Neural Network. Computer vision with Convolutional Neural Networks (CNNs) has received much attention in recent years, as it is well suited for processing large amounts of data and providing outstanding performances in classical problems like object [146] and image style [136] recognition. In fact, our approach fine-tunes CNNs pre-trained for image classification with the intention of co-opting their effective representational power to indirectly capture the aesthetic attributes of photographs, with the ultimate goal of predicting the personality traits associated with them. This allows us to discover more entangled attributes, unattainable by simple hand-crafted CA features because of the interaction between many different factors, and to better generalize the patterns that identify a trait. In practice, whereas CA features are explicitly crafted to reveal information about the style of an image, remaining agnostic w.r.t. the content of the image, CNNs exhibit no such limitation, capturing both the aesthetic patterns in the pictures and their content, unveiling semantic information (for example, capturing possible recurrent objects preferred by a user).

Experiments have been focused on the *PsychoFlickr* corpus of Chapter 9. The experimental results show that the proposed method sufficiently captures what characterizes a certain trait: on a quantitative level, it performs around 10% better on attributed traits than on self-assessed ones, with a best accuracy of 68% on attributed Neuroticism; on a qualitative level, ranking the test images by confidence shows a clear distinction of features, patterns and content between low and high values in a given trait. We also compare with Chapter 9, by suitably re-casting their results from a regression to a classification framework similar to our own, and largely outperform these results in all but one trait. We also introduce an online application demo that uses our trained classifiers to predict personality traits given a proposed set of uploaded or selected pictures liked by a subject.

In the following sections we introduce our approach based on processing the *PsychoFlickr* corpus using Convolutional Neural Networks, followed by a section discussing the results. Finally, we present our demo.

10.2 The Proposed Approach

The ILSVRC challenges [146] have clearly demonstrated the CNN aptitude at deconstructing the elements and features contained within photographs. Most importantly, they share some basic primitive components analyzed in the first works of personality inference (which essentially applied standard CA features): color, composition, textural properties, etc. More in the detail, CNNs represent a more effective way to combine such primitive components into holistic representations, aimed not only at highlighting stylistic patterns (as done for example in [136], and in general by all the computational aesthetics works), but also to unveil information about the content of the image. To investigate this idea, we adopt CNNs pre-trained for image classification and fine-tune them to predict the personality traits contained in the *PsychoFlickr* dataset. In this way, we leverage the deconstructing power of networks trained on millions of images in order to learn visual representations correlated with psychological profiles. The upside of this approach is that, since pre-trained networks are learned on a large set of images, the intermediate layers capture the "semantics" of the general visual appearance in a way superior to hand-crafted features. The downside is represented by the "black box" nature of the technique: the insights discovered and codified by the network are locked within its connectivity structure, only revealed approximately by considerable analysis work [308]. However, this black box can be effectively exploited by adding a few layers on top, tailored for our specific problem.

10.2.1 Personality Classification by Fine-tuning Convolutional Neural Networks

In our approach, we decide to simplify the task of predicting the personality traits in the *PsychoFlickr* corpus into five distinct binary classification problems, one for each trait. As they result from the application of Factor Analysis to behavioral evidence [52], the Big-Five traits are independent. The range of values for each trait is partitioned into three sections: low set, for values below the first quartile; high set, for values above the third quartile, and middle set (see Figureure 10.1). We select only the users with values included in the low and high sets for our binary classification problems in order to get a greater separation between the two classes, and proceed to collect all the related images. Figure 10.1 shows the distributions of values for both self and attributed traits.

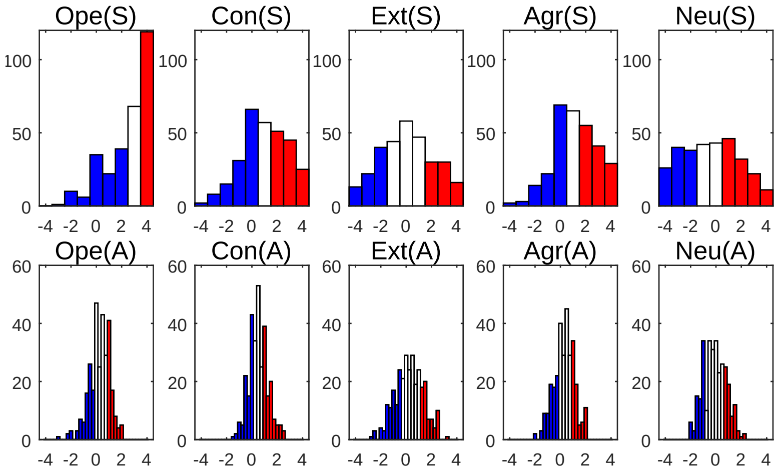


Fig. 10.1: The distribution of values in the self-assessed (top row) and attributed (bottom row) personality traits, with the low and high classes highlighted in blue and red, respectively.

Considering both self-assessed and attributed traits, we thus have 10 independent datasets of images with binary labels. As protocol, we keep 75% of each dataset for training and 25% for testing.

Fine-tuning a CNN is the process of training the classifier not from a blank, random, initialization, but from the model parameters previously trained on a related, but well-defined problem. It also involves modifying the model structure to a certain extent, to suit the new problem (for example, a difference in the number of output classes). As a reference, we use the eight-layer network that won the ImageNet challenge in 2012 (see [146] for details), and change the last layer in order to adapt it to our binary classification problems. This is an ideal candidate for fine-tuning because that CNN has been trained on a large number of images (1.2 million) and a wide number of classes (1000 object categories), providing for a very solid acquisition of representational power.

We employ the open source software implementation of Caffe [126] for our fine-tuning problem. The development of our approach was not immediately successful, as we encountered several challenges. Chief amongst them, we initially found that the CNN would either fail to learn, performing not better than random guessing, or tend to overfit, causing the testing accuracy to drop significantly as the training performance rose.

We unsuccessfully tried different configurations of the learning parameters to avoid the first problem, and we only managed to get it under control by fixing the convolutional layers of the network. The eight layers in the network are conceptually split into two blocks, the first five convolutional layers (with some max-pooling layers interleaved) and the last three fully connected layers: the first block processes an input image in terms of visual patterns, and the second block seeks a higher-level semantic representation based on those visual patterns. By fixing the first block, we exclusively relied on the same visual patterns of the ImageNet problem, which luckily cover a very wide spectrum due to the huge number of images and classes. This allowed the network to focus just on finding the right mix of visual patterns to characterize the psychological traits. One possible drawback is an understandable bias towards objects rather than styles within the images. Moreover, we controlled the overfitting tendency by raising the value of the weight decay parameter, responsible for regularizing the weights and reducing the model's training error [146].

In detail, Caffe was run on a linux pc with an NVIDIA graphical card to exploit GPU speed-ups, and we fine-tuned our models using stochastic gradient descent with a batch size of 50 images, momentum of 0.9, weight decay of 0.1, base learning rate of 0.0001, with one step-down in learning rate to 0.00001 halfway down the 60 cycles through the training set required to complete the learning process.

10.3 Experiments and Results

During our experiments, we found a stark difference between the self-assessed and the attributed traits. The former consistently performed worse than the latter, as shown in Table 10.1: the test accuracy is around 10% lower on all self-assessed traits and reached a plateau very early for each trait.

Overall, the attributed traits achieve good results: up to 0.68 for Neuroticism, 0.66 Conscientiousness and 0.64 for Extroversion and Agreeableness, while the hardest trait to classify is Openness.

	SELF		ATTRIBUTED	
	TRAIN	TEST	TRAIN	TEST
O	0.57	0.53	0.73	0.61
C	0.59	0.54	0.81	0.66
E	0.60	0.54	0.76	0.64
A	0.57	0.54	0.76	0.64
N	0.55	0.52	0.81	0.68

Table 10.1: Accuracies on the training and test sets.

As in the case of Chapter 9, the reason is that the judges are unacquainted with the users. Therefore, the pictures dominate the personality impressions that the judges develop. Furthermore, the consensus across the judges is statistically significant. These two conditions help the classification 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, being the correlation between visual features and trait scores low, this does not allows the classification to achieve high performances.

In Table 10.2, we show the confidence matrices for the test set, with the dominant class highlighted: in the attributed traits, the low class is dominant for Openness, Extraversion and Agreeableness, with the high class for Conscientiousness and Neuroticism.

<table>
<tr><td colspan="10">SELF</td></tr>
<tr><td>0.30</td><td>0.19</td><td>0.29</td><td>0.21</td><td>0.27</td><td>0.23</td><td>0.21</td><td>0.26</td><td>0.10</td><td>0.39</td></tr>
<tr><td>0.28</td><td>0.23</td><td>0.25</td><td>0.25</td><td>0.23</td><td>0.27</td><td>0.20</td><td>0.33</td><td>0.09</td><td>0.42</td></tr>
<tr><td colspan="2">O</td><td colspan="2">C</td><td colspan="2">E</td><td colspan="2">A</td><td colspan="2">N</td></tr>
<tr><td>0.36</td><td>0.16</td><td>0.27</td><td>0.22</td><td>0.36</td><td>0.15</td><td>0.34</td><td>0.18</td><td>0.29</td><td>0.22</td></tr>
<tr><td>0.22</td><td>0.26</td><td>0.11</td><td>0.40</td><td>0.20</td><td>0.29</td><td>0.18</td><td>0.30</td><td>0.09</td><td>0.40</td></tr>
<tr><td colspan="10">ATTRIBUTED</td></tr>
</table>

Table 10.2: Confusion matrices for both self and attributed traits.

In Table 10.3, we show the results of the classification test using the oversampling method in Caffe, which uses the average response on 10 crops taken for each image. Also in this case the attributed classification achieves better results than the self one, reaching an accuracy of 0.69 for Neuroticism and 0.66 for Conscientiousness.

	SELF	ATTRIBUTED
O	0.53	0.62
C	0.54	0.66
E	0.54	0.65
A	0.54	0.64
N	0.53	0.69

Table 10.3: Accuracy of classification test.

As a comparative test, we consider the work in Chapter 9, here named *PCMIL*. To obtain a fair comparison with our approach, we modify the framework of that work with our classification pipeline: we selected only people whose traits where in the same quartile range described above, and once regression was carried out by LASSO, we compare the regression result with the population mean score for a given trait. A regression score higher than the population mean is labeled as high and the opposite as low. The results are portrayed in Table 10.4: except for the attributed trait of Neuroticism, all the other results shows the superiority of the deep learning approach.

	SELF		ATTRIBUTED	
	PCMIL	our	*PCMIL*	our
O	0.49	0.53	0.59	0.62
C	0.50	0.54	0.51	0.66
E	0.52	0.54	0.64	0.65
A	0.47	0.54	0.56	0.64
N	0.51	0.53	0.75	0.69

Table 10.4: Results comparisons in terms of accuracies between Chapter 9 (*PCMIL*) and our approach.

Generally, the classifiers for the attributed traits display stark differences between the *low* and the *high* classes, and those for the self-assessed traits show weaker distinctions, but manage to keep some similarities with the former ones, despite their lower accuracy.

As last experiment of this section, we evaluated the classification performances when varying the number of images per user used in the test phase, in the range between 30 and 50. In Figure10.2 we report a curve for each trait considering the self-assessed and attributed scores, respectively. We can notice a almost perfect linear increase in the performances while increasing the images. This range was selected to individuate the minimum number of images useful for having a result above the chance: for the self-assessed classification we need at least 48 images, while for the attributed classification at least 38. This is in line with the fact that self assessed traits are harder to encode.

