



Influence of image features on face portraits social context interpretation: experimental methods, crowdsourcing based studies and models

Filippo Mazza

► To cite this version:

Filippo Mazza. Influence of image features on face portraits social context interpretation: experimental methods, crowdsourcing based studies and models. Signal and Image Processing. Ecole Centrale de Nantes (ECN), 2015. English. NNT : . tel-01291459

HAL Id: tel-01291459

<https://hal.science/tel-01291459>

Submitted on 21 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse de Doctorat

Filippo MAZZA

*Mémoire présenté en vue de l'obtention du
grade de Docteur de l'École centrale de Nantes
Label européen
sous le label de l'Université de Nantes Angers Le Mans*

École doctorale : Sciences et technologies de l'information, et mathématiques

Unité de recherche : Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN)

Soutenue le 11 décembre 2015

Influence of image features on face portraits social context interpretation: experimental methods, crowdsourcing based studies and models

JURY

Président :	M. Guillaume MOREAU , Professeur des Universités, Ecole Centrale de Nantes
Rapporteurs :	M. Vincent COURBOULAY , Maitre de Conférences HDR, Université de la Rochelle M. Tobias HOSSFELD , Professeur, University of Duisburg-Essen
Examineurs :	M^{me} Alice CAPLIER , Professeur des Universités, Université Grenoble Alpes, Saint-Martin d'Hères M. Marco CARLI , Professeur, Università UniRoma TRE
Directeur de thèse :	M. Patrick LE CALLET , Professeur des Universités, Université de Nantes
Co-encadrant de thèse :	M. Matthieu PERREIRA DA SILVA , Maitre de Conférences, Université de Nantes

**Influence of image features on face
portraits social context
interpretation: experimental
methods, crowdsourcing
assessments and models**

Abstract

With the great evolution of Internet and consumer technology, online social relationships grew exponentially: it is nowadays natural for people to meet, to talk, to work and play completely online. This behavior has been greatly boosted by social networks, that hugely expanded reaching the majority of the global population and differentiated between multiple specific purposes. In fact, while at the beginning social networks were basically meant for maintaining relationships (i.e. Facebook was born to maintain contact between people within the university), they later evolved allowing to meet new people and establish friendly relationships; with time, different social networks have been made in order to separate the different kinds of connections: some of them are made for work while others are for dating purposes, Everyone can be present on these networks by creating an online profile. This profile represents the user and essentially contains basic personal information. While this information can be fake - supposing this is allowed, i.e. to protect privacy - in some contexts having a fake profile is completely useless and actually counter productive for obvious reasons, as in work related social networks. Online profiles can be seen as an "online business card" and these will often form the first impression of profile owner. This is true when meeting new people but also for established relationships, as people will continue to see these information: it is then crucial to maximize profiles' effectiveness. In particular, in our digital era, our profile can be filled up with pictures and especially with a personal profile main picture. A common adagio says "a picture is worth a thousand words": this is the case also for our profile picture, especially considering that it is the first noticeable thing in a profile and that the image will be usually shown aside each interaction. This picture is usually a portrait image depicting the subject owning the profile. However, there is a huge variety of possible shots of a same subject and these can convey very different messages. Today the selection of the best picture for a profile is driven mostly by personal tastes, experience and empirical suggestions.

This thesis focuses on how the personal profile portrait picture is perceived respect to the different possible purposes existing on current social networks. In particular, the work done investigates which are the influential factors within an

image that shift its perception for a purpose more than another. We started developing a methodology to address this problem, considering current research in Quality of Experience assessment and research fields related to our topic. In particular, we investigated electrophysiology and crowdsourcing. After choosing the more reliable crowdsourcing technique, we conducted online subjective experiments to both validate our methodology and our software framework. A first experience aimed at checking if the adoption of a different portrait for the same subject can have a statistically measurable effect on users' perception; for this part we used specific in laboratory made portraits. Successively, the experiments focused on influential factors within the portraits. For this purpose we retrieved a larger amount of real online available portraits and from them we extracted both low and high level features. While the first ones are related to basic properties of an image, such as brightness or contrast, the high level ones are related to content interpretation by humans, such as which kind of background is depicted or the kind of clothes of depicted subject. The choice of which elements to consider has been done both considering existing literature on multimedia assessment studies as well as considering social psychology findings. In order to extract low level features we relied on existing computer vision algorithms, but the high level ones have been evaluated by humans in crowdsourcing, as they represent a much bigger challenge for computer vision approaches.

We statistically evaluated the influence of each feature on context perception asking experiment participants which social context would be the most appropriate for a given portrait. With these evaluations as ground truth, we adopted different mathematical models to achieve a good fit while being able to understand the contribution of each feature. We adopted powerful black box like methods (Neural Networks, Support Vector Machines) as well as white box ones (Linear Regression, Decision Trees). Analysis underlined many high level features as well as some low level ones, as statistically relevant, as they shift the perception of a portrait from "for friendly relationships" to work or dating purposes. In particular, the dress of the portrayed person and the color saturation are shown to be discriminant for the likelihood of a portrait to be perceived as work related. The gender of the portrayed person appeared to be influential for the likelihood of a portrait to be perceived as dating related. However this was also dependent on the gender of the rater. The final part of the work deals with our first steps of a computer vision approach to completely automate the portrait evaluation.

The work and its results can give birth to practical applications (i.e. portrait classification) also considering the consumer point of view, as building automatic portrait composition or recommendation systems for a given social context.

Keywords: Portrait images, social context, subjective perception, crowdsourcing, influential factors, low and high level image features, image processing.

Résumé

L'essor important de l'Internet et technologies de l'information et de la communication ont favorisé la croissance exponentielle des relation sociales en ligne. Il est en effet maintenant naturel pour la plupart des gens de se rencontrer, discuter, travailler ou jouer en ligne. L'évolution de ces comportements ont été catalysés par l'apparition et le développement des réseaux sociaux. Originellement conçus afin de maintenir des relations existantes (Facebook) ils ont rapidement évolué afin de permettre de rencontrer de nouveaux amis potentiels. Les réseaux sociaux se sont également diversifiés et spécialisés dans différents types de relations : amis (Facebook, Google+), travail (Linkedin, Viadeo) ou rencontre amoureuse (Meetic, Tinder). Sur chacun de ces réseau, l'utilisateur est représenté par son profil en ligne. Ces profils, contenant de nombreuses informations personnelles, peuvent être vus comme des cartes de visite numériques. Les informations qu'ils contiennent détermineront la première impression que l'on pourra se faire de la personne en ligne. Il n'est par exemple pas rare qu'un recruteur consulte le profil en ligne d'une personne pour compléter son opinion[Manant 14] sur celle-ci. Optimiser l'impact de son profils en ligne peut donc se révéler utile. Une attention tout particulière doit être portée à l'image principale du profil qui sera le premier élément à être vu, car comme le disait Confucius : « une image vaut mille mots ». Cette image est généralement un portrait représentant le propriétaire du profile. Le choix de celui-ci est crucial, car différents portraits peuvent transmettre des messages très différents. Actuellement, le choix de cette image de profile est effectué sur la base des goûts personnel de l'utilisateur, de son expérience et de quelques règles empiriques.

L'objectif de cette thèse est de comprendre comment un portrait est perçu, en fonction du contexte dans lequel il est présenté. En particulier, nous analysons quels sont les facteurs d'influence qui changent la perception d'un portrait et font en sorte que celui-ci soit plus adapté à un réseau social particulier. Par exemple, poser devant un décor bucolique rend-il notre portrait plus efficace sur un site de rencontre ? Porter une cravate rend-il notre portrait plus professionnel ?

Pour notre étude, nous avons adopté le concept de **contexte social**, qui limite la portée de notre recherche à des cas particuliers (cf. section 1.3). Ce large concept, issu de la psychologie sociale, peut être vu comme l'union d'une situa-

tion, d'un rôle social et d'un ensemble de normes culturelles et sociales. On peut aussi plus simplement le définir comme un «environnement social» dans lequel les gens interagissent. Dans le cadre de l'évaluation de portraits auquel nous nous intéressons, nous nous référons au contexte social **comme perception générale de la situation dans une scène, permettant de juger de la pertinence d'une image pour un usage particulier**. Nous supposons que ce jugement est principalement liée à la perception de ce qui se passe dans une image et que de nombreux éléments peuvent l'influencer. En particulier, la culture et les souvenirs du spectateur peuvent avoir un effet. Dans le cas des portraits des visages, de nombreux contextes sociaux sont possibles. Sur la base des tendances actuelles dans les réseaux sociaux, nous en avons sélectionné trois : le contexte de travail, le contexte de rencontre amoureuse et celui des relations amicales.

La première étape de cette thèse a été la recherche d'une méthodologie appropriée pour réaliser l'évaluation du contexte sociale des images recueillies. Pour ce faire, nous avons, entre autres, étudié la littérature scientifique relative à l'évaluation de la Qualité d'Experience (QoE, ref. [Le Callet 12]), domaine proche de notre sujet. Une première stratégie analysée a été **l'électrophysiologie** (chapitre 2). Cette technique, d'abord abordée dans d'autres domaines de recherche, a été exploitée en informatique après l'introduction du concept d'"Affective Computing" par Rosalind Picard [Picard 97]. Ce concept met l'accent sur la prise en compte des émotions. En effet, la connaissance de l'état émotionnel des utilisateurs peut être très utile, en particulier dans les recherches où l'utilisateur joue un rôle important dans les modèles d'évaluation. C'est le cas dans la recherche sur la Qualité de l'Expérience, où les facteurs humains sont pris en compte. Au vu des premiers résultats positifs avec l'électrophysiologie (ref. [Bos 06, Zhong 08]) et des études liées à la Qualité d'Expérience (i.e. [Haese, Arndt 11]), nous avons évalué les possibilités de ce technique pour mesurer les réactions suscitées par des stimuli multimédias. Après deux études préliminaires, nous avons compris que de nombreux problèmes étaient encore à résoudre pour obtenir des résultats exploitables dans notre domaine. Nous avons alors décidé de donner la priorité à d'autres méthodes pour nos évaluations subjectives. Nous avons en particulier étudié le **crowdsourcing**, une technique ayant fait ses preuves pour les évaluations multimédias. Cette nouvelle technique se concentre sur l'externalisation de petites tâches simples à un large public. Cette foule est généralement anonyme et en ligne. Cette méthodologie alternative fournit rapidement un grand nombre de participants et elle est généralement plus économique que les études en laboratoire. De plus, avec le crowdsourcing, l'expérience est lancée en ligne dans le monde entier : il est alors possible d'atteindre un public bien plus large, plus riche en termes démographiques. Tous ces points positifs sont obtenus au prix d'une préparation de l'expérience plus longue (i.e. design et logiciels particuliers) et une analyse de résultats plus délicate, due aux

nombreux comportements indésirables des participants (« outliers »).

La seconde étape de cette thèse a été la mise en oeuvre pratique du crowdsourcing afin de répondre à nos questions de recherche. Pour mettre en pratique cette méthodologie, trois éléments importants sont nécessaires : 1) une plateforme pour recueillir les participants, 2) un logiciel (« framework ») pour l'expérience en ligne, 3) des stimuli. Sur la base de la littérature du domaine, nous avons choisi une plateforme et construit un framework pour nos expériences, comme expliqué dans le chapitre 4. **En parallèle, nous avons construit notre base de portraits, considérant plusieurs sources disponibles** (les détails de cette partie sont présents dans l'Annexe F). Différentes bases d'images existent dans la communauté scientifique, mais celles-ci - en focalisant sur des buts différents (i.e. reconnaissance des visages) - ne sont pas optimales pour notre recherche. Nous sommes intéressés par des photos de visage contenant également des informations de contexte (i.e. une partie du torse, un arrière plan, etc.). Ces caractéristiques sont typiques des photos de portraits amateur. Par conséquent, nous avons surtout cherché des photos amateurs ou semi-professionnels en ligne, représentant peu de portraits officiels et « posés » - une caractéristique que nous supposons avoir une influence sur la perception subjective du contexte sociale. Pour cette tâche Flickr, la plateforme de partage photos en ligne, nous a permis de collecter de nombreux portraits libres de droits.

Une quantité mineure de portraits a été réalisée en laboratoire afin de mieux contrôler les différents facteurs (section 4.3.1). Ces images ont été utilisées dans une étude pilote qui nous a permis de tester la méthodologie du crowdsourcing dans notre contexte. Ceci avait le double but d'évaluer la méthodologie / les logiciels développés mais aussi de répondre à une question de recherche préliminaire à notre travail. L'étude a confirmé que différents portraits d'un même sujet peuvent avoir des effets statistiquement mesurables sur la perception de son profil. Cette étude a également montré que la technique du crowdsourcing était utilisable et adaptée à nos recherches.

Nous avons donc choisi le crowdsourcing pour poursuivre nos expériences subjectives concernant le contexte sociale, but principale de nos recherches. Nous avons **recueilli des évaluations subjectives du contexte sociale** pour notre base des portraits (chapitre 5) : pour chaque portrait présenté, les participants de l'expérience devaient déterminer quel contexte social, parmi les trois choix fixés (travail, rencontres amoureuses, relations amicales) était le plus approprié. Pour faire évaluer les images en crowdsourcing, nous avons préparé une interface graphique simple après avoir étudié les stratégies déjà adoptées dans la littérature pour étiqueter des bases de données. L'expérience a permis d'avoir un nombre suffisant d'évaluations et a souligné le fait que la perception du contexte sociale est particulièrement subjective et influencée par des facteurs liés à la démographie des

participants (i.e. le genre).

L'étape suivante de notre travail a consisté en l'étude des facteurs pouvant influencer la perception des photos de portrait. Nous avons donc **extrait de nos portraits deux différent types de caractéristiques : celles de bas niveau et celles de haut niveau**. Alors que les premières sont liées aux propriétés de base de l'image, comme la luminance et le contraste, les caractéristiques de haut niveau sont liées à l'interprétation humaine du contenu de l'image, comme le type de fond de l'image (i.e. un bureau, un paysage naturel, ...) ou le type de vêtements du sujet. Comme indiqué par les auteurs de [Joshi 11], il est important de considérer la sémantique de l'image où « la subjectivité humaine » joue un rôle important. À cet égard, les recherches actuelles liées à l'évaluation des images s'accordent sur le fait que considérer seulement les caractéristiques de bas niveau est tout simplement insuffisant. Ainsi, nous devons déplacer notre attention vers des fonctionnalités de haut niveau, plus axé sur "le domaine de la psychologie"[Fedorovskaya 13]. C'est pour cela que le choix des éléments à évaluer a été fait en considérant la littérature existante dans les domaines liés à l'esthétique des images mais aussi en considérant des notions provenant de la psychologie. Cet état de l'art a souligné des éléments à considérer par rapport à la personne dans le portrait, comme l'orientation du visage, l'habillement ou la présence de lunettes, mais aussi par rapport à l'environnement où il se trouve. Dans ce travail, les caractéristiques de bas niveau ont été évaluées par des algorithmes automatiques alors que celles de haut niveau ont été évaluées par des personnes, toujours via le crowdsourcing. Nous avons effectué ce choix afin d'éviter toute influence néfaste des erreurs de classification des caractéristiques de haut niveau, inhérentes à l'utilisation d'algorithmes de vision par ordinateur.

Pour **évaluer l'importance statistique de chaque caractéristique dans la perception du contexte sociale**, nous avons utilisé différents outils d'apprentissage automatique. Nous avons en particulier utilisé des techniques de type 'white box' et 'black box' (chapitre 6) afin d'obtenir des modèles différents, capables respectivement d'expliquer clairement la contribution de chaque caractéristique dans le résultat ou de fournir les meilleures performances possibles. L'analyse a souligné plusieurs caractéristiques de haut niveau, et aussi certaines de bas niveau, comme statistiquement influentes dans la perception du contexte social. En particulier, l'habillement de la personne représentée, l'arrière plan du portrait et une saturation des couleurs plus faible dans l'image sont des facteurs importants pour augmenter la probabilité d'un portrait d'être perçu comme lié au travail. L'orientation du visage et le genre de la personne représentée sont également influents pour un portrait dans le contexte de rencontres. Une analyse plus détaillée souligne que ce dernier facteur est lié au genre des évaluateurs. Alors que construire deux modèles différents pour les hommes et les femmes serait la solution idéale, le

manque de femmes parmi les participants ne nous a pas permis d'explorer cette possibilité.

La partie finale de cette thèse focalise sur les conclusions et perspectives ainsi que sur nos premiers pas vers l'automatisation de cette évaluation, en utilisant cette fois la vision par ordinateur. Ce travail ouvre la porte à plusieurs applications pratiques, en particulier en considérant le point de vue d'un utilisateur. Une première idée est par exemple de construire un système de recommandations automatique pour les photos de profil, à partir d'une base d'images (de la même personne), étant donné un but spécifique. Si la vision par ordinateur nous permet d'évaluer de façon fiable les caractéristiques de haut niveau, un pareil système pourrait être directement implémenté dans les appareils photos / téléphone mobiles.

Mots clés : photos de portrait, contexte sociale, crowdsourcing, caractéristiques d'images, apprentissage automatique.

Acknowledgments

This PhD has been for me a huge improvement. While during the first years all the work to do was (seemingly) very demanding, it appears now very normal – and really interesting. This result is also due to all the people that kindly helped me while I was stuck or demotivated; that gave me advice or encouragement.

First of all, I want to say thanks to my supervisors, Professor Patrick Le Callet and Matthieu Perreira Da Silva, that both guided me along the research path as well as in this experience. Thanks for all the help and for giving me this opportunity. Even if we argued from time to time, these experiences too were useful to grow.

I am grateful to all my friends in the IVC team, where I worked during these years. It has been a pleasure to integrate myself into the team and evolve with it; a special thanks goes to current PhD students: Romain, Lukas, Dimitri, Yashas, Karam, Ahmed ... as well as to former ones: Emilie, Jing, Cedric, ... but also to all professors with whom I worked: Marcus, Vincent, Nicolas, Benoit, Jean Pierre, Harold, Christian and Antoine. To the people that worked or are still working in IVC: Romuald, Josselin, Florent, Laurent ... It has been a pleasure to meet you all.

I want to thanks the people that helped me to review this manuscript: Marianne, Jag and Ana. Thanks for patiently discovering how many English mistakes I can make in a written page.

Another big thanks goes also to all of my friends outside the lab, that shared with me the good and the bad moments of these years, and with whom I lived many other beautiful moments. A special thanks goes to all my Italian friends, that helped me to feel less far from home: Marco, Giuseppe, Melania, Libera, Filomena ... I have to mention also all the latinos and the dance community, that helped me to lose hours of sleep and with whom I spent hours of dancing; thanks Carlos and all.

Of course, I definitely want to say thanks to my family, that supported me all along these years, especially from the very moment I knew that I would have left

them for many years. I really appreciated how you were close even when I was distant.

A last important thanks goes to the person who gave the first input to start all this trip when, in the parking lot of the university he proposed to me to check for internships abroad: Simone, all of this is also your fault! Thanks a lot for this.

My gratitude goes also to all the people in Polytech Nantes and Ecole Centrale - to the EDSTIM in particular - that made this PhD possible. A special thanks goes also to the European Qualinet COST Action, that I joined during this PhD and allowed me to be part of a very interesting active community of researchers and to make many great international experiences.

I met so many special people during this PhD, from all over the world, that I would need another manuscript just to mention all of them. I will conclude then here, say thanks to all that helped me in making this experience productive and relieving the pressure of hard work. In the end, until you try to do something beyond what you have already mastered, you will never grow [R.W. Emerson].

List of Publications related to experiments described in this thesis

- Mazza, Filippo, Matthieu Perreira Da Silva, Patrick Le Callet, and Ingrid E. J. Heynderickx. 2015. “What Do You Think of My Picture? Investigating Factors of Influence in Profile Images Context Perception.” In Proc. SPIE 9394, Human Vision and Electronic Imaging XX, <http://spie.org/Publications/Proceedings/Paper/10.1117/12.2082817>.
- Mazza, Filippo Perreira Da Silva, Matthieu Le Callet, Patrick. 2014. “Would You Hire Me? Selfie Portrait Images Perception in a Recruitment Context.” In Proc. SPIE 9014, Human Vision and Electronic Imaging XIX, 90140X, eds. Bernice E Rogowitz, Thrasyvoulos N Pappas, and Huib de Ridder. , 2–6. <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2042411> (February 27, 2014).
- De Moor, Katrien, Filippo Mazza, Isabelle Hupont, Miguel Ríos Quintero, Toni Mäki, and Martín Varela. 2014. “Chamber QoE: A Multi-Instrumental Approach to Explore Affective Aspects in Relation to Quality of Experience.” In Proc. SPIE 9014, Human Vision and Electronic Imaging XIX, 90140U, eds. Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder, . <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2042243> (February 27, 2014).
- Mazza, Filippo, Matthieu Perreira Da Silva, and Patrick Le Callet. 2013. “Investigating Electrophysiology for Measuring Emotions Triggered by Audio Stimuli.” In Asilomar Conference on Signals, Systems and Computers 2013.

Contents

Abstract	iii
Acknowledgments	xiii
List of Publications related to experiments described in this thesis	xv
Contents	xvii
List of Tables	xxiii
1 Introduction	1
1.1 Context	1
1.2 Motivations	4
1.2.1 The importance of face portraits in everyday life interactions	4
1.2.2 Online profiles influence interactions	6
1.2.3 The right portrait for the right online profile	6
1.3 Social context of face portraits: definitions in this thesis	7
1.3.1 Working definition of face portrait	8
1.4 Scientific positioning	9
1.4.1 The multifaceted panorama of image evaluation	9
1.4.1.1 Objective and subjective media quality	9
1.4.1.2 Quality of Experience	10
1.4.1.3 Aesthetics	11
1.4.1.4 Affective Computing	11
1.4.1.5 Memorability	12
1.4.2 Research on faces	12
1.4.3 Positioning our work	13
1.5 Proposed approach and goals	13
1.6 Thesis outline	15
Keypoints	18

NOVEL METHODOLOGIES FOR MEASURING QUALITY OF EXPERIENCE **19**

2	Exploring the potential of electrophysiology for measuring the emotional impact of multimedia content	21
2.1	Introduction	21
2.2	Motivations: the role of emotions	22
2.2.1	Emotions influence on decision processes	23
2.3	Measuring emotions	24
2.3.1	Conventional measurements	26
2.3.2	Electrophysiological affective measurements: « your body can't lie »	30
2.3.3	Discussion	32
2.4	Pilot studies	32
2.4.1	Study 1: Investigating Electrophysiology for Measuring Emotions Triggered by Audio Stimuli	33
2.4.2	Study 2: Chamber QoE – A Multi-instrumental Approach to Explore Affective Aspects in relation to Quality of Experience	39
2.5	Challenges with electrophysiology	45
2.6	Conclusions	46
	Keypoints	48
3	Exploring the potential of crowdsourcing as an alternative to in-laboratory experiments with questionnaires	49
3.1	Introduction	49
3.2	The technique of Crowdsourcing	50
3.2.1	Generalities	50
3.2.2	Commercial projects exploiting unpaid crowdsourcing	54
3.2.3	Adoption of crowdsourcing in scientific research	56
3.3	Crowdsourcing negative points	59
3.3.1	Reliability checks	62
3.3.2	Evaluating collected data	64
3.4	Motivating participants	65
3.4.1	Economical incentives	65
3.4.2	Gamification	66
3.5	Conclusion	70
	Keypoints	71

APPLYING CROWDSOURCING FOR COLLECTING PORTRAITS AND THEIR CONTEXT PERCEPTION	73
4 Adapting crowdsourcing for social context evaluation	75
4.1 Introduction	75
4.2 Platforms and frameworks: tools to run crowdsourcing tasks in practice	76
4.2.1 Online commercial platforms for crowdsourcing	76
4.2.2 Frameworks	79
4.2.2.1 Available frameworks	80
4.2.2.2 Design and realization of a dedicated CS framework	82
4.3 The data set: portrait sources for research purposes	84
4.3.1 Building a data set for our crowdsourcing pilot study : professional shots versus selfies	86
4.4 Would you hire me? Selfie portrait images perception in a recruitment context	87
4.4.1 Experiment description	87
4.4.2 Demographic considerations on crowdsourced portrait evaluation	96
4.5 Conclusion	101
Keypoints	103
5 Collecting perceived social context of portrait images	105
5.1 Introduction	105
5.2 Crowdsourcing for labeling purposes	106
5.2.1 Classic interfaces	107
5.2.2 Graphic appealing interfaces	108
5.2.3 Free-form interfaces	110
5.2.4 Conclusion	111
5.3 Building a data set to investigate features influence on social context perception	111
5.4 Applying crowdsourcing portrait social context labeling	112
5.5 Analysis of gathered social context evaluations	121
5.5.1 General results	121
5.5.2 Do observers agree on social contexts of portraits?	124
5.5.3 Are chosen social contexts too strict? Hierarchical clustering approach to visualize social context mixtures	128
5.5.4 Demographic issues with social context evaluations	129
5.6 Conclusion	132
Keypoints	134

UNDERSTANDING AND MODELING PORTRAIT FEATURES INFLUENCE IN PERCEIVED SOCIAL CONTEXT 135

6 Understanding the importance of image features in portraits social context classification 137

6.1	Introduction	137
6.2	Portrait image features	138
6.2.1	Low level features	139
6.2.2	High level features	142
6.2.3	Assessing high level portrait features in crowdsourcing . . .	146
6.2.3.1	Challenges with high level features and analysis of uncertainty	150
6.3	Analysis of influential features in portraits social context perception	152
6.3.1	Black box approaches: Neural Networks and Support Vector Machines	156
6.3.1.1	Neural Network	156
6.3.1.2	Support Vector Machines	156
6.3.1.3	Discussion	158
6.3.2	White box approaches	158
6.3.2.1	Logistic Regression	158
6.3.2.2	Decision Tree	161
6.3.3	Discussion	164
6.4	Conclusion	165
	Keypoints	167

7 Conclusions and perspectives 169

7.1	Ongoing work with Computer Vision tools to automate portrait analysis	169
7.1.1	High level features assessment with computer vision	170
7.1.2	Practical implementation	171
7.2	Summary and Contributions	172
7.3	Discussion: limitations and improvements	176

Annexes 179

A Qualinet COST Action 181

B Theories of Emotions: external stimuli and elicited emotions 183

C Laser Doppler Perfusion Monitoring 185

D	Details regarding first electrophysiology pilot study: Investigating Electrophysiology for Measuring Emotions Triggered by Audio Stimuli	187
E	Details regarding pilot study 2: Chamber QoE – A Multi-instrumental Approach to Explore Affective Aspects in relation to Quality of Experience	191
F	Portrait sources for research purposes	193
	F.0.1 Public image databases	193
	F.0.2 Personal collections	197
	F.0.3 Online portraits	202
	F.0.4 Issues in adopting online resources	206
	F.0.5 Shooting a portrait set in laboratory	208
	F.0.6 Our data sets	208
	F.0.6.1 First data set: professional shots versus selfies . . .	209
	F.0.6.2 Second data set: real online portraits for social context evaluation	210
	Bibliography	215
	Proprietary images adopted with author permission	

List of Tables

2.1	Types of emotion assessments	26
2.2	Pros and cons of reviewed methodologies	32
2.3	Confusion matrix for LDPM classification; percentage of correct classification, for stimuli group low (A) and high (B) impact. . . .	37
2.4	Confusion matrix for video typology Neural Network classification based on EEG band power.	43
3.1	Summary of crowdsourcing problems and possible countermeasures.	62
4.1	Summary of CS platforms.	79
4.2	Pros and Cons of reviewed portrait sources.	85
4.3	Mode and Variance of the three regions evaluations' timings. Values expressed in seconds, half a second precision.	101
5.1	An extract of subjective preferences expressed for social context of portraits. First context chosen is shown, preferences are expressed in percentage. Similar results have been gathered for the context of second and third choice.	124
6.1	A non-exhaustive list of low level features found in recent literature on image-related researches, that helped us with our feature subset choice.	143
6.3	Low level features adopted.	147
6.4	High level features adopted, with possible values aside.	148
6.5	High level features adopted (continued).	149
6.6	Features significance with Logistic Regression.*= $p < 0.05$, **= $p < 0.01$, ***= $p < 10^{-3}$	162
F.1	Summary of reviewed publicly available databases	198
F.2	Summary of reviewed databases - continued from table 2.1.	199
F.3	Summary of reviewed databases - continued from table 2.2.	200

Chapter 1

Introduction

1.1 Context

For years the way we produce, work and consume digital pictures has changed radically. Multimedia is nowadays omnipresent as part of communications. Around fifteen years ago, people were both amused and stressed when someone was sending a picture by email, waiting quite some time in front of the screen for the whole image to be downloaded - just to see a small low quality picture. Today, on the opposite, people are surprised if in the morning they don't find any friends' pictures upon opening Facebook on their mobile. The main factors of this change are twofold: the increasing pervasivity of the web and the evolution of consumer devices. Both contributed to a huge expansion of image production and distribution, changing the main paradigm that held for years; today the main producer of multimedia contents is the consumer - that we can now call "prosumer" - while before multimedia contents production was mainly done by fewer experts. An important example is given by consumer produced footages appearing today on news channels, shot by people that took part in an event and posted the material online.

Thousands of pictures are uploaded daily on social networks and online communities (see figure 1.1.1)¹. These communities grew very fast over last decade, differentiating between the disparate purposes that they are meant for: some of them are just for sharing a moment with their friends, others just to archive personal pictures, others are for dating purposes, others again for professional purposes. Some communities are even more specific, like those showing photography skills and sharing suggestions. Facebook, Linkedin, Meetic and Flickr are some of the most popular ones.

¹While many considerations are valid for multimedia in general, we focus from this moment on images only.