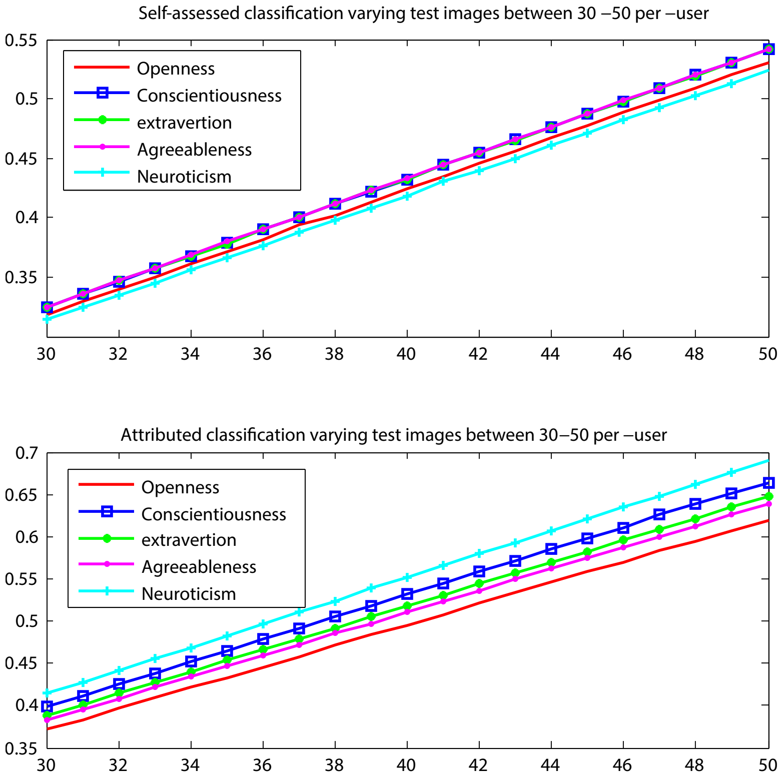
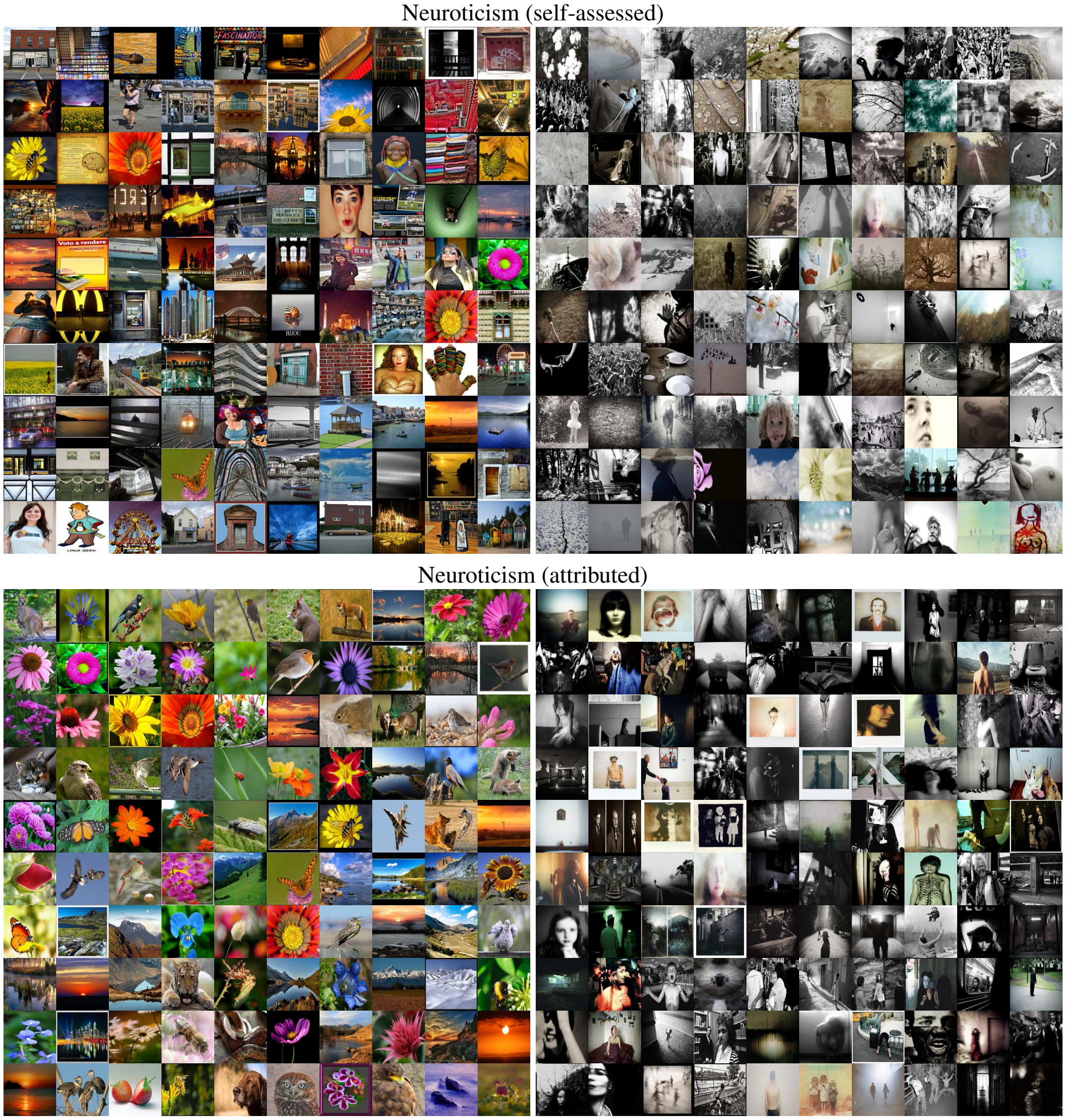


Fig. 10.2: Classification results varying the nuber of test images per user.

10.3.1 Attributes Learned for each Personality Trait

In this section we presents two different experiments for interpreting the semantics that each network related to a particular trait learn.

In the first experiment we first collected the output values of the softmax unit at the end of the network, and then used these values to rank correctly classified test images in decreasing order of confidence. This procedure is similar to collecting the images with the highest probability of falling into the specified class. In Figure 10.3 - 10.7 we show some representative images for each self and attributed trait for both low and high level.

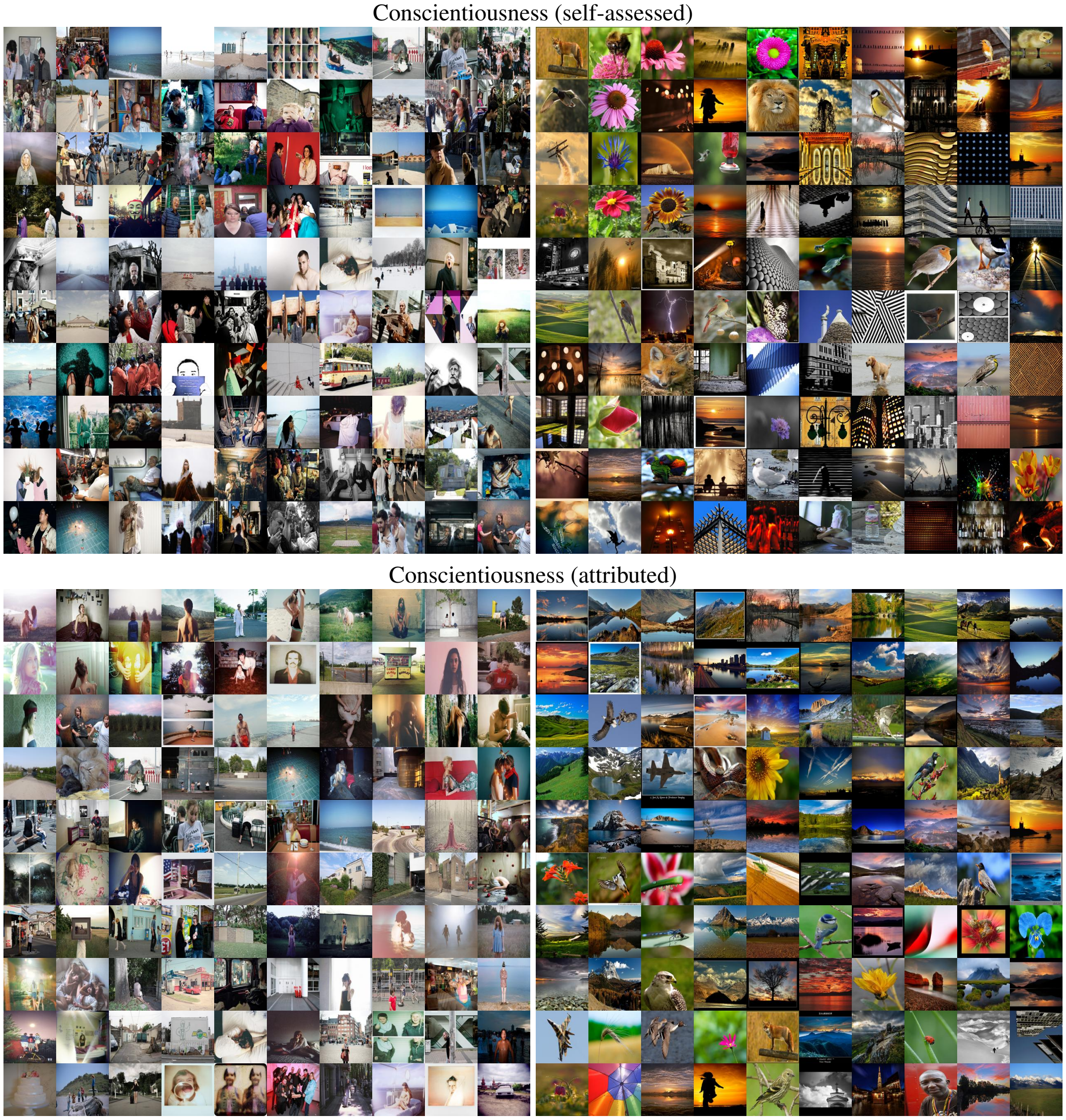


Neuroticism (self-assessed)

Neuroticism (attributed)

Fig. 10.3: Representative images in the low (left) and high (right) value classes for the Neuroticism trait.

In the specific case of Neuroticism, shown in Figure 10.3, black and white images with sharp contrasts seem to be regarded as typical of highly neurotic personalities, the presence of faces and indoor environments, while flowers and nature belong to the opposite tendency.