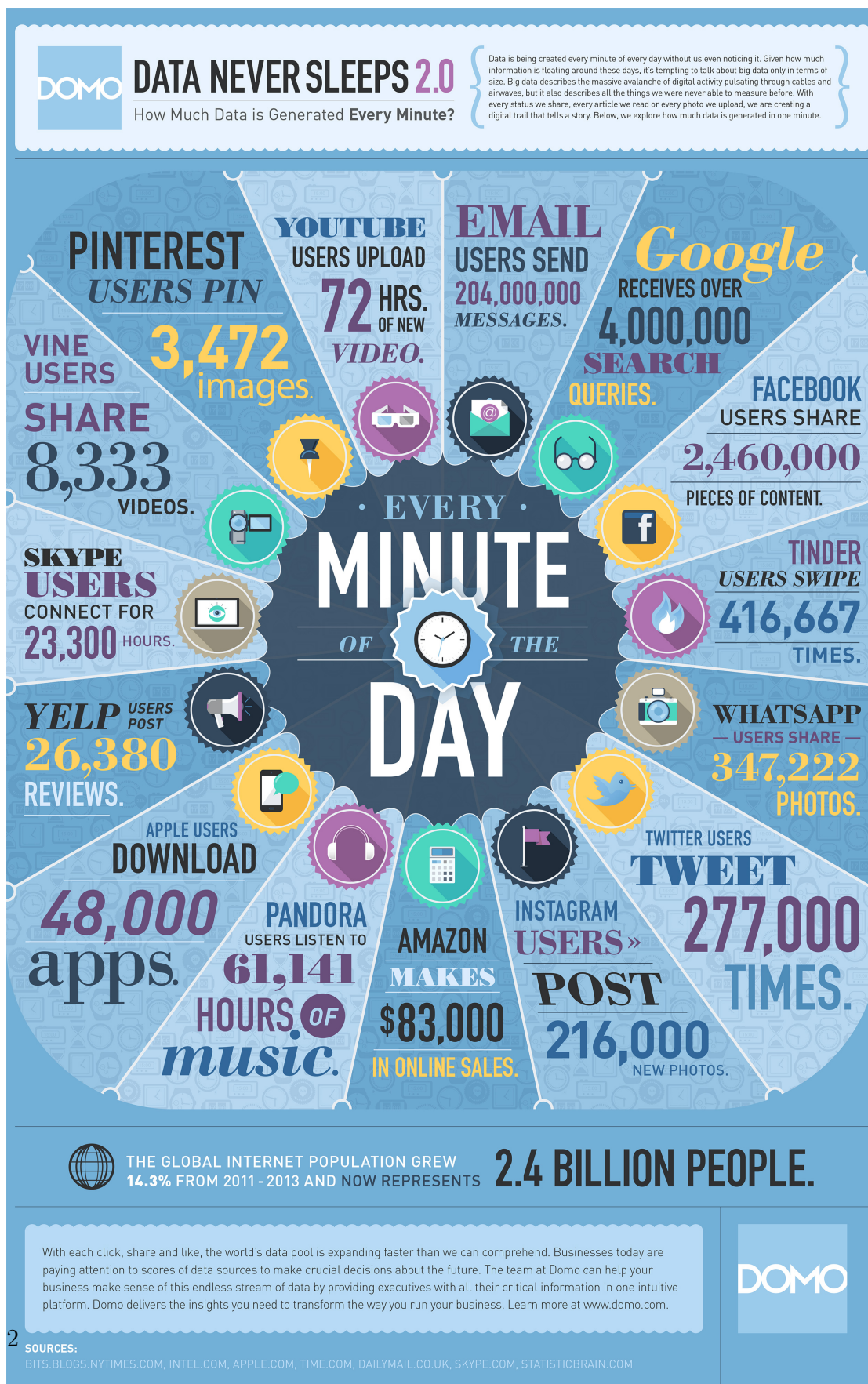


Figure 1.1.1: Begin 2014: data upload per minute, on popular online communities and services. SOURCE: Domo.inc²

In these virtual spaces, people create their own profiles to show to the world. This online showcase and what is published inside are the main elements representing their identity in that particular network. The profile image in particular, which usually is a portrait, is probably the most critical one: it is the first element that will create an impression of ourselves to other online people and will usually appear for every interaction. However other portrait pictures on profiles will also contribute to complement our appearance. Portraits are nowadays overflowing social networks; the development of mobile devices and the introduction of frontal cameras led to a new phenomenon: selfies. While the concept is not new³, there has recently been an explosion of these shots online, especially on social networks. Young and less young people cannot help letting the whole world know how great was their last activity, posting a picture of them and waiting for comments. This trend allowed to personalize our online identity to show to the world even more than before. Even in social networks, a picture is worth a thousand words.

The impact that these portrait pictures can produce is very different as different photos of the same person can deliver very different messages. This is even more true online, as there is nothing but online information to form an impression of the subject. This is particularly important if we consider that different social networks may have very different aims: it is a safe guess to imagine that a professional portrait would be more appropriate than a funny one within a work related network. However, the same won't hold i.e. for a dating network, where a professional profile probably won't maximize the objective in such a context. Consequently it becomes important to learn how to control the messages hidden in a portrait. **The aim of this thesis is exactly to make a step forward in this direction: to understand the influential elements in a portrait which will modify the perception of the portrait context.** This task is very complicated as the messages that can be investigated are multiple - as the influential elements that can bias the perception. Moreover many psychological / socio-cultural considerations come into play and the topic opens the way to a huge research work. For these reasons, we restricted our aim to the perception of what we called "social context" as explained in 1.3 and restricted the possible influential elements to some low and high level image features as described in 6. Psychological considerations are out of the scope of this work and we did not carry research on these; instead we just considered already achieved results in the field to design our methodology and make our choices.

²<https://www.domo.com/blog/2014/04/data-never-sleeps-2-0/>, retrieved July 2014

³The first selfies almost date the birth of photography (en.wikipedia.org/wiki/Robert_Cornelius).

1.2 Motivations

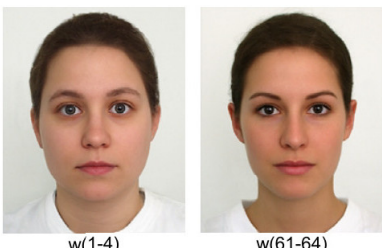
In this section we underline the motivations that drive the interest for our research. We underline the importance of our face in personality perception, the importance of online profiles in social interaction and consecutively the need of the right portrait picture in different online profiles.

1.2.1 The importance of face portraits in everyday life interactions

Before formalizing our aim and scope in following paragraphs, we underline the motivations that led to this study. As we define in 1.5, in this research we consider portraits clearly depicting subjects' faces. Face pictures must be considered as a particular kind of multimedia content, because socio-psychological implications come into play: when people see a content in the image, many thoughts come to the mind – memories, previous experiences, opinions, ... - and this is much more true for people interactions. A face in a picture is not only perceived as a face, but as a person. When we recognize a face – our brain activates special brain areas [Rossion 12] - we perceive not only the image but also the person depicted and involuntary we have an impression of it. This comes from natural evolution, as sensing and quickly understanding the situation we are facing has proven to be crucial for survival: “rapid recognition of familiar individuals and communication cues (such as expressions of emotion) is critical for successful social interaction” [Todorov 05]. Our brain evolved for ages in a way that is capable of having a first impression of a person we are meeting for the first time in less than a second [Willis 06] - even at an unconscious level. This impression demonstrated to be difficult to overcome, also because later during the interaction our mind is pushed to find confirmation of that first impression, that is to say we have a confirmatory bias [Rabin 99].

Our perception of other people impacts decision-making regarding many aspects of life: normal social interactions, work decisions and mate selection process just to name a few important examples. In large part, people perception depends on face traits. The more attractive people are, the more positively they are perceived. Correlations between moral judgments and physical attractiveness have been demonstrated in [Braun 99] (fig. 1.2.1), that underlined how people's opinions regarding intelligence, success, honesty and social life are positively influenced by beautiful faces. Trust has been demonstrated to be related to facial similarity [Vugt 10]. The mating process is especially influenced by our face: facial traits reveal a lot of information about our quality as a potential mate, like hormone levels or health: “facial attractiveness is then a prelude to sexual selection” [Edler 01].

While our face traits are part of ourselves and cannot (easily) be changed, *other elements in a portrait can influence how people perceive us*. Researches in psychology underlined that gaze can positively influence perceived persuasiveness, even if a prolonged gaze can be detrimental [Chen 13]. A simple smile too can influence attractiveness perception, even if empirical research shows that the effect is subtle and context dependent [Whitehill 08]. For women, makeup is an important “signal” - related to biology phenotype - that relates with likability and trustworthiness [Etcoff 11]. Elements outside our face are important too: dress for example is a very important element in impression formation⁴ [Damhorst 90, Behling 91]. Posture can be influential too, influencing especially personality perception - “power” is “likely to occur non verbally” [Carney 05].



	w(1-4)	w(61-64)
successful	2,79	5,50
satisfied	3,43	5,64
sympathetic	3,50	5,21
intelligent	4,21	5,29
available	3,77	5,38
sociable	3,50	5,57
honest	4,62	5,54
exciting	3,00	4,93
diligent	3,93	5,07
creative	3,14	4,86

Figure 1.2.1: Face traits can bias personality perception: some results in [Braun 99] showing average subjective opinions based only on faces (scale: 1-7).

⁴A.k.a. “person perception or social perception”, <http://www.bergfashionlibrary.com/page/The-Social-Psychology-of-Dress/the-social-psychology-of-dress>, retrieved July 3rd, 2015.

1.2.2 Online profiles influence interactions

Previous considerations underline how important are our face and portraits depicting our face in social interactions. With the proliferation of social networks, our aspect influence people's impressions about us even more than before, especially considering two facts. First the basic objective of these platforms is to establish new relationships. Online profiles are really important but they are often overlooked by people: our first impressions do not only influence online interactions, but also real life communications. For example job recruiters may look up online profiles to complement their opinions, as demonstrated by Manant et Al.[Manant 14]. Second, no other information except posted material is available to form a first impression. Even if online communities are used to keep already existing relationships, the online profile may be influential, as usually no 'offline' interaction is possible (i.e. due to distance ...). Interactions will definitely be impacted by profiles; this is not only common sense but also a subject of study in psychology too [Gosling 07]. The "e-Perception" has been considered by psychologists even before the invention of social networks, when personal webpages were more common [Vazire 04]. Research confirmed that clear impressions can be elicited in observers. Interactions are also influenced by profile pictures, and the majority of users will tend to appear as best as possible. A good picture quality will definitely be important, but especially a good shot, a good posture and an alluring expression are likely to be fundamental. Moreover, if it was not enough, in some cases, mobile profile pictures in contact lists started to be updated automatically with the ones on the social networks, showing the image of the person at every call or message. Again, our aspect comes first.

1.2.3 The right portrait for the right online profile

Social networks relationships range in various categories, from friends and family to public life and work. As previously said, Facebook is a very popular example of social networks for friendly interactions, as LinkedIn is for work and Meetic for dating. However, many other networks exist and will probably be developed over time. At the same time, personal online profiles have become full of all types of portraits: some of them are funny, some are serious, some are made to impress someone special. As different profile pictures can convey very different messages, it is then reasonable that *profile pictures should differ between social networks as the objectives and the message we want to convey are different*. Clearly, some portraits do not fit everywhere. A good example is provided by profile images in social networks meant for personal friendly relationships (i.e. Facebook): these are usually funny amateur pictures. Many people adopt selfies with friends or pictures of themselves doing some amusing activity. While these pictures can be

fine for such networks, they won't fit at all a work-related social network profile, as they won't be perceived as conveying a professional attitude (i.e. fig. 1.2.2). This perception is of course subjective and influenced by the elements in the picture and maybe culture / education. In this last context an inadequate message conveyed by the image can mean a loss of money in terms of loss of opportunities. However the same holds for "offline" portraits: a social network profile picture may not fit a resume picture at all. A very different context is a dating one, in which the aim - and then the purpose of a portrait picture - is completely different.

In this scope it is important to understand that the "goodness of fit" of an image is really context dependent. Understanding how people perceive our portrait is important in order to maximize the impact of the picture on the viewer, no matter what the objective is. Moreover, it would be important to maximize the effectiveness of a picture depending on the purpose we want to attend. This is the aim of this work, as explained in the next paragraph.

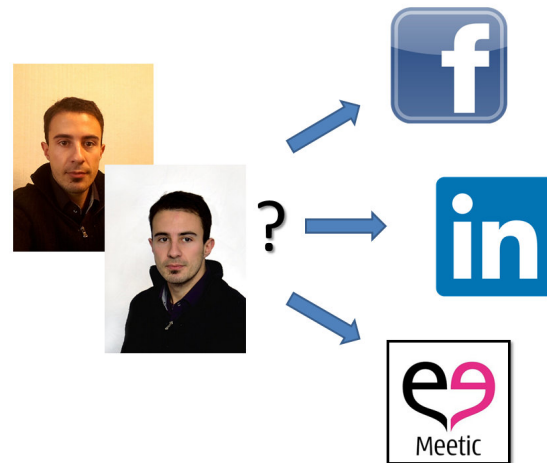


Figure 1.2.2: Choosing the right picture given a purpose in a Social Network can give important advantages.

1.3 Social context of face portraits: definitions in this thesis

With previous considerations in mind, we can introduce the main goal of this work. We will successively place it within the panorama of different image research branches (par. 1.4.1). The previous section underlined that being able to understand and manage the message we want to convey with an image is crucial.

This work focuses on understanding how to control messages given by portraits in order to maximize their effectiveness within a context. We focus on the concept of context, and more specifically on what we call “*social context*”, as we are dealing with images showing a person. This broad concept, studied in psychology and neuroscience, involves the processing of high-level contextual information and can be roughly⁵ seen as the combination of a situation, a social role and set of cultural and social norms [O’Connor 15]. More broadly, it is conceived as a “social environment”, as a setting in which people interact [Barnett 01]. It is not our purpose to rigorously define the concept from a point of view of social sciences as in our work we will just adopt this concept without any other implication except those that we explicit here. In light of previous definitions and the topic of portrait assessment in which we are interested, we refer to *social context as the perceived overall feeling of the situation in a scene, allowing to judge the adequacy of a portrait picture for a purpose*. We suppose that it is mainly related to the perception of what is going on in a picture, and many elements may influence it, even culture and memories of the viewer. Very recently, detecting a social context within multimedia has been an object of interest as in previously cited research [O’Connor 15].

Even focusing on face portraits, many are the possible social contexts. Moreover, these can be seen at multiple scales [Barnett 01] (i.e. overlapping groups). For example, we can mention broad contexts like “work” or be more specific by defining “office work” or “industrial work”. Clearly, the level of detail is arbitrary. In our work we preferred as a first step to focus on three broad social contexts based on the existing main categories of social networks; therefore we chose working purposes, dating purposes and friend interactions as social contexts under study.

Still focusing on face portraits, there is a large variety of possible portrait shots. To conduct our research we defined face portraits to consider as follows:

1.3.1 Working definition of face portrait

Having our purpose in mind, we consider a portrait image *an image in which one person is clearly the subject of the shot, showing his face clearly and a variable part of his body* (torso, legs even on full body portraits) *and some elements useful to understand the context in which the shot has been taken, i.e. a part of background*. We do not consider portraits that are taken from a far distance as they do not clearly show the face, that is to say in which the presence of the person is only marginal. While another possible distinction is between frontal pictures or side/half side face pictures, we will consider them all. However we do not consider portraits the shots in which the subject is not facing the camera. In the end a

⁵We are not interested in details regarding psychology definition of the topic, out of the scope of this work.

very important distinction to be made is between a person alone or in a group: as it is not clear who is the subject of the shot, we won't consider these pictures.

1.4 Scientific positioning

While carrying this kind of work in the perspective of informatics engineering can be surprising, it must be considered that multimedia related research is really broad. Many aspects have been contemplated in order to fully evaluate the impact of multimedia in today's society. Starting from the well known topic of quality assessment - evidently an important aspect of research in multimedia evaluation where a huge amount of scientific research has been produced - research evolved greatly with time, considering more and more the user and semantic interpretation, opening the way to a large panorama of research areas. In the next section we will briefly outline the broad field of research in image evaluation to position the work of this thesis.

1.4.1 The multifaceted panorama of image evaluation

Over the last decades research in multimedia assessment evolved greatly, considering broader and broader concepts related to human subjectivity and perception. In particular, the importance of considering the user also comes from the fact that brain is implicitly processing incoming information, biasing the perception of incoming stimuli. Today the research field of image assessment is very broad and involves many aspects in common with the field of psychology. We then briefly outline the different aspects of scientific research on image assessment in order to give the reader the context and place our work in it.

1.4.1.1 Objective and subjective media quality

At the beginning image assessment research focused on media quality; in particular the focus was on objective quality, considering measurable factors related to the degradation of the media itself in the content delivery chain, from content production to delivery. The concept is closely related to another one called Quality of Service (QoS), that considers technical factors impacting the overall performance of a system. Considering for example a telecommunication network, technical factors could be the transmission delay or the error rate. Similarly, in image quality research technical factors to consider can be image resolution, color depth or compression artifacts that can degrade an image from the original one. Fully objective measures were adopted, as Peak Signal-to-Noise Ratio or Mean Squared Error. Later on research noted that while technically it was correct to measure objective

quality, it did not reflect the actual perception from the user's point of view. For example, a naive viewer may not even perceive a difference between two versions of the same media slightly differing only for technical parameters, i.e the resolution (i.e. fig. 1.4.1). Research then started to consider the perceived subjective quality, underlining the importance of how the user perceives the stimuli; measures like SSIM and VQM start to take into account human perception. A remarkable review of Image Quality Assessment has been done by Chandler in [Chandler 13]. Introducing the user into the evaluation loop opened the way to a much broader scope for multimedia assessments.

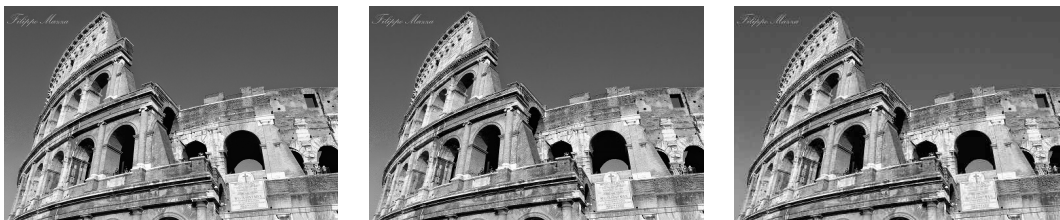


Figure 1.4.1: Same picture shown with different quality levels, best at the left, worse at right. Objectively, the first image has much higher resolution of the second image, but subjectively this may not be perceived. The third image instead will be perceived as a lower quality one. The third SOURCE: personal collection.

1.4.1.2 Quality of Experience

A broader concept that evolved especially during the last decade is the one of Quality of Experience (QoE). This concept, appeared for the first time within the domain of multimedia within the IEEE community with [Jain 04], incorporates elements that encompass the concept of overall quality in a much broader way: research underlined that inherent media quality is not everything and there are other aspects to consider in the evaluation. The most remarkable effort in the field is given by the research community called Qualinet - the European Network on Quality of Experience in Multimedia Systems and Services, an Action of COST framework - that conducted impressive research work on the subject, helping in defining the QoE and its components. Qualinet defines the QoE as “the degree of delight or annoyance of the user of an application or service” and underlines that three are the main characteristics that influence QoE (influential factors): the user, the system and the context [Le Callet 12]. These belong to three areas that must be considered to “better express everything involved in a [...] service”. A small presentation of Qualinet is given in Annexes (see A). We are greatly interested in the QoE domain especially for the novel methodologies applied in this

research. As described in chapters 2 and 3, alternative methodologies such as electrophysiology and crowdsourcing replace the traditional in laboratory assessments with questionnaires. We investigated these novel approaches in our research as described in mentioned chapters.

1.4.1.3 Aesthetics

Broadening the scope, research expanded considering also the interpretation of the media content itself. A branch of research in image assessment focused then on the aesthetic quality, as subjective perception is inherently affected by information processing made in our brain (ex. figure 1.4.1.3). Our opinion can be influenced not only by the technical quality but also by the content, as several studies demonstrated [Lassalle 12]. Among the pioneers in this field is Datta, that defined the “aesthetic gap”, underlining how to characterize the high-level human perception of aesthetics only with low-level features (adopted until then for image assessment) was an interesting question [Datta 06].



Figure 1.4.2: Similar shots, but very different aesthetics results. Inspired by [Ke].
SOURCES: left, B. Andersen, Golden Gate Bridge at Sunset; right,
D. Ronan, Beautiful Day at the Golden Gate Bridge⁶

1.4.1.4 Affective Computing

Emotions too are an important element to consider within multimedia assessment. Media contents, processed by our brain, can elicit emotions due to the semantics of the content itself as well as from our memories. Even low quality shots may have a huge impact, as the message they convey may be sometimes more important

⁶[flickr.com/photos/ldandersen/3431825/](https://www.flickr.com/photos/ldandersen/3431825/) and [flickr.com/photos/worldsurfer/149113516/](https://www.flickr.com/photos/worldsurfer/149113516/)

than their inherent quality. A clear example is given by family photo galleries, often containing bad quality pictures that however mean something for us, or shots from photo reporters that can be invaluable, regardless their technical quality. A branch of research focuses specifically on multimedia affective evaluation in order to understand how to include the effect of emotions in subjective evaluations as well as how we can measure the emotions for research purposes. In order to measure emotions, research measured elicited body reactions as these are 'hard-wired' to emotions through the nervous system. For this purpose research relies on electrophysiology, measuring - for example - heart rhythms or EEG. While now this branch is actually affirmed and well known, at the beginning raised many doubts and laughs, as the pioneer of this field, Rosalind Picard, underlined [Picard 10]. Electrophysiology brought affective computing closer to another practical application: to help people affected from disabilities to communicate.

1.4.1.5 Memorability

Another branch of image related research has recently focused on the memorability of stimuli itself. This research too is multidisciplinary, merging psychology for what it relates to the memorability and engineering for image analysis. The pioneer in the field is the MIT team of Aude Oliva. Their studies confirmed that some images are intrinsically more memorable "independently than subject past experiences or biases" [Isola 11b]. The team successively investigated which elements in a picture make it more memorable than others, focusing on image characteristics and regions [Isola 11a, Khosla 12], but also on memorability of faces (ref. next paragraph). This branch is however very recent and few research studies have been conducted out of MIT. the study of [Mancas 13] that linked memorability with visual attention is remarkable.

1.4.2 Research on faces

Portrait images have already been the object of scientific research. Studies on objective guidelines for face portraits in official documents have been conducted in [Castillo 06], where there has been an emphasis on image quality. Within the aesthetics branch, computer vision efforts focused on visual aesthetics for photographic portraiture considering face features, as input in machine learning algorithms to predict aesthetic assessment, as done by Li [Li 10b], or more recently by Khan [Khan 12] and [Xue 13]. Computer science also joined with neurobiology to investigate if and how face aesthetics can be assessed with machine learning [Eisenthal 06]. Joint studies with the psychology field have been conducted modeling personality perception based on face beauty [Braun 99] or, more recently, considering emotions with facial expressions in portraits, with the objective of

helping the user to select the best portrait in a collection [Lienhard 15a]: i.e. the user can ask for the portrait that gives the most friendly impression. The MIT team working on image memorability focused on faces too, after finding that face presence and enclosed spaces are influential elements. They underlined that face memorability is an intrinsic property, that is to say that some faces are simply more memorable than others [Bainbridge 13b]. Following research aimed at biasing the memorability by modifying the face characteristics [Khosla 13].

1.4.3 Positioning our work

Face portraits have been investigated under the light of aesthetics, memorability, emotions and psychology. To the best of our knowledge we notice that a research focusing specifically on the perception of social context in portraits is lacking. This is particular important considering previously described motivations - importance of both face pictures and online interactions. We see in this lack an opportunity to carry on, not only from a theoretical point of view but also for the potential practical applications (i.e. automatic recommendation systems or photo enhancing software). The goal of this thesis is a first step in this direction. Our positioning is close for certain aspects to psychological studies investigating social relationships and perception; however, we do not work within this field but we take some psychological findings as clues to conduct our research. With multimedia quality research we share a similar principle in the methodology - we try to understand factors influencing a perception in order to maximize an objective. However we underline that we are not interested in quality assessment, still we review some work and methodologies adopted in the field in order to design our research. We are also close to affective computing if we consider that the social context evaluation can be an implicit perception, like a feeling, that sometimes cannot be easily explained rationally. For this reason, as we see in chapter 3, we try to adopt electrophysiology, the most used assessment methodology in that research area. Still, we are close to the aesthetic branch because we are interested in studying which elements in the picture are influential. With other multimedia research branches we share also crowdsourcing, a methodology replacing classical in-laboratory studies that retrieves many subjective assessments, needed for statistical analysis on many factors at the same time - as in multimedia quality assessment.

1.5 Proposed approach and goals

After having described motivations and positioning of our study, in this part we define our approach underlining our goals.

The first step is related to the design of a proper methodology to collect perceived context evaluations: this is a subjective opinion, due to its intrinsic nature. For this reason we investigate and evaluate methodologies adopted in other multimedia assessments research branches to gather subjective assessments. Notably, we investigated the field of Quality of Experience in particular, encompassing a broader evaluation considering human factors and the context. In this panorama, two novel methodologies have been investigated, emotional assessments with electrophysiology and online crowdsourcing. **Our goal resides in finding the best suitable methodology, considering the State of the Art but also pros and cons underlined by preliminary experiments:** we need to test strategies with simpler pilot studies to really acquire the methodologies.

Once a proper methodology is chosen, the second step is to adapt it to our purposes and put it in practice for the experiments. In our research, we've opted for crowdsourcing, more effective and efficient than electrophysiology or traditional in laboratory assessments. **Our contribution in this phase is the development of a dedicated software framework to run experiments and the review of platforms to use.**

At the same time, to run experiments, we need to retrieve portraits to exploit. We then investigate available image sources to adopt, starting with databases used in research. As most of them are not featuring enough social context information as we would like (i.e. they feature a neutral face over a white background for face recognition purposes), we mainly rely on online communities and social networks, constantly filled with a huge variety of portraits taken in real conditions. **Our contribution in this phase is the review of available image sources and the development of personal data sets.**

The third step is to use it to collect social context evaluations of collected portraits. **Our goal is to gather a reliable portrait social context labeling.** We need in fact to analyze context evaluations to check if assessments are reliable and dependent on factors out of our control (i.e. participants demographics). To conduct our analysis, we decided to restrict the possible social context to focus on, as a large panorama of contexts can be the possible for a given portrait. Considering modern social networks, we focus on three contexts that we believe are the most important for a portrait: context of friendship, of working purposes and of dating.

Successively we evaluate image features of our assessed portraits and run mathematical analysis to understand the influence of features on the social context perception. As stated by authors of [Joshi 11], it is important to consider image semantics where "human subjectivity" plays an important role. In this respect, current research related to image assessment agrees that considering only low level technical features (pixel related features such as luminance or contrast) is simply

not enough, and so, we have to shift the attention towards high level features, more focused on "the domain of psychology" [Fedorovskaya 13], considering the cognition of the content of a scene. Given previous considerations, we consider both low level and high level features, since we believe that the latter category greatly influences the evaluation of a portrait. High level features have been chosen considering also results coming from social psychology, that underlined how attractiveness and personality perception can be influenced by smile, gaze, glasses and more. These being very hard to be computed via computer vision tools, we relied again on crowdsourcing to label them manually. This stage goal is mainly **the review of psychology results to select high level features, the methodology to label them in our portraits and the labeling itself.**

In order to understand which features are influential in the perception of social context, we adopt white box approaches as statistical analysis to obtain an interpretable predictive model, linking features to social context. We also adopt other black box methodologies (i.e. Neural Networks) to compare the results. **However, our goal resides also in the design and evaluation of the methodological approach to adopt.**

1.6 Thesis outline

Three main sections compose the thesis outline, shown as a diagram in figure 1.6.

I NOVEL METHODOLOGIES FOR MEASURING QUALITY OF EXPERIENCE

In this section we describe which and how face portraits we retrieved and how we collected subjective assessments of the perceived social context. It describes also the analysis of the two possible strategies that we investigated to assess the perceived social context. It contains chapters:

- “Exploring the potential of electrophysiology for measuring the emotional impact of multimedia content”, dedicated to the evaluation of electrophysiology for affective assessments. Two pilot studies have been conducted for this purpose, described in this chapter.
- “Exploring the potential of crowdsourcing as an alternative to in-laboratory experiments”, in which we describe the evaluation of online crowdsourcing in our field.

II APPLYING CROWDSOURCING FOR COLLECTING PERCEIVED SOCIAL CONTEXT

In this part we focus on the practical implementation of crowdsourcing for our purposes and to the gathering of social context evaluations of portraits. We also describe adopted data sets. It contains chapters:

- “Adapting crowdsourcing for social context evaluation”, in which we discuss how we apply crowdsourcing in practice and we describe a pilot study aimed at testing the methodology.
- “Collecting perceived Social Context of portrait images”, where we describe how we gather subjective social context evaluations for collected portraits, exploiting crowdsourcing.

III UNDERSTANDING & MODELING PORTRAIT FEATURES INFLUENCE ON PERCEIVED SOCIAL CONTEXT

This part is dedicated to: i) the selection and evaluation of low / high level portrait features; ii) statistical analysis to understand features influence and iii) conclusions and tentative approaches to automate the portrait evaluation, based on the developed model.

- “Understanding the importance of image features in portraits social context classification”, dedicated to the design and evaluation of a set of meaningful portrait features, as well as to the mathematical approaches to analyze features influence on social context perception.
- “Conclusions and Perspectives”, in which we summarize our work and explain our first tentative approaches with computer vision to automate portrait social context evaluation.