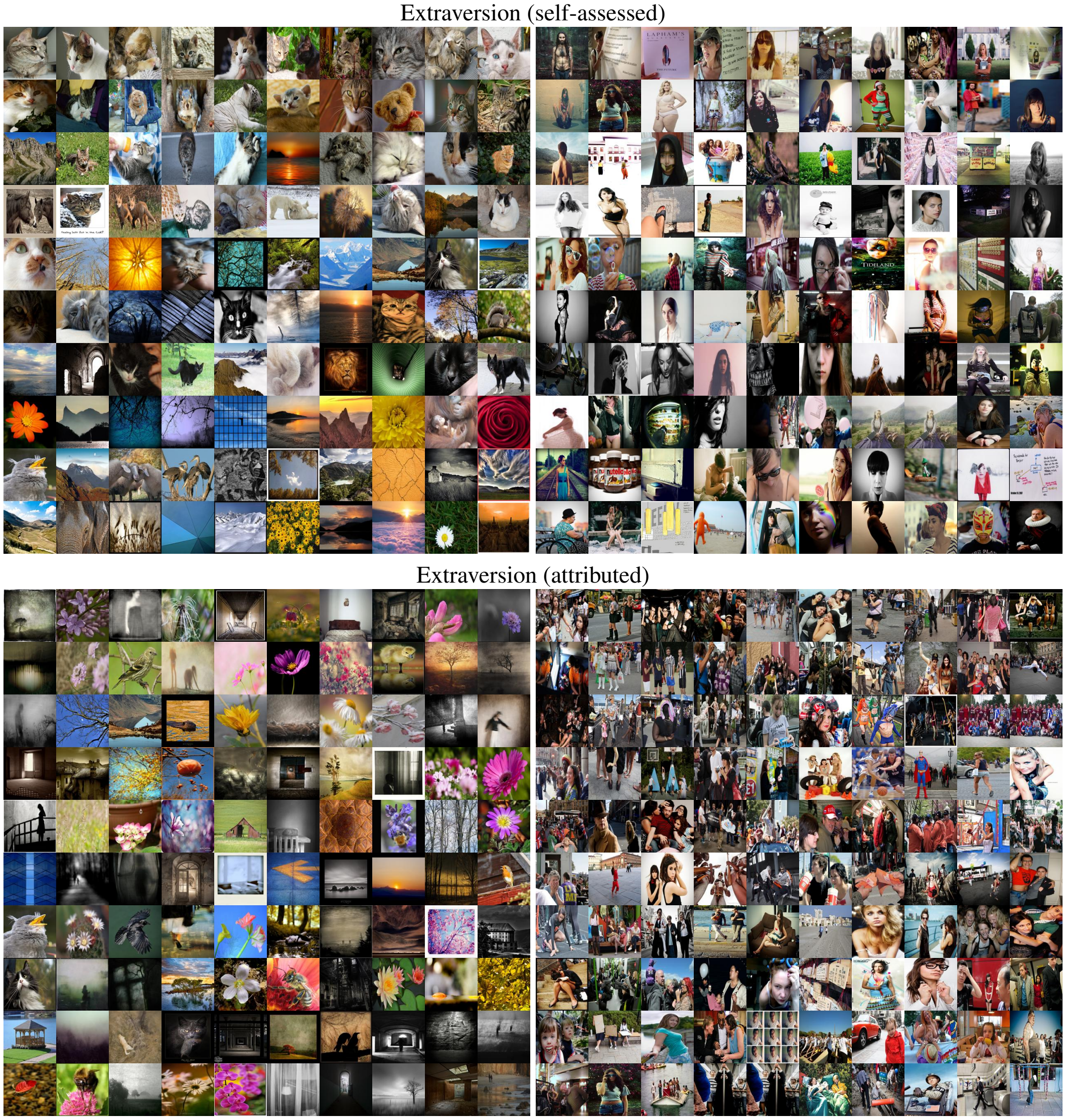


Conscientiousness (self-assessed)

Conscientiousness (attributed)

Fig. 10.4: Representative images in the low (left) and high (right) value classes for the Consciousness trait.

When it comes to Conscientiousness, shown in Figure 10.4, we can observe for the *high* class the presence of orderly images, scene composed of landscapes, mountains where the line between the sky and floor is sharp or the presence of buildings as matter of appreciation for regular geometry, otherwise if there is an object in the image, it appears in the foreground in front of a blurred background. In the *low* class we can see pictures where colors are faded, with pastel colors.

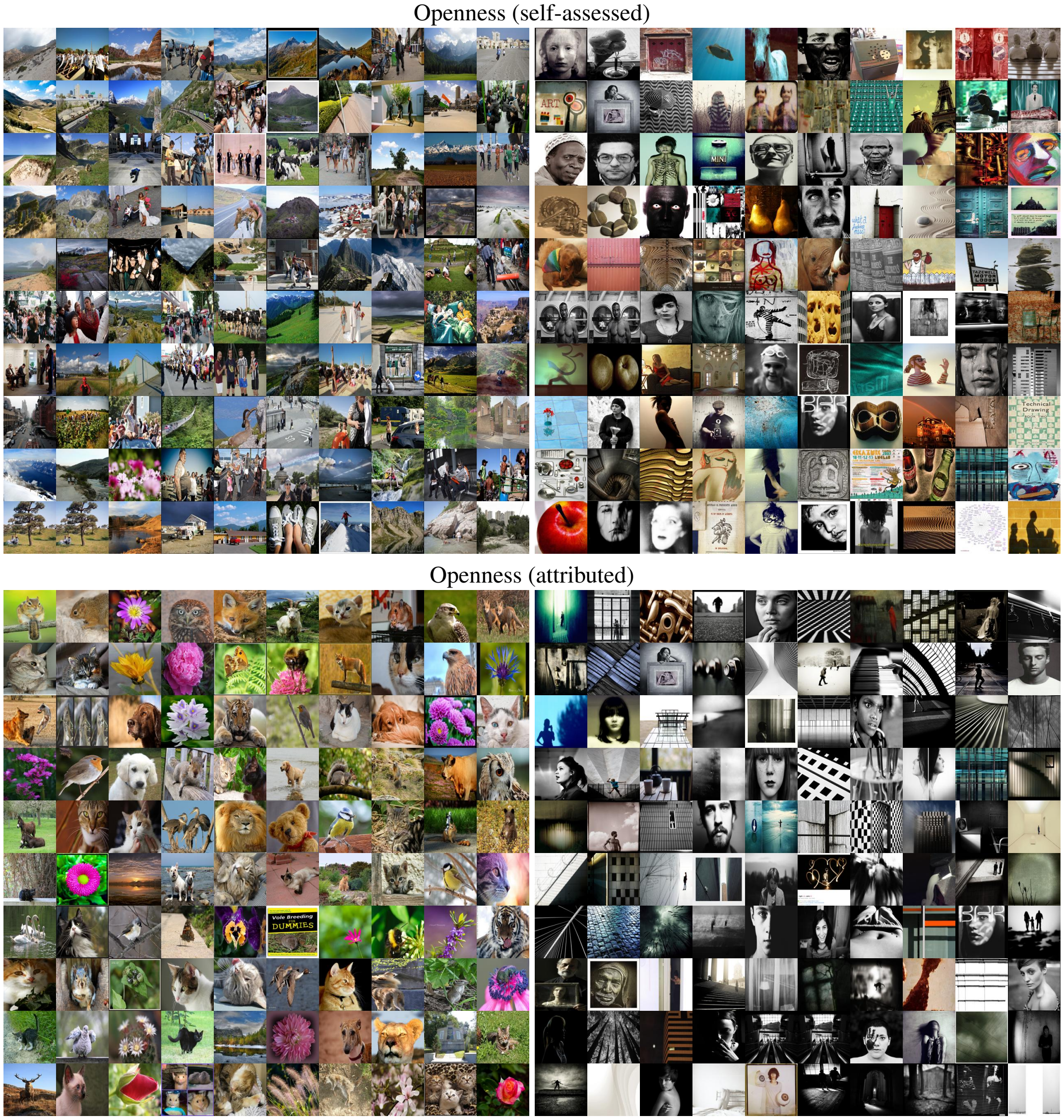


Extraversion (self-assessed)

Extraversion (attributed)

Fig. 10.5: Representative images in the low (left) and high (right) value classes for the Extraversion trait.

The Extraversion trait is the most impressive, shown in Figure 10.5. For the *high* class we can see that all pictures are crowded of people, while the *low* class represents images where there are just flowers, plants, and indoor scenes, showing the tendency of introvert people to not interact as extrovert ones.

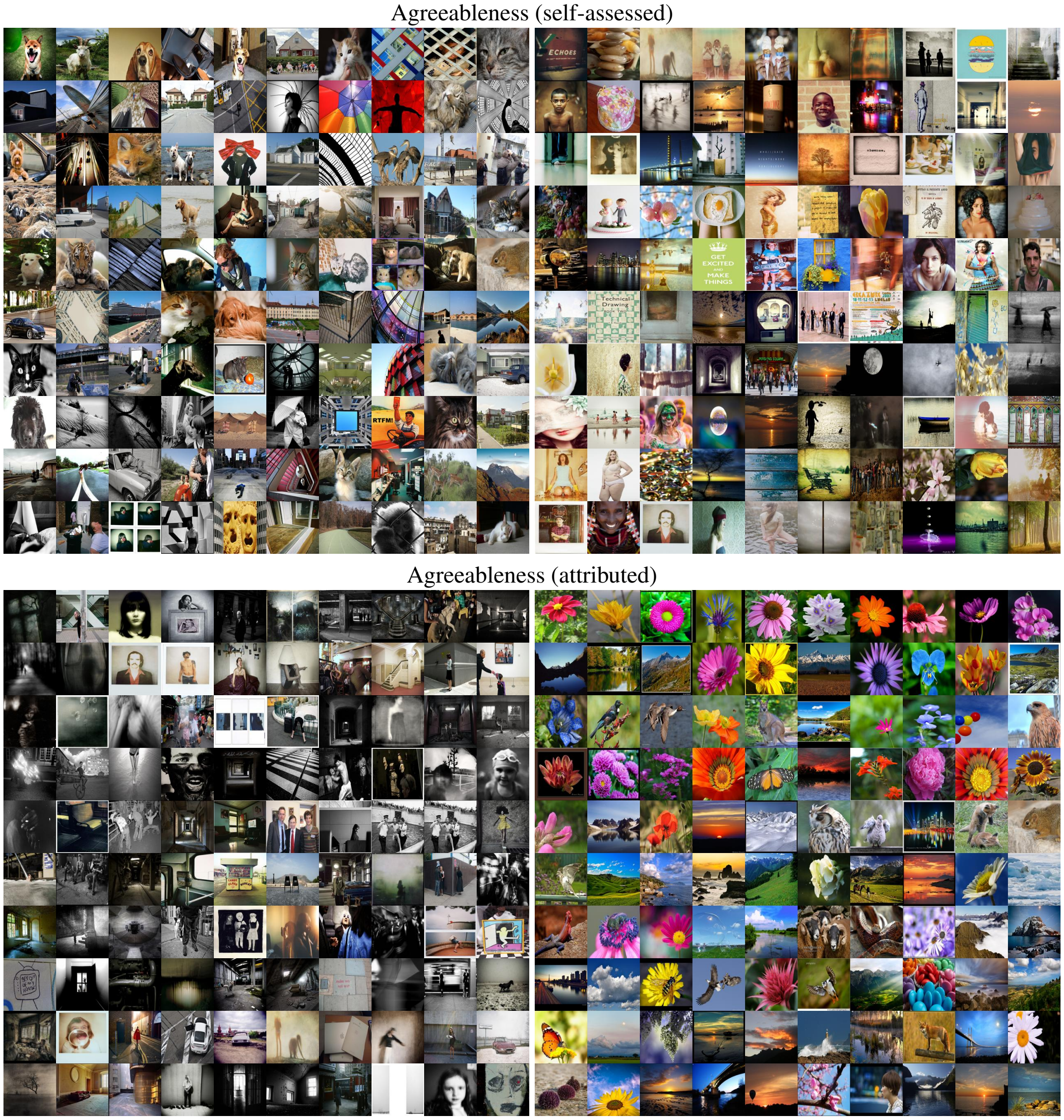


Openness (self-assessed)

Openness (attributed)

Fig. 10.6: Representative images in the low (left) and high (right) value classes for the Openness trait.

The Openness trait also shows really interesting emerging characteristics, shown in Figure 10.6: *high* class show really complex and artistic pictures, ranging from black/white pictures to portraits and complex shapes and scene to abstract and surrealist pictures. *Low* class shows pictures related to animals, most of them are felines and flowers, highlighting the fact that this people tend to prefer simple, real life pictures.



Agreeableness (self-assessed)

Agreeableness (attributed)

Fig. 10.7: Representative images in the low (left) and high (right) value classes for the Agreeableness trait.

Agreeableness *high* class show really colorful, high contrast, warm color pictures. On the opposite side *low* class is represented by images mostly in black/white, and not friendly environments. Notice how there are quite a few images in common between attributed Neuroticism and Agreeableness, but in opposite classes. This is a confirmation of the strong negative correlation between the two attributed traits.