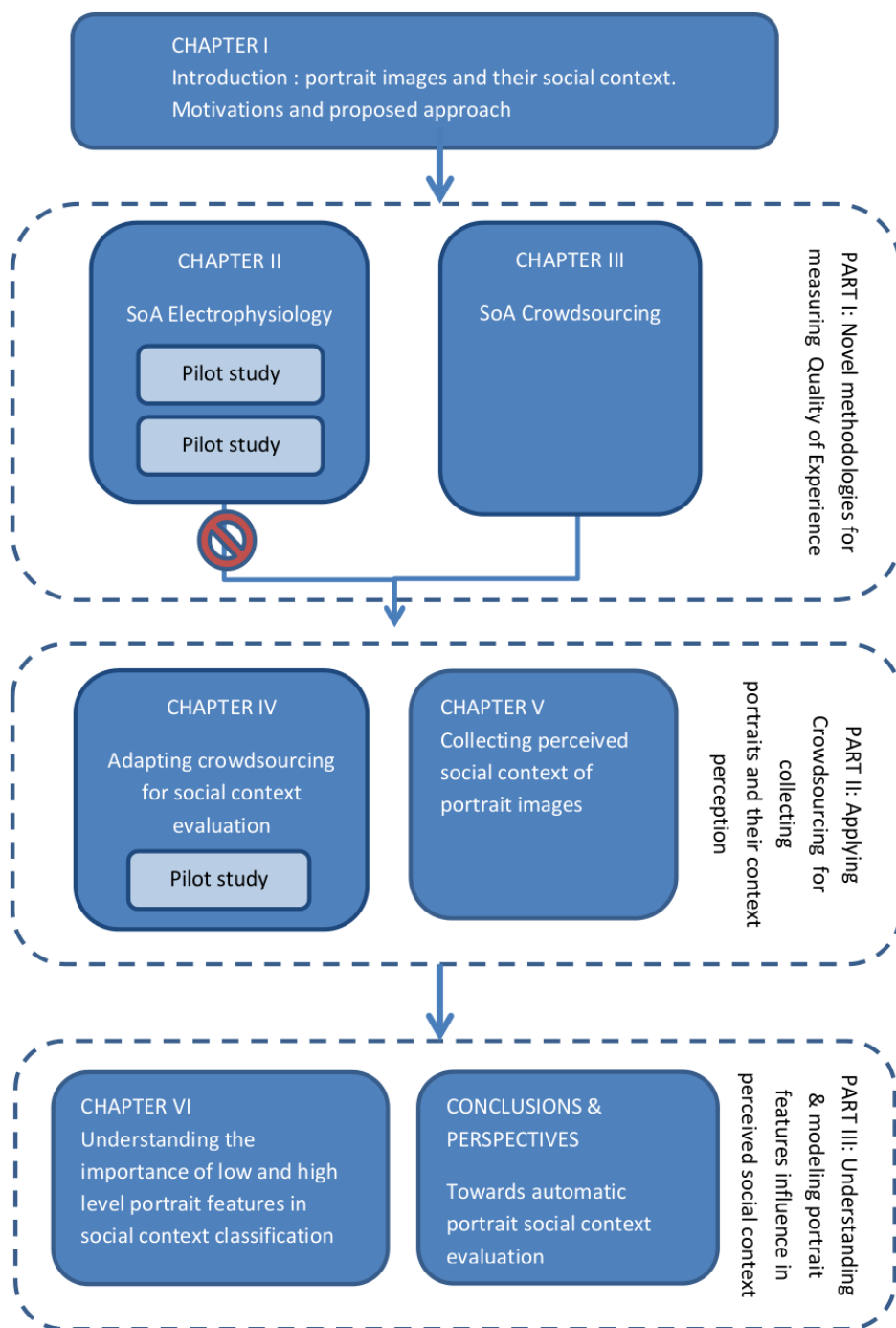


Figure 1.6.1: Schematic of thesis plan.

Keypoints

Context

- ❑ Different online communities and social networks exist today for different purposes; in this context online profiles are our “business card”. These networks have a huge diffusion and can influence also real world interactions.
- ❑ Portrait images have a big impact in making first impression and bias subject perception even after.
- ❑ As different portraits of the same subject can convey different messages, it is important to manage these messages understanding which elements are influential within the picture.

Contributions

- ❑ We positioned the problem as a special case of image assessment and formalized it adopting the concept of “social context”.
- ❑ We reviewed methodologies currently used in multimedia research and proposed a methodology to address our research problem.
- ❑ Methodology proven to be effective and we underlined few low and high level influential features for social context perception.

“Research is to see what everybody else has seen, and to think what nobody else has thought.”

(Albert Szent-Gyorgyi)

**NOVEL METHODOLOGIES FOR
MEASURING QUALITY OF
EXPERIENCE**

Chapter 2

Exploring the potential of electrophysiology for measuring the emotional impact of multimedia content

We describe here our investigation of the potential of electrophysiology to measure subjective reactions elicited by multimedia stimuli, as with portrait social context assessments. Recently this novel methodology has been investigated within the Quality of Experience field, showing positive results in some specific cases. Subjective perception can elicit emotions and physical reactions in the body: these can be measured with electrophysiology. Motivations are described in detail, together with a state of the art on electrophysiology within the field of affective computing for multimedia assessments. Two case studies have been investigated, conducting experiments to test electrophysiology potential. Results show that while this technology can detect reactions to stimuli, there's still a lot of work to do in practice to make it reliable and effective.

2.1 Introduction

Human emotions have been a subject of study in the field of psychology since a long time. Emotions have been addressed also in related fields, such as neurobiology, where electrophysiology approaches have been attempted in order to clinically measure the effect of perceived emotions on the nervous system (i.e. [Rada 95]). However, during the past few decades this topic has been addressed even in other research fields, notably in electric and electronic engineering, especially after the introduction of the concept of «Affective Computing» by Rosalind

Picard [Picard 97]. This concept focuses on computing what «relates to, or arises from, or influences emotions», and it represents probably the first long-term research joining electrophysiology for emotions and engineering. Some studies within this scope have been carried out previously - i.e. «The truth machine» [Bunn 14] - however these did not produce a very widespread interest and following researches such as the 'Affective Computing' field. In fact this topic has become from abstract theoretical discussion one of the most active topic in recent research [Picard 10]. At the beginning even the potential of this kind of research was unclear and the scientific community was skeptical of its utility. Nevertheless, electrophysiology emotion evaluations were revealed not only to be possible but also to be useful for many aspects. After a decade of work, a dedicated IEEE Transactions has been started on the subject¹. Such research opens up to many applications. First of all, as Picard underlined in her works, from a social point of view it allows many new possibilities within psychological research. Moreover, knowledge on user's emotional state opens up many possibilities also in other fields, i.e. in consumer studies. Wearable devices today are taking more and more electrophysiological measurements.

In general, understanding a user's state of mind can be very valuable in research where the user itself plays an important role in evaluation models, as in research on Quality of Experience [Le Callet 12]: perceived quality is related to a broader evaluation taking into account multiple influential factors. Among them there are human factors, taking into account user related context and notably a user's emotional state. Users' affective state evaluations have thus started to be included in perceptual models, as emotions play an important role in perception. While our research is not focused on multimedia quality assessment, emotions are an important element linking that research branch and ours. As underlined in the next paragraph, emotions can both bias human perception and reflect perception itself, being «hardwired» to the nervous system. The chapter continues (sec. 2.3) underlining the different ways to measure emotions, outlining the state of the art in affective computer research related to multimedia. In section 2.4 two pilot studies are described, conducted to evaluate the use of electrophysiology for our research.

2.2 Motivations: the role of emotions

Emotions impact many aspects of human life. This conclusion comes not only from common life, but also from scientific findings. Evolutionary psychological theories of emotions claim they are as driving forces of actions that «direct the activities and interactions [...] governing perception; attention; inference; learning; [...] physio-

¹IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, 1st Volume edited in January 2010

logical reactions» [Cosmides 00]. Different theories have been made regarding how emotions originate, underlining that they are deeply linked in our brain. Details regarding theories of elicited emotions by external stimuli are given in annex B. Even today, studies in cognitive psychology have in fact demonstrated that decision making processes are “much more influenced by intuition and emotional responses than it was previously thought” [Markic 09]. As evaluations during subjective tests are a decision process, researchers started to evaluate the mental state of participants before making tests, to better understand emotions felt and estimate their impact on evaluations. However affective evaluations for subjective assessments may have a double purpose. Emotions can bias evaluations (ref. section 2.2.1) and it would be worth evaluating them to correct this possible bias. Second, as emotions are closely related to the nervous system (see B) and physiological reactions may be measured, emotions elicited by stimuli may be exploited to have an indirect measure of subjective perceptions. Such technology would not only allow faster and more spontaneous evaluations, but also eliminate the problems linked to assessments bias. This is true for every kind of subjective perception, be it related to i.e. quality or aesthetics assessment, memorability or - maybe - context evaluations as in our case: **the validity of this approach as an alternative measure for context subjective perception tests is the research question of this chapter.**

We do not investigate the first purpose, evaluating emotions bias on perception, but we focus on a common problem that both purposes have: how to adopt electrophysiology measurements to quantify emotions. The reason is that, as explained in next paragraph, emotions are influential factors in user decisions².

2.2.1 Emotions influence on decision processes

Emotions play an important role in how we perceive the world around us and how we react to it as emotions have been shown to influence decision processes. As many research’s findings confirm, emotions impact perception and decision making as well as many aspects of life of course [Picard 97, Markic 09]. For example, during the evolutionary process, the powerful emotion of fear has played a very important role in the survival of the species, allowing easier - and probably correct - decisions, such as avoiding approaching predators. Neurological studies have demonstrated this fact in practical cases, as in [Damasio 94]: different patients which had suffered serious non fatal injuries in cerebral zones related to emotions have been the object of study. This research demonstrated intact mental faculties

²Different terms have been adopted to describe the different affective reactions. We refer here to «emotions», considering the effect of the stimuli, or to the overall user «mood», considering the whole process linked to the content fruition (details in annex B).

but underlined a substantial change in personality. In another study researchers found that injured subjects may be completely unable to take decisions, even when facing all the possible rational solutions. In some cases, only an external factor - a proposed suggestion - can lead to decision. The injured brain area of the brain was the particular area supposed to integrate “emotional responses with higher logical thought” [O’Hagan 98]. Mood also may have been proven to influence people judgments [Forgas 87]: researchers investigated how a positive or negative mood can lead to different impressions on people. Manipulating mood of experiment participants they demonstrated that positive moods helped to form more favorable judgments regarding presented people. Studies have been carried on in different quality perception contexts, as in [Jang 09]: impact of emotions plays an important role in users quality perception and overall experience in restaurants. As authors say the results are “meaningful because they address the relationships among [...] types of perceived quality (product, atmospherics, and service), customer emotions (positive/negative) [...] in the restaurant consumption experience”.

Social context evaluation are likely to be affected too, as it is a decision process: content is cognitively processed by the user and personal opinions or memories arise spontaneously. A simple yet extreme example may be a dramatic scene in a picture, that can elicit an emotion in the user. Elicited emotions can have an impact on content interpretation and bias subjective evaluations; to this extent, measurements of emotional reactions on the nervous system can be a valuable complementary information for subjective evaluations.

These premises underline a sort of chicken and egg problem: external stimuli perception induce physiological reactions and can elicit emotions, and the latter influence decision processes and perception. This point is not investigated deeper in this work, and instead focus is given to emotions and their physiological expressions arising from stimuli perception.

2.3 Measuring emotions

As said previously, even if science is still conducting research on emotion generation in the brain, external stimuli can elicit emotions - no matter what is their real purpose. The question now shifts on how can we measure them.

Different ways of quantifying emotions have been studied, each one with pros and cons. A thorough review can be found in the psychology work of Mauss and Robinson [Mauss 09]. What is proposed here is a shorter review to give an insight on the work done. It is remarkable to notice that every time the user is aware that we survey his reactions, we have a problem similar to the one described by the *observer effect principle*: the simple act of observing a phenomenon may affect it

and bias the measurement³. We do not know in fact if the emotions of subjects undergoing an electrophysiological measurement are biased by the act of taking the measurement itself, for example because of the new experience they are making with this novel methodology. We do not even know what the effect of this eventual bias is, i.e. augmenting either positive emotions because they are entertained or instead negative emotions because they feel stressed. No affective research focused explicitly on this point only, at the best of our knowledge.



Figure 2.3.1: A participant in [Koelstra 12] before starting the experiment. It is clear that this situation entertains the subject, as she is smiling. Will this bias the affective experiment? Source: original research article, courtesy of the authors.

While it would be very interesting to further investigate this aspect and conduct experiments to discover if this effect is present - and eventually find the best methodology to mitigate it - this deviates too much from the main scope of this work. We thus prefer to focus on emotions measurements, starting with a description of different methodologies used to measure them. Table 2.1 organizes the different typologies of assessments, reflecting the format we use in successive paragraphs.

³It has been debated that this principle has been sometimes confused in literature with the Heisenberg's uncertainty principle. It is not our purpose here to dig into the topic and we refer to the cited one at the best of our knowledge.

	Electrophysiology	Conventional
Explicit	EEG, GSR, ECG, LDPM, ...	Questionnaires
Implicit	-	Behavioral

Table 2.1: Types of emotion assessments

2.3.1 Conventional measurements

This paragraph describes conventional affective measurements unrelated to electrophysiology. These are the most common adopted methods in psychology literature. Two typologies can be distinguished, based on the fact that the affective status is explicitly asked about or implicitly understood.

Explicit self assessments

This category of methods is probably the simplest and most direct one. The methodology implements questionnaires given to users, usually before and immediately after the stimuli ends. The questionnaires adopt different emotional models theorized in psychology to estimate both the nature and the intensity of the emotion experienced. This simplicity however poses two problems: (i) the user must be well trained on questionnaire use, (ii) he may be unable to express his feelings (or he does not want to).

Questionnaires can may ask about emotions explicitly or indirectly by asking more general questions. In the second case revealing details about a subject's mental state can be a difficult task for experts in psychology field. As an example, Mental Status Examination test is an explicit way to evaluate and describe a person's current state of mind but the evaluation is the real core task [Trzepacz 93]. The fewer approaches with these methodologies in our field is more likely to be due to the intrinsic difficulty to evaluate results and the need to conduct those approaches jointly with psychologists.

Regarding assessments asking explicitly about emotions, different strategies and modeling have been adopted. A initial somewhat naïve method is to directly ask the user to describe his emotions with adjectives; although simple, this strategy opens to a large variability and uncertainty of results, as same words can be differently interpreted by different people. Reviews by professional psychologist should be useful in that case. Moreover user's affective assessment can be dependent of other factors, such as the relationship with the experimenter or the difficulty to express their feelings overtly [Picard 01].

In order to evaluate the emotional state, different representation systems and emotional dimensions have been theorized [Borod 00]: pleasure, arousal and dom-

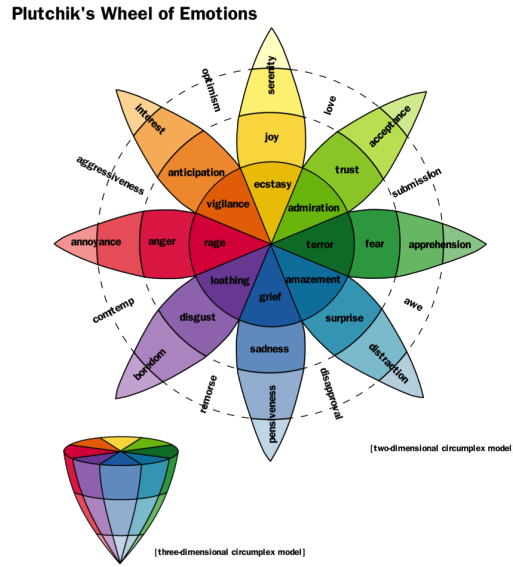


Figure 2.3.2: The Plutchik's Wheel of emotions, depicting eight basic emotions and their compositions [Plutchik 01]

inance (PAD) emotional state model is one of the most well known, describing emotions in those three different dimensions [Lang 80]. Other models used are the Plutchik's Wheel of emotions [Plutchik 01] (fig. 2.3.2) and the Lövheim Cube of emotion [Lövheim 12].

In the PAD model, dimensions are labelled as pleasure, arousal and dominance; in order to allow a simpler adoption of these scales and to avoid linguistic problems (i.e. translations) a manikin system has been proposed. It first appeared under the name of Self Assessment Manikin (SAM) in [Lang 80] (fig. 2.3.3). This method allows users to rate their feelings in a non-verbal way, as it depicts the three main affective dimensions through graphical figures of a “stylized man” with different representations for the three scales and their intensities. In any circumstance, a clear explication of the model must be provided to the user to make it usable.

It is commonly accepted that valence represents the judgment of pleasantness, while arousal expresses the degree of excitement, ranging from calm to excited [Soleymani 08]. Dominance value expresses instead how much one is in control of the situation. While Pleasure and Arousal have been validated in more studies, it has been argued that dominance evaluation can in some cases be equivocated and it's significance on behaviour has been discussed [Jang 09], [Russell 78]. In our study, we assume that all three dimensions are valid as self assessment methods in

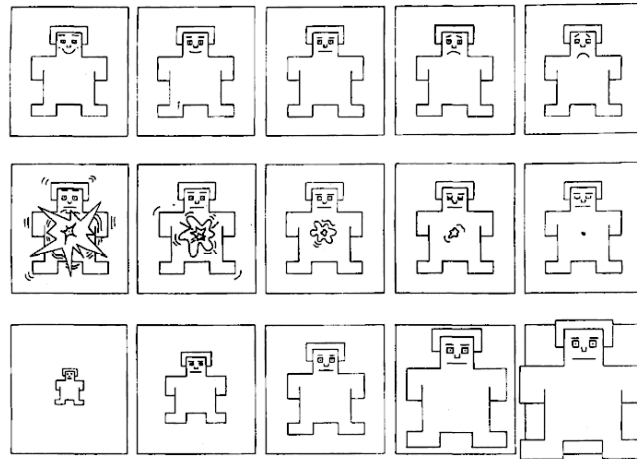


Figure 2.3.3: The SAM as in Lang original work. From top to bottom, valence arousal and dominance scales. Without proper explanation its use is not likely to be easy.

PAD space, also with mannikin representation, as positively used in different researches and databases [Lang 97], [Takahashi 04]. Subjective biases in evaluations can be noticed and corrected through techniques already used for other discrete scale's ratings [Rossi 13], [Greenleaf 92]; for example, the use of training stimuli ranging the whole emotional space can be an option. To minimize the misunderstanding of the scale detailed, simple description must be given to users before tests. Regarding the difference between PAD scales meaning, a research study [Koelstra 12] showed some correlations, probably indicating that strong valence stimuli are usually perceived as high arousal.

Recently another pictorial affective instrument has been developed by Desmet et Al. [Desmet 12] (fig. 2.3.4). This system called Pick a Mood is meant to be a reporting instrument for affective states that “require little time and effort of the respondents”. People can communicate their mood choosing between different avatars depicting a mood in particular. Desmet studies demonstrate also that users of different nationalities can correctly use it after the explanation.

Other evaluation methods and models have been proposed and a detailed review is present in the same work of Desmet previously cited.

For a long time moods and emotions have been evaluated through questionnaires developed by psychologists or other non verbal instruments such as behavioural analysis [Desmet 03]. The problem with those systems is inherent to the highly subjective nature of the topic but also to the less obvious problem of describing correctly what we feel. Current literature still largely adopts self-assessments given to users, but it implements them with facial expressions or electrophysiological



Figure 2.3.4: Pick a Mood self assessment, as done by [Desmet 12].

analysis, discussed in following paragraphs.

Implicit measures

Emotions impact the body on an unconscious level. Reflections of our mental state and reactions to stimuli can be determined by analyzing behavioral features, such as facial expressions, speech or gestures [Gunes 06, Koelstra 12]. These methodologies take advantage mostly of behavioral psychology.

Regarding facial expressions an in depth study was conducted by P.Ekman, that demonstrated the universality of emotions in facial expressions [Ekman 87]. This implication lead following studies and is still going on today with the creation of online databases with facial expressions for affective purposes [Gunes 06, Koelstra 12, Soleymani 12]. Automatic face detection and expression analysis systems have been implemented [Stathopoulou], even for Quality of Experience purposes; Aragon Institute of Technology developed a recognition system called Emotracker [Mateo 13] - more details are given in 2.4.2, where we report the second pilot study. Other implicit methodologies exploit speech [Pathak 11, Yang 12] or posture analysis [Hatfield 93]. In this panorama no research applied emotion analysis to social context evaluation.

Another possible implicit measurement is to propose participants some kind of activity to perform before tests; the activity can reveal a user's emotional state, like i.e. listening to music or playing a game (with these we can see current mood by its preferences and attention by skill level).

2.3.2 Electrophysiological affective measurements: « your body can't lie »

Describing perceived emotions can be inherently difficult for many reasons and it can be even more difficult for subjects undergoing an experiment. First of all, we have to consider that disclosing emotions to a complete stranger - such as the scientist conducting the experiment - can be frustrating. Subjects describing emotions may be impaired by this fact and may not be completely truthful. Second, it is not easy at all to fully understand our own emotions and describe them, especially if they are subtle. Third, this can be even more complicated by the use of different methodologies to describe emotions.

Emotions affect not only our perception but also our physiology, having effects on the nervous system. Noticeable evidence of emotion on physiological activity has been found in both central and peripheral nervous system. Distinctive patterns of autonomic nervous system (ANS) activity have been found for some emotions [Ekman 99]. These reactions are probably impossible to be voluntarily controlled. As «our body can't lie», electrophysiology has been adopted to evaluate emotional impact achieving remarkable results. Many different kinds of electrophysiology measures have been studied in the field, both related to the Central (CNS) or Peripheral Nervous System (PNS) [Chanel 11]. A large amount of studies exist documenting electrophysiology measures related to emotions felt presenting multimedia stimuli. Often research focuses on correlations between physiological patterns and emotion self assessments, used as ground truth.

Brain activity, related to CNS, is directly related to emotions and many studies have been carried out in this field. *EEG* is among the most used measurements in affective research and even if mechanisms related to emotions in the brain are yet to be completely understood, research on EEG is now quite advanced. Psychophysiological studies underline how frontal brain lobes are directly related to emotions [Niemic 02, Bos 06]. EEG band power are the main features used as input for machine learning systems. Four main bands are commonly used, called alpha, beta, delta, and gamma, based on different signal patterns appearing during different brain activities. Signal classifications using these and other features have been adopted successfully, in particular Neural Networks [Bos 06] or Gaussian processes [Zhong 08]. Remarkable results have been obtained in [Nie 11], where positive and negative emotions are classified with Space Vector Machines achieving an accuracy slightly below 90%. Magnetic Resonance Imaging has also been adopted to evaluate brain reactions to affective audio stimuli [Viinikainen 12]: strong relationships have been found between sound emotional value and blood oxygenation in areas of the brain related to emotional reactions, such as amygdala and prefrontal cortex. Notable are the recent results adopting EEG to assess video quality [Arndt 11].

As CNS manages all biological basic functions supporting life like cardiovascular

and respiration activity, emotions' influence on biosignal is likely to be found also on PNS. Evidence of this assumption has already been found and different biological signals related to PNS have been measured in relation to emotions. These measure usually present the advantage of being less invasive than those of CNS. *Cardiorespiratory activity* is probably the most basic physiological activity related to emotions. Distinct patterns have been found and associated to basic emotions felt. In [Rainville 06] researchers investigated cardiorespiratory activity during experiences of different emotions. Those emotions have been elicited asking subjects to disclose a personal strongly emotional experience. Differences between electrophysiological patterns have been found using a multivariate analysis approach and the study underlines how peripheral physiological activity is associated with different emotions. Many other research studies have focused on cardiovascular activity related to emotions, especially in the field of medicine. A comprehensive review can be found in [Kreibig 07]. Another measure used is *Galvanic Skin Resistance* (GSR), especially used for reactions to stress [Wu 11] but also related to cognitive load during certain tasks [Nourbakhsh 12]. *Electromyography* has also been used in conjunction with GSR as in [Nakasone 05] or with facial electromyography in [Niedenthal 09] while participants judged emotional and neutral concepts. Recently *pupil diameter* has been used as well [Ren 13]. Multimodal approaches in affective research have been approached too, using multiple electrophysiology measurements at the same time, such as in [Koelstra 12]. This research also provides the collection of physiological signals online, supplying a data set consisting of multiple measurements as EEG, GSR, respiration amplitude, blood volume, face recordings and electrooculogram taken while users were watching audio-video stimuli.

As mentioned before, temperature is another monitored parameter. *Infrared thermography* has been used to demonstrate reflexes due to emotional video stimuli [Kistler 98]. In this case, the sympathetic vasoconstriction in forearms was monitored. A correlation with stimuli presentation was found, demonstrating local temperature changes. In detail, the video stimuli presented a thriller scene and the temperature decreased during this stimuli. The same principle has been recently adopted with fingertip temperature to implement the first steps toward an automatic recognition system [Shivakumar 12].

Evaluations related to the cardiovascular system for emotional impact are mainly heart pulse and blood pressure of systemic circulation; however blood fluxometry is another adopted measure. Also known as blood perfusion, this measure focuses on blood perfusion of peripheral tissues. Recently skin blood flow has been monitored with the help of Laser Doppler Perfusion in affective research for other kind of stimuli, notably tastes and odors of water [Haese]. At the best of our knowledge until now blood flowmetry alone has never been used for multimedia affective

research, as it has used only in coordination with thermal imaging [Kistler 98]. We detail this methodology, adopted in our first pilot study as later described, in annex C.

Based on these results, we wanted to explore the ability of these techniques for our purpose: to understand social context of portrait images exploiting subjective affective reactions.

2.3.3 Discussion

Different methodologies of emotions assessment have been underlined. Our review underlined that no best option is present, as each methodology offer pros and cons to consider in order to chose the best one for a particular study. From the point of view of simplicity, conventional explicit measures are most definitely a fast and easy solution, especially when adopting a direct self-assessment; provided answers do not need to be interpreted or treated. However, a subject under study is fully aware to be under evaluation and - moreover - he can be unable to correctly describe his/her feelings. From this point of view behavioral assessments are the best, as users are unaware of analysis but assessments require skilled personnel. Electrophysiology seems to be a valid alternative even if data analysis can be complicated and subjects are aware of undergoing affective assessment. Table 2.2 summarizes our review.

	PROS	CONS
Self assessments	Simplicity	User awareness, hard to describe emotions
Behavioral assessments	User unawareness	Complicated interpretation
Electrophysiology	«Body can't lie»	User awareness, data analysis

Table 2.2: Pros and cons of reviewed methodologies

2.4 Pilot studies

After studying what has been done in literature, we conducted two pilot studies to check the feasibility of the affective analysis track. Two case studies adopting different measures have been considered. The first one is related to laboratory experiments with blood perfusion measurements, a methodology recently used for affective research, while the second is related to the multimodal experiment carried out during a Short Term Scientific Mission (STMS) supported by COST Action

Qualinet. While these studies provided some partially positive results, they especially outlined that many factors influence the effectiveness of electrophysiology affective assessments. This fact underlined that a lot of work is still needed to effectively implement this methodology; for these reasons we preferred to put electrophysiology aside for now, in favor of crowdsourcing (ref. next chapter).

2.4.1 Study 1: Investigating Electrophysiology for Measuring Emotions Triggered by Audio Stimuli

Given the positive findings using LDPM for sympathetic responses [Kistler 98] and considering that the use of LDPM alone has showed interesting results for affective research with stimuli other than multimedia, notably water taste [Haese], we conducted a pilot study to evaluate the use of this technology alone within multimedia research and check how it relates with other better known alternatives. This technology measures variations in blood peripheral micro circulation, due to vasoconstriction, through Doppler analysis of a laser light reflections. Details are given in Annex C.

Participants in this study listened to widely adopted affective sounds. We made the choice of using sounds instead of portraits as we preferred to acquire the methodology with well known calibrated affective stimuli before trying to measure social context evaluations that can produce more subtle variations. While calibrated affective images also exist, the sound database was already available in our laboratory and this fact allowed us to quickly start investigating electrophysiology. Participants physiological reactions were recorded through LDPM. We resume here the description of the methodology and the main results (details in Annex D).

Method

A compact format is adopted here to describe the pilot study. The same format will be reused with successive studies in this work, notably the one described later in this chapter as well as in following chapters.

Participants

26 people, aged between 23 and 30, participated in the experiment. Before participation people have been informed about the experiment methodology, the kind of stimuli used, the time required and the anonymity of results. All participants were volunteers.



Figure 2.4.1: Experiment room in pilot study 1

Materials

A single modality stimuli - only audio - has been chosen to make hypothesis testing as simple as possible and avoid complications (i.e. we still don't know if there is any effect regarding multimodality audio video). A subset of the International Affective Digitized Sounds (IADS) database has been adopted as a stimuli source [Bradley 07]. A total of 15 stimuli has been chosen, representing all the different elicited emotions. These stimuli have been divided by clustering in two main groups, related to low and high affective value. The experiment relies completely on the effectiveness of chosen stimuli, and the hypothesis that the user actually perceives the content as expected, eliciting emotions, is demonstrated.

The experiment took place in a dedicated experiment room under controlled conditions; a noise free environment and a high quality headset made by AKG have been chosen. Audio volume was fixed on an audible level for all users and was not normalized (hypothesis of complete reliability of prepared sounds in the IADS database). Ambient temperature has been monitored, adopting a warm ambient with a temperature of 22 ± 1 C, as thermal factors such as core and skin temperatures deeply impact dermal blood flow and perspiration [Kondo 09, Maniewski 99].



Figure 2.4.2: The PERIMED LDPM device in our laboratory, with it's calibration device on top.

Measures

Blood perfusion has been measured as a peripheral electrophysiological parameter. The principle behind this measure is that the emotional state has an influence on blood circulation, i.e. making it faster/slower, and blood perfusion is an easy way to measure blood circulation. It has been measured only via Laser Doppler fluxometry, using a medical grade device made by Perimed, the PeriFlux System

5000 equipped with a PF 5010 LDPM unit (fig. 2.4.2). Data has been acquired through the official software provided, sampled at 125 Hz. Blood perfusion has been measured on the left index fingertip, as done similarly by [Kistler 98].

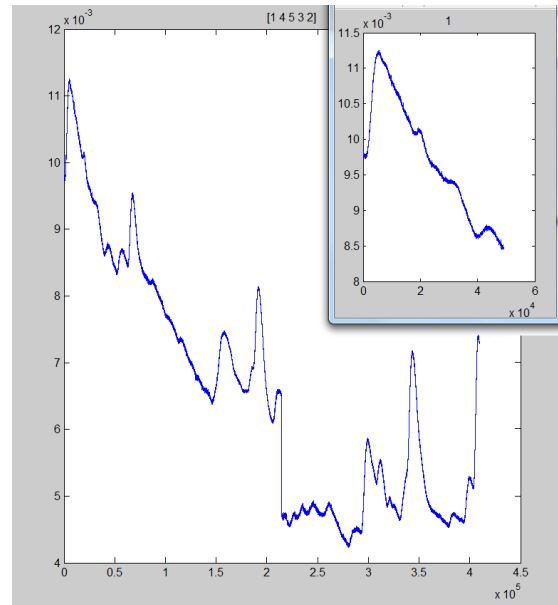


Figure 2.4.3: Some readings from LDPM measurements, with a detail of a rise fall pattern during a measurement. The X-axis represents time in seconds while the Y-axis represents Perfusion Units as defined by Perimed for the LDPM measurement.