In the second experiment we propose a deeper analysis of the semantic learned for each trait. We kept the collected images with the highest probability of falling into the specified class and performed an unsupervised clustering using t-SNE [275]. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets by giving each datapoint a location in a two or three-dimensional map. It is a way of converting a high-dimensional data set into a matrix of pairwise similarities visualizing them and capable of capturing much of the local structure of the high-dimensional data very well, while also revealing global structure such as the presence of clusters at several scales. This applies well in our case where the features of the CNN are 4096 dimensional. The aim of dimensionality reduction is to preserve as much of the significant structure of the high-dimensional data as possible in the low-dimensional map.

In our case, we used it to cluster and visualize how the most influent images of a given trait and level are grouped. This allow us to exploit the semantics and attributes that describe the trait. To this purpose we retained the CNN features in the fully connected layer before the classification and gave them as input to t-SNE. The code preprocesses the data using PCA, reducing its dimensionality to a chosen dimensions (we crossvalidated this parameter keeping values between 5 and 10). The perplexity basically sets how many near neighbors each point is trying to stay close to in the map. Therefore, small values for the perplexity will result in large numbers of small clusters, whereas large values will produce a smaller number of larger clusters. In our case we crossvalidated this parameter as well achieving better and more interpretable results with a value of 5. The function returns a matrix that specifies the coordinates of the 2 low-dimensional data points. With this map we can build a plot where we can project each image on 2-dimensional space, allowing us to visualize the clusters.

In the specific for a high level of Neuroticism (see Figure 10.8 top) we found three clusters about portrait of a person with blurred background, indoor scene and alone person in the dark. When it comes to a low level of Neuroticism (see Figure 10.8 bottom) we found five clusters about birds, flowers, sunset/sunrise scene, landscape of mountains and animals in more general. In the case of high level of Openness (see Figure 10.9 top) we found six clusters about portraits where the person is in a certain pose and nude, pastel colors paintings, pastel color shapes, artificial and digital painting, complex and twisted shapes similar to tissues, other artistic paintings.

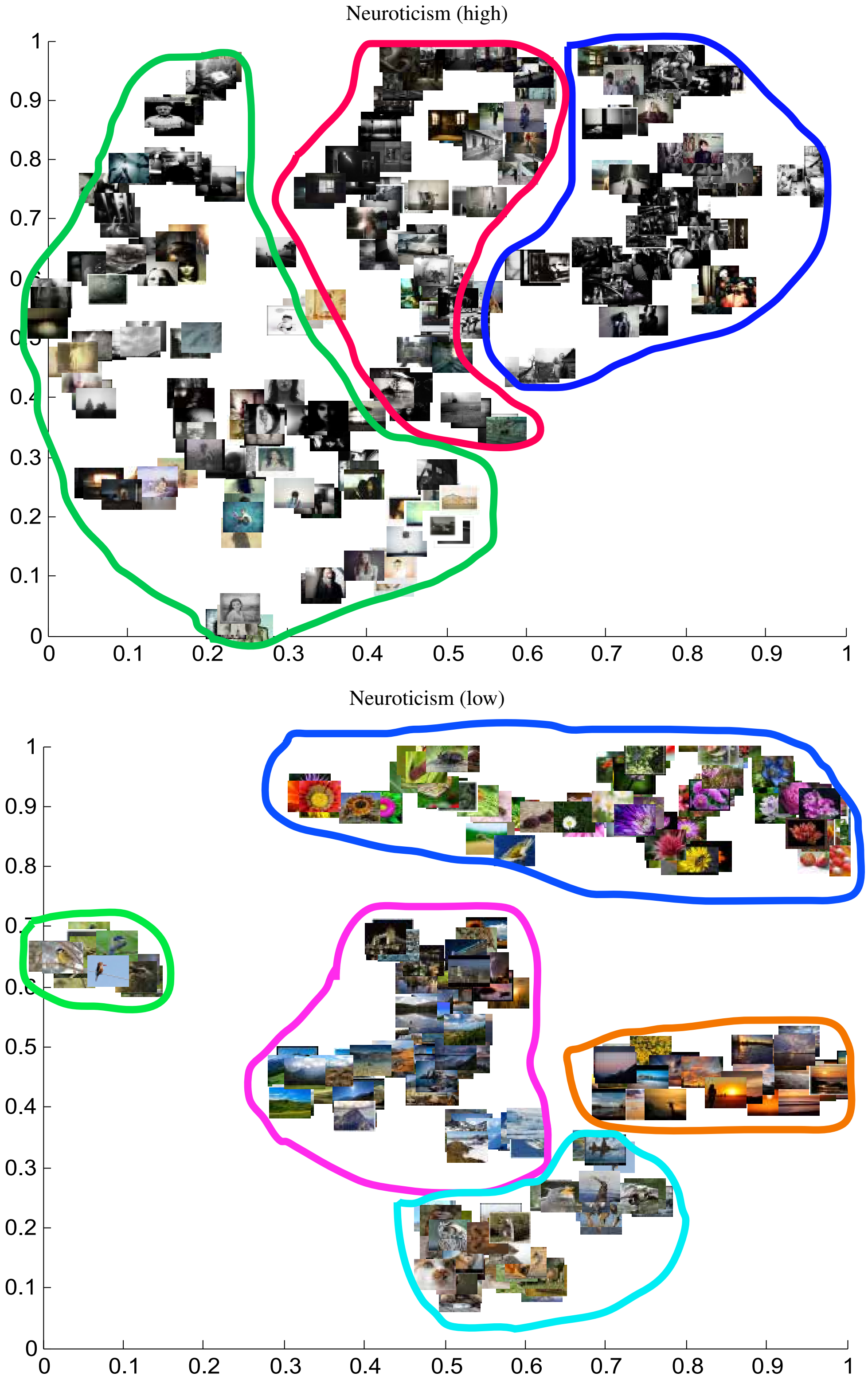
Instead for a low level of Openness (see Figure 10.9 bottom) we found four clusters about flower/plants, felines, outside activity/sport and landscapes.

10.4 Online Demo

As a matter of proof that our proposed method effectively works, we present in this section a possible demo application [1]. To this aim we developed a web interface where a subject can upload an image, a set of images, paste a web address linked to a picture or select an image form a list already stored in the server that he/she likes. The proposed demo loads the Caffe models of the aesthetic preferences related to the attributed traits and it classifies the pictures assigning them to the *low* or *high* class of each trait. If a set of images is provided, then it computes a majority vote to decide the level of the trait. In Figure 10.10, the interface of the demo and the page result are shown.

As we can notice and compare with respect to the pictures in Figure 10.3 - 10.7, the image selected by a possible subject reflects the *high* level of Neuroticism and Extraversion for the dark colors and the presence of a person, and *low* level of Openness, Conscientiousness and Agreeableness as the picture displays a landscape with no regular shape and faded colors.

[1] The demo is available at `http://psychoflickr.di.univr.it:8000/demo/`



Neuroticism (high)

Neuroticism (low)

Fig. 10.8: Projection of the most probable images for Neuroticism trait, on a low-dimensional space.

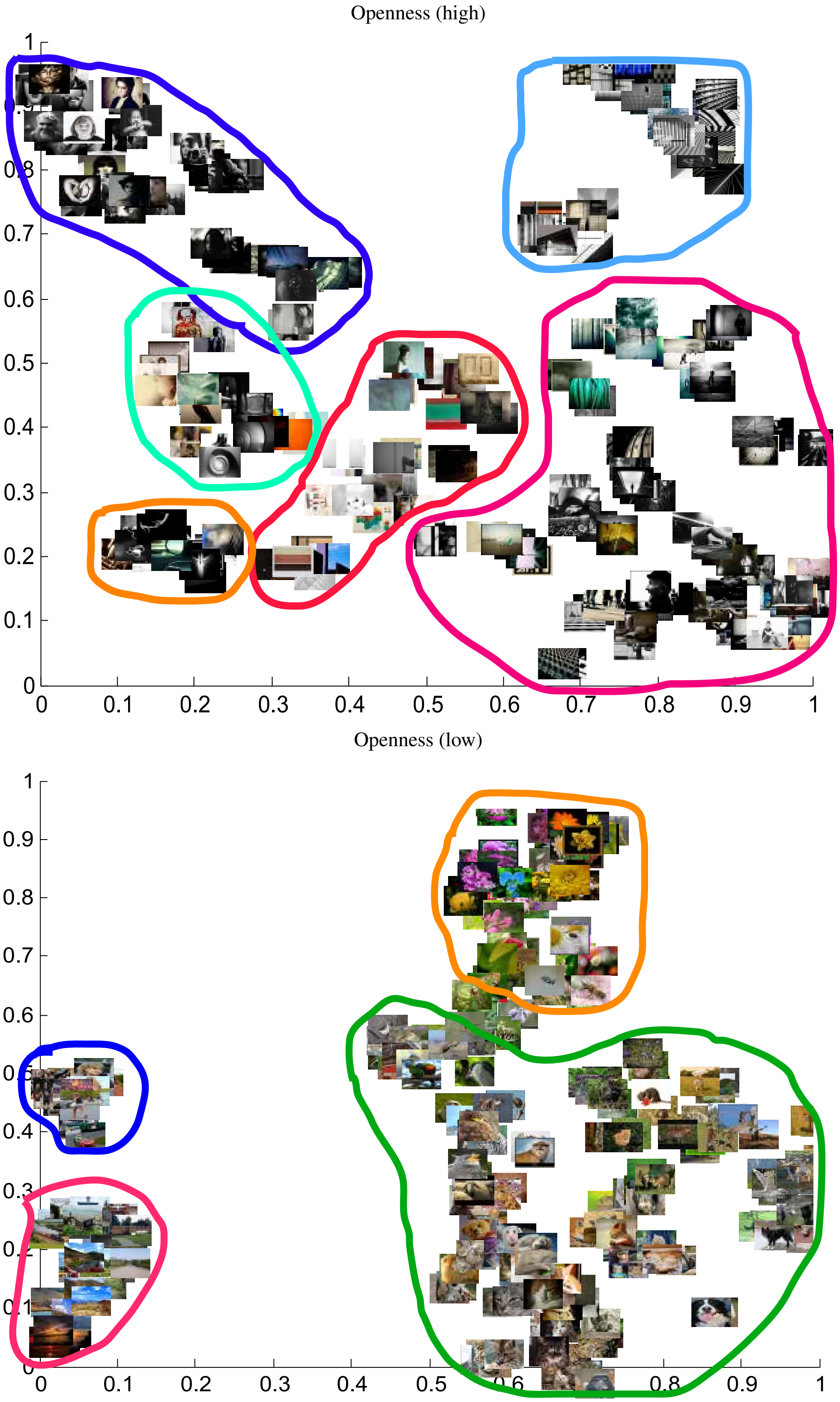
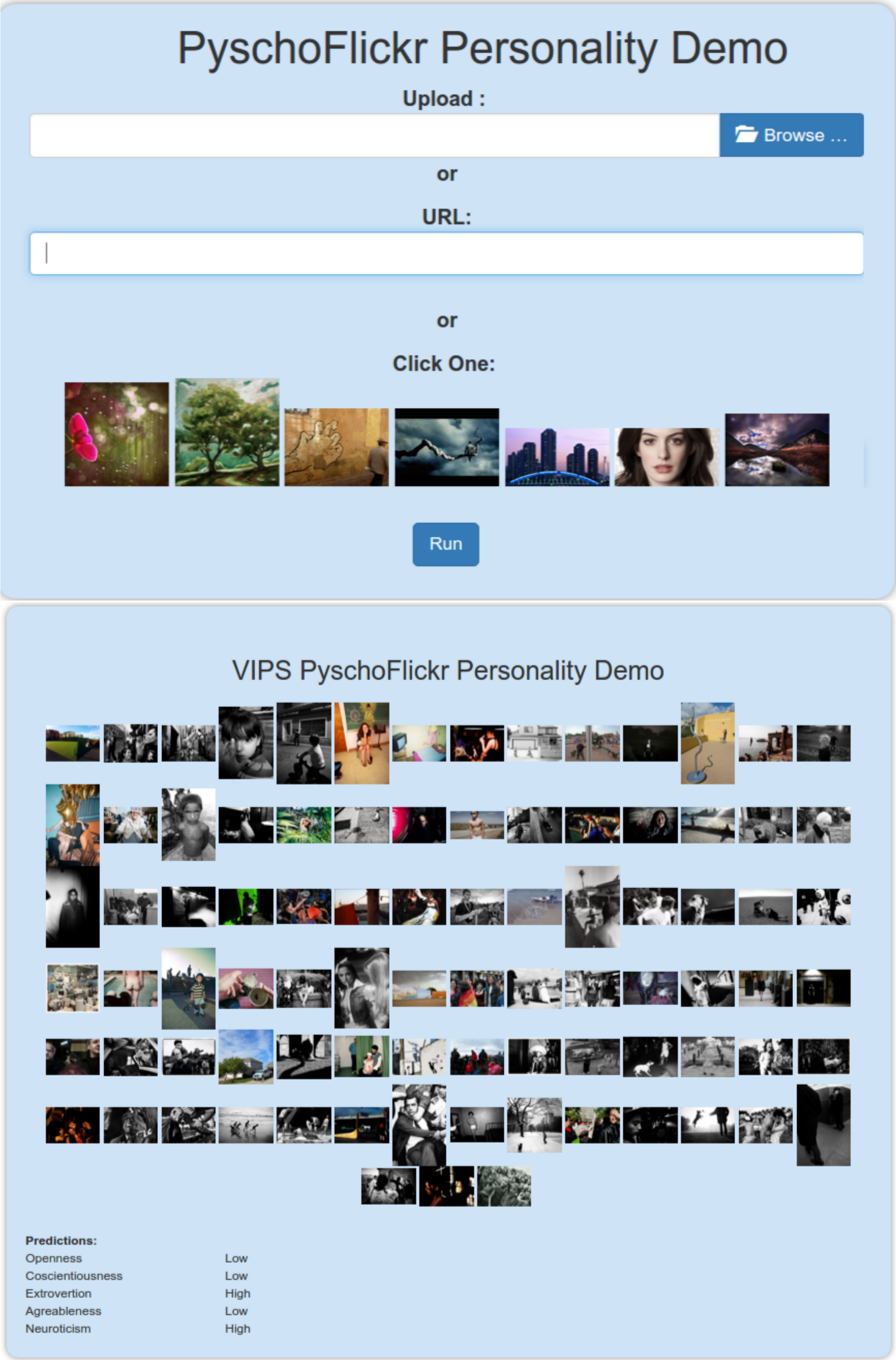


Fig. 10.9: Projection of the most probable images for Openness trait, on a low-dimensional space.


PyschoFlickr Personality Demo
Upload :
Browse …
or
URL:
or
Click One:
Run
VIPS PyschoFlickr Personality Demo

```
Predictions:
Openness                    Low
Coscientiousness            Low
Extrovertion                High
Agreableness                Low
Neuroticism                 High
```

Fig. 10.10: Start (top) and result (bottom) of the web demo application.

Part IV

Conclusions

11

Conclusions and Future Perspectives

Digital images can be understood by computers in many ways: the first, simpler way, amounts to extracting low level information from the pixel values in the form of color histograms, frequency responses etc., and to use it to create a representation in a vectorial space, where tasks of clustering or classification can be carried out [43, 270]. The second way is to extract the semantic content of the image by means of segmentation/classification/detection approaches, and to adopt structured higher-level representations for tasks such as content-based indexing and retrieval [159, 251]. These two directions have shaped the image understanding field of the last 30 years [165], accepting the images as a given representation of diverse objects or scenes, without investigating the authorship of the images. The advent of Internet, the capability of dealing with Big Data, and the diffusion of social media, gave rise to a third way of dealing with images [127, 286]; specifically, images started being associated with *people*: in facts, images are now digital objects that could be easily uploaded *by a certain user* into social platforms often showing excerpts of her personal life.

Both of these activities (uploading and tagging pictures) indicate a substantial revolution in how images are used: from means to represent visual aspects of reality, where the ownership of the photo is neglected, they have become *personal messages*, from the sender (the subject which uploads the photos into a social network, or that selects some shots as favorite) to her receiver(s) (the user of the social network that see the uploaded or the preferred pictures). In this fresh new perspective, uploading or "preferring" images will communicate something, that is, personal messages as the kind of subjects that one may like (cars, landscapes, people) or the life experiences one is going through. But images communicate more than this, and this fact does represent a true revolution in the image understanding field, with a new layer of image interpretation which has started to be unveiled; to explain this new perspective, the sender/receiver communication perspective becomes invaluable.

In face-to-face communications, people share their opinions, experiences and impressions of life by using explicit verbal signals (that are, spoken sentences) and non-verbal signals (for example, by how they deliver the sentences, or by assuming bodily expressions). Many social psychology studies highlight the fundamental importance of both aspects, the verbal content and the non-verbal signals, for the successful exchange of messages.

The rationale behind the focus on nonverbal behavior is twofold. On one hand non-verbal communication (facial expressions, body gestures, posture, vocal behavior, eye gaze etc.) was shown to influence, to a significant extent, the perception of socially relevant characteristics [141]. On the other hand, domains like Affective Computing [217] and Social Signal Processing [287] have shown that nonverbal behavior cues work effectively as an evidence for technologies dealing with emotional and social phenomena. When it comes to technologies, we attributes traits not only to people, but also to machines that exhibit human-like features and behaviors, including robots, embodied conversational agents, animated characters etc. [194].

In particular, these signals trigger unconscious cognitive processes in the receiver, aimed at building a mental image of the sender's character. For example, a dominant person typically takes a high number of floorgrabs and turns [233]. In other words, face-to-face communications are characterized by the exchange of various social signals, such as dominance, social status [287] and personality [285, 286]. Knowledge and awareness of the existence and meaning of these non-verbal signals, both for the sender and for the receiver, determine the success of a communication, while a misunderstanding of these signals makes a communication unclear and difficult. This two-body communication paradigm is modeled by the Brunswick Lens, in the field of social psychology [36]. In the same way that an email, a blog or a chat reveals something of its author, so now the images authored by a person may embed some of her individual traits, such as her personal aesthetic preferences or her personality traits.

This thesis argues that taking into account cognitive effects is possible and it can also improve multimedia approaches. For this purpose we take into account Computational Aesthetics and Social Signal Processing principles under a computational point of view to study the phenomenon.

The key idea of Part II is that the cognitive mechanisms that regulate the appreciation of an image are personal and unique, and that their distillation can provide an interesting soft-biometrical trait. To this aim, we used a typical learning approach, considering images tagged as “favorites” by a certain person as training data, incorporating the expression of her aesthetic preferences. In the experiments we validated this intuition with a consistent amount of pictures taken from Flickr, showing thus that personal aesthetic trait may be collected and managed easily from the social web. This makes “personal aesthetics” a very hot topic for soft biometrics.

To the best of our knowledge, such a perspective has never been adopted in a forensic technology context before. The probable reason is that multimedia data became an interaction channel only recently, when the diffusion of appropriate technologies for data production (cameras, smartphones, tablets, etc.) and consumption (social media, digital libraries, etc.) made it possible to exchange multimedia data as easily as we previously exchanged written material (letters, messages, etc.) [37].

Our work adopted also the counting grid to convincingly combine content and aesthetics for capturing image preferences of users. A novel inference on CG allows to visualize the user tastes as elevation maps over the counting grid manifold. In this case, a generative embedding strategy: a generative step (the mapping on the Counting Grids) is followed by a discriminative step (the SVM training). This way, exploiting the advantages of hybrid generative/discriminative approaches, the compact and interpretable CG representation becomes a feature for a discriminative classifier, resulting in the new state of the art on a dataset of 200 users and 40K images. This work presented also one of the main limitation of our approach, that is, the images selected by the user have to come from a large (potentially infinite) set of images, with no images shared among users; when this assumption holds no more, we have two problems emerging: the first is that the variability of the test signatures irremediably diminishes, as the number of images to select from is smaller, and the second is that the number of images which can be chosen by more than a user is no more negligible. Other limitations of the approach are that we considered only the positive preferences of a given subject, that are what he/she likes, without taking into account the possible negative ones, that are what he/she dislikes or does not like to see or observe in an image, and that, of course, can help to draft a more detailed and accurate model of an individual aesthetic preferences; this suggests that, for a valid biometric application (and not solely a soft biometric one), different (and more structured) multimodal interaction paradigms would be necessary. Finally the content features do not allow to reach high performances in recognition and identification tasks.

The last chapter of the part also showed that with the help of the new machine learning framework dubbed deep learning we overcome one of the limitations of the previous approach. These results lead us to believe that a more abstract, higher-level representation that analyze data in a similar fashion as our brain can help us to better understand what kind of cues are important when dealing with aesthetic preferences of individuals. At the state-of-the-art the system gives 97% of probability of guessing the correct user using 5 preferred images as biometric template; as for the verification capability, the equal error rate is 0.11

Our works could be interesting for several aims: apart from the design of a classical biometric system, where personal aesthetics may support other stronger biometrical cues, other applications can be thought of: for example, exploiting eye tracking devices, the spatio-temporal patterns with which preferred images are explored may enrich and reinforce the biometrical traits. Even more pioneering, analyzing the subjective preferences of a person may unveil the interplay between personality traits and image features (see Part III), which can bring in ethical aspects ignored so far.

Part III proposes experiments aimed at mapping aesthetic preferences into personality traits. To this purpose the Brunswick Lens model has been customized for this new kind of communication by images: in this new setting, personality traits have been considered as the social signals sent with the uploaded images, and whose inference is one of the most intriguing challenges. In this respect, the work of Chapter 9 focused on inferring with a regressor the real personality traits of the sender (collected by self-assessed tests), but also those traits that unacquainted people (the *assessors*) associate with the sender by looking at her images. In particular, we have shown that low-level features, automatically extracted from images tagged as favorite, allow one to infer the personality traits of Flickr users. The experiments have required the development of new regression approaches based on Multiple Instance Regression. The adoption of topic models to represent the instances of a bag has been the main advancement with respect to the state-of-the-art. One of the new proposed approaches, the Multiple Instance Generative LDA, has achieved the highest overall performance reaching a correlation up to 0.68 between actual and predicted traits. This work showed also that the assessors' evaluations were 1) consistently similar, 2) in partial disagreement with the self-assessed evaluations, and 3) more easily predicted by machine learning techniques. In other words, the act of sharing images online may evoke a common psychological response in the receiving crowd, and this can be reasonably predicted by automatic approaches. Thus, it is possible to build a *wisdom of the crowds* model of personality profiles from collections of images, based on the impressions these may generate on a general hypothetical audience.
These results pave the way to algorithms that, given a certain image upload, will tell the personality profile of the sender and which kind of impressions he would generate about himself in a general audience. In a way, this amounts to learn the wisdom of crowd.

There are several directions for future work. The results of this work 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 [302] or to infer the content of videos and pictures from the behavioural reactions of people that watch them [209]. 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*" [218]. 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 [286].