Procedure

After welcoming each participant, subjects were brought to the controlled room approximately 6 minutes before starting the experiment in order to allow them i) to relax, as physical activity has been shown to impact on blood perfusion [Kvernmo 98] and ii) to adapt to this temperature, as many parameters impact on thermoregulatory time, and previous works show adaptation times inferior to this one in similar conditions [Kondo 09]. During this time frame the subject was instructed and instruments have been prepared and set up. The subject was informed only about the exact content of stimuli; purpose of the experiment has not been revealed in order not to false reactions during listening.

Successively subjects listened quietly to the sounds without interruptions, except for the pauses between different stimuli as detailed in D. Stimuli order has been randomized to minimize the bias introduced by presentation order or by any

cumulative effect on emotion that can arise. Afterwards the listening instrumental setup has been taken down. Participants were asked not to disclose stimuli contents to other future participants. Before dismissing the participant, free-form comments regarding the experiment have been asked, in order to improve experiment methodology and get helpful insights.

Data analysis and results

As a standard reference methodology for emotional state assessment doesn't exist to the best of our knowledge, methods suggested in literature for LDPM and EEG signals were followed. The LDPM signal was inspected to remove clear artifacts⁴, pre-processed to standardize it applying the z-score transformation as:

$$z = \frac{x - \mu}{\sigma}$$

Heart rate, present in the signal, has been filtered removing frequencies between 1 and 2 Hz, considering analysis done by [Kvandal 06]. It has then been windowed, to filter the signal and syncing it with proposed audio stimuli. The LDPM signal excerpts have been extracted - corresponding to instants when a stimuli was played - and machine learning has been implemented (ref. fig. 2.4.4). To test the research question, machine learning has been used to classify signal excerpts within two classes, depending on the type of stimuli: either low or high affective valence. A neural network approach - feedforward backpropagation, one hidden layer, 15% of data set for validation - achieves approximately 70% classification accuracy (table 2.3). It seems then that with LDPM it is possible to see physiological reactions to affective sounds. Attempts to achieve a finer classification - adopting the stimuli subjective assessments as ground truth to understand also the affective strength of each sound - produced results no higher than fate, adopting the same machine learning approach and LDPM signal excerpts.

Actual Class	Predicted Class		
		A	B
	A	73%	26%
	B	33%	66%

Table 2.3: Confusion matrix for LDPM classification; percentage of correct classification, for stimuli group low (A) and high (B) impact.

⁴i.e. user movements.

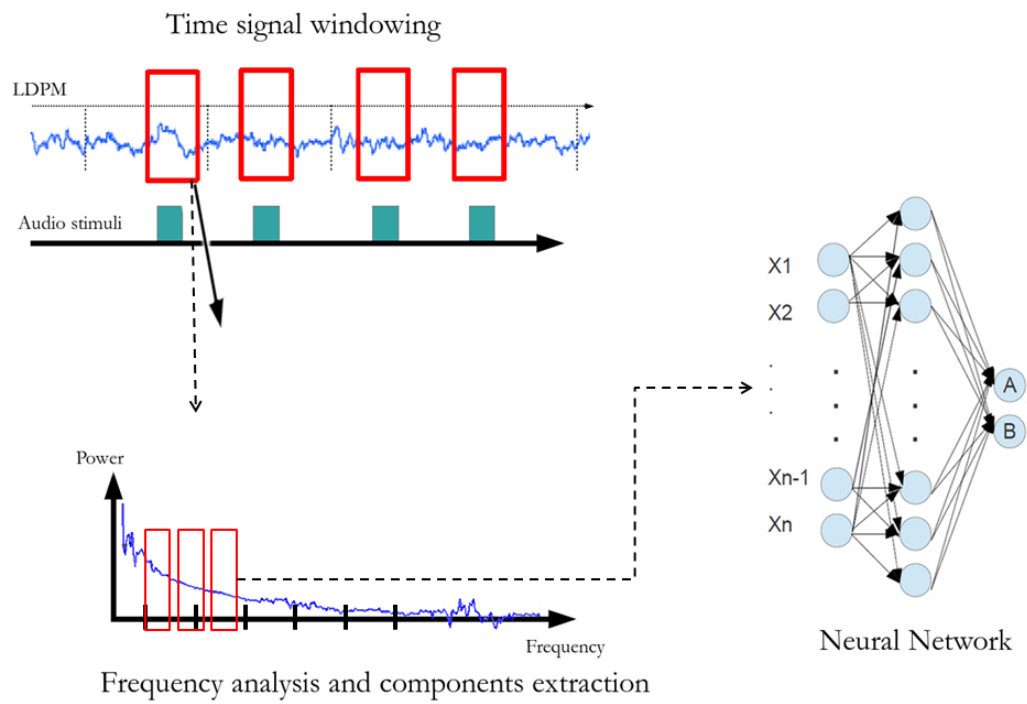


Figure 2.4.4: The schema of adopted data analysis.

Discussion

While classification accuracy is higher than fate for classifying high or low affective sounds, error rate is still important. Errors are mainly misclassifications of stimuli with high affective valence, classified as low ones. Multiple elements in our experiment can contribute to this outcome: the choice of probe position and small variabilities of its position due to differences in users fingertips; circadian rhythms influencing blood pressure baseline; valence of affective content used, previously evaluated from a different group of individuals, as those opinions may not reflect the actual perception of our subjects; stimuli length may be too restrictive, etc. Measured reactions to stimuli showed high variability from user to user; this can be caused also by the fact that some stimuli elicit more powerful emotions if related to previous users experiences.

Some comments given by participants after the test underlined that some stimuli were perceived as much more relevant as they made them remember personal facts.

While results obtained are encouraging, a lot of work is still to be done to reliably adopt LDPM alone with affective multimedia stimuli. This is even more important if stimuli have lower affective strength as portrait images will likely be.

2.4.2 Study 2: Chamber QoE – A Multi-instrumental Approach to Explore Affective Aspects in relation to Quality of Experience

Our previous study underlined that electrophysiology measures can be adopted to discriminate which kind of affective stimuli was proposed to a user, even if many challenges are present. To investigate if this methodology can also be adopted with non-affective stimuli - such as portraits - we conducted a second pilot study. In this case we did not use calibrated affective stimuli as before, but normal video sequences that might elicit emotions. This generalization suits our research with portrait images well, because these are not calibrated affective stimuli neither. In this case however, we preferred to use a better known method and adopted electrophysiology measures, the EEG, partially motivated by the positive approaches within Quality of Experience research (i.e. [Arndt 11]) and by the presence of different approaches in literature.

Participants in this experiment watched video excerpts impaired by artifacts. In this study we are interested only in users' affective reactions - i.e. in terms of delight or engagement - more than correlating artifacts presence with users reactions. Multiple techniques have been used to measure participants affective reactions. Our main purpose here is to understand if we can reliably assess with electrophysiology which kind of stimuli has been shown. The experiment has been carried out within the Qualinet Community as a joint effort, and research

institution partners took care of stimuli and complementary measures to EEG: traditional self-assessing questionnaires and facial expressions - recorded on video and analyzed through dedicated software⁵. We describe here the whole experiment, including affective measures alternative to electrophysiology. However, we will focus only on data analysis and results that are related to our main electrophysiology topic, for the sake of clarity. Analysis of complementary measures was carried out by research institutes participating this joint effort⁶. More information are in joint publication [De Moor 14]. As done for 1st pilot study, we resume here the methodology description and main results, (details in Annex E).

Despite our efforts, electrophysiology measures allowed us only to reliably discriminate between the first un-impaired relaxing video sequence and movie excerpts.

Method

Participants

27 people, aged between 25 and 35, participated in the experiment. Almost all participants were men, Finnish VTT employees. Just as for the first pilot study, people were informed before participating about the experiment methodology, the kind of stimuli used, the time required and the anonymity of results. Participants were compensated with cinema tickets, for a cost of approximately 15€.

Materials

Three video sequences from a famous action movie were selected as stimuli. Streaming impairments (i.e. IPTV-like) with three intensity levels were introduced into sequences. Another video sequence, made of relaxing music and neutral affective images, has been adopted as relaxing sequence. These sequences were provided by TU Berlin and we adopted them as provided.

Participants' emotional responses were recorded with different methodologies. Traditional self assessment questionnaires on paper were proposed after each movie sequence. These adopted a 10 point Absolute Category Rating for rating quality of videos, both SAM and PAM scales for affective assessments and also free text answers to ask for previous personal experiences regarding videos content. EEG

⁵Experiment took place in VTT Technical Research Centre of Finland, in Oulu, Finland. Research as been funded as Short Term Scientific Mission (STMS) from Qualinet COST Action, participants are in the Emotion Task Force, part of the 2nd working group. The core of this STSM application was to conduct a large-scale multi-lab joint research effort between IRCCyN IVC, VTT, the Aragon Institute of Technology (ITA), NTNU and TU Berlin / T-Labs.

⁶ NTNU examined the subjective assessments, ITA the facial expressions

measurements have been taken with an Emotiv Headset, a consumer EEG wireless device. Raw EEG signals have been recorded at a frequency of 128 Hz from the 14 electrodes of the device all along the experiment. Facial expressions have been recorded by Emotracker, a dedicated device developed by Aragon Institute of Technology. A personal computer has been used as a common time reference, shared between the devices.

Experiment took place in an experiment room in VTT institute. Room light, display, seat and stimuli audio video quality were controlled as detailed in E.



Figure 2.4.5: From right to left: the device adopted (Emotiv EEG) with its USB receiver, electrodes and saline solution.

Measures

Procedure

After an initial welcoming we described the experiment to the subject and let him sign the disclaimer to use his measurements and recordings. During this phase we showed self assessments questionnaires and explained how to use them. We continued placing the EEG device on participant's head, taking care of electrodes positions and contacts. The device's software (fig. 2.4.6) helped us to monitor the

connections' quality. Successively we adjusted the participant's seat and screen in order to have the same position for each participant: this was mandatory due to Emotracker limitations. After starting the recordings, we played the first relaxing video. Its purpose is to allow the user to get accustomed to the test environment. We then played the movie excerpts. The excerpts order followed the original scene order in the movie. Impairments strength was randomized. After showing each video sequence we proposed the self assessments questionnaires and then re-verified EEG electrodes signal quality. At the end of the test we dismissed the participant, after removing the EEG device and compensating him for his participation.

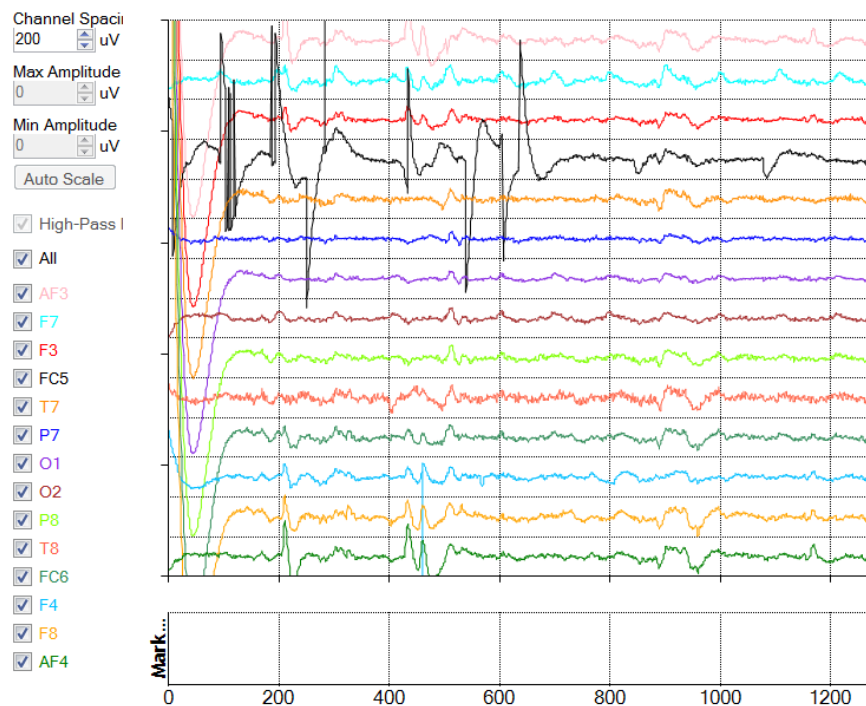


Figure 2.4.6: Reading of EEG measurement, for the first seconds of an experiment session. Re-positioning the fourth electrode produced artifacts that can be seen in the corresponding trace (4th from top).

Data analysis and results

We focus here only on our EEG analysis. Nevertheless, it is important to underline the main conclusions from facial expression analysis and self assessments, as these would have been useful as ground truth for affective evaluation. Unfortunately, both analysis underlined low affective reactions from participants. Facial expression in particular showed affective activations only in a minimal part. Only

self assessments underlined statistical correlations between scales of engagement and pleasure with videos perceived quality. Without a continuous affective ground truth for every video we then focused on affective differences between them. Videos and measurements were temporally synchronized thanks to the shared clock. To analyze EEG signal we adopted a band power approach, following methodology in [Bos 06]; only signals from electrodes related to frontal temporal lobes were used. Some portions were rejected as impaired from transmission errors due to RF interference - our device being wireless - and subjects' movements. In fact, it is likely to find in EEG signals activations due to head movements as with eyelids, raised eyebrows, etc. . Successively, signals were filtered and windowed to extract excerpts of 2 seconds in length. For each excerpt we extracted the alpha and beta frequencies from each electrode, that should be related to states of activation/engagement and relax, as done by Bos. We then computed the power of these components for each video, and for each participant. We discarded the first seconds from each recording in order to avoid considering the reactions to the beginning of a video. Neural network classification on EEG signals excerpts achieves around 78% recognition accuracy discriminating those belonging to moments where was played a relaxing or an action video.

	Predicted Class		
Actual Class		Relax	Action
	Relax	82%	18%
	Action	23%	77%

Table 2.4: Confusion matrix for video typology Neural Network classification based on EEG band power.

Discussion

Despite our efforts results are still limited and partial. In particular, the lack of a precise ground truth regarding elicited emotions seriously impacted the effectiveness of our study. Moreover we had to deal with different impairments analyzing the EEG signal. While RF interferences could have been eliminated by using a cabled device instead of a wireless one, user freedom would have been limited strongly. Bothersome artifacts coming from facial movements (i.e. eyes closed and opened) are instead inevitable. From time to time signals coming from electrodes weakened due to normal drying of saline solution used for electrodes contact.

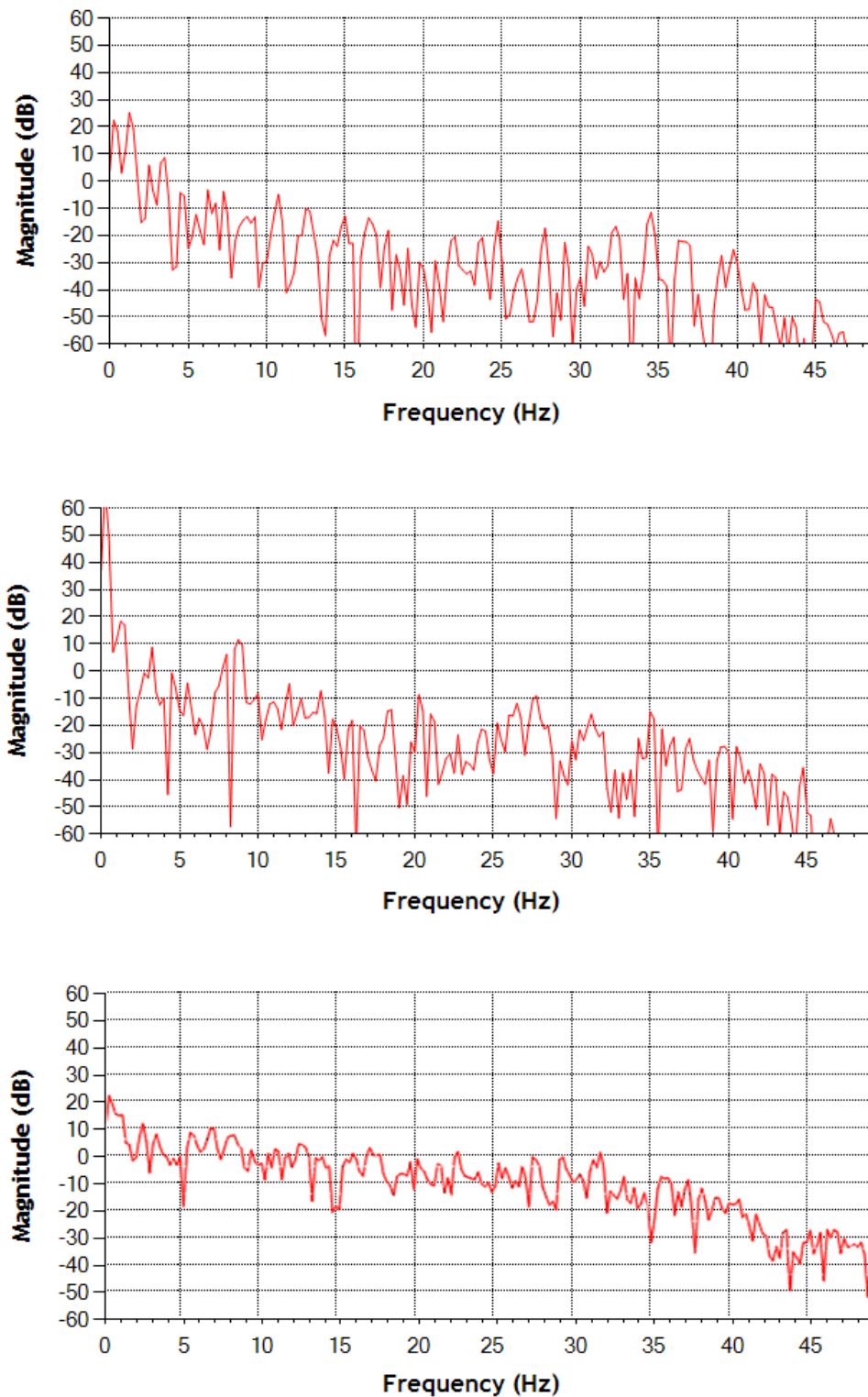


Figure 2.4.7: An example of EEG signal excerpts spectrum, relative to a neutral, low and high impairment video (from bottom to top, 4 seconds window). Differences are small, but present.

2.5 Challenges with electrophysiology

Research on affective research, adopting different electrophysiology measurements, demonstrated some positive results in literature and in our pilot studies. However many challenges still arise: finding «exact relationships between affect and physiology is problematic» [Janssen 09]. During case studies we discovered many small details, both methodological and practical, that greatly influence the outcome of an experiment.

A primary factor of influence is the test subject itself. Physiological reactions vary greatly between subjects and even for the same subject they vary with time, conditions, etc. . Between the most important hypothesized variables were the probes positioning, the room temperature, the personal fatigue/stress levels, other personal biological factors. While some suggestions about these elements are present in literature, a complete “good practices” list is still missing at the best of my knowledge, without referring of course to studies in the medical field. Although some effects can be taken into account as user position, many others are difficult to control, such as the comfort degree felt by the subject during the test and the current fatigue level and mood of the participant. In particular, the recall of previous personal experiences from affective stimuli - as outlined in Pilot Study 1 - is a very interesting and important factor to take into account in this kind of research.

Another critical factor relies in electrophysiology technologies. Different methodologies may be suitable with different stimuli or subjects, as reactions may be inherently different. Regarding this aspect, multiple technologies can even be used at the same time, as they can be complementary. However, it is not always feasible both in terms of costs but especially in terms of impact on test subject. To limit his freedom or make him uncomfortable (or tired, as happens with many subjective questionnaires) impacts of course his feelings by definition. Technologies adopted can also pose problems regarding technical aspects. For example an element to consider carefully is probes positioning. Small differences from optimal position can sensibly lower the effectiveness of the measurement; the «optimal [electrodes] position may vary across subjects» [Li 09]. However in our case even taking great care positioning the probes we had low signal to noise ratio in a few cases, both in the first and second experiment. This fact is unfortunately normal somehow, as electrodes should be placed firmly but at the same time they should leave some freedom to the user, otherwise they will be disturbed by their presence, biasing their emotional response. While other technologies as thermal imaging may solve this issue, they are usually much more expensive and have other limiting factors that greatly influence accuracy (i.e. subject position, external factors such as temperature).

This point brings us to a third important factor of influence, the environment. Not only can the environment influence the user itself, but also impair measurements. For example the room temperature can influence subject homeostasis and micro-circulation. If electrophysiology measures rely on this aspect, we have to pay great attention to this. Ambient sound should be taken into account too. Unexpected sounds may induce sudden subject reactions. While this factor should be controlled also in normal non-affective tests, in our case this element is even more important. Still related to the environment, radiofrequency interference can influence measurements. In some cases currents measured by electrodes are in the order of few milliAmperes, it is easy to affect this kind of measures, if radio signals are in the same frequency range.

All these elements underline how fragile affective research can be without taking into account all these elements. A «good practices» research study in the field is to the best of our knowledge still missing. Moreover, considering our two pilot studies as well as reviewed SoA, it seems that every experiment requires a different methodology and adroitness.

2.6 Conclusions

In this chapter we underlined motivations for our interest in emotional assessments, we reviewed the SoA regarding electrophysiology for affective studies and showed preliminary results on affective research for multimedia evaluation purposes. We conducted two pilot studies recording affective reactions, allowing us to take confidence with the instruments, the protocols and to competently gather sets of data measurements to work with. Moreover they helped us in identifying main challenges.

The SoA clearly underlined that this is a tentative and innovative methodology, bringing up many really interesting points as well as many unknown variables: emotional effects are inherently subjective. However even if people show different reactions, we were expecting some common findings that could have helped in considering emotions' effects in evaluations. Results that describe clear, links between emotional state via electrophysiological measurements and quality of multimedia experience have not been reached yet. While this opens possibilities to further investigate the topic, it does not constitute a main objective of our research and other strategies seem more promising. For these reasons, rather than continuing to dig into these problems, it has been decided to give priority to a different kind of evaluation, crowdsourcing, as explained in next chapter.

However, the work done on electrophysiology has not been a waste of time. As said, pilot studies provided some first positive results and underlined many critical point to consider. Useful lessons have been learned from them, giving birth to

scientific publications (ref.). Two interesting points have been raised and would be worth investigating them in further research studies. The first one is related to the influence of participants moods before the experiment on assessments. The second one is if emotions inducted by a stimuli bias successive evaluations. If these biases are present and how they can be taken into account has never been investigated at the best of our knowledge, at least in this particular field. More than these two points, we leave also for future research the possibility to deepen data analysis, exploiting information unexploited in pilot studies.

So to conclude, it is true what the psychologist Watzlawick says: «One cannot not communicate»⁷. However, the problem of being able to listen and understand reliably still remains.

⁷Axiom of communication, www.wanterfall.com/Communication-Watzlawick's-Axioms.htm, retrieved July 2015.

Keypoints

Context

- ❑ Electrophysiology has been used to measure physiological body reactions in subjects exposed to multimedia stimuli. Research attempted to link these reactions with multimedia characteristics (i.e. technical quality), in order to find an alternative to classical explicit assessments.
- ❑ Different technologies, some more invasive than others, have been adopted; EEG and thermal imaging are two widely known examples. Moreover, even with the same technology, many different approaches and methodologies are present.
- ❑ Results are highly variable. Some studies found good results in very particular and specific cases, suggesting that this methodology can be adopted within well defined and controlled cases.
- ❑ Theories of emotions coming from the field of psychology suggest that there is a strong link between subjective opinions and emotions.

Contributions

- ❑ Analysis of electrophysiology SoA within the particular context of affective multimedia assessment.
- ❑ Attempt to adopt electrophysiology in our field through two specific use cases linked to multimedia assessments.
- ❑ Recommendations coming from case studies, that revealed inherent problems to address in order to effectively adopt this strategy.

“Le cœur a ses raisons que la raison ne connaît point.”

(Blaise Pascal)

Chapter 3

Exploring the potential of crowdsourcing as an alternative to in-laboratory experiments with questionnaires

In this chapter we discuss the adoption of crowdsourcing as an alternative methodology to carry subjective experiments involving questionnaires. This relatively new technique exploits the power of the web to outsource small tasks to people gathered online. It has already been investigated intensively and positively exploited in multimedia quality assessments, notably within the Quality of Experience research field. We review here existing literature, underlining the pros and the cons. The validity of crowdsourcing being clear, we decided to use it for our research. The next chapter describes how we adapted it for our purposes.

3.1 Introduction

Previous chapter underlined the importance of emotions in the decisional process. Encouraged by preliminary positive results in research - notably for QoE - we investigated electrophysiology as an alternative measure. However, many problems influencing results have been found, related to external factors as well as highly subjective differences between people. For these reasons, instead of continuing to dig into these problems, we decided to give priority to alternative methodologies for subjective evaluations.

We then investigated crowdsourcing, already proven to be effective and reliable in research; this novel technique is the topic of present chapter. General description of the methodology and adoption in scientific research are described in next

sections (ref. 3.2.1-3.2.3). Crowdsourcing outsources small simple tasks to a large audience, usually anonymous crowd gathered online. This possibility can be exploited to propose the test questionnaires that are already adopted in laboratory environments, but online. Attention is shifted to this alternative methodology for different reasons. First, it quickly provides a lot of participants; this can be very useful while investigating many influential factors at the same time. We discuss about this point in chapter 5. Secondly, it is much cheaper in economical terms compared to usual tests in laboratory studies; we avoid many costs to run the experiment (i.e. engineers and materials), as it runs online, and the regular payment for a participant is usually a tenth of that paid to participants coming to the laboratory. Lastly, participants' demographic diversity is another positive point to underline: experiments run worldwide online and then we can reach a much broader audience, richer in terms of nationalities as well as participant profiles. However all these positive points come at a price: more attention and longer data analysis are needed to avoid undesired/outlying behaviors (ref. 3.3) and a longer preparation of the experiment; we discuss about practical implementation in next chapter. Next section deals with the SoA of this technique.

3.2 The technique of Crowdsourcing

Objective of this section is to describe the practice of crowdsourcing. We will outline the State-of-the-Art dividing it into different sections, outlining generalities, adoption in scientific research, negative points and countermeasures. We want to underline that while in theory crowdsourcing by itself can be both run online or offline, we will refer only to the online crowdsourcing if not otherwise specified.

3.2.1 Generalities

Crowdsourcing (CS) is a recent practice adopted in many fields and it appeared around fifteen years ago. The name, given by Jeff Howe in an article of Wire magazine [Howe 06] comes from the union of the words «crowd» and «outsourcing». The main concept is to outsource a task commonly done in a controlled context or by a dedicated worker to a crowd of unknown uncoordinated people. Today, crowdsourcing is exploited for four main objectives: accomplish numerous repeated small uncoordinated tasks, dispatch different parts of work, collect ideas and raise money (ref. fig. 3.2.1). We focus here on the first two use cases, especially on the first one. In scientific research, the term is usually adopted to simply designate the exploitation of a large number of workers to accomplish a task. The technique is very powerful when a lot of simple tasks are needed to be done simultaneously or when an harder task can be decomposed in many simple serialized tasks. A



Figure 3.2.1: Infographic realized to resume Crowdsourcing uses.

typical example is the translation of a book: one approach is to give the book to a translator, another one is to give each page to a different interpreter. The advantage of the second solution - crowdsourcing the job - is clear, as it would require *less time*. It will probably require *less money* too, as it is easier to find a lot of workers willing to perform small tasks, and these are less expensive than paying much bigger effort (translating an entire book demands a bigger investment, given the same quality¹). Moreover, CS jobs are usually very small and paid in small amounts of money: nowadays they require less than five minutes and are paid less than a dollar in commercial platforms. As a reference price Ribeiro indicates a cost of 0.6 \$ to have enough subjective MOS evaluations per image [Ribeiro 11] while Snow with 1\$ obtained 3500 useful annotations [Snow 08]. In our CS experiments we paid every participation around half a dollar. Compared to common laboratory experiments in general, where every participation can be rewarded with an equivalent of around 10\$, CS is much cheaper. Crowdsourcing has also been underlined to offer much more flexibility as it offers solutions for on-demand operations without long-term commitments [Kuikkaniemi 11]. The smaller works addressed in CS are commonly named simply as «tasks» or HITs - Human Intelligence Tasks - in crowdsourcing context, whereas the group of allowed positions for accomplishing a task is called «campaign».

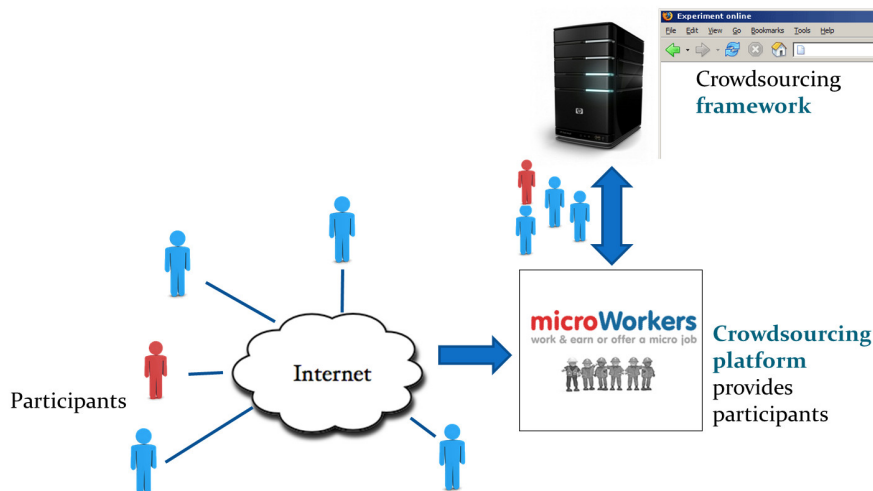


Figure 3.2.2: General schema of online Crowdsourcing.

While crowdsourcing campaigns are usually paid, different examples of non-paid crowdsourcing are present, adopting explicit as well as implicit jobs: for example,

¹We do not consider here CS initial costs, supposing them to be not considerable.

when crowdsourcing is employing games to motivate participants (see 3.4.2) or is calling for volunteering: some other services rely on communities that work for a common objective, that becomes the real commercial service. We discuss about this point in 3.2.2.

In theory, CS can be also performed offline, i.e. exploiting regular mail for the tasks. However with the increasing power of Internet, crowdsourcing has seen a huge growth and nowadays the outsourcing is *done through the Internet*: people participate connecting through their devices to dedicated online platforms. These platforms propose a list of available jobs posted by employers, that is to say people having a task to be solved. We talk about some of these platforms in paragraph 4.2.1. Jobs are proposed in the form of small tasks to be accomplished through a normal web browser, following the instructions provided by employer. Upon participant work acceptance, CS platforms redirect the participant to the employer website, where the actual work is done. A general outline is provided in figure 3.2.2. Online there is a huge variety of tasks, in the form of an open call: we can find examples of very simple and independent ones, as clicking «Like» on a Facebook page, as well as more complex ones like the book translation in our previous example. Tasks can even be part of bigger projects, as done by [Kawrykow 12] for DNA sequences alignment. The possibility to *exploit human intelligence*, much more powerful than any other computer algorithm, is another great advantage of crowdsourcing, as it helps in solving very complex problems. To make these problems simpler and to motivate participants at the same time, gamification has been proposed; we discuss this point in section 3.4.2. However, boosting participation with a game requires a lot of effort in making an entertaining game, that is not as simple as it seems. For the sake of completion, it is important to underline that unfortunately the power of an intelligent crowd has been adopted also for malicious purposes, as for massive spam generation and unfaithful reviews for marketing purposes [Eaton 10, Wang 12]. However we do not dig into this topic.

Scientific experiments can be proposed in the same way. This strategy has already been positively adopted, especially where large amounts of data are needed, as in multimedia quality research where subjective assessments are needed. Crowdsourcing showed to be in certain cases a valid cheap methodology [Gardlo 12a, Snow 08] and consistent with laboratory setups [Figuerola Salas 13]. We conduct an analysis of the State of the Art in 3.2.3.

The internet offers another big advantage, a much *richer demography of participants*, as they usually come from all over the world. Audience demonstrates big differences in terms of culture but also in terms of age and level of instruction. Of course, internet users are much more likely to be reached, but audience is not limited to them. As we underline later with our experiments, participants back-

ground is very different :students, engineers, lawyers and freelancers participated in our experiments, just to name a few. Figure 4.4.11 shows part of works carried out by participants. In particular, many young students from Eastern Asia are likely to participate, due to the monetary value of CS pay.

It has to be noted that even if crowd is anonymous and spread worldwide, it uses blogs to communicate². This allows the crowd to share information about works to be accomplished, as give and get suggestions regarding most interesting works or tasks which malfunction to avoid. The presence of blogs is also very useful for employers as they can monitor discussions there and retrieve indirect feedback to improve both their jobs and workers satisfaction.

To conclude, even if CS has many positive points and has been positively adopted many times, some factors can't be controlled at all and can cause problems, as we see in section 3.3.

3.2.2 Commercial projects exploiting unpaid crowdsourcing

By itself, the concept of crowdsourcing can be very broad: also the adoption of communities that work towards a common objective can be considered as such. In some cases, community's objective becomes also a commercial service but the crowdsourcing comes for free from the users. For example, online projects are positively adopting this business model [Times 15], like with popular networks to rate restaurants and hotels: Forsquare³ and TripAdvisor⁴ contain reviews of people who have been to a place and want to publicly share their opinions. Later, this data is used after to rank places and suggest itineraries, providing a service to users planning a trip. The revenue comes with the commercial exploitation of the service itself (i.e. advertising). However, the work of the community may also come for free. This is the case of Wikipedia⁵, the free encyclopedia, where every user is a contributor and the encyclopedia is actually written, reviewed and maintained by crowdsourcing. Less explicit forms of crowdsourcing are exploiting the behavior of users as an indirect measure for a service, as Google did for partially ranking the web search engine results [Times 15]. Waze⁶ navigation software exploits crowdsourcing, both explicitly and implicitly. The former is based on the explicit feedback that users give when using the navigator while on a trip: user is questioned when software detects a stop in a trip, i.e. to ask gas prices at gas stations or presence of incidents / roadworks. However, even just the ongoing journey is an implicit piece of information gathered in crowdsourcing: speed, direction, path and

²For example in <http://www.crowdsourcing.org/directory> and related links.

³<https://it.foursquare.com/>

⁴<http://www.tripadvisor.com/>

⁵<https://en.wikipedia.org>

⁶<https://www.waze.com/>

more are used to build and maintain road maps (fig. 3.2.3). Another important example is with Questions and Answers communities, like Stack Overflow⁷ and Quora⁸. In these communities the knowledge of the crowd is adopted to answer questions input on the platform. A particular remark should be made for social networks, that under certain aspects they share characteristics of crowdsourcing: they also gather a large crowd of individuals, providing many outputs. However, this crowd is usually not anonymous and the common task is not defined. We then consider social network as a special case but we won't dig into this aspect. Still, it is worth to mention some projects between social networks and crowdsourcing communities as «Photo.net»⁹. Here the crowd is gathered around the objective of providing and rating pictures taken by the crowd itself. Pictures are also labeled and commented by users.

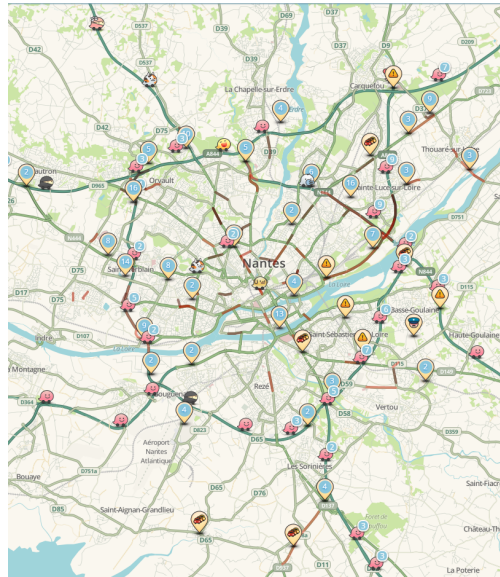


Figure 3.2.3: screenshot of Waze online livemap¹⁰, showing the city of Nantes in the afternoon and real drivers traveling on the ring road. Many user provided reports are displayed, grouped together (small yellow bubbles on the map).

The forms of crowdsourcing that we mentioned are voluntary and free: people participate to get a free service or to improve the service itself. Section 4.2.1 is dedicated instead to commercial platforms in which tasks are paid. However,

⁷<http://stackoverflow.com/>

⁸<http://www.quora.com/>

⁹www.photo.net

¹⁰<https://www.waze.com/livemap>

participation is huge even in free crowdsourcing services: Wikipedia in English accounts for more than 25k new articles monthly, reviewed by than 33k active editors¹¹, StackOverflow provides almost 16 millions answers¹² and Tripadvisor accounts for almost 140 contributions every minute¹³. The idea of gathering volunteers as participants led researchers to focus on motivation strategies such as gamification for entertaining participants: we see this point later on in section 3.4.2.

3.2.3 Adoption of crowdsourcing in scientific research

Crowdsourcing has been adopted in scientific research too, providing positive results. In particular CS has been proven to be very useful when dealing with large amounts of data, for example in researches that rely on a big ground truth to train models or with subjective assessments. This last point has been intensively investigated especially in the domain of Quality of Experience, where human factors are believed to play an important role in the overall research scope. In this section we briefly outline first works that adopt this technique, before reviewing the most remarkable ones that helped us for our research. Analysis of State of the Art continues also in following sections.

Between the first works addressing CS-like strategies (even if the name crowdsourcing came later in time) was the Open Mind initiative, proposed by Stork in [Stork 99]. In this work the author proposes an initiative based on open source to build a framework for «large scale collaborative efforts» to address «document and language understanding, speech and character recognition, and so on». The initiative considered the participation of non specialist «e-citizens» to contribute in training and data labeling. Later on, the concept of «distributed knowledge acquisition» online was under development by von Ahn et Al., who were studying a way to get a large amount of human labels for images. They developed different approaches based on gamification: we review these approaches later in section 3.4.2. However, the term crowdsourcing was not yet coined: it appeared only when Jeff Howe talked about it later on. At the same time Amazon Mechanical Turk (AMT), a commercial CS platform opened the year before, was becoming popular. Scientific research started using it few years later; during the time this platform became one of the most important and is still highly active today. In [Kittur 08] authors adopted AMT and investigated the utility of a «micro-task market» - provided by CS - in order to collect quick and cheap quality assessments of Wikipedia articles. Their experiments aimed to compare subjective ratings of

¹¹<http://stats.wikimedia.org/EN/SummaryEN.htm>

¹²<http://data.stackexchange.com/>

¹³http://www.tripadvisor.com/PressCenter-c4-Fact_Sheet.html

CS versus normal reliable raters: they found poor correlations when not adopting reliability checks, suggesting that the latter removed outliers effectively. Authors concluded that this strategy was a promising technique and that special care must be given to task design. Questions about the reliability of CS were raised at the same time from other researches: in particular Snow et Al. perfectly underlined the problem with the title of their work «Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks» [Snow 08]. This work focus on quality of non-expert annotations on five tasks related to natural language understanding. Their conclusion is that a careful design of the experiment as well as a careful screening of gathered data can greatly mitigate the problems and allow them to run reliable tests online.

In the following years CS became really popular and an impressive number of research publications adopted it: more than 800 are reported only by IEEE considering conferences and journals. More than 100 are focused especially on multimedia¹⁴. To make an extensive review of these works is out of our scope and we focus on those that helped us in developing our methodology. For a deeper review we suggest PhD thesis of B. Gardlo [Gardlo 12a] and the more recent review of Hosseini [Hosseini 14]. We underline that CS has been adopted for both objective and subjective tasks; while tasks are different, from CS technical point of view there is no clear difference between the two. Moreover, as underlined in next paragraph, some strategies can be used for both purposes.

Adoption of CS for labeling

Social context perception of portraits is subjective. However, as explained in next chapter, we can propose this task as labeling: participants can express their opinion adopting labels among a fixed set of possible values. This strategy has already been adopted in CS, where labeling has been addressed many times both for objective tasks and to express subjective opinions. In [von Ahn 04] image labeling has been proposed as game (ref. in 3.4.2) and then proposed aside famous anti-bot CAPTCHAs for text transcription [von Ahn 08]; in [Welinder 10] labels considering uncertainty have been adopted for large image databases; in [Loni 13] crowdsourcing has been attempted for assessment of high level labeling of social multimedia fashion-related content. The technique has been also adopted to annotate videos, as in [Steiner 11]: here the focus is on semantic annotation so that users are unaware of taking part in a crowdsourcing task (fig. 3.2.4). In Galaxy Zoo, participants helped in classifying the huge collection of galaxies observed with space telescopes around the world [Lintott 08]. The project, accessible online, provides a nice interface to guide the labeling process (see fig. 5.2.4 in section 5.2).

¹⁴IEEE Xplore Digital Library, querying «crowdsourcing», retrieved online June 3rd, 2015.

CS has been proposed also to label stimuli other than images or videos. This is the



Figure 3.2.4: The interface of the browser extension adopted to label video entities in [Steiner 11].

case of emotional speech assets, labeled in order to gather training samples for emotional speech recognition as done in [Tarasov 10]. Similarly, it has been adopted to summarize relevant entities in texts, such as news articles [Demartini 10]. Sentiment analysis has been addressed with CS, especially where non expert ratings are sufficient. This is the case of research carried out in [Brew 10], where sentiment analysis of online media is addressed with a large scale annotation. Labeling based on a simple ternary choice (positive, negative or irrelevant feeling) is collected with the objective of producing aggregated statistics for large collections of news articles. Binary labeling has been adopted instead in [Grady 10] to evaluate results relevance in search engines. This work investigates different factors influencing CS assessments, notably time, influencing cost and terminology adopted in the study. Their results are however «largely inconclusive» due to a number of encountered problems: we see in next section that while CS is very powerful has many drawbacks to consider in order to make this technique successful.

Adoption of CS for subjective studies

Another use of CS in scientific research is for subjective tests, where a large number of participants are welcome. This technique is also interesting as it allows to collect subjective evaluations from a much broader and heterogeneous audience than the one accessible within a normal laboratory environment. Among the first

to address the topic there is Tobias Hossfeld et Al.; their research is a reference in the field. For many years they addressed different aspects of CS. In [Hoßfeld 11c] they proposed a crowdsourcing based QoE assessment methodology and applied it to YouTube QoE assessment, validating again CS also for multimedia subjective assessments. It also underlined how CS allows the study of demographic factors differently than in lab studies. Also, their works conducting statistical analysis on CS commercial platforms are important ([Hoßfeld 11b], [Hirth 11b]). In this last joint study they underlined the difference between posted jobs on platforms - in terms of time and cost - but also important differences especially between provided workers, regarding demographics. Their study produced guidelines for all the actors of CS: platforms operators, employers but also workers. Hossfeld and Hirth also addressed the problem of unreliable participants in CS, proposing cheat-detection mechanisms [Hirth 10]: we discuss more about this work in section 3.3.1.

3.3 Crowdsourcing negative points

While crowdsourcing has many positive points, it also has many negative points. In our previous example of crowdsourcing a book translation, we made a fundamental implicit assumption: that the achieved quality was equal to or at least comparable between the normal translation and the CS solution. However especially this may not be the case in crowdsourcing. Table 3.1 summarizes the main problem found with possible countermeasures. Current section summarizes main problems in CS, underlined in literature or found in our practical experiences. Part of this work has been adopted as input to the joint Qualinet white paper on Crowdsourcing Good Practices ([Hoßfeld 14]).

Crowdsourcing is mainly affected by two important negative points: firstly, *participants are unreliable*, as they perform proposed tests without supervision, and secondly because *participants' context is unknown*, both in terms of hardware or connection and environment around. Participants can be unreliable as they make mistakes due to lack of attention or misunderstanding of the instructions given, especially if these are not in their native language, or by a dishonest behavior ([Gadiraju 15]). In fact, participants are paid by task and people can try to maximize the revenue minimizing the time per task. As said by Gardlo in previous cited work, “While workers from under developed countries are very likely to depend on the money earned [...], workers from very high developed countries normally work “for fun” or to earn a little extra money”. However, control systems have been increasingly put in place by both employers and platforms. Workers are also paid by correctness of results. For example in Microworkers, participants may be asked to provide a proof of task completion to get paid (proof of completion),

and this is given at the end of the task itself under certain conditions. These conditions are imposed and measured by employer, supposing the correctness of a work is measurable. If possible, this is also done in real time through software. To this extent some researchers proposed cheat detection mechanisms in order to measure the attention the worker is paying to the work. We talk about them in next section.

Another point to underline is that participants can prematurely withdraw from the test: participants are usually much less motivated in CS than in laboratory, so if jobs last too long for the pay they will be likely to withdraw and participate to more rewarding jobs. A remarkable example is given by [Keimel 12], where paying only 0.08\$ to evaluate 28 videos resulted in having «only 7% of all workers assessed the complete test set and 83% of all workers finished less than half of all videos». Pay and stimuli number/required time for completing the experiment must be considered carefully. Moreover, a fatigue effect can impair results, so long experiments should be avoided in any case. Crowdsourcing experiments should be no longer than 5 to 10 minutes [Redi 13b, Hoßfeld 14].

Without supervision, users can easily skip non-compulsory tasks: if users are free not to perform a particular action in a CS job, they probably will skip it. For example, this is what happened in [Keimel 12], where users weren't checked actually viewing videos before rating, leading to incorrect user behavior. We found the same problem in our first pilot study: helping users with pre-compiled forms leaves the opportunity to use the default value even if wrong (ref. sect. 4.4.2).

However reliable participants alone may not guarantee 100% reliable results. While in-laboratory participants have a well-defined environment, their context and hardware adopted to participate in crowdsourcing can vary greatly from user to user. Moreover people can participate from wherever they want and some environments can be unsuitable to participate. For example, in the case of users participating with small screens unable to display our contents. Very different devices are used by participants; some parameters can be detected and checked but for others we need to rely on other elements, as «a posteriori» controls on results useful to detect outliers.

Crowdsourcing can also pose problems regarding the difference in both devices and Internet connections adopted to participate. This point can be both a positive feature as we have participants from anywhere in the world, but is possible that they have much slower connections - that can impair the normal execution of a test¹⁵. Differences exist also in terms of devices: screen size, gamma, resolution, luminance and contrast, audio quality will change, especially if we consider mobile devices. These factors can bias experiment results especially when dealing

¹⁵A factor of 10 has been shown on average internet bandwidth available between CS users from Asia and Europe [Hoßfeld 11a].

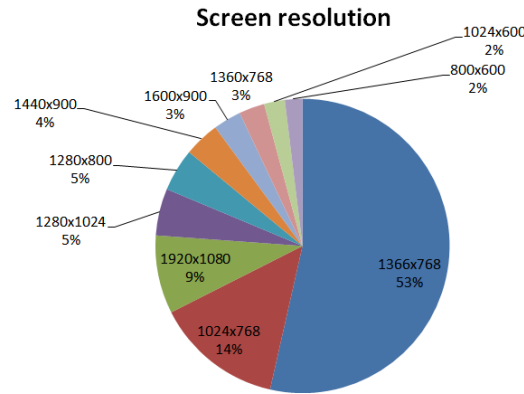


Figure 3.3.1: Detected screen resolution of a part our tests' participants, N~1800 (preliminary tests plus study described in chap.5).

with multimedia researches. On the contrary, if these factors are properly taken into account (i.e. measuring device characteristics with automated software during an experiment), they can be very useful in research. Interesting statistics on connections and devices differences between different regions of the world are reported in CS by [Gardlo 12a]. Our measurements for preliminary study and main experiments (ref. chap. 4 - 5) are shown in figure 3.3.1.

World wide participation leads to another possible problem: *participants demographic unbalance*. A huge amount of participants come nowadays from East-Asian Countries near India, Bangladesh and Nepal. This fact has been noted by different researches, as in [Hirth 11b], and also by our studies (ref. 4.4.2)¹⁶. While this fact is not a problem per se, it can become a problem in experiments where cultural biases play a role. Research on multimedia quality assessment started to consider these as influential factors, without detecting any significant impact of other factors like «age, level of internet usage or content type»[Hoßfeld 11c]. The main cause of imbalance is money; the small economical rewards in CS are indeed important in poor Countries. To mitigate this problem, some platforms allow to exclude/select Countries for experiments. Another possibility is to modulate allowance rate, that is the number of participants allowed per unit of time¹⁷, taking into account timezones. In particular, as those in developing countries are mainly in Eastern Asia, their activity will be prominent during hours that correspond to night in Europe. This time zone difference can then become important

¹⁶Participants' distribution strongly depends on practical factors such as the platform and the timezone too. We refer here to the case of European CS platforms (i.e. in US Mechanical Turk only allows US residents nowadays).

¹⁷Microworkers platform calls it "Campaign Speed" and allows to modify it dynamically.

to help in selecting a particular area of the world. However this factor alone is not enough; developing Countries' users have been reported to participate slightly even during the night [Hirth 11b]. Particular attention is required to choose when to launch an experiment online, not only in terms of timezone but also considering holiday periods, different from Country to Country, that can affect participation demographics.

These factors bring the need to check CS gathered data and participants during and after the experiment, as we will see in next section. However, some problems are related to practical implementation; the choice of CS platform and software framework adopted, topic of section 4.2, is particularly important.

PROBLEM	COUNTERMEASURE
Unreliability of participants	Reliability tests / honeypots
Unknown device adopted	Preliminary technical checks in frameworks
Poor motivation	Adoption of motivation strategies
Need of shorter experiment length	Cut experiments in parts
Participants demographic unbalance	Paying attention to experiment pay and start time; actively limit participations

Table 3.1: Summary of crowdsourcing problems and possible countermeasures.

3.3.1 Reliability checks

Previous considerations underline that proper reliability checks should be designed. In the simplest form, they are based on delivered jobs analysis. More intelligent strategies involve content related questions or traps within tests. While an official nomenclature is missing, many times literature addresses hidden traps as 'honeypots'¹⁸, referring implicitly to the fact that people are 'lured' into something that is instead forbidden. A clear example can be asking to fill a questionnaire but leaving the possibility to skip it.

Content related questions have been adopted since the first adoptions of CS, when [Kittur 08] found improvements in results asking participants to answer verifiable questions regarding stimuli before rating them. As proposed stimuli were

¹⁸In this work we will adopt the term with the general meaning of reliability check if not differently specified.

The interface is divided into two main sections for video comparison. The left section, titled "Undistorted original video, perfect quality", shows a video player with a colorful, cartoonish scene featuring a calendar, a red mushroom, and various animals. The right section, titled "Distorted video for evaluation", shows the same scene but with visible distortions. Below the video players, a question asks: "How do you rate the distortions in the right, distorted video compared to the original video?". There are five radio button options: "Imperceptible", "Perceptible, but not annoying", "Slightly annoying", "Annoying", and "Very annoying". Below the question, a note states: "The following both fields will be filled in automatically after answering the question or watching the video respectively." There are two input fields: "Video watched completely?" and "Question answered?", both with a "No" value. At the bottom, there is a blue "Submit task" button.

Undistorted original video, perfect quality Distorted video for evaluation

How do you rate the distortions in the right, distorted video compared to the original video?

☐ Imperceptible
☐ Perceptible, but not annoying
☐ Slightly annoying
☐ Annoying
☐ Very annoying

The following both fields will be filled in automatically after answering the question or watching the video respectively.

Video watched completely?
No

Question answered?
No

Submit task

Figure 3.3.2: Interface of study [Keimel 12]: videos are presented aside and should be played by participants. However the task can be skipped («Submit» below). SOURCE original paper, courtesy of authors.

news articles, researchers added four content related questions. Not only they found a bigger correlation with normal evaluations, but they also found a bigger withdrawal rate: it is likely that participants who just wanted to earn some quick money ended the test as soon as they discovered that indeed they must pay attention to the task to earn the wage. As we will detail later in 4.4.1, we adopted the same technique of content questions to screen participants in our pilot study. Instead in [Isola 11b] authors adopted 'vigilance tasks' proposing repeated assessments of the same images. This strategy allows to continuously monitor participants. In [Kuikkaniemi 11] consistency checks between different portions of same task are considered: tasks staked in levels, in which upper tiers are dedicated to validation of completed tasks. This 'peer-review' achieves better results quality exploiting redundancy and the large number of participants available at a low price. Also Hirth et Al. [Hirth 10] addressed the problem of unreliable participants. Notably, they proposed two strategies: a majority decision based algorithm and an approach with a control group. The first method just considers the results provided for the same task by different workers and compares them. The second one instead adopts a two stage work, where in the second stage a group of workers control the work provided by a first worker. However, in their models the hypothesis is that only cheaters submit wrong results, as their methodologies do not consider accidental mistakes. Both approaches are useful to detect outliers in objective assessments, i.e. in image classification. Indeed in case of subjective assessments we have some problems with their approach: a majority decision will definitely threshold the assessment while a control group won't be possible, as the two groups of workers may disagree; this would be perfectly possible due to the inherent subjectivity of the task. Nevertheless, the limits of subjective preferences variability are not straightforward to chose. Their choice could sensibly change the final outcome.

3.3.2 Evaluating collected data

Even with proper reliability checks, multiple factors influence reliability of results. Data evaluation techniques have then been proposed for CS, either re-adopting existing techniques for general data analysis or exploiting the large amount of cheap data available in CS.

In CS data labeling high uncertainty is usually present as it is affected by the presence of bad annotators and because evaluators are a non-expert untrained crowd¹⁹. The challenge is then to identify «good annotators» [Tarasov 10]. For this reason raters evaluation has been subject of study: self-confidence scores have been proposed to improve labeling quality [Oyama 13], as well as strategies to

¹⁹We consider here only objective clear labels.

learn labeling accuracy from provided CS data [Donmez 09]. Nevertheless, these strategies are based on the hypothesis that labels are objective and that we can define golden standards to infer reliability. We cannot apply these when adopting labeling to measure a subjective quantity. Other reliability evaluation strategies proposed for CS focus on raters consensus [Brew 10] or adopt repeated labeling [Sheng 08]; the latter is particularly interesting in CS due to the very low price of labels. Especially [Karger 13] addressed the problem of finding a good trade-off between reliability and redundancy adopting inference algorithms. Their mathematical algorithms are completely data-driven and useful to find good quality data. Also proper noise filtering in CS collected data has demonstrated to be effective: in [Nowak 10] authors investigate the reliability of CS image annotations making a comparison with labels provided by experts. Authors have proven that the overall influence of different label source on dataset evaluation is small, while the agreement varies a little depending on metric used. They adopted different statistical tests, such as Kendall Tau and Kolmogorov-Smirnoff correlation, to evaluate the different groups of data and adopted a majority vote to filter noise.

In order to improve collected data quality, research investigated the fact that participants are not really motivated to provide good data. Mainly it's because they have no personal interest in the work, but also because of the low price that is usually given in CS: willingness to provide a good impression is low. Moreover participants feel to be much less controlled than in normal laboratory conditions as we underlined before. This is why another point investigated for improving CS is related to engaging the user to motivate him, as discussed in next section.

3.4 Motivating participants

Motivating the participants in crowdsourcing is fundamental, as they are much less committed than regular dedicated workers. For this purpose, economical rewards are important, however one of the positive points of crowdsourcing is the economical advantage in respect to traditional studies. As an alternative, researchers also investigated other incentives in CS, such as gamification strategies, that have proven to be important incentives in CS²⁰.

3.4.1 Economical incentives

Monetary compensations are the simplest way to reward good workers in CS [Kuikkaniemi 11]. This is done with higher wages but also with bonuses; platforms

²⁰Another possible motivating point is the CS participant «fame» on a platform, like a raking that rewards good workers. However not every platform provides this possibility and it is not as much adopted. For these reasons we do not discuss about it here.

like AMT and Microworkers already give this opportunity after a job review. This strategy can be valuable, as underlined by the experiments of [Grady 10]: «higher-paying HITs or HITs with bonus opportunities may correlate with greater Worker effort». A better paid worker is (usually) a happier worker. However successive studies questioned this point, studying productivity adopting stringent margin of profit for workers: this factor can push workers to lower the quality of their contributions [Kazai 12]. Even if they confirmed that workers were more satisfied with a higher pay, they also underlined that low pay tasks do not attract participants who just seek money from CS, thus being less affected by malicious non-interested workers. Recently the importance of adopting alternatives to economical rewards has been investigated in [Redi 14]; in this work authors investigate if is a paid crowd is better than a volunteer crowd. Within the scope of rating aesthetic appeal of images, they recruit volunteers from their Facebook friends and paid workers on a commercial platform. Results show that monetary reward do not imply at all a more reliable work but only push workers to complete it - to get paid. A better quality is given by volunteers instead, as the less motivated can withdraw more easily. This result is confirmed by the work in [Gardlo 12b], where authors made a comparison between Microworkers and Facebook: non-paid users of Facebook seems to better understand the task. It is not possible to say if this is due to a higher volunteer motivation.

This small review underlines also that there is not a well-defined guideline leading this strategy but more empirical experiences and golden rules. To conclude, economical incentives can play an important role as incentives but we have to keep in mind that CS is interesting also for its economical advantage. If costs are higher it won't be more useful - economically speaking - than a laboratory study.

3.4.2 Gamification

Very early in the development of CS, gamification has been suggested as a powerful incentive: games have always channeled huge amount of energy and efforts of players. To focus these energies into useful activities, the so called Games with a Purpose (GWAPS) have been proposed and realized for scientific researches [von Ahn 06]. The very first effort in this direction was done by Luis Von Ahn with the *ESP Game* [von Ahn 04]. Adopting human computation to label images, the ambitious objective was to «label the majority of images on the World Wide Web». The game was meant to be played between pairs and the objective was to guess what was the label that the partner has given to a common image. The fact of guessing the thought of the opponent, as with an «extrasensory perception», gave the name to the game. Users were provided with a labeling interface online showing the image itself with game elements (i.e. scores and time elapsed) as well as some 'taboo words' that should not be used to label the image. Those were there

to ensure that each image would receive as many tags as possible. Taboo words were taken by previous game sessions automatically, based on previous adopted tags. Successively the ESP game was acquired by Google, with the goal to build better image search engines exploiting metadata. Figure 3.4.2 shows its interface. The game, called Google Image Labeler, ran for five years before being shut down with the whole connected project²¹. While Google labeling project ended, the same concept has been renewed for art (i.e. fig. 3.4.1), in the Artigo Project²². Unfortunately initial project suffered from the presence of outliers; however these



Figure 3.4.1: screenshot of an Artigo session, showing the gamification applied to an art piece labeling. SOURCE: our screenshot of a session from Artigo website.

latter were real malicious workers providing wrong tags with the purpose of – probably – influencing search results. Google noticed this while testing the game worldwide. After few months, Google changed the playing strategy in order to avoid this behavior. Results provided before had to be filtered out and many images targeted from malicious workers had to be removed. However Von Ahn team planned in the original game some outliers detection mechanisms, but those were mainly focused on avoiding cheating in game. This is a very good example of how important it is to forecast all scenarios in crowdsourcing as gamification by itself is not a solution for all problems related to participants. A prevention strategy must be adopted in these systems to avoid outlying behaviors. Later on, the same team developed another game, *Pekaboom*, following the same approach from the ESP game. The aim was to locate objects in images for labelization

²¹Official webpage removed, official information on <http://googleblog.blogspot.fr/2011/09/fall-spring-clean.html>, retrieved Sept. 2015

²²Artigo, <http://www.artigo.org/>, retrieved on June 2, 2015.

purposes; even if the game was not focused on CS in particular, it was meant to be played online, providing top scores and ranking lists as an incentive for participation.



Figure 3.4.2: screenshot of a Google Image Labeler session, showing the gamification principle applied to labeling. SOURCE: Wikipedia, Google Image Labeler, fair use copyright policy.

Another remarkable example of gamification is the work done by Borsboom [Borsboom 12]. In this work the author recreated the *Guess Who* game by Hasbro in order to label emotions on face images. In brief, the original game consists in guessing which person, between a fixed set of people, was selected by the opponent. The first player to guess the other's choice wins. In each round participants ask a question regarding the physical aspect of selected person, like if he has glasses or not. In this case researchers replaced the physical aspect with facial expression, with the aim of labeling pictures. Laboratory experiments showed that the game could provide useful results, as expert and non-expert assessments were comparable. No result is however given for an online CS campaign; we do not know if future trials in CS have been proposed in literature. However, some problems may affect the game, as such a strategy imply the presence of two subjects at the same time or of an automated computer opponent²³. While these alternatives are feasible, they limit the easiness of CS.

Games have been proposed also in order to solve complex problems, addressing these as puzzles. Project Phylo²⁴ has the aim of solving complex DNA alignments adopting a game in which participants align colored blocks following simple rules[Kawrykow 12]. In two years, more than 12000 participants participated in the game, improving the accuracy of alignment by 70%. Figure 3.4.3 shows a part of the interface.

²³Even if VonAhn proposed in previous citation also pre-recorded games to play in solo.