In Chapter 10 we examined the problem of relating a set of image preferences to personality traits by using a deep learning framework. We cast this recently introduced application problem as a new level of image understanding that enhances the role of images through considerations on the social aspects of contemporary online activities. The role of social platforms like Flickr, Facebook, Instagram, etc., in building online social personas where most activities are shared to a wide audience creates a unique opportunity to study image-based activities (like authoring, uploading and preferring images) as social messages, embedded with characteristics resembling verbal and non-verbal signals in face-to-face spoken communications. Moreover, the presence of an author-audience paradigm imbues the messages with an extra layer of significance: there is the communication intended by the author, and there is the communication assumed by the audience. Thus, our problem becomes inferring both self-assessed and attributed personality traits, and our results reinforce previous work in considering the latter easier to approach than the former. Overall, we demonstrate the viability of using a recent, powerful methodology like Convolutional Neural Networks, in tackling these new types of image understanding applications. The experimental results are promising, outperform previous results and point towards a mixture of stylistic aspects (typical of computational aesthetics) and content-based aspects (typical of object detectors) as crucial for building reliable predictors.

We believe there are many impactful rewards for this type of research: an immediate application might be providing social networks with tools to soft-profile users, suggesting compatible users to connect with, or indicating the most suitable groups to join. It could also be used as a marketing evaluation tool to help predict the impact of a set of images on an hypothetical audience of customers. Our online demo is a step along this direction, evidence that this new kind of image understanding is not only a mere academic research endeavor, but a potential groundbreaking market application.

From an application point of view, the main interest of this work is twofold. On the one hand - whether it is an opportunity ("*Personal data is the new oil of the Internet and the new currency of the digital world*" [85]) or a risk of intrusion in private life [143] - the experiments show that aesthetic preferences allow the inference of data that people do not necessarily intend to share (the self-assessed traits in this case). On the other hand, the experiments show that it is possible to predict the personality impression that people convey to unacquainted others through aesthetic preferences expressed online. This problem is going to become increasingly more important if it is true that online "*many meet their social, emotional and economic needs*" [227]. A typical example is when potential employers *google* candidates before an interview, a widespread practice where impressions based on online traces make the difference between getting a job or not [54]. Overall, 67% of the US Internet users access at least one social networking platform [70]. Therefore, managing our digital impression seems to emerge as a new social skill, as important as taking care of appearance in everyday life. Understanding the interplay between online activities and personality attribution is a step in such a direction.

Future work can involve the improvement of technical aspects (e.g., feature extraction and inference approaches) and the extension of the methodologies to other online activities. Furthermore, the results of this work open the possibility of investigating whether digital appearance can work as a social signal, i.e. an act capable of engaging and involving others in an interaction, possibly leading to the same effects we observe in face-to-face social exchanges like, e.g., the *halo effect* (the tendency to attribute socially desirable characteristics to attractive people) [199] or the *similarity attraction effect* (the tendency to like people similar to us) [39]. Understanding the phenomena above might lead, e.g., to better social media strategies, the production of viral content, or the development of better interfaces.

Others possible future perspectives follow:

- *Social Network*: an immediate application might be providing social networks with tools to soft-profile users, suggesting compatible users to connect with, or indicating the most suitable groups to join using also other demographic information.
- *Marketing*: the results of this thesis could also be used as a marketing evaluation tool to help predict the impact of a set of images on an hypothetical audience of customers and tailor personalized advertising.

- *Viral marketing*: the "*diffusion of information about the product and its adoption over the network*" [155], is an advertisement technique aimed at spreading information as widely as possible through (mostly online) *word of mouth* mechanisms. As previous stated the exchange of multimedia data, being a form of human-human communication, can be thought of as a form of word of mouth. Therefore, implicit cognitive processes might contribute to explain and enhance virality. In the same vein, communication strategies based on social media can benefit from the prediction of perceptual judgments likely to be attributed to a given multimedia message diffused through online social platforms [134].
- *Big Data*: the perspective adopted in this thesis can be useful in Big Data Analytics - the domain aimed at making sense of large amounts of unstructured data [184] - one of the most important challenges technology faces today. In particular, there is consensus among Big Data experts that no useful information can be extracted from large databases without associating automatic mining approaches and human interpretation [205]. This latter is likely to be influenced by cognitive processes similar to those illustrated in the experiments of this thesis.
- *Social Science*: this thesis open the possibility of investigating whether digital appearance can work as a social signal, i.e. an act capable of engaging and involving others in an interaction, possibly leading to the same effects we observe in face-to-face social exchanges like, e.g., the halo effect (the tendency to attribute socially desirable characteristics to attractive people) or the similarity attraction effect (the tendency to like people similar to us).Understanding the phenomena above might lead, e.g., to better social media strategies, the production of viral content, or the development of better interfaces.
- *Virtual agents*: improve machines that exhibit human-like features and behavior like robots, animated character, embodied conversation agents.
- *Neuroscience*: understand the role/ activation of neurons when appreciating images or the brain activity occurring during human-human interactions showing social signals.
- *Online Games*: a peculiar form of communication through multimedia material is the participation in online games where several participants interact via avatars or animated characters. The choice of a particular character or particular gaming strategies and options is likely to convey information about the player "states" (see, *e.g.*, the approaches in [128, 300, 303] for the case of personality). In a similar way, computer mediated communication can be influenced by implicit cognitive processes via interface characteristics like the profile picture of Skype users.
- *Social Surveillance*: we introduced an important problem raised by the use of social network and multimedia content in our everyday life. One of the main worries about social networking platforms is that every trace we leave online might reveal more than what we think and possibly disclose information we prefer to keep private. This can be used for social surveillance purposes.
- *Generative Art*: use the findings of this works to improve algorithms and techniques used in generative and evolutionary art [173], as well as consididering cognitive processes to embedded within them.

We hope that this thesis can be of inspiration to many researchers and engineers for further improvements and extensions, aiming at discovering more reliable solutions and models of people's visual preferences considering as well, other technologies, social interactions and behaviors, multimedia data and interdisciplinary fields.

References

[1] American psychological association (2013). glossary of psychological terms.

[2] *Bags of words Models of empitome sets: HIV viral load regression with counting grids*. World Scientific Publishing, 2014.

[3] Brett Adams. Where does computational media aesthetics fit? *IEEE Multimedia*, 10:18–27, 2003.

[4] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

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

[6] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[7] Michael Argyle. *The Psychology of Interpersonal Behaviour*. Penguin books. Penguin Adult, 1994.

[8] B. Babenko. Multiple instance learning: Algorithms and applications.

[9] Shuotian Bai, Tingshao Zhu, and Li Cheng. Big-five personality prediction based on user behaviors at social network sites. Technical report, Cornell University, 2012.

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

[11] Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 255–262. ACM, 2011.

[12] Christian Bauckhage and Kristian Kersting. Can computers learn from the aesthetic wisdom of the crowd? *KI-Künstliche Intelligenz*, 27(1):25–35, 2013.

[13] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, February 2013.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[15] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising autoencoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013.

[16] Alexander C Berg, Tamara L Berg, Hal Daume III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3562–3569. IEEE, 2012.

[17] Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. University of California Press, 1991.

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

[19] I. Biederman and E. Vessel. Perceptual pleasure and the brain. *American Scientist*, 94(3):1–8, 2006.

[20] Joan-Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013.

[21] Joan-Isaac Biel and Daniel Gatica-Perez. Mining crowdsourced first impressions in online social video. *IEEE Transactions on Multimedia*, 16(7):2062–2074, 2014.

[22] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. Hi youtube!: Personality impressions and verbal content in social video. In *International Conference on Multimodal Interaction*, pages 119–126. ACM, 2013.

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

[24] G.D. Birkhoff. *Aesthetic measure*. Harvard University Press, 1933.

[25] A Peterson Bishop, Nancy A Van House, and BP Buttenfields. Digital library use. *Social Practice in Design and Evaluation*, 2003.

[26] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[27] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[28] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Scene classification via plsa. In *Proceedings of the European Conference on Computer Vision*, pages 517–530. Springer, 2006.

[29] Anna Bosch, Andrew Zisserman, and Xavier Muoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[30] Guillaume Bouchard and Bill Triggs. The tradeoff between generative and discriminative classifiers. In *Proceedinng on the International Symposium on Computational Statistics*, pages 721–728, 2004.

[31] Gregory J Boyle and Edward Helmes. Methods of personality assessment. 2009.

[32] Alessandro Bozzon, Marco Brambilla, and Stefano Ceri. Answering search queries with crowdsearcher. In *Proceedings of the ACM International Conference on World Wide Web*, pages 1009–1018, 2012.

[33] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[34] P.M. Bronstad and R. Russell. Beauty is in the "we" of the beholder: greater agreement on facial attractiveness among close relations. *Perception*, 36(11):1674–1681, 2007.

[35] James Dean Brown. *Testing in language programs*. Prentice Hall Regents New Jersey, 1996.

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

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

[38] M.J. Burge and K. Bowyer. *Handbook of Iris Recognition*. Springer, 2013.

[39] D. Byrne, W. Griffitt, and D. Stefaniak. Attraction and similarity of personality characteristics. *Journal of Personality and Social Psychology*, 5(1):82, 1967.

[40] Allan Campbell, Vic Ciesielksi, and AK Qin. Feature discovery by deep learning for aesthetic analysis of evolved abstract images. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, pages 27–38. Springer, 2015.

[41] Arnett Campbell, Vic Ciesielski, and Karen Trist. A self organizing map based method for understanding features associated with high aesthetic value evolved abstract images. In *IEEE Congress on Evolutionary Computation*, pages 2274–2281. IEEE, 2014.

[42] Joao Carreira and Cristian Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1312–1328, 2012.

[43] Chad Carson, Megan Thomas, Serge Belongie, Joseph M Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Visual Information and Information Systems*, pages 509–517. Springer, 1999.

[44] F. Celli. Unsupervised personality recognition for social network sites. In *Proceeedings of the International Conference on Digital Society*, pages 59–62, 2012.

[45] 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.

[46] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[47] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *International Symposium on Wearable Computers*, pages 29–36. IEEE, 2011.

[48] Vic Ciesielski, Perry Barile, and Karen Trist. *Finding image features associated with high aesthetic value by machine learning*. Springer, 2013.

[49] S. Cloninger. Conceptual issues in personality theory. In P.J. Corr and G. Matthews, editors, *The Cambridge handbook of personality psychology*, pages 3–26. Cambridge University Press, 2009.

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

[51] Philip J Corr and Gerald Matthews. *The Cambridge handbook of personality psychology*. Cambridge University Press Cambridge, 2009.

[52] P.T. Costa, R.R. MacCrae, and Inc Psychological Assessment Resources. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual*. The SAGE handbook of personality theory and assessment, 1992.

[53] Paul T Costa Jr and Robert R McCrae. Domains and facets: Hierarchical personality assessment using the revised neo personality inventory. *Journal of Personality Assessment*, 64(1):21–50, 1995.

[54] D. Coutu. We googled you. *Harvard Business Review*, 85(6):1–8, 2007.

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

[56] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[57] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Proceedings of the European Conference on Computer Vision*, volume 1, pages 1–2, 2004.

[58] William Curran, Travis Moore, Todd Kulesza, Weng-Keen Wong, Sinisa Todorovic, Simone Stumpf, Rachel White, and Margaret Burnett. Towards recognizing cool: can end users help computer vision recognize subjective attributes of objects in images? In *Proceedings of the ACM international conference on Intelligent User Interfaces*, pages 285–288. ACM, 2012.

[59] Florin Cutzu, Riad Hammoud, and Alex Leykin. Estimating the photorealism of images: Distinguishing paintings from photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–305. IEEE, 2003.

[60] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[61] Antitza Dantcheva, Carmelo Velardo, Angela D'angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011.

[62] 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*, pages 288–301, 2006.