²⁴Project web page, <http://phylo.cs.mcgill.ca/>, retrieved June 2nd, 2015.

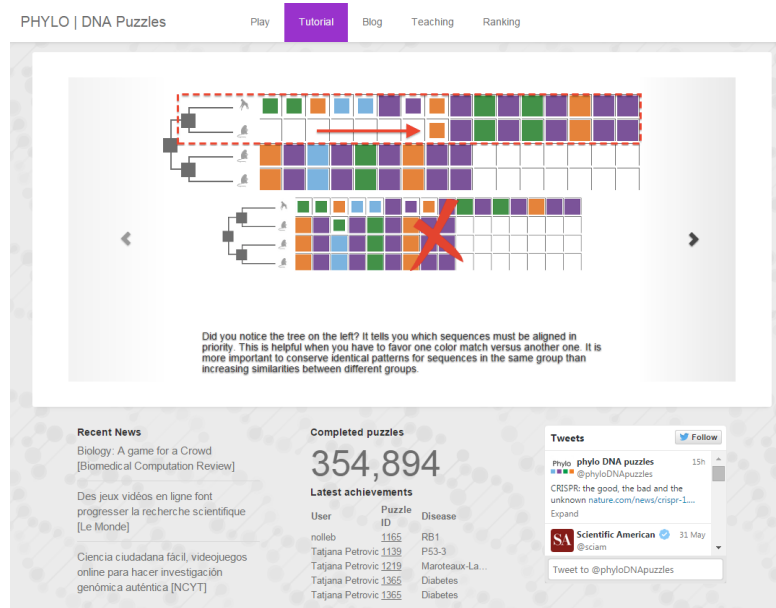


Figure 3.4.3: a screenshot of project Phylo tutorial, showing how to align the blocks - that are DNA sequences in reality. SOURCE: official website project

Games for crowdsourcing enhancement have also been proposed for social networks, to boost participation. In [Riek 11] authors propose a Facebook-based game to label multimodal affective video data. The game, called *Guess What?*, showed videos to participants and then asked a question about it. The main aim was to earn as many points as possible giving correct answers in fixed choice questionnaires. To gain even more points, participants have to provide an answer that they think most people gave. Questions were aimed at labeling video context, so for example possible questions were about the time of day or location of the video. The game was tested with an initial pilot study where only 33 people participated in the labeling, that data was adopted four years later in [O'Connor 15]. At the best of our knowledge the game did not gain popularity and wasn't reused in future. We cannot know if it was the game that did not prove to be so entertaining to boost participation as expected.

This literature underlines that, for making gamification a winning strategy, a lot of effort must be made. Moreover, Social Networks integration may help, but it needs a good advertising strategy. Making a game that really entertains participants is not a simple thing; it may happen that more effort is required by the game design than by the CS experiment itself. Thus, gamification strategies have been left aside at first.

3.5 Conclusion

Reviewed literature underlined that crowdsourcing is effective and efficient also in the image assessment field, notably for Quality of Experience. While many factors must be taken into account, as some negative points are present, some known and useful countermeasures have been proposed. In particular, task (experiment) complexity, required time and participants pay are among the critical factors underlined. Gamification strategies can be planned, but it is not an easy task. Proper reliability checks and data quality evaluation strategies must be adopted to avoid unreliable participants, even if participants are more engaged by a game. In the end crowdsourcing appears much more valid to carry this type of research in respect to electrophysiology measurements.

The review described in these pages allowed us to acquire the methodology in theory, but practical considerations too have to be taken into account in order to adopt crowdsourcing; these are described in the next chapter, together with our crowdsourcing pilot study.

Keypoints

Context

- ❑ Crowdsourcing has been underlined to be a useful tool both for commercial purposes and scientific research.
- ❑ Reliability problems have been however underlined.
- ❑ Countermeasures to problems have been proposed, both in the form of reliability check and incentive strategies.

Contributions

- ❑ State of Art of crowdsourcing for scientific purposes, underlining both positive and negative points.

"Every solution breeds new problems.""

(Murphy's Law)

**APPLYING CROWDSOURCING
FOR COLLECTING PORTRAITS
AND THEIR CONTEXT
PERCEPTION**

Chapter 4

Adapting crowdsourcing for social context evaluation

In this chapter we discuss how to exploit crowdsourcing in practice for our research purposes. After the literature review in previous chapter, we decided to use crowdsourcing methodology: we outline here the practical needs - stimuli, platforms and frameworks - and we explain the design of our dedicated tools for running experiments. This chapter contains also the description of our pilot study with portrait images carried out in crowdsourcing in order to validate the methodology. Work done has outlined many pros and cons that have been partly adopted as input for the Qualinet task force on Crowdsourcing (ref. Annex A).

4.1 Introduction

In previous chapter we investigated crowdsourcing, a relatively new technique already proven to be effective in research. This novel technique offers many positive points at a reasonable price in terms of efforts. As said, in order to run crowdsourcing experiments many practical considerations must be done. This chapter deals with these points, describing three elements that must be examined in order to run crowdsourcing experiments: a platform to gather participants, a software framework to propose the actual experiment online and of course a data set for our purposes. We start reviewing available commercial platforms to collect, manage and pay experiment participants (ref. next section). Successively (ref. 4.2.2.2) we discuss about experiment implementation talking about available software frameworks and our developed solution. Later on we focus on the data set to use considering our research on face portraits. In fact, important scientific considerations (i.e. the kind of stimuli to use) but also legal considerations (i.e. licenses, privacy policies) must be done as the experiment runs online and worldwide (ref. 4.3). In

this chapter we explain also a pilot study we run to test our design and methodology in section 4.4. This experiment confirmed the usefulness of the crowdsourcing strategy. Next chapter deals with the adoption of this methodology to collect social context evaluations.

4.2 Platforms and frameworks: tools to run crowdsourcing tasks in practice

As explained in previous chapter (ref. 3.2.1), participants of a crowdsourcing task are gathered online and then redirected to the task in exchange of a pay. Consequently, to crowdsource a job in practice, two elements must be considered other than the needed data itself: a platform to recruit participants and a software framework to host the task. These two elements are the topic of next two subsections. Of course, physical requirements in terms of equipment (i.e. a server) must be considered too, but are not part of this dissertation.

4.2.1 Online commercial platforms for crowdsourcing

Specific platforms have been developed for online paid crowdsourcing. In this section we will briefly outline the most important; it is not our purpose to describe them in detail but to introduce the CS technique in practice.

Platforms collect small tasks provided by employers and gather huge crowds of people willing to participate. Adopting Amazon Mechanical Turk self-definition, they are «marketplace for work that requires human intelligence»¹. Platforms take care about paying workers after successful completion of tasks: employers pay directly the platform in advance, who takes part of money for itself and redistribute the other part to workers.

In this case tasks are explicit, clearly defined and paid. Moreover they are usually simple and short, but these factors may vary with the associated reward. Typical examples are related to online engagement, as liking a Facebook page, providing positive reviews or generating traffic as clicking on banners. Some of these activities are clearly against Terms of Services of some internet services (i.e. clicking on advertising banners for money) and are usually banned by some platforms. However many other activities are allowed in practice, even if they are malicious practices (i.e. bias product reviews).

Between the most important platforms for crowdsourcing there are Microworkers and Amazon Mechanical Turk. Their commercial model is simple: they charge a

¹Amazon Mechanical Turk general question and answers, <https://www.mturk.com/mturk/help?helpPage=overview>

fee for every job posted by employers on their website and sometimes host also advertisement. For example, AMT collects a 10% commission on top of the amount paid for a worker. The choice of the platform to adopt depends on many factors, as legal constraints (i.e. the residence of the employer for fiscal reasons), the kind of work to carry on as well the audience that should be addressed for it. In fact while the crowd is anonymous, some restrictions are possible to target specific geographical areas.

Considering current scientific literature, AMT and Microworkers are by far the most adopted ones. *Amazon Mechanical Turk* (AMT or MTurk) is the oldest platform, and currently offers more than 300.000 jobs². The job posting procedure is very easy, suggesting employers how to use the service with clear guidelines. The service proposes CS for different purposes, as collection and verification, items classification (products, images ...), surveys completion and content creation and moderation. While this platform is very active and reliable, the main drawback is that some restrictions are present for payments due to the fact that the service is hosted in US. Employers and workers have limitations on payment accounts, as bank accounts must be enabled for particular transactions on US - so most of times a US bank account is required - otherwise the only way to pay is with Amazon gift certificates. This limits somehow the demographic composition of the platform: in an interesting demographic study on AMT platform Ross et Al. found that almost 60% of workers are from US, while the other part was mostly from Eastern Countries of Asia and a small part from UK [Ross 10]. Authors discovered also that almost half of workers are students, and the average worker in AMT is a young woman with a bachelor degree and low income. Mainly a worker there gains around 5 dollars a week working up to 5 hours per week. Ross et Al concluded that however people on AMT are not representative of US population. Still, AMT has been adopted in a very large amount of works; in [Maji 11] for annotations, in [Alonso 09] and in [Grady 10] for relevance assessment as well as in [Nowak 10] for CS data evaluation.

Microworkers is another very powerful platform, especially in Europe³. Different scientific studies helped the platform itself; these also contributed to its development [Hirth 11b, Gardlo 12b, Redi 14]. Mainly participants are from eastern Asia, as confirmed by B. Gardlo and also by our studies. Between its advantages, there is the lack of limitations on money transfer. The platform offers also an improved job posting to aim for a selected group of workers. With this kind of campaign, called 'Hire Group', the employer can select a group of workers to accomplish a task. This group can be built for example with a basic campaign after reviewing results and selecting most efficient/reliable workers. It is not purpose of this doc-

²<https://www.mturk.com/mturk/>

³<http://www.microworkers.com>

ument to review this platform in detail; interested reader can find a good review in [Hirth 11b].

Privacy issues are also to consider as CS platforms may forbid some data collection practices. For example, Microworkers explicit it in the terms of service: it's forbidden to "harvest, collect or use addresses, phone numbers or email addresses or other contact information of users". Explicit authorization request to participants is then required. This point poses problems regarding user localization and the possible cultural studies with this info. Asking directly the user is possible, but it is not possible to be sure of data reliability.

The screenshot shows the Microworkers website interface. At the top, the logo 'microWorkers' is displayed with the tagline 'work & earn or offer a micro job'. Navigation links include 'Blog', 'API', 'Success rate', 'Reputation', and 'Support'. A user profile bar shows 'Filippo Mazza' with a nickname 'Omegafil' and email 'mazza.networking@gmail.com', along with a 'Logout' button. Below this, a section titled 'Available jobs' indicates '195 jobs available to you'. A message states: 'You should only accept jobs you are capable of finishing.' There are filters for 'running & available' and a 'remove from the list' option. Sorting options include 'Most paying', 'Latest', 'Best rating', and 'Time To Rate (TTR)'. A grid of job categories is shown, including 'All jobs', 'Testing', 'Mobile Applications', 'Surveys', 'Sign up', 'Click, Search', 'Bookmark', 'Google', 'Youtube', 'Facebook', 'Twitter', 'Promotion', 'Yahoo Answers', 'Forums', 'Download-Install', 'Comment on blogs', 'Write a review', 'Write an Article', 'Blog/Websites', 'Leads', and 'Other'. A table of available jobs is displayed with columns: Job name, Payment, Success %, TTR, TTF, Done, and Remove. The table lists various tasks such as 'Describe the Highlights in the Soccer Game', 'Youtube: Comment 3x (#PANL)', 'Gmail: Sign up + Bonus', etc., with corresponding payment amounts, success rates, and completion counts.

Job name	Payment	Success %	TTR	TTF	Done	Remove
Describe the Highlights in the Soccer Game	\$0.11	N/A	14	10	0/30	[X]
Youtube: Comment 3x (#PANL) [X]	\$0.12	72	7	3	278/300	[X]
Youtube: Comment 3x (#SKYA) [X]	\$0.12	89	2	5	247/270	[X]
Gmail: Sign up + Bonus	\$0.11	6	1	5	17/60	[X]
Twitter Post: GoWild [X]	\$0.30	0	7	3	129/300	[X]
Youtube: Comment 3x (#RENK) [X]	\$0.12	72	2	4	82/110	[X]
Youtube: Comment 3x (#GRAT) [X]	\$0.12	79	2	4	342/370	[X]
Google+ Post: GoWild [X]	\$0.20	0	7	3	206/300	[X]
... [X]	\$0.07	0	1	4	.../1000	[X]

Figure 4.2.1: main page of Microworkers, showing available jobs.

For the sake of completion, we mention also less adopted platforms, at least in multimedia related researches. ShortTask⁴ is a quite recent CS platform born in 2011. It is until now mainly adopted for exploiting human intelligence to carry online researches and articles writing. Apparently they do not enforce any restriction against malicious jobs (i.e. spam) [Wang 12]. No other remarkable uses have been found in literature. Crowdfunder⁵ is another worldwide known platform, gathering workers from more than 200 Countries. Based in US, Crowdfunder focuses

⁴www.shorttask.com

⁵<http://www.crowdfunder.com/>

mostly on tasks related to data collection, review and labeling, integrating content questions mechanisms: recently it has been positively adopted for Twitter content analysis [André 12].

All these platform allow the employer to review submitted jobs. This is useful of course in order to check results at the end of a campaign to refuse improper submissions and also refuse to pay respective workers. Review allows also to check ongoing campaigns (i.e. detecting errors in job design) and to note good workers to contact again in the future.

Based on the researches reviewed, the possibility to contact Hossfeld team and the lack of limitations on bank accounts with Microworkers platform, we finally adopted this platform for our study. Table 4.1 summarizes main pros and cons of platforms.

PLATFORM	PROS	CONS	NOTES
Amazon MT	widely adopted	problem without US bank account	
Microworkers	widely adopted; tasks are well structured		the most versatile and used in EU
Crowdfower	well known for data collection and labeling	still focused on US	
ShortTask	focused at a wide variety of human intelligence tasks	no malicious jobs restrictions	

Table 4.1: Summary of CS platforms.

4.2.2 Frameworks

Participants from the crowdsourcing commercial platforms will then be redirected to the task. Then, in order to run particular tasks, a dedicated system should be done by the employer. We call this system «framework», even if «CS engine» is also another possible name. However, a clear nomenclature is missing still at the best of our knowledge⁶. As we are interested in running online research experiments in crowdsourcing, we need a framework to host our task. Here we review existing ones before outlining our personal solution.

⁶In some cases also these systems are called with term platform, as in [Figuerola Salas 13].

4.2.2.1 Available frameworks

We underlined before that in online crowdsourcing people participate connecting to the CS platform through their devices. The worker is usually redirected to a web page describing the job and after having accepted the work he must accomplish the task and provide at the end a proof to the employer; this proof is sometimes mandatory as it helps in checking the work accomplishment. Focusing on the job to accomplish, when this is more complex than carrying out some activities on existing websites (i.e. posting a comment on a forum), it is usually done directly on a dedicated web page made by the employer. To continue with our book translation example, the CS participant must be provided with a method to translate one page, like a series of steps to follow in order to read and send the translation or a tool in which directly input the translated text. While many possibilities may exist, the simplest is always the best option, especially in crowdsourcing⁷. We have to consider in fact that an automated system would be preferred, as if we need a lot of participants a manual strategy (i.e. sending via email the pages one by one) would become quickly unfeasible. However, a complex automatic methodology has its disadvantages, especially because can take more time to be ready and it is error prone.

For these reasons the common solution is the adoption of a single online web page to do all the tasks. This strategy is today the most commonly adopted also for research purposes and sometimes required for the employer to propose particular tasks in CS platforms. This web page must be able to manage all the processes related to the job itself, from the acceptance of participations to the results submissions and proofs check as said previously. It then needs to perform a much more complex task than a simple web page. We will then refer to this complex system as framework to indicate that is a software solution in which different elements interact between them to perform the task. However, it has to be noted that an official nomenclature is still missing at the best of our knowledge.

Different crowdsourcing frameworks already exist in research. They are mostly related to QoE research, due to the interest that this branch has additional influential factors (i.e. user context) on multimedia assessments. Many of these frameworks are private and have been adopted only by developers for particular research works. This is the case for example of first CS studies, like with the Java Applet running the ESP Game [von Ahn 04]. Between the oldest publicly proposed framework there is *Quadrant of Euphoria*, realized by Chen et Al [Chen 10] in 2010 but already developing since 2009 [Chen 09]. This is probably the very first public framework for these purposes. It is focused on Quality of Experience

⁷Supposing that it provides effectively the result, of course.



Figure 4.2.2: Quadrant of Euphoria interface, as displayed to users. SOURCE: original work, [Chen 10].

assessments, and it allows to run online experiments through authors' online page, adopting their server and interfaces. However, possibilities are quite limited in terms of methodologies that can be adopted and it is not customizable at all. In particular, it does not adopt honey pots and reliability checks are based only on pair comparison transitive property. Other honeypots features have been instead added in authors' newer work, that proposes a trusted framework focused on cheat detection mechanisms [Wu 13]. However, those are focused on pair comparisons and other testing methodologies are believed not to allow a trustable framework as with this methodology. *QualityCrowd* framework [Horch 11] was instead developed keeping in mind the possibility of technical issues with internet video evaluations and results comparison with laboratory environments. As underlined by Figuerola in [Figuerola Salas 13], this framework offers the advantage of an hybrid approach stimuli rendering based on participant's browser capabilities. Another web framework for subjective tests has been done recently by National Telecommunication and Information Administration (NTIA) in USA. The Web-Enabled Subjective Test software (*WEST*) has been developed last year with the objective to allow researchers to «conduct subjective tests on multiple devices with aggregated data collection and reporting »[Catellier 14]. It focuses on technical aspects, allowing tests on both personal computers and mobile devices, adopting modern web browsers functionalities. This OpenSource framework is currently available online for everyone to install it on a server and run experiments⁸. One interesting point

⁸Available on GitHub, <https://github.com/NTIA/WEST>, retrieved on 02 June 2015.

in which it differs from other frameworks is that instead of following as much as possible protocols imposed by international ITU standards for testing conditions, they prefer to drop some constraints by purpose in favor of offering an experience that is closer to real viewing conditions out of the lab. Only four video testing methodologies are implemented at this time.

Other generic web based data collection tools exist too. Those are not focused specifically on large scale online CS but they can be adopted also for this purpose. This is the example of *Tally*, published in [Jain 13], which source code is available online. This framework is a web based tool to automate subjective video experiments. It focuses on video displaying issues, as synchronizations issues, resolution limitations and interfaces design. However many features that crowdsourcing requires, as previously mentioned completion codes, are missing in these multi purpose generic tools. To complicate the situation, experiment configurations have to be done through configuration files hard to understand for who's not into original development. Only QualityCrowd offers a web interface to interact with the test, but it's limited and stimuli to be displayed must be managed manually per session. Experiment sessions have to be manually prepared and separated before launching the job online. While this may seem a normal execution, in some cases where many intra-observer conditions must be respected (as in our experiment, ref. later) this is practically unfeasible. As all customizations must be coded, sometimes it's faster to rewrite the whole system than add something to it.

4.2.2.2 Design and realization of a dedicated CS framework

We developed a dedicated framework, considering already existing frameworks as well as problems underlined in previous chapter. This choice offers more flexibility, more control over the whole process and freedom to run the particular kind of experiments. This tool has also the advantage of being expandable for successive experiments. However this choice takes more time than adopting an existent tool.

Developed framework allows to manage essential CS features (notably run experiments online through a web browser) as well as to easily expand or add functions. It is composed of different «modules» interacting between them on a server⁹. The aim is that each part of the software can be seen from the other parts as a black box, and adopted knowing just its inputs/outputs. These modules can be enabled or disabled depending on needs.

Main functions are modules too, even if are not meant to be disabled. The first basic functions are the system initialization and welcome screen, that run when

⁹The framework is written in HTML, PHP and JAVASCRIPT. This design allows to let the framework run on a simple web server such as APACHE. All tools used are completely free of charge, allowing a laboratory to run crowdsourcing tests at no cost (except paying participants).

participants arrive from a crowdsourcing platform (i.e. Microworkers). These modules take care of saving basic information about the user arrival and its acceptance to participate the test. They also perform needed functions to retrieve stimuli for next experiment phases. The second part of modular functions is related to propose the user a questionnaire, in order to gather demographic information. At the same time it collects information regarding participants' hardware (notably screen resolution and Internet browser). The third part is instead dependent on the methodology that a particular test requires, as it deals with the experiment to carry. Different modules can collect data with different methodologies (i.e. Absolute Category Rating, Pair Comparisons). A module showing on screen instructions is common to every experiment; the experimenter must prepare a graphic overlay to show on user interface. Data is stored in a database¹⁰, allowing to retrieve data easily through queries and to adapt experiment in real time where needed. For example, this allows to detect sessions that have been ended prematurely (i.e. users withdrawing from the experiment) and react accordingly. Data regarding previous users is stored in the same way, allowing us to track participants between different experiments and easily conduct statistics on participants. Reliability checks can be still on a modular base but in our work it has been easier to integrate them inside the other "methodology" modules, being really dependent on the ongoing research. Another module is meant to distribute proof of completion accordingly, needed to screen results.

Modules required to allow connections with social networks are only drafted as real time connection between our experiments and social networks is not interesting at the moment. Figure 4.2.3 shows the general scheme.

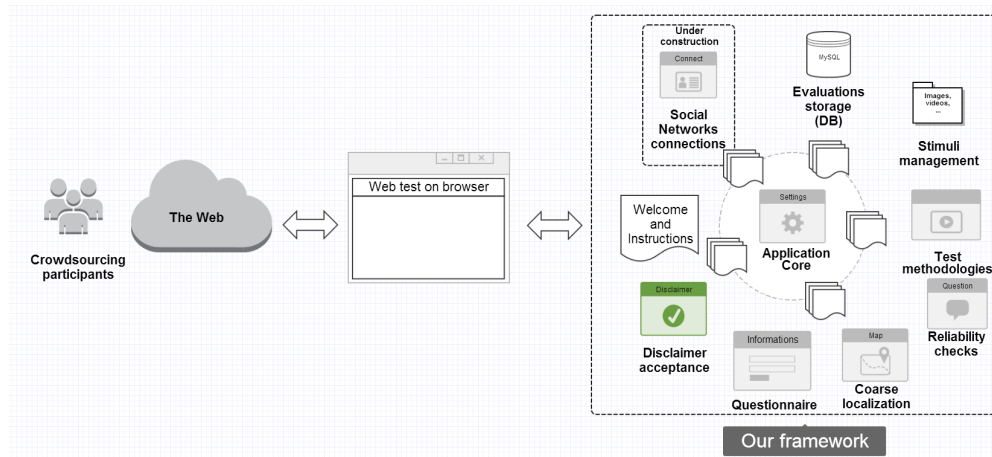


Figure 4.2.3: schema of our crowdsourcing framework.

¹⁰MySQL on the same server

The framework has been tested and positively adopted in our research. The first study adopting it is explained in next section.

4.3 The data set: portrait sources for research purposes

The third element to consider is of course the data set of stimuli to evaluate. We discuss here about this point because adopting crowdsourcing imposes more attention especially in terms of pictures licenses and privacy: there is no control over experiment participants and their behavior (i.e. they can download and use the pictures elsewhere). We summarize here the review of available portrait sources, leaving the extensive discussion in Annex F. choices Second, because we use the two data set we built immediately after for the experiments. The first data set is described in this chapter (ref. next section) while the second is described in the next chapter.

Section 1.3.1 underlined our interest in face pictures including complementary information regarding the context, like a part of a person's torso and partly showing a background. Hence, we looked for amateur or semi-professional pictures, reflecting less formal and "posed" portraits - a characteristic that we suppose influence subjective context perception. We also investigated if different portraits are available depicting the same subject, as this would help in the study of influential factors. Scientific research already addressed the need of image databases for different purposes, as to study image classification, automatic labeling and quality assessment just to name a few. Many image sources containing portraits are available, each with pros and cons. In our review, we outlined four big categories adopted in research: public image data sets, personal collections, online open resources and in-laboratory shootings. The first category features image databases adopted in scientific literature that contain face images. This category has plenty of examples; table F.1 and following in Annex summarize the most important ones. On one hand with available data sets we can have a large number of depicted subjects but with very few different contexts for each one. Some sets offer more pictures of few subjects, but pose some problems (i.e. containing famous people pictures or featuring small portraits). More important, databases are mainly focused on the face only as their purpose is mainly for face detection, recognition and pose estimation. As such, they provide e.g., a neutral facial expression against a white background, or on full body portraits. In order to focus on the social context evaluation, we cannot take these as stimuli source. The second category on other hand, features a small number of personal collections providing a large number of pictures of the same subject in different contexts, as in daily

shot selfies. Private collections offer the advantage of featuring multiple shots of the same subject in different contexts. Moreover, no privacy issues are present, as users who adopt it are authorized by the owner¹¹. Legal issues must instead be considered adopting the third category, online communities. In order to adopt these we must pay attention to licenses and privacy problems, as the simple fact that a picture is online and downloadable does not allow at all to use it for our purposes. This is particularly important when the experiment must run online and there is no control over the participants behavior. However, these sources are rich in terms of faces, contexts and image formats. The last option we underlined, self made portraits, is useful while a small number of specific stimuli are needed, but it becomes unfeasible when a large variety of pictures are needed. Table 4.2 summarizes main pros and cons of our sources.

With these considerations in mind, we then decided to use few portraits made in laboratory for our first pilot study, as explained in next section. For the main analysis of influential factors instead we relied on online communities, as explained in next chapter, for their availability and richness of portraits.

COLLECTION	PROs	CONs
Public image data sets	Economical, tested, large variety	Require a careful selection of useful stimuli between the many present
Personal Collections	Feature a large amount of shots for few subjects, in different contexts	Many collections are needed to have a large number of depicted subjects. Not immediately available.
Online Communities	Free and feature a large variety of both subjects and contexts	Legal issues to consider carefully, technical issues to address for retrieval
Portraits shot in lab	Complete freedom in making stimuli	Time and cost consuming, requiring effort to find contexts/subjects

Table 4.2: Pros and Cons of reviewed portrait sources.

¹¹We do not consider private shots that author do not want to share (i.e. containing children).

4.3.1 Building a data set for our crowdsourcing pilot study : professional shots versus selfies

First chapter underlined that a problem to solve in this research is to find a way to practically address the social context evaluation. So, as a first step we investigated if we can measure social context influence on the perceived “message” conveyed by a portrait. We then run a simple pilot study to test both this hypothesis and our crowdsourcing methodology. For this reason we limited the context to a specific use case, both regarding the context and the typology of the image: we set up a simulated hiring process for an invented company. The message that can be influenced by the picture is the resume of proposed candidates, which has a profile picture attached. Regarding portraits, we decided to focus on two picture versions of the same subject, one being a professional shot and one a selfie. We based this choice also on current phenomenon of selfies, that become recently really popular, especially online. Experiment is discussed in detail in section 4.4; our idea is that social context is an influential factor in candidate choice, as probably the professional portrait will be more suitable for work purposes. As our purpose is very specific and we do not need a large amount of stimuli, we opted for making portraits in laboratory. This allowed both to focus on our methodology and avoid a combinatorial explosion due to multiple factors of the experiment. Moreover, this strategy leaves the complete control over portrait characteristics, and has proven to be relatively fast and problem free.

Twelve portrait images have been adopted as resume pictures: two different portrait versions (a selfie and a professional shot) have been realized for each of our 6 fake candidates. Both kind of pictures has been taken in our laboratory, with subjects from different countries and face traits - aged between 23 and 40 - participated in the photo shooting as models. The first version was a professional portrait, an high quality shot taken in controlled conditions. We took great care of subject position, expression and photo quality, adopting professional equipment.

For the second version instead we asked subjects to take themselves a selfie with a mobile phone. We instructed and guided subjects as little as possible, just in order to have comparable selfies (i.e. having the same phone to face distance), but they have been left free as much as possible to have real selfies. Images’ resolution and size have been taken into account to avoid misbehaving during the visualization, due to participant equipment (i.e. browsers re-sizing images). Pictures have been manually post processed via software to improve luminance, contrast and increase the level of detail ¹². Pictures have also been retouched to reduce imperfections and make these portraits even more professional, leaving however unaltered as

¹²As we are not professional photographers, we followed software guidelines and references.

much as possible the naturalness of the subject¹³. The full data set is provided in mentioned Annex; four samples are given in figure 4.3.1.

4.4 Would you hire me? Selfie portrait images perception in a recruitment context

This section describes our first study run in crowdsourcing. It's aim is double: first, it allowed to evaluate if portrait pictures suggesting different social contexts can bias the perception of depicted subject. We did this adopting different candidate portraits within a resume selection case. Secondly, other than the scientific question addressed, it allowed to evaluate crowdsourcing for our purposes and the developed framework: we adopted this technique to collect subjective assessments and later analyzed data to assess portrait bias on messages given aside. Study has been published in [Mazza 14].

4.4.1 Experiment description

In research path discussed in first chapter one point was to find of a way to practically address the evaluation of the influence of a face picture. As the problem would be too complicated considering all possible cases, we limited the scope to the social context. At first, to develop our methodology as well as to get experience in the field, we addressed a specific case both regarding the message related and the typology of the image. The same methodology can be applied later to broaden the scope. Regarding portrait social context, we set up a simulated hiring process for an informatics company. The message that can be influenced is the resume of proposed candidates, which has a profile picture attached. It is interesting to underline that context's choice comes also from a practical consideration: recruiters may look for candidates profiles also online, having access to their pictures, as in social networks. The importance of this point has already been raised in social networks research [Hum 11]: people should consider carefully the picture based on the social network and the objective.

Our research question is if a professional portraits gives more importance to a message, even unconsciously, respect a selfie one. We expect professional portraits to give more importance to message associated with them, as this kind of portraits are usually conceived as more valuable.

Working positions have been restricted to software developing area and resumes have been provided with minimal information to minimize complexity and influential factor for candidate choice. This last element has been justified to participants

¹³We wanted also to avoid photo retouch to be too visible not to bias evaluations.

saying that hiring company was at the end of selection process, filtering unneeded information. We left on resumes only candidates degree, specialization and age. This last detail was in a small fixed range for all candidates, accordingly with their degree. Degrees were all related to software developing field and taken from real world courses. To consider the influence of degree for each candidate we created two resumes, differing between them for the degree achieved (PhD or Master) and we then changed accordingly the age. Each participant was provided with either one or the other version. Different picture having same subject with few factors varying is instead a bigger issue. We then decided to manually create our pictures to control different aspects. Portraits are the ones detailed in previous chapter section 4.3.1.