[63] Ritendra Datta. *Semantics and Aesthetics Inference for Image Search: Statistical Learning Approaches*. PhD thesis, University Park, PA, USA, 2009.

[64] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[65] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[67] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1664, 2011.

[68] A.P. Dijksterhuis and J. A. Bargh. The perception - behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, 33:1–40, 2001.

[69] Daniel R. Dooly, Qi Zhang, Sally A. Goldman, and Robert A. Amar. Multiple instance learning of real valued data. *The Journal of Machine Learning Research*, 3:651–678, 2003.

[70] Maeve Duggan and Lee Rainie. Cell phone activities 2012. *Pew Research Center*, 2012.

[71] A.H. Eagly, R.D. Ashmore, M.G. Makhijani, and L.C. Longo. What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin*, 110(1):109, 1991.

[72] P. Eckersley. How unique is your web browser? In *Proceeding of Privacy Enhancing Technologies*, pages 1–18, 2010.

[73] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, and Klaus Scherer. Relative importance of face, body, and speech in judgments of personality and affect. *Journal of Personality and Social Psychology*, 38(2):270, 1980.

[74] Arkady Epshteyn and Gerald DeJong. Generative prior knowledge for discriminative classification. *Jorunal of Artificial Intellegence Research*, 27:25–53, 2006.

[75] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660, 2010.

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

[77] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[78] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531. IEEE, 2005.

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

[80] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[81] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[82] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from google's image search. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1816–1823. IEEE, 2005.

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

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

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

[86] Brendan J. Frey and Nebojsa Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transaction on Pattern Analalysis Machine Intelligence*, 27(9):1392–1416, 2005.

[87] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[88] David C Funder. On the accuracy of personality judgment: a realistic approach. *Psychological review*, 102(4):652, 1995.

[89] David C. Funder. Personality. *Annual Review of Psychology*, 52(1):197–221, 2001.

[90] Adrian Furnham and Margaret Avison. Personality and preference for surreal paintings. *Personality and Individual Differences*, 23(6):923 – 935, 1997.

[91] Adrian Furnham and John Walker. The influence of personality traits, previous experience of art, and demographic variables on artistic preference. *Personality and Individual Differences*, 31(6):997–1017, 2001.

[92] Adrian Furnham and John Walker. Personality and judgements of abstract, pop art, and representational paintings. *Journal of Personality*, 15:57–72, 2001.

[93] P. Galanter. Computational aesthetic evaluation: Past and future. In J. McCormack and M. d'Inverno, editors, *Computers and Creativity*, pages 255–293. Springer, 2012.

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

[95] Alastair J Gill, Scott Nowson, and Jon Oberlander. What are they blogging about? personality, topic and motivation in blogs. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2009.

[96] Shiry Ginosar, Daniel Haas, Timothy Brown, and Jitendra Malik. Detecting people in cubist art. In *Proceedings of the European Conference on Computer Vision*, pages 101–116. Springer, 2014.

[97] Mark Girolami and Ata Kabán. On an equivalence between plsi and lda. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 433–434. ACM, 2003.

[98] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[99] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *IEEE Third International Conference on Privacy, Security, Risk and Trust and Social Computing*, pages 149–156. IEEE, 2011.

[100] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems*, pages 253–262, 2011.

[101] Lewis R Goldberg. The structure of phenotypic personality traits. *American psychologist*, 48(1):26, 1993.

[102] Enhao Gong. Deep aesthetic learning.

[103] Samuel D Gosling, Sam Gaddis, Simine Vazire, et al. Personality impressions based on facebook profiles. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 7:1–4, 2007.

[104] Samuel D Gosling, Sei Jin Ko, Thomas Mannarelli, and Margaret E Morris. A room with a cue: personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82(3):379, 2002.

[105] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the ACM workshop on Privacy in the Electronic Society*, pages 71–80. ACM, 2005.

[106] Chunhui Gu, Jasmine J Lim, Pablo Arbeláez, and Jagannath Malik. Recognition using regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1030–1037. IEEE, 2009.

[107] Rosanna E Guadagno, Bradley M Okdie, and Cassie A Eno. Who blogs? personality predictors of blogging. *Computers in Human Behavior*, 24(5):1993–2004, 2008.

[108] Martin Guha. Oxford dictionary of psychology. *Reference Reviews*, 20(6):16–16, 2006.

[109] Sharath Chandra Guntuku, Joey Tianyi Zhou, Sujoy Roy, Lin Weisi, and Ivor W Tsang. Deep representations to model user 'likes'. In *Proceedings of the Asian Conference on Computer Vision*, pages 3–18. Springer, 2015.

[110] Kilem Li Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* Advanced Analytics Press, 2012.

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

[112] GG Harrap. Alphonse bertillon: Father of scientific detection. *New York: Abelard-Schuman*, 1956.

[113] Albert H Hastorf, David J Schneider, and Judith Polefka. Person perception. 1970.

[114] Geoffrey Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.

[115] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[116] Florian Hoenig. Defining computational aesthetics. In *Proceedings of the Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, pages 13–18. Eurographics Association, 2005.

[117] David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[118] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–152. IEEE, 2011.

[119] Tommi Jaakkola, David Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, pages 487–493, 1999.

[120] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Can soft biometric traits assist user recognition? In *Defense and Security*, pages 561–572. International Society for Optics and Photonics, 2004.

[121] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *Biometric Authentication*, pages 731–738. Springer, 2004.

[122] Anil K Jain, Robert PW Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[123] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of biometrics*. Springer Science & Business Media, 2007.

[124] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2005.

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

[126] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[127] Xin Jin, Chi Wang, Jiebo Luo, Xiao Yu, and Jiawei Han. Likeminer: a system for mining the power of 'like' in social media networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 753–756. ACM, 2011.

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

[129] M.H. Johnson. Subcortical face processing. *Nature Reviews Neuroscience*, 6(10):766–774, 2005.

[130] Nebojsa Jojic, Nemanja Petrovic, Brendan J Frey, and Thomas S Huang. Transformed hidden markov models: Estimating mixture models of images and inferring spatial transformations in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 26–33. IEEE, 2000.

[131] A Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, 14:841, 2002.

[132] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

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

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

[135] R. Kaplan and S. Kaplan. *The Experience of Nature: A Psychological Perspective*. Cambridge University Press, 1989.

[136] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013.

[137] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 419–426. IEEE Computer Society, 2006.

[138] D.A. Kenny. PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, 8(3):265–280, 2004.

[139] D.A. Kenny, L. Albright, T.E. Malloy, and D.A. Kashy. Consensus in interpersonal perception: acquaintance and the Big Five. *Psychological Bulletin*, 116(2):245–258, 1994.

[140] John F Kihlstrom, Terrence M Barnhardt, Douglas J Tataryn, et al. Implicit perception. *Perception without awareness: Cognitive, clinical, and social perspectives*, pages 17–54, 1992.

[141] Mark Knapp, Judith Hall, and Terrence Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[142] M. Kosinski and D. Stillwell. mypersonality research wiki. [online]. available: http://mypersonality.org/wiki.

[143] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

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

[145] Klaus Krippendorff. *Content Analysis. An Introduction to its Methodology, 3rd ed.* Thousand Oaks, CA: Sage Publications Inc., 2013.

[146] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[147] Ziva Kunda. *Social cognition: Making sense of people*. MIT press, 1999.

[148] Quoc V Le. Building high-level features using large scale unsupervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8595–8598. IEEE, 2013.

[149] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10), 1995.

[150] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[151] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[152] Yann LeCun et al. Lenet-5, convolutional neural networks. *Internet: http://yann. lecun. com/exdb/lenet*, 2013.

[153] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 253–256. IEEE, 2010.

[154] H. Leder, B. Belke, A. Oeberst, and D. Augustin. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4):489–508, 2004.

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

[156] Joseph Lev. The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20(1):125–126, 1949.

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

[158] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009.

[159] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems*, pages 1378–1386, 2010.

[160] Wentian Li. Dna segmentation as a model selection process. In *Proceedings of the annual International Conference on Computational Biology*, pages 204–210. ACM, 2001.

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

[162] David Liu, Datong Chen, and Tsuhan Chen. Latent layout analysis for discovering objects in images. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 468–471. IEEE, 2006.

[163] David Liu and Tsuhan Chen. Unsupervised image categorization and object localization using topic models and correspondences between images. In *Proceedings of the International Conference on Computer Vision*, pages 1–7. IEEE, 2007.

[164] H. Liu. Social network profiles as taste performances. *Journal of Computer-Mediated Communication*, 13(1):252–275, 2007.

[165] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[166] PH Lodhi, Savita Deo, and Vivek M Belhekar. The five-factor model of personality. In *The five-factor model of personality across cultures*, pages 227–248. Springer, 2002.

[167] 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.

[168] Pietro Lovato, Manuele Bicego, Cristina Segalin, Alessandro Perina, Nicu Sebe, and Matteo Cristani. Faved! biometrics: Tell me which image you like and i'll tell you who you are. *IEEE Transactions on Information Forensics and Security*, 9(3):364–374, 2014.

[169] Steve Love and Joanne Kewley. Does personality affect peoples' attitude towards mobile phone use in public places? In *Mobile Communications*, pages 273–284. Springer, 2005.

[170] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[171] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 457–466. ACM, 2014.

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

[173] Penousal Machado and Amílcar Cardoso. Generation and evaluation of artworks. In *Proceedings of the European Workshop on Cognitive Modeling*, volume 96, pages 96–39, 2010.

[174] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. MM '10, pages 83–92, 2010.

[175] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Eesearch*, pages 457–500, 2007.

[176] Giovanni Malerba, Linda Schaeffer, Luciano Xumerle, Norman Klopp, Elisabetta Trabetti, Michele Biscuola, Ugo Cavallari, Roberta Galavotti, Nicola Martinelli, Patrizia Guarini, et al. SNPs of the FADS gene cluster are associated with polyunsaturated fatty acids in a cohort of patients with cardiovascular disease. *Lipids*, 43(4):289–299, 2008.

[177] D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar. *Handbook of Fingerprint recognition*. Springer, second edition, 2009.

[178] 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*, pages 1784–1791, 2011.

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

[180] K.V. Mardia and P.E. Jupp. *Directional Statistics*. Wiley, 2009.

[181] C. Martindale, K. Moore, and J. Borkum. Aesthetic preference: Anomalous findings for berlyne's psychobiological theory. *American Journal of Psychology*, 103(1):53–80, 1990.

[182] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.

[183] IC McManusU, Amanda L Jones, and Jill Cottrell. The aesthetics of colour. *Perception*, 10:651–666, 1981.

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

[185] Monika Mital, D. Israel, and Shailja Agarwal. Information exchange and information disclosure in social networking web sites: Mediating role of trust. *Learning Organization*, 17(6):479–490, 2010.

[186] Gelareh Mohammadi and Alessandro Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. 2012.

[187] Gelareh Mohammadi, Alessandro Vinciarelli, and Marcello Mortillaro. The voice of personality: mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the International Workshop on Social Signal Processing*, pages 17–20. ACM, 2010.

[188] H. Moon and P.J. Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30:303 – 321, 2001.

[189] Margaret E Morris, Carl S Marshall, Mira Calix, Murad Al Haj, James S MacDougall, and Douglas M Carmean. Pixee: pictures, interaction and emotional expression. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 2277–2286. ACM, 2013.

[190] Gordon B Moskowitz. *Social cognition: Understanding self and others*. Guilford Press, 2005.

[191] Carol A. Mullen. The media equation: How people treat computers, television, and new media like real people and places. *International Journal of Instructional Media*, 26(1):117, 1999.

[192] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[193] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

[194] C.I. Nass and S. Brave. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, 2005.

[195] Clifford Nass and Kwan Min Lee. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171, 2001.