Adopted methodology has been pair comparisons, and participants had to choose one of the two proposed candidates each time. Each participant was provided with only one resume version per candidate, in order not to make them understand the real purpose of the experiment.

We designed our experiment taking care to divert participants attention from the real purpose, in order to focus as much as possible to the eventual unconscious bias coming from portrait images. We paid attention not to show two versions of portraits/resumes for the same candidate to the same observer. This fact augmented the number of comparisons needed in the design. Moreover, the different possible versions of a resume (4 as we have portrait type and resume degree as factors, each with two possible values), multiplied by the number of different candidates we proposed (6), increased notably the number of possible combinations to propose in pair comparisons (24 possible cases). The number of needed participants increased substantially too, as for every combination we wanted a sufficient minimum number of participants (we targeted at least 20) and at the same time limit the number of comparisons per participant¹⁴. This fact, joint with the possibility of conducting demographic considerations, justified the adoption of crowdsourcing. To run the experiment in laboratory would have had a much higher cost and time requirements.

Method

Participants

Almost 1000 subjects participated in the experiment through crowdsourcing. We adopted the popular Microworkers platform. Experiment lasted for six days, in which we tuned acceptance rate to avoid people from only some parts of the world. Details about demographics are given in section 4.4.2. Considering current prices

¹⁴As in CS participants are likely to withdraw in longer tests.

4.4 Would you hire me? Selfie portrait images perception in a recruitment context



Figure 4.3.1: Examples from data set created for our pilot study.

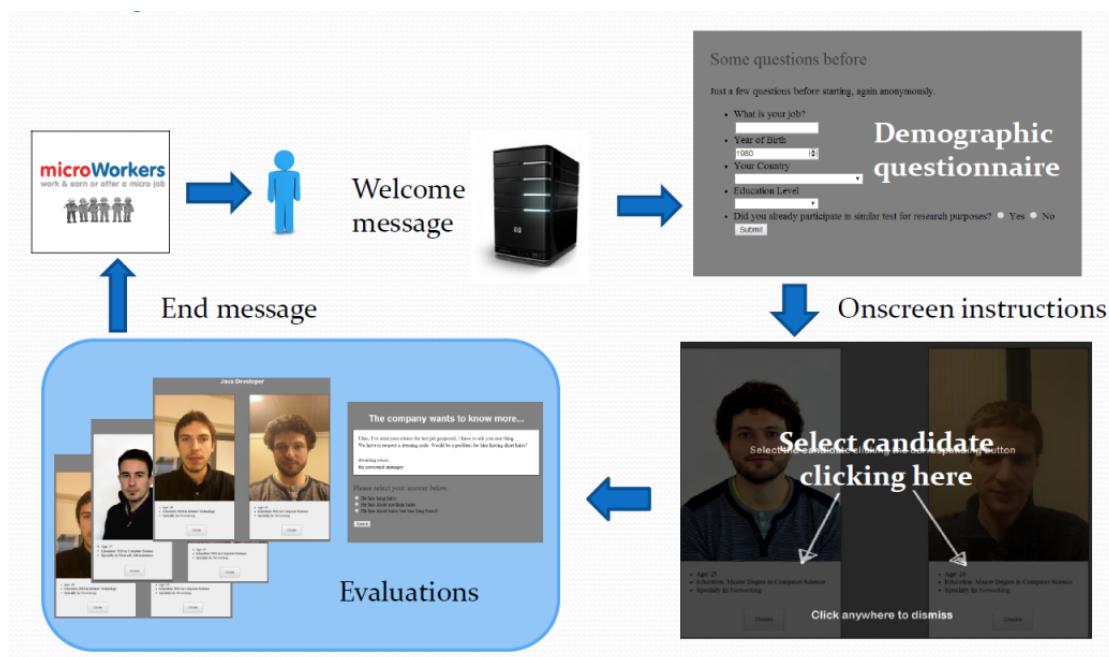


Figure 4.4.1: experiment outline, starting from top left corner

found in Microworkers for similar survey tasks and experiment duration, as well as previous literature review, we decided to pay 0.50€ each participation.

Materials

Twelve portrait images have been adopted as resume pictures. Images come from our first data set described in Annex F. We selected a smaller set in order to have portrait as much as possible homogeneous between them. Figure 4.4.1 shows an example of pictures taken for a subject.

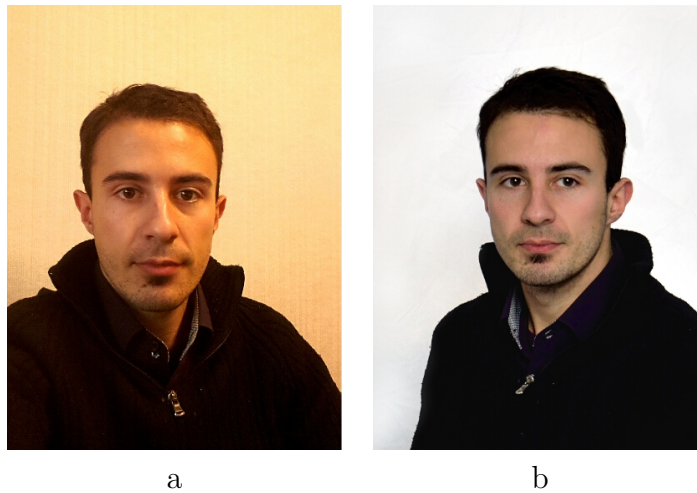


Figure 4.4.2: example of the two different portraits taken for same subject. Unbalanced tones, contrast and light of left (selfie) create a different effect than the right (professional) picture.

Resumes excerpts have been proposed with created portraits. To simplify this part, we told participants that we were at the end of the selection process and information on resumes were summarized. We did this in order to have minimal resumes information, simple enough for us to deal with - create, modify, propose in comparisons - but also for participants to evaluate. Details proposed in the resume are candidate age and instruction level. These are invented on the proposed context of an informatic company seeking employees. Each created portrait has been associated with two versions of essential resume details; both have same specialty but different degrees: one is PhD and one Master Degree.

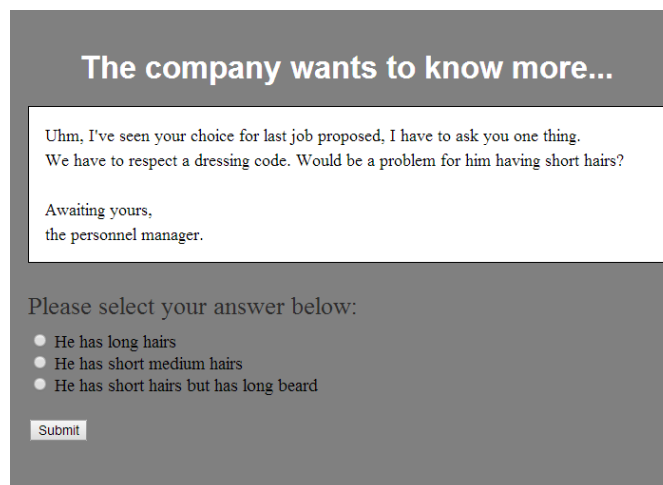
Measures

Adopting a pair comparison methodology, CS participants were faced with a couple of resumes each time, from which they must select either the left or the right one.

We constructed all possible combinations of resume degree and portrait typology for each candidate in our experiment. However, differently from a normal pair comparison configuration, we had more constraints regarding which stimuli to show to a particular observer¹⁵. We adopted a nested design, avoiding to show to the same participants the same portrait subject twice, either showing different pictures or showing different resumes. These constraints inevitably led us to need more participants. Considering general time constraints for a CS experiment, we proposed 15 couples each session to stay under a duration of 5 minutes considering also the honey pots.

A demographic questionnaire has been proposed to participants. We collected the participants' gender, job, nationality, birth year, education level and if they were new to this kind of experiments. Answers were mandatory to continue the test.

We screened participants' reliability with the controls we designed in framework, honeypots in form of content questions and timings analysis. Five content questions were asked to each participant, aimed at asking a characteristic of previous selected candidate, either regarding his picture or his face (i.e. fig. 4.4.1). Questions have been proposed in the same context of candidate selection, explaining that the human resources manager wanted to know more about last candidate chosen. Timings have been recorded on the server side and are related to page loadings and pair comparisons answers.



The company wants to know more...

Uhm, I've seen your choice for last job proposed, I have to ask you one thing.
We have to respect a dressing code. Would be a problem for him having short hairs?

Awaiting yours,
the personnel manager.

Please select your answer below:

- ☐ He has long hairs
- ☐ He has short medium hairs
- ☐ He has short hairs but has long beard

Figure 4.4.3: example of "honey pot" proposed during the experiment. They have been made to follow the same simulated context

¹⁵We underline again that we did not show twice the same subject with different picture/resume versions.

Procedure

After task acceptance on Microworkers platform, participants have been redirected to our CS framework, hosted on a server in our laboratory. First page welcomed them, showing experiment description and disclaimer to accept before starting the actual test. The description was related to the overall experiment scope and the fake candidate selection context provided, without giving details regarding the real objective. Instructions given were also explaining how to adopt the interface to follow our methodology. After accepting our disclaimer (focused on data collection and anonymity), we proposed the demographic questionnaire. After it, we proposed a first pair comparison with on-screen instructions. Dismissing these, participants can start evaluating pairs. Each pair showed both resumes and candidates portrait. Presentation order was randomized, still following previously mentioned constraints. Pair comparison preference was given with a mouse click. Each participant rated three pairs before being proposed with an honeypot. This procedure was repeated 5 times, for a total of 15 pairs evaluated by each participant. A warning message was shown to participants which were providing answers too fast. At the end of the test participants received the confirmation code, to give back to Microworkers in order to be paid, and thanked. Figure 4.4.1 shows the experiment scheme.

Data analysis and results

Participants' reliability: our analysis started with timings inquiry; it revealed that around 12% of provided answers have been given before the web page containing the resumes was fully loaded at user's side. While in some cases this can be due to really slow internet connections and it is possible that participants chose a candidate while resumes were almost fully loaded, we preferred to consider them as outliers. Analysis of content questions' answers revealed that around 50% of participants mistaken more than one out of five asked questions. In our trials in laboratory before the CS experiment, paying attention to the test, we committed at most one error in different trials. We then considered this behavior to be normal and considered making more errors suggesting poor attention. We considered these participants outliers. Figure 4.4.1 shows results of the analysis.

Portrait typology influence on candidate choice: we investigated statistical significance of comparisons where same candidates appeared, with all conditions being equal except portrait typology. To this extent, we adopted Barnard test, proven to be reliable and efficient for contingency tables 2x2 [Barnard 47]. Only 10% of combinations proposed (portrait/resume) underlined portrait typology as statistically significant for the result. However, to check if this outcome is due to fate or it is a factor of influence on the overall result, we checked the likeli-

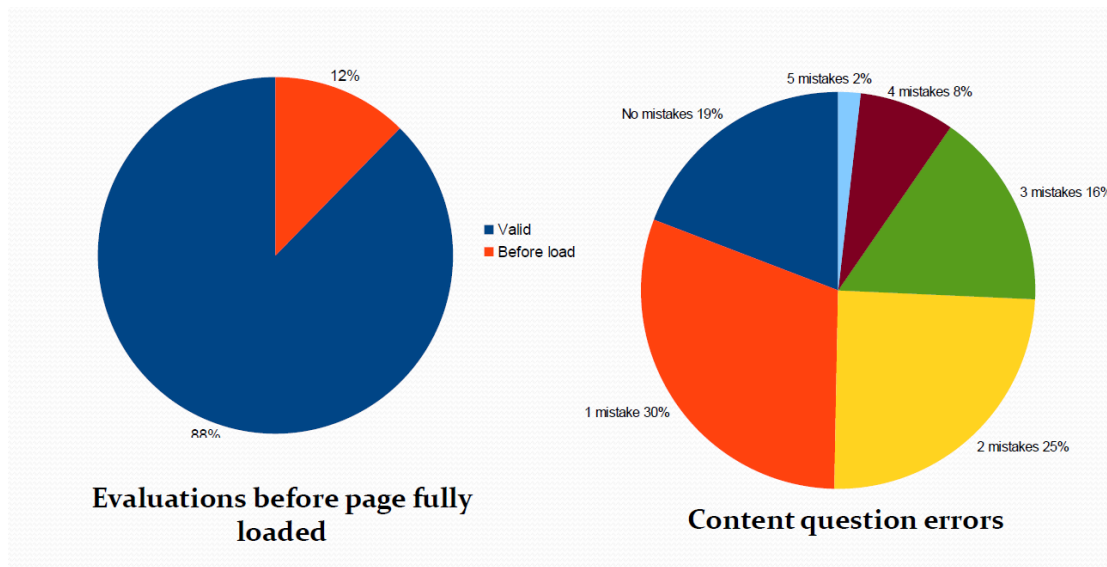


Figure 4.4.4: Data screening results

hood of having the same result by fate with permutation tests. This methodology implies random inversions of some expressed preferences, all conditions being equal but the factor under investigation. We adopted the same method of [Li 13], where the method is described. Algorithm 4.1 details the method while figure 4.4.5 shows a simplified schema of this methodology.

Shortly, the number of significant cases is computed after each permutation; repeating it a large number of times allows to evaluate the distribution of outcomes. If the original result before permutation appears to be an outlier in this distribution, we can then understand that it's not due to pure chance. Our data analysis revealed that influence of portrait typology was not due to pure chance: the result lies on the 95th percentile of distribution (significance level 5%). We reconsidered previously removed participations for abnormal behaviors (timings and content questions) and we run again same methodology. In this case, no evidence of statistical significance can be found, suggesting that outliers indeed did behave differently and maybe did not pay attention enough to the test. Figure 4.4.6 shows mentioned results of permutation tests.

Discussion

This study analyzed image psychology biases given by portrait contents. We analyzed the bias given by the difference of self shots and professional portraits through a simulated resume selection for a work profile. Factors of influence have been outlined and controlled; in this work we focused on portrait typology.

Algorithm 4.1 Permutation test algorithm.

Inputs:

Subjective comparisons;

F ;

▷ Factor to test

$Loop_num$ number of loops;

Output:

Vector Sig_ratio

▷ # of significant pairs every iteration

$Sig_ratio \leftarrow$

▷ Initialization

$Group_1, Group_2 \leftarrow$

▷ Two groups division based on F

for $n \leftarrow Loop_num$ **do**

repeat

if $n \neq 1$ **then** randomly swap m quotes between the two conditions

end if

$p(n) \leftarrow \text{barnard test}()$

▷ Barnard's test p-value

if $p(n) < 0.05$ **then**

$Sig_ratio(n) \leftarrow Sig_ratio(n) + 1$

end if

until every possible pair e has been evaluated

$Sig_ratio(n) \leftarrow Sig_ratio(n)/N$

end for

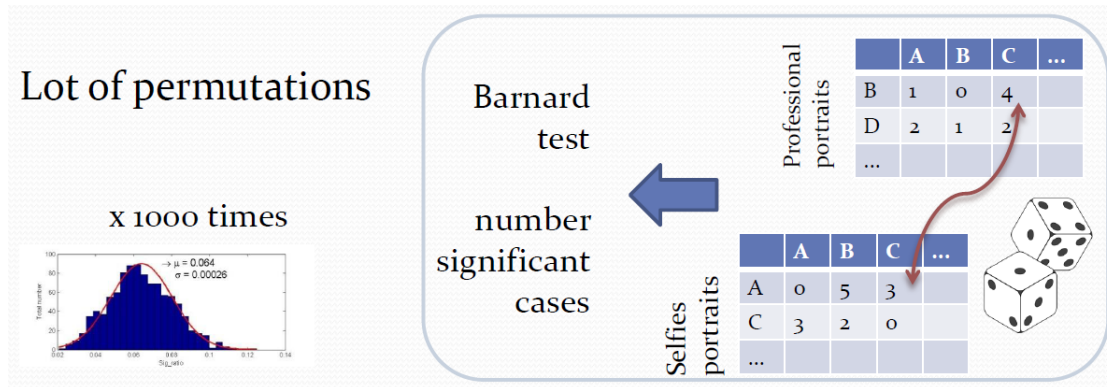


Figure 4.4.5: simplified schema of permutation tests. Procedure in right block is repeated multiple times, getting an histogram of results. Preferences are splitted in two based on factor of influence investigated. Dices represent the random permutations.

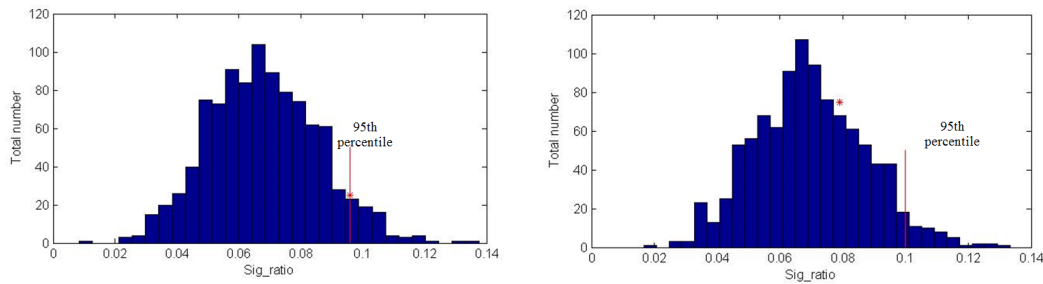


Figure 4.4.6: histograms of significant outcomes after permutations, for both cases with and without data screening (left and right respectively). Asterisks indicates outcomes before permutations.

Barnard exact test has been adopted to check statistical significance of this factor for each pair. Permutation tests have been run to check the overall influence on all the stimuli proposed. Evidence of preference for professional portraits has been shown to be an influent factor for part of proposed stimuli. Many analysis are still possible with the data gathered; these concern demographic data collected with our pre-test questionnaire proposed to participants but also analysis regarding the impact of resume itself. We will not dig further for what concerns resume bias as not really interesting for our research. Instead next section explains work done on demographic factors.

However, this data only is not sufficient to derive a conclusion and deeper investigation is needed. Anyway all these elements underlined that this methodology is suitable to research social biases impact. Mastering those biases is important especially for evaluations done remotely via Internet. Furthermore from a practical point of view it can open the path to further research in product advertising as it can influence social impact.

To conclude, the study has been useful mainly for two different aspects. First to find a methodology to run online experiments regarding portrait images factors of influence, in which data screening seems to be crucial. Crowdsourcing methodology has been adopted, outlining some problems and adopting control strategies; outlier participants have been screened based on answer timings analysis and content questions during the experiment. The main outcome is that CS is a suitable strategy also for this kind of picture evaluation, although a careful planned experiment is needed as well as reliability measures. Secondly it has been useful to demonstrate that an influence is present even on messages that should not be influenced by a portrait image.

4.4.2 Demographic considerations on crowdsourced portrait evaluation

This part is dedicated to the demographic analysis of study data. As said, crowdsourcing allows a much richer variety of participants than in normal laboratory studies. This fact, joint to the possibility of having a large amount of test subject, allows to carry statistical analysis on demographic differences.

Collecting demographic data

Commercial CS platforms usually do not allow to select only a Country for participants' provenance, but mostly a group of Countries. Alternatively they allow to block individual Countries from participating a job. This is the case for example of Microworkers, adopted for our study. In our research we wanted the most possible difference in participations and then we allowed all possible provenances. We controlled the arrival of participants only through the "campaign speed" parameter, in order to allow a fixed rate of people per hour. We allowed the same amount of people during each hour of the day, so that demographics were not biased by time zones.

As said, we also included a small anonymous questionnaire; between provided pieces of information there was also participant Nationality. This is done through web interfaces provided by our framework. The a posteriori analysis outlined that even if participants come from all over the world, the majority of them (around 75%) come from Eastern Asia Countries (fig. 4.4.7). Notably half of them come from Bangladesh (around 48%), a quarter come from Nepal India, Sri Lanka and Pakistan (11%, 8%, 4%, 4%). The other quarter come from western Countries (US and Europe). However none of the latter two has enough participants to have a representative sample for each Country; we will consider this part aggregated in our analysis.

While it would be possible that provided information are false - as they were compulsory participants may have been tempted to put garbage data just to continue - we are confident about the truthfulness of the majority of them for two reasons. First, there is simply no point in lying in this questionnaire. We told in instructions that those were for research purposes, no reward / punishment mechanisms were enforced on data; many other CS jobs are asking similar information. Moreover, if they really wanted to put garbage data, they would have put a Country on top of the alphabetical list provided instead of scrolling down to select i.e. India. Second, provided answers confirm some findings outlined in mentioned references, especially regarding Countries [Gardlo 12a]. This unbalanced behavior has been detected also in other CS researches and some motivations outlined, mainly related to economical differences and the importance of the pay [Hirth 11a].

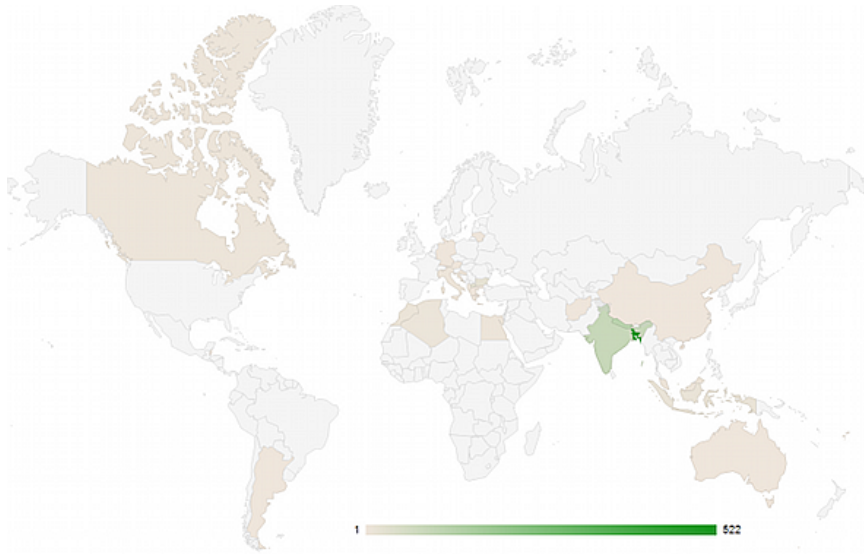


Figure 4.4.7: world wide distribution of pilot study participants (ascending order from brown to green)

This demographic diversity of users, mostly in terms of geographical distributions but also of ages and education, would have been much more difficult to achieve with a normal laboratory setup. This element can be valuable in our research as social biases can be influenced by participants' social background.

We have instead clues that many people reported a fake age. In fact, age distribution shows a big peak around 33 years old, that corresponds to the default value we provided for the field “Year of Birth”. Unfortunately instead of doing something useful for the easiness of answers, we opened the possibility to avoid the answer, and seemingly many participants did that. Users were able to just left the field as it was, as it was an admitted value. We did not conduct successive analysis on age influence on results and modified the interface to have a default value corresponding to an impossible age. This fact also opened the opportunity for another “honeypot” to check participants reliability: in future experiment leaving the field unchanged will point out an unreliable participant. The default value corresponds now to an age of 133 years.

Data Analysis

To investigate if participants' provenance was an influential factor, we divided participations' Countries in three macro regions, R1 R2 and R3. As Bangladesh alone constitutes around half of participations, we considered it by itself, in what we called Region 1. Then we grouped others Asian Countries between them in Re-

gion 2 and EU&US joint in Region 3. Figure 4.4.8 shows the groups we described.

Our research questions can now be summarised as follows. Is there a global statistical significant difference between evaluations of:

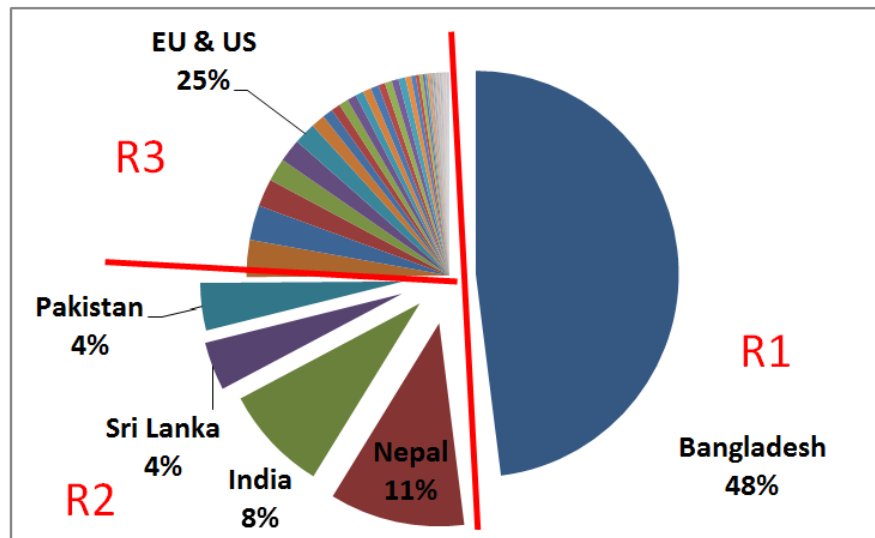


Figure 4.4.8: participants' demographic composition and groups done

- Bangladesh and the other Countries (R1 vs R2 & R3)?
- Bangladesh and other Eastern Countries (R1 vs R2)?
- Eastern and Western Countries (R1 & R2 vs R3)?

We repeated the same methodology applied for evaluating previously described portrait typology influence, starting with discarding outliers' evaluations. This time evaluations have been differentiated on the base of Regions while running permutation tests. We run tests considering the different possible combinations with the outlined regions; tests have been repeated for each combination. So to answer first question our two groups on which run permutations are evaluations from R1 and evaluations from R2 joint with R3. For the second answer instead, groups are evaluations from R1 one and evaluations from R2. Consequently, for the third question, R1's evaluations were part of first group while R2 and R3 of the second.

All three permutation test rejected the hypothesis of significant difference under selected significance threshold. Figure 4.4.9 shows outcomes. We can conclude then that for all the three questions, for these evaluations different demographics are not a factor of influence. As done in the original work for studying portrait

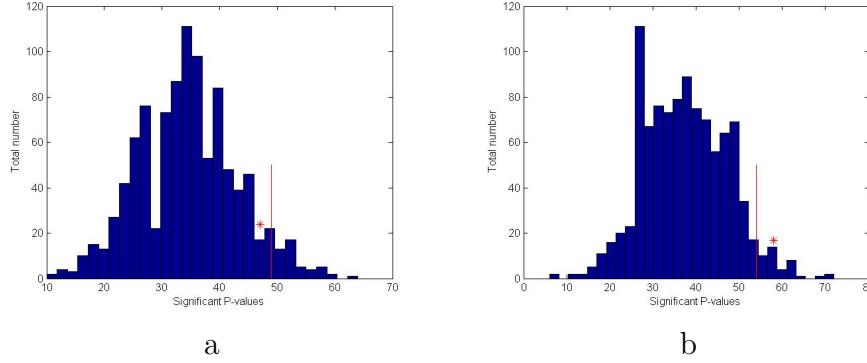


Figure 4.4.9: permutation tests outcomes for first research question - R1 VS R2 & R3 - removing (a) and considering (b) outliers. Histogram of significant p-values while permutating is plotted. In (b) outcome before iterations - marked with * - falls over significance threshold, underlining factor of influence statistical relevance.

typology influence, we run again the tests reconsidering outliers. Interestingly, only for the first research question - R1 vs R2 & R3 - permutation test underlined a significant difference between the two groups.

It is interesting to investigate further this point. While we cannot consider in this analysis uncontrollable technical factors (i.e. participants' screen parameters), we analyzed evaluations' timings measured during evaluations by adopted crowd-sourcing framework. For all the three regions we found a non-normal unimodal distribution: figure 4.4.10 shows evaluations' timings histogram for R1. Similar distribution have been found R2 and R3.

Normality has been tested with Jarque-Bera test [Jarque 87] null hypothesis of normality has been rejected for all three distributions at 5% significance level. To check if these distributions are significantly different we preferred adopting non-parametric tests instead of adopting a log transformations on data to obtain a log-normal distribution on which run different tests. As non-parametric test we adopted a two-sample Kolmogorov-Smirnov test [Massey 51], to check if the data comes from two different unknown distributions. We run the test for all the three combinations (R1-R2, R2-R3, R1-R3); tests do not reject the hypothesis of equal distribution for R1 against R2 ($p=0.0643$, $\alpha = 0.05$) while this hypothesis is rejected for R1 against R3 and for R2 against R3 (p-values of $5.33e-013$ and $2.59e-008$ respectively). It is then interesting to check statistics of these distributions. Due to distributions shape, mode has been calculated, representative of most frequent value in datasets. In table 4.4.2 mode and standard deviations are indicated for the three regions.

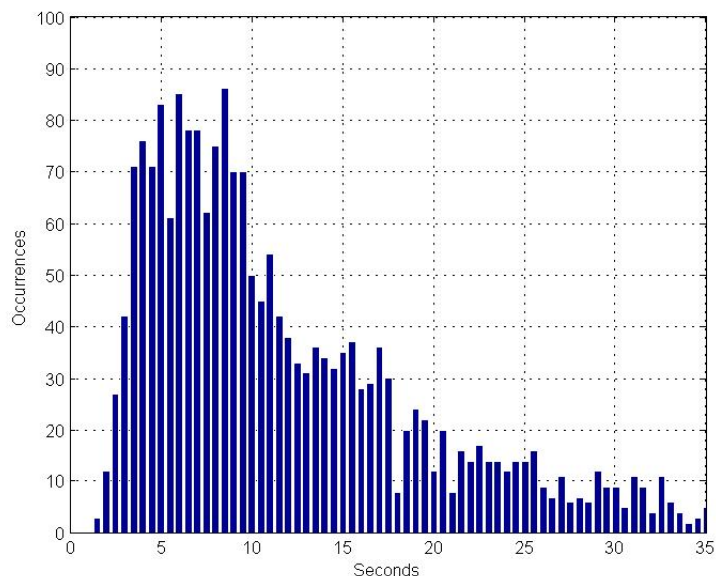


Figure 4.4.10: histogram of evaluations' timings for R1. Values under 90th percentile are plotted for clarity. Similar distributions hold for R2 and R3.

Data underlines that evaluations from R1 require on average more time but have also a higher variability respect R2 and R3. The cause cannot be underlined with this data only; this difference can be due to different behaviors (perhaps from outlier participants) but also to other factors, notably technical factors at participants end side.