[196] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 689–696, 2011.

[197] 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*, pages 566–569, 2011.

[198] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *Workshop on Machine Learning for Information Filtering*, volume 1, pages 61–67, 1999.

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

[200] Tianhua Niu, Zhaohui S Qin, Xiping Xu, and Jun S Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 70(1):157–169, 2002.

[201] Scott Nowson. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *In Proceedings of the International Conference on Weblogs and Social*. Citeseer, 2007.

[202] L. Olejnik, C. Castelluccia, and A. Janc. Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. In *Proceedings of Workshop on Hot Topics in Privacy Enhancing Technologies*, 2012.

[203] Stephen Olejnik and James Algina. Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, 8(4):434, 2003.

[204] Daniel Olguın Olguın, Peter A Gloor, and Alex Sandy Pentland. Capturing individual and group behavior with wearable sensors. In *Proceedings of the AAAI Spring Symposium on Human Behavior Modeling*, volume 9, 2009.

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

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

[207] Christopher Y. Olivola and Alexander Todorov. Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, 34(2):83–110, 2010.

[208] Daniel J Ozer and Veronica Benet-Martinez. Personality and the prediction of consequential outcomes. *Annual Review Psychology*, 57:401–421, 2006.

[209] M. Pantic and A. Vinciarelli. Implicit human-centered tagging. *IEEE Signal Processing Magazine*, 26(6):173–180, 2009.

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

[211] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[212] Alex Pentland. Social signal processing [exploratory dsp]. *IEE Signal Processing Magazine*, 24(4):108–111, 2007.

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

[214] Alessandro Perina, Marco Cristani, Umberto Castellani, Vittorio Murino, and Nebojsa Jojic. A hybrid generative/discriminative classification framework based on free-energy terms. In *Proceedings of the International Conference on Computer Vision*, pages 2058–2065. IEEE, 2009.

[215] Katherine Peterson and Katie A Siek. Analysis of information disclosure on a social networking site. In *Online Communities and Social Computing*, pages 256–264. Springer, 2009.

[216] F. Pianesi. Searching for personality [social sciences]. *IEEE Signal Processing Magazine*, 30(1):146–158, 2013.

[217] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[218] I. Poggi and F. D'Errico. Social signals: a framework in terms of goals and beliefs. *Cognitive Processing*, 13(2):427–445, 2012.

[219] Tim Polzehl, Sebastian Moller, and Florian Metze. Automatically assessing personality from speech. In *IEEE Fourth International Conference on Semantic Computing*, pages 134–140. IEEE, 2010.

[220] M. Pusara and C.E. Brodley. User re-authentication via mouse movements. In *ACM Workshop on Visualization and Data Mining for Computer Security*, pages 1–8. ACM, 2004.

[221] Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718, 2012.

[222] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *IEEE Inernational Conference on Social Computing Privacy, Security, Risk and Trust*, pages 180–185. IEEE, 2011.

[223] Daniele Quercia, Renaud Lambiotte, David Stillwell, Michal Kosinski, and Jon Crowcroft. The personality of popular facebook users. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 955–964. ACM, 2012.

[224] Daniele Quercia, Diego B Las Casas, Joao Paulo Pesce, David Stillwell, Michal Kosinski, Virgilio Almeida, and Jon Crowcroft. Facebook and privacy: The balancing act of personality, gender, and relationship currency. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2012.

[225] Mika Raento, Antti Oulasvirta, and Nathan Eagle. Smartphones an emerging tool for social scientists. *Sociological methods & research*, 37(3):426–454, 2009.

[226] Harrison Rainie and Barry Wellman. *Networked: The new social operating system.* Mit Press Cambridge, MA, 2012.

[227] L. Rainie, J. Brenner, and K. Purcell. Photos and videos as social currency online. Technical report, Pew Research Center, 2012.

[228] B. Rammstedt and O.P.P 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*, 41(1):203–212, 2007.

[229] Nikhil Rasiwasia and Nuno Vasconcelos. Latent dirichlet allocation models for image classification. 2013.

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

[231] Byron Reeves and Clifford Nass. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.

[232] Daniel A Reid and Mark S Nixon. Using comparative human descriptions for soft biometrics. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6. IEEE, 2011.

[233] Rutger Rienks and Dirk Heylen. Dominance detection in meetings using easily obtainable features. In *Machine Learning for Multimodal Interaction*, pages 76–86. Springer, 2006.

[234] Salah Rifai, Yoshua Bengio, Yann Dauphin, and Pascal Vincent. A generative process for sampling contractive auto-encoders. *arXiv preprint arXiv:1206.6434*, 2012.

[235] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the International Conference on Machine Learning*, pages 833–840, 2011.

[236] Daniel Rosenberg. 1 data before the fact. 2013.

[237] Robert Rosenthal. Conducting judgment studies: Some methodological issues. *The new handbook of methods in nonverbal behavior research*, pages 199–234, 2005.

[238] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[239] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 1988.

[240] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, April 2015.

[241] M. Rybnik, M. Tabedzki, and K. Saeed. A keystroke dynamics based system for user identification. In *International Conference on Computer Information Systems and Industrial Management Applications*, pages 225–230, 2008.

[242] A.A. Salah, H. Hung, O. Aran, H. Gunes, and M. Turk. Behavior understanding for arts and entertainment. *ACM Transactions on Interactive Intelligent Systems*, 5(3):12:1–12:10, 2015.

[243] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[244] Ruslan Salakhutdinov, Joshua B Tenenbaum, and Antonio Torralba. Learning with hierarchical-deep models. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 35(8):1958–1971, 2013.

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

[246] K. R. Scherer. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, 8:467–487, 1978.

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

[248] Johann Schrammel, Christina Köffel, and Manfred Tscheligi. How much do you tell?: information disclosure behaviour indifferent types of online communities. In *Proceedings of the International Conference on Communities and Technologies*, pages 275–284. ACM, 2009.

[249] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob Van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al. The interspeech 2012 speaker trait challenge. In *INTERSPEECH*, 2012.

[250] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[251] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

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

[253] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[254] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2012.

[255] Ruchir Srivastava, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim. Don't ask me what i'm like, just watch and listen. In *Proceedings of the ACM International Conference on Multimedia*, MM '12, pages 329–338, New York, NY, USA, 2012. ACM.

[256] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the ACM Conference on Ubiquitous Computing*, pages 321–330. ACM, 2012.

[257] Hsiao-Hang Su, Tse-Wei Chen, Chieh-Chi Kao, Winston H. Hsu, and Shao-Yi Chien. Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features. *IEEE Transactions on Multimedia*, 14(3-2):833–843, 2012.

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

[259] John Suler. Image, word, action: Interpersonal dynamics in a photo-sharing community. *CyberPsychology & Behavior*, 11(5):555–560, 2008.

[260] John Suler. Photographic psychology: Image and psyche, 2012.

[261] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *TSMC*, 8(6), 1978.

[262] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[263] Dogancan Temel and Ghassan AlRegib. A comparative study of computational aesthetics. In *IEEE International Conference on Image Processing*, pages 590–594. IEEE, 2014.

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

[265] Howard E Tinsley and David J Weiss. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22(4):358, 1975.

[266] David Rawlings Fionnuala Twomey, Elizabeth Burns, and Sharon Morris. Personality, creativity, and aesthetic preference: Comparing psychoticism, sensation seeking, schizotypy, and openness to experience. *Empirical Studies of the Arts*, 16(2):153–178, 1998.

[267] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[268] James S Uleman, S Adil Saribay, and Celia M Gonzalez. Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review Psychology*, 59:329–360, 2008.

[269] James S Uleman, Leonard S Newman, and Gordon B Moskowitz. People as flexible interpreters: Evidence and issues from spontaneous trait inference. *Advances in Experimental Social Psychology*, 28:211–279, 1996.

[270] Aditya Vailaya, Mário AT Figueiredo, Anil K Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001.

[271] M. Valafar, R. Rejaie, and W. Willinger. Beyond friednship graphs: a study of user interactions in flickr. In *Proceedings of ACM Workshop on Online Social Networks*, pages 25–30, 2009.

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

[273] 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*, pages 1–8, 2007.

[274] Egon L van den Broek. Ubiquitous emotion-aware computing. *Personal and Ubiquitous Computing*, 17(1):53–67, 2013.

[275] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

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

[277] Nancy A Van House. Flickr and public image-sharing: distant closeness and photo exhibition. In *Human Factors in Computing Systems CHI*, pages 2717–2722. ACM, 2007.

[278] Nancy A Van House. Personal photography, digital technologies and the uses of the visual. *Visual Studies*, 26(2):125–134, 2011.

[279] Roelof van Zwol, Adam Rae, and Lluis Garcia Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *Proceedings of the International Conference on Multimedia*, pages 1015–1018. ACM, 2010.

[280] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[281] Simine Vazire and Samuel D Gosling. e-perceptions: personality impressions based on personal websites. *Journal of Personality and Social Psychology*, 87(1):123, 2004.

[282] Prasanth Veerina. Learning good taste: Classifying aesthetic images.

[283] Edward A Vessel and Nava Rubin. Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of vision*, 10(2), 2010.

[284] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.

[285] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[286] A. Vinciarelli and A. Pentland. New social signals in a new interaction world: The next frontier for social signal processing. *IEEE Systems, Management, and Cybernetics Magazine*, 1(2):10–17, 2015.

[287] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[288] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signals, their function, and automatic analysis: a survey. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 61–68. ACM, 2008.

[289] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

[290] Alessandro Vinciarelli, Hugues Salamin, Anna Polychroniou, Gelareh Mohammadi, and Antonio Origlia. From nonverbal cues to perception: personality and social attractiveness. In *Cognitive Behavioural Systems*, pages 60–72. Springer, 2012.

[291] 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*, pages 511–518 vol.1, 2001.

[292] Kiri L. Wagstaff, Terran Lane, and Alex Roper. Multiple-instance regression with structured data. In *ICDM Workshops*, pages 291–300, 2008.

[293] Jin Wang, M. She, S. Nahavandi, and A. Kouzani. A review of vision-based gait recognition methods for human identification. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 320–327, 2010.

[294] Jun Wang, et Jean-Daniel Zucker, and Jean daniel Zucker. Solving the multiple-instance problem: A lazy learning approach. In *International Conference in Machine Learning*, pages 1119–1125. Morgan Kaufmann, 2000.

[295] David Watson. Strangers' ratings of the five robust personality factors: Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology*, 57(1):120, 1989.

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

[297] JS Wiggings. *The five factor model of personality: Theoretical perspectives*. Guilford Press, 1996.

[298] L.K. Wong and K.L. Low. Saliency-enhanced image aesthetics class prediction. In *IEEE International Conference on Image Processing*, pages 997–1000, 2009.

[299] A. Wright. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, 5(3):292–296, 2014.

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

[301] R.V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *International Journal on Biometrics*, 1(1):81–113, 2008.

[302] 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*, pages 349–358, 2012.

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

[304] Che-Hua Yeh, Yuan-Chen Ho, Brian A. Barsky, and Ming Ouhyoung. Personalized photograph ranking and selection system. In *International Conference on Multimedia*, pages 211–220. ACM, 2010.

[305] T. E. Dominic Yeo. Modeling personality influences on youtube usage. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2010.

[306] Matthew D Zeiler. *Hierarchical convolutional deep learning in computer vision*. PhD thesis, NEW YORK UNIVERSITY, 2013.

[307] Xiaohua Zeng and Liyuan Wei. Social ties and user content generation: Evidence from flickr. *Information Systems Research*, 24(1):71–87, 2013.

[308] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[309] Y. Zhu, T. Tan, and Y. Wang. Biometric personal identification based on handwriting. In *Proceedings of the International Conference on Pattern Recognition*, volume 2, page 2797, 2000.