No evidence of statistical effect has been underlined. To dig further, we reconsidered evaluations removed after data screening as described for portrait typology analysis. Interestingly, tests underlined statistical difference between the evaluations of R1 against R2 and R3. This may point out that in fact – at least for our experiment – the difference in evaluations between regions lies more on a different behavior of outlier participants in some particular regions instead of cultural factors. Other analyses are possible with information we asked to participants. In particular we asked also their work, through an open field. Also in this case, the uselessness of lying about this answer and provided answers' diversity makes us confident about reliability of gathered data. This can be used to investigate if for example people having an Internet related work have a different behavior during tests than other people. I decided to leave an open field instead of leaving a pre-defined set to choose from as I had no information at all regarding possible options. A possibility that has been identified after the test is to create dynamic

	R1	R2	R3
Mode	8.5	6	4.5
Standard Deviation	39	20	17

Table 4.3: Mode and Variance of the three regions evaluations' timings. Values expressed in seconds, half a second precision.

set of possibilities, built on real time data, where people can either select a work type that someone has already provided before or provide a new one. Interestingly leaving the open field also allowed us to spot some participants misbehavior due to lack of attention or poor question understanding. We found in fact many people that inputted their age inside the field instead of their work. Mainly participants are either students or having a work in informatics field. In figure 4.4.11 is shown the word cloud showing participants works. We decided not to take into account this data at first, while a posteriori data clustering is still possible.

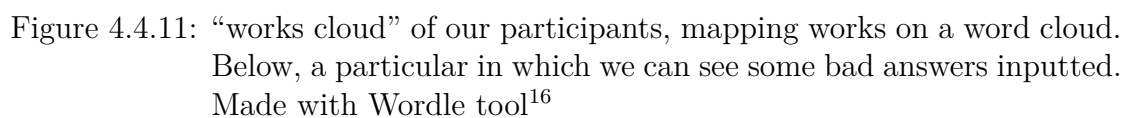
4.5 Conclusion

Practical considerations too have to be done in order to adopt CS. The review described in these pages allowed us to acquire a methodology and skills to exploit CS to conduct research. For our particular purposes a personalized framework has been developed.

The study carried out underlined that CS can be adopted also for portrait image assessment, once proper considerations on outlying participants and cultural differences have been addressed. Two research questions have been addressed; first, if it is possible to detect portrait images biasing effect on a message given aside the picture itself. Second, if this effect is demographic dependent, as people online come from all over the world and on our research this can have an influence. The experiment underlined both that portrait images aside a message have an influence even if the message should be uncorrelated with the picture and that demographic factors seem to play a significant role. Experiment underlined also that a lot of unreliable participants are present and that a careful design and data screening are needed to get reliable results. The experiment allowed also to acquire new skills to exploit CS.

Pilot study validated this methodology also for our type of research: we then adopted it to investigate which social context is perceived for portrait pictures as explained in next chapter.

¹⁶<http://www.wordle.net/>



Keypoints

Context

- ❑ Practical considerations must be done in order to run crowdsourcing experiments in practice.
- ❑ Different platforms and frameworks exist, each with positive and negative points.

Contributions

- ❑ State of Art of crowdsourcing platforms, frameworks and related scientific research works.
- ❑ Design and development of a dedicated expandable framework, including reliability checks.
- ❑ Realization of a small controlled data set to run face portrait pilot study
- ❑ Study to evaluate both developed framework and feasibility of our research.

“Don’t find fault, find a
remedy.”

(Henry Ford)

Chapter 5

Collecting perceived social context of portrait images

In this chapter we explain how we collected subjective evaluations of portrait social context through crowdsourcing. Adopted portrait images have been taken from online resources, conveying more complementary information compared to already existing databases. We exploited crowdsourcing to gather a large number of evaluations adopting our framework, described in previous chapter. Gathered evaluations are needed as ground truth for successive analysis of influential image features, subject of next chapter.

5.1 Introduction

We said that social networks allow us to find good examples of portraits belonging to different social contexts, as it may be i.e. with Facebook or LinkedIn. However we underlined why those profile images cannot be used as ground truth: we know only that the profile owner - that most likely is also the person in the portrait - thinks that chosen portrait is appropriate for that social network. His opinion may not be shared by others, that may perceive differently its portrait. In practice, these portraits have been evaluated regarding their social context only by one subject - the profile owner. In order to construct a more solid ground truth and run further analysis we need more subjective assessments of social context perception. In this chapter we explain how we collect subjective assessments of portraits. Portraits have been retrieved online as described in section 5.3.

In order to run statistical analysis considering many factors at the same time - as we suppose that many elements influence context perception - we need a lot of evaluations. After the discussion in last two chapters, we discarded the use of alternative methodologies to gather subjective assessments but crowdsourcing. In

order to gather social context evaluations we then designed, implemented and run a crowdsourcing campaign. Crowdsourcing seems a perfect candidate methodology, considering both the fact that it is relatively fast, inexpensive and provides many participants. Moreover in crowdsourcing it is easy to repeat the experiment once a campaign has been prepared, to tune the subjective test based on real time results or just to recruit more participants if needed. Repeatability is a valuable feature because there are many aspects that we still not master: quick and cost effective preliminary tests have been very useful for our research. As we show later in this chapter, some outlying behaviors were expected but we understood how to deal properly with them only after the first unexpected behaviors arose.

While the concept of social context is quite clear for us¹, we have to define and limit the number of possible contexts to deal with and find a way to practically allow participant to express their choices. As it is the first time that this problem has been addressed to the best of our knowledge, we limited the possible social contexts to very few options; we've chosen those that are most clear and well known nowadays, based on current trends on social networks online. The main purposes of today's social networks are three: friend interactions, work-related relationships and dating purposes. For example, the most known and adopted today are in fact Facebook, Linkedin and Meetic². While this choice is definitely limiting the possibilities, we prefer to start with these three options in order to master proposed methodology and further analysis. With the crowdsourcing campaign we implemented, it will be easy to repeat tests with other social contexts in the future, if needed.

In this chapter we then review the research literature adopting crowdsourcing for labeling purposes in order to cover the State of the Art on the topic; this part is the subject of next section. After this part, we explain how we applied crowdsourcing to collect subjective social context labeling for our portraits. The analysis of obtained assessments is following that part. Next chapter describes instead the analysis that links collected context assessments to image features.

5.2 Crowdsourcing for labeling purposes

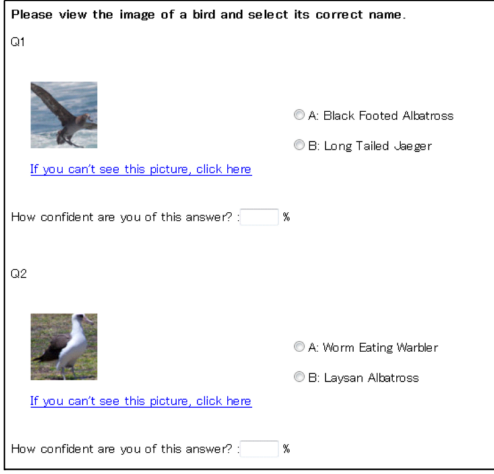
In this section we will briefly discuss the adoption in literature of crowdsourcing for classification purposes, reviewed in order to best design our social context labeling experiment. In particular, we analyzed interfaces that have been designed for the task, as we adopt a labeling-like method to evaluate portraits social context. We

¹we defined it in first chapter as *the perceived overall feeling of the situation in a scene, otherwise the use for a portrait picture that would best fit its purpose*

²It can be argued that this is true only for certain world regions. However, alternative social networks with same purposes are known worldwide.


focus on image stimuli, due to our research topic; however different stimuli typologies have been labeled, as i.e. textual entities [Bragg 13], or sounds [Shamir 14]. It has to be underlined that crowdsourcing labeling is usually adopted for objective assessments, i.e. image databases labeling for object detection. While different evaluation methodologies would have been possible for our purpose³, we preferred a labeling-like method to make the task as easy and clear as possible. The idea is that our participants must be able to clearly state the best social context for a portrait and a second one that would fit only as alternative, mutually exclusive.

Many research works positively exploited crowdsourcing to apply labels; we review here those that mostly inspired us. Adopted structure divide the literature based on the kind of interface adopted.



Please view the image of a bird and select its correct name.

Q1




☐ A: Black Footed Albatross

☐ B: Long Tailed Jaeger

[If you can't see this picture, click here](#)

How confident are you of this answer? : %

Q2



☐ A: Worm Eating Warbler

☐ B: Laysan Albatross

[If you can't see this picture, click here](#)

How confident are you of this answer? : %

Figure 5.2.1: Labeling interface as done by [Oyama 13] [SOURCE: original article].

5.2.1 Classic interfaces

A first and easy technique to associate images with labels in crowdsourcing is questionnaire-like interfaces. A good example given by [Welinder 10]. In this work authors simply show the whole set of images that a rater has to evaluate, asking to click the images that contain a certain object. This process is repeated for every label in which researchers are interested in. For example, if there is the need to apply labels as 'apple' and 'pear' to images, first all images must be shown and only apples must be clicked, then again we have to repeat the procedure for pears. While this technique is really easy for a researcher to implement and for a participant to comply with it, the main inconvenient is that the whole set of

³i.e. absolute category scales to express portrait suitability for a particular context.

stimuli must be evaluated for every single label. This procedure become then less convenient for a big number of labels to apply; this is particularly true in crowdsourcing, where short tests are largely preferred.

Another approach that has been adopted in literature is to show each image with possible labels for that image aside [Oyama 13]. An example is given in figure 5.2.1. As their interest was focused on integrating self-confidence scores for improving labeling, authors also asked to participants their level of confidence for each labeled image, in the same interface. The interface is very clean, however such approach can be very repetitive and boring for the user as duplicates the same task for every image, while instead a more interactive approach (i.e. asking to click on images for which the level of confidence is high or low) would probably have been more interesting.

Multi-class labeling strategies have been proposed in crowdsourcing too. In particular, authors of [Bragg 13] exploited crowdsourcing to create classification taxonomies of generic entities. Entities may be text, visual stimuli or other. In their research each crowdworker is presented with an interface in which he's able to select one or more labels for each proposed entity, before confirming and evaluating the next one. Their work however focuses on the optimization of proposed labels for each stimuli and taxonomy creation, more than optimizing the labeling process from the participant point of view.



The screenshot shows a web-based interface for labeling a stimulus. At the top, the text 'The Boston Globe' is displayed in a large, bold font. To its right, a small red link says '(click here if you don't know what this is)'. Below this, a line of text reads: 'is an example of which of the following (check **all** that apply, or select "none"):'.

Below the text, there is a list of labels, each preceded by a checkbox:

- ☐ football player
- ☐ boat
- ☐ city
- ☐ creative work
- ☐ person
- ☐ military person
- ☐ country

To the right of the 'none' label, there is a checkbox labeled 'none'.

At the bottom of the form, there is a red button labeled 'Submit'.

Figure 5.2.2: Example of textual multi-label approach from [Bragg 13]. More than one label can be applied. [SOURCE: original paper]

5.2.2 Graphic appealing interfaces

Another kind of interfaces is the one adopting much more appealing graphics, usually providing an entertaining experience to the user. Often, these try to make the labeling process as much interesting and easy as possible. This is the case for example of the interface adopted for image aesthetics evaluation on Facebook described in [Povoa 14]. This study adopts a simple absolute category scale for

image rating, but requires the user to drag and drop the images on the scale points to both avoid repeating the scale and at the same time make the rating more appealing for the user (fig. 5.2.3). An interesting approach is obtained

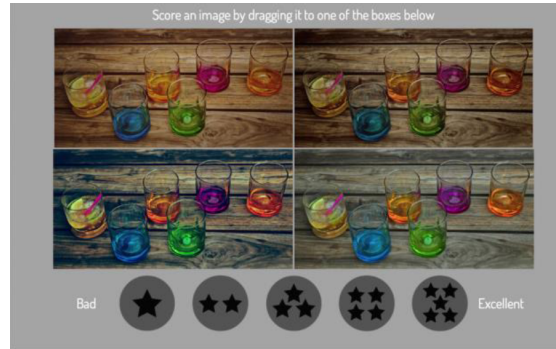


Figure 5.2.3: The interface adopted in [Povoa 14] for aesthetic labeling. Courtesy of author.

when such interfaces are applied to less «traditional» strategies. A very good example is given by [Lintott 08] for the project Galaxy ZOO, aimed at taxonomies generation. Researchers aim to classify galaxies showing their telescope images and asking questions that are successively refined (i.e. more precise) based on previous answers. Not only their methodology provided useful results, but labeling made by general public has been found to be consistent with the one made by professional astronomers. This result may point out that even complicated task can be crowdsourced once they can be decomposed in simpler smaller tasks well designed for online participation: their interface is very neat and contains many graphics element to guide the participant. An example is provided in figure 5.2.4. The project is online and is freely accessible without providing any information⁴.

A crowdsourcing gamification approach has been proposed by [Borsboom 12] for labeling. In this research crowdsourcing has been adopted to label facial expressions constructing a game similar to 'Guess Who'. As in the original game, two players must guess the portrait that the opponent has chosen, asking questions regarding it. In this research, questions are only related to the facial expression of chosen portrait. This shrewdness allows to gather portrait labels. While this strategy can be powerful under many aspects - mainly the increased interest in participants and economical savings if participants are volunteers - it requires objective labels for the game to work.

⁴<http://www.galaxyzoo.org/#/classify>, retrieved on 11 May 2015

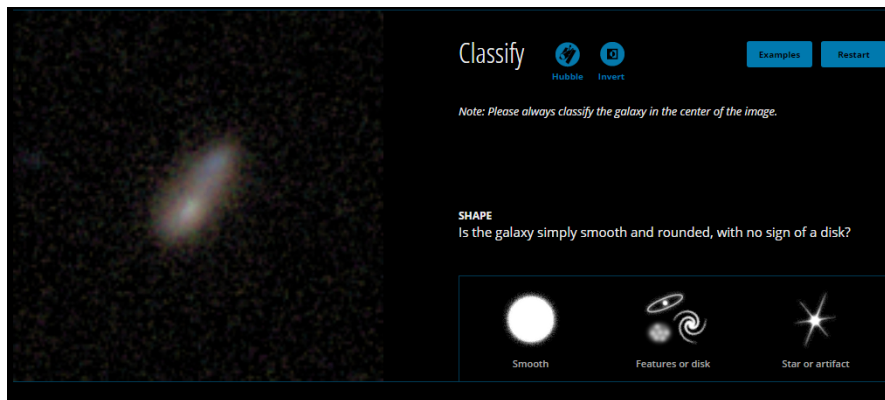


Figure 5.2.4: Interface of Galaxy ZOO crowdsourcing project. Graphical elements greatly help answering the questions. SOURCE: Galaxy ZOO platform website

5.2.3 Free-form interfaces

A more complex kind of interface is the one leaving freedom regarding labels to the user. This can be done leaving the possibility to propose new labels - that were not proposed by the researchers - or to leave freedom to propose more complex label hierarchies or different levels of detail (i.e. pixel based labels in an image). In this category, a well known labeling tool in the computer vision community is LabelMe, proposed by Russel and Torralba[Russell 08]. However in this case the labeling is done on areas inside an image more than labeling the entire image. For this purpose the authors developed an online tool, providing a graphical interface in Javascript to draw polygons around parts inside images and to provide the correspondent label for that area. Figure shows the interface during the labeling process. There is no limit on the number of objects that can be labeled in an image; this leaves the freedom to the user to put as much effort as he wants. Such freedom raises also some problems, as without proper motivation crowdworkers tend to provide work with as less effort as possible. At the moment of their publication, authors remarked that a vast majority of proposed stimuli have been labeled with very few labels (1 to 5 objects labeled per image). With the approach provided by LabelMe users collaborate refining the labeling, as previous applied labels on an image can be seen while providing an annotation: if there are mistakes users can correct them and redraw polygons borders. However quality control is still a problem as there is no other check than collaborative work. The problem is partly solved by Google, that adopts a similar approach for its MapMaker⁵. This partly free-form tool allows to add elements (i.e. places and paths) into Google Maps through a

⁵Google Map Maker, <http://www.google.com/mapmaker>, retrieved August 2015.

simple graphic interface. Participants can add edits similarly LabelMe. However, due to the collaborative approach, people can review and modify other users' edits.

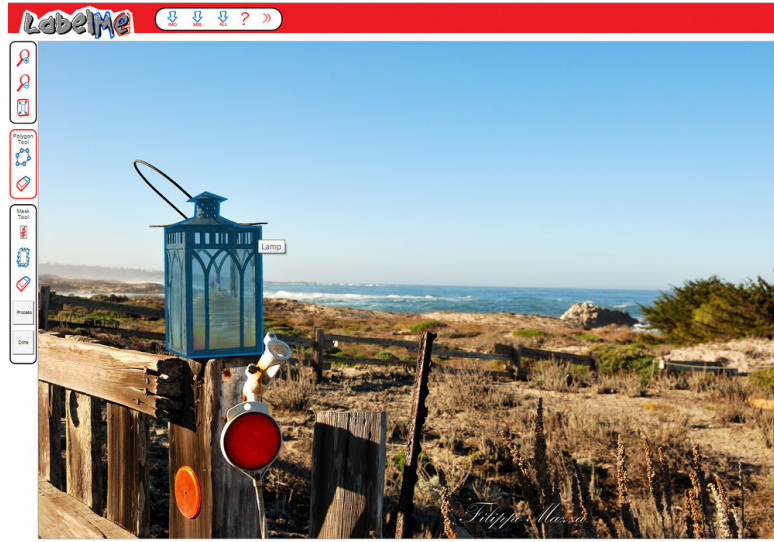


Figure 5.2.5: Labeling through LabelMe online tool. Only one simple object has been labeled as example. [SOURCE: personal picture]

5.2.4 Conclusion

These examples helped us to have an idea about pros and cons of different strategies and interfaces. The most important points underlined are task simplicity and clarity. As said, one of our objectives is to make the labeling task as easy as possible for the user. For this reason, we did not consider the possibility of leaving participants the freedom to propose labels, i.e. free-text. At the same time, we need to minimize required time and costs for the experiment. With these premises in mind we developed a simple interface showing all the labels and images at once in a single screen; however we wanted to avoid to bore participants, so we tried as possible to make an appealing interface requiring drag and drop interactions as described in next paragraph.

5.3 Building a data set to investigate features influence on social context perception

Here we briefly describe the data set created for the main purpose of our research; the details are provided in annex F. In previous chapter we briefly discussed about

portrait sources used in research (see section 4.3). For this second data set, as we aim at having a large number of different subjects and contexts, we decided to look for portraits on online resources, as it was the best solution in terms of speed and costs. Adopting personal collections was not available as we lack of any sufficiently big and varied personal collection to be used and asking permission to each author of previously cited ones would have probably required too much time. We avoided also existing portrait databases for research purposes, as they provide posed images, mostly lacking complementary information on social context.

Regarding the possible social contexts, we decided to focus on three categories, notably for friendship, for working and dating purposes (as said in 1.1.5). We then looked for pictures that are related to these purposes in our opinion. Being this only our opinion, social context perception has been assessed with subjective tests as explained in the next section. We mainly retrieved real online portraits in online networks and image sharing sites, taking care about licenses' restrictions. Based on our review of portrait sources, we opted for the online community for photo sharing «Flickr». Here we found many portraits that we believe fit the categories friends and dating purposes. To retrieve images, we used the public software library that Flickr provides, considering tags and keywords on images, given by uploaders. However, a manual selection on obtained images was required, as user provided keywords are not fully reliable. Many images suitable to «friends» context were found. To have more pertinent work and dating context images, we added some images made for our previous data set (ref. 4.3.1) and the Labeled Faces Wild database [Huang 07]. Details about photo gathering are given in Annex F. Figure 5.3.1 shows some examples of our data set. In the end, a total of 216 portraits have been collected. This number has been found to be a good trade off between a sufficient number for the analysis and feasibility.

5.4 Applying crowdsourcing portrait social context labeling

In this section we will describe the experiment we designed and carried out in crowdsourcing to gather portraits social context labeling. For the sake of clarity, we will adopt here the same structure of previously described studies, and later discuss the results.

In order to gather portraits social context subjective evaluations we run a sub-

⁹<https://www.flickr.com/photos/125303894@N06/14202199100>

⁹<https://www.flickr.com/photos/125303894@N06/14408940363>

⁹<https://www.flickr.com/photos/125303894@N06/14387365942>

⁹<https://www.flickr.com/photos/roland/14038308487/in/photostream/>

LFW database



Flickr



S. Wilson, in “Smiling businessman”⁶



S. Wilson, in “Woman at work”⁷



S. Wilson, in “Business”⁸



R. Tanglao, in “Dana.io portraits”⁹

Figure 5.3.1: Some portraits of our second data set, created for social context study, showing examples of what we believe to be good representatives of work and dating categories.

jective test. We adopted as stimuli a subset of previously collected real online portraits. Chosen subset is related to three different contexts, namely to friends, work or dating purposes. We based this choice on current trends in social networks, nowadays focused on these three kinds of interactions (e.g., Facebook, LinkedIn and Meetic). We then collected subjective assessments of the perceived context of each portrait. We asked for the contexts that suits the most, as well as a second choice. To have enough subjective evaluations of content category, we opted for a large scale subjective campaign via crowdsourcing.

As underlined in chapter 3, it is well known that crowdsourcing participants are much less committed than participants in laboratory environments. Thus, particular attention on experiment duration and price paid is required, otherwise people may withdraw prematurely the test. Still participants can provide unreliable evaluations, both in good or bad faith (in the latter case, for example, to collect easy money). Reliability strategies or gold standards to improve labeling quality (i.e. rejecting 'spammers') have been proposed [Raykar 11, Kazai 12]. However we can't adopt reliability measures based on participants' behaviors with training stimuli, as there is no ground truth for the social context being it a subjective opinion. To have some control on that, we included three hidden reliability checks, partially under the form of honeypots¹⁰. A general scheme of the experiment is given in figure 5.4.1. In case of failure in any of them participants were excluded from analysis and not paid. Crowdsourcing demonstrated to be an effective and efficient methodology. While almost 40% of participants has been excluded for honeypots failure, we gathered more than 8000 valid context evaluations from 216 portraits. Study, together with next chapter data analysis, has been published in [Mazza 15].

Method

Participants

Participants have been recruited through the popular crowdsourcing platform Microworkers. With crowdsourcing, almost 500 subjects spread worldwide have been recruited. Participants were mostly aged between 20 and 45 years old and men (80%). They are mostly students. Participants belong to over 41 Countries; around 33% of participants came from Bangladesh, around 20% from Nepal, a little less than 16% from India, Sri Lanka and Pakistan, the rest (30%) from European Countries and US. These demographic information come from the initial questionnaire we proposed, as explained below. A posteriori internet addresses analysis reveals that only in few cases users participated from the same network, up to

¹⁰As said in previous chapter, while it may be argued about the adoption of this term for all our checks, we will use it for the sake of simplicity and explain in detail which checks we adopted.

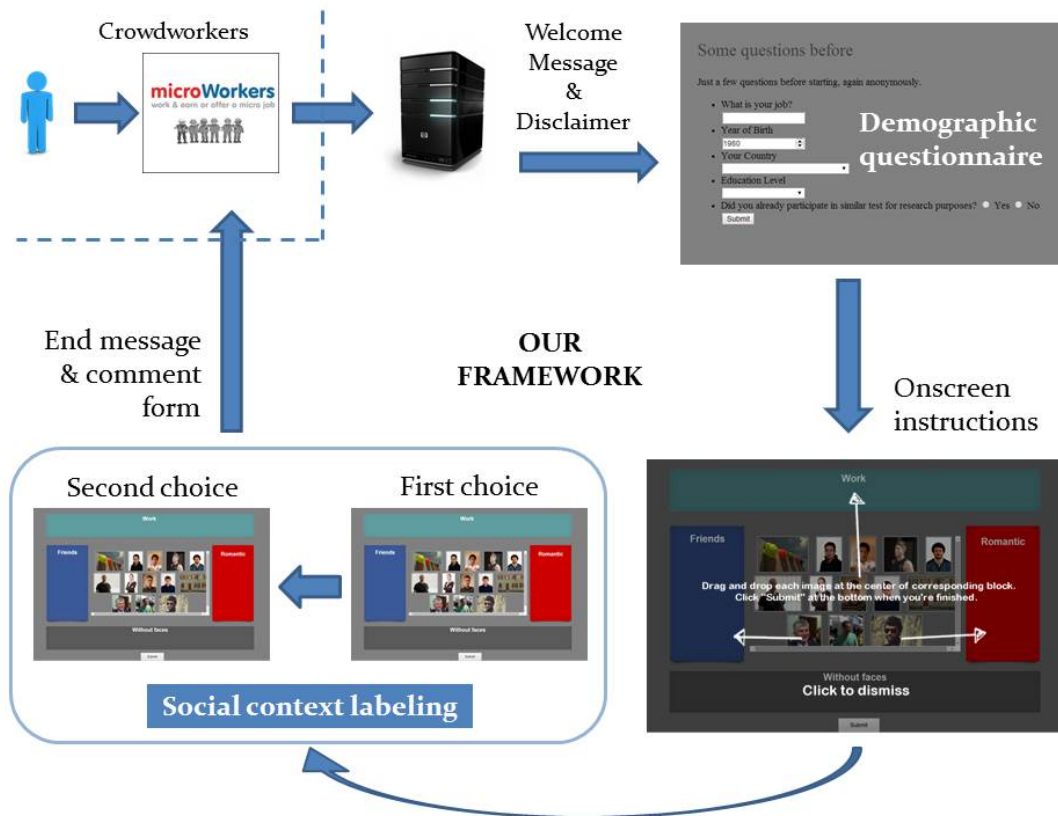


Figure 5.4.1: Schema social context labeling evaluation we implemented. Different sessions will show different images, but follow the same procedure.

five people per location (i.e. same place as a University or Internet Point that share a common outbound IP). Regarding the economic reward for participating the experiment, we decided to pay each participation 0.50\$, basing our choice with our previous experience in crowdsourcing and considering current prices paid for similar work.

Materials

We adopted as image source our database of portraits, collected online as explained in Annex F. We tried to select pictures related to the three chosen context categories: friends, work and dating purposes.

For the work related purpose, we took from our image database especially images taken from Elance.com, as that social network requires explicitly a portrait for professional purposes in their Terms of Service. In addition, we adopted 35 pictures from the LFW data set[Huang 07], suitable for the work related category. We avoided portraits of world famous people (e.g., world known politicians), since knowing their profession might influence the assessment of the fit to a category. A careful image selection was required. Extreme close-ups or too small pictures were also not selected, as content information was actually lacking. On the other hand, also pictures showing the context too clearly (e.g., showing signing a contract in a working environment) were discarded. To have more stimuli, we also used the best portraits we created for our previous pilot study (see 4.4.1), including both selfies and professional portraits. While it would have been possible to have a huge number of portraits, we limited the actual number for the experiment, since all pictures had to be assessed also on their high level features manually. Thus, as a compromise in terms of accuracy and reasonable time/cost for the experiment, we decided to use 216 collected pictures.

Two images of the Toyama database[Toy 10], that did not contain any face and then by definition are not portraits, were included in each session, for outlier detection purposes as explained later.

Asking participants to rate all 216 images at once would have been unfeasible in crowdsourcing, as the experiment would have been too long. We then split the portrait subset in smaller subsets, preparing different sessions for the experiment. Preliminary experiments showed that a good compromise to maximize efficiency while avoiding participants' withdrawal is providing 25 images per session. This number corresponds approximately to a test of ten minutes max.

Measures

The same demographic questionnaire of previous pilot study (ref. 4.4) has been proposed to participants, to collect participants' gender, job, nationality, birth

year, education level and if they were new to this kind of experiments (figure 5.4.2). This time birth year field was providing a default value (birth year=1880). This was our first honeypot:

Figure 5.4.2: The proposed demographic questionnaire.

Social context ranking for each portrait has been asked to participants. Participants were asked to express for each portrait to which context it fits best. Subjective assessments have been given through the web interface we created for the experiment, as shown in figure 5.4.3. Classification was done simply with "drag and drop", as was explained in the provided instructions. They were given three categories, referring to friends, work and dating purposes. A special fourth option named "without faces" was added for reliability check purposes: images without any person inside must be labeled as such. Categories were displayed on user interface as colored boxes.

Three honeypots have been adopted, as reliability measures. The first one is in the demographic questionnaire: people not paying attention to the questionnaire, leaving unmodified the default value for birth year, would report an impossible age. The second one is the presence of non-portrait images that have to be labeled as such, labeled with the fourth option given as previously described. Measures of time required by participants to label images are our third reliability check. We measured required time to evaluate images through our crowdsourcing framework. Empirically, we found that to conclude the test as fast as possible - without actually looking at the images and just randomly giving assessments - would require no more than 20 seconds (10s for the first choice, 10s for the second one). We then considered such a timing an outlier behavior. We gave participants with such behavior a second chance: the first time they rate images too quick, they are visually warned by an on-screen message.

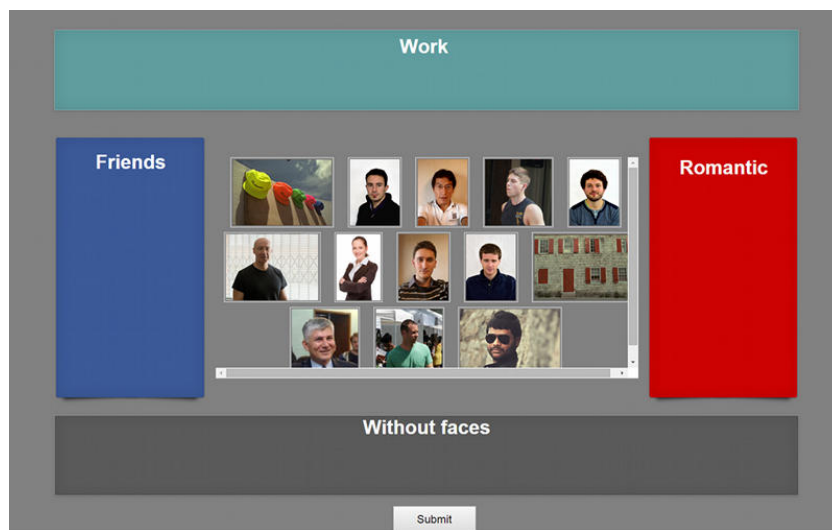


Figure 5.4.3: The interface we developed for participants to classify portraits within the three social contexts. Images shown are those in our second data set.

Procedure

After participants accepted the work on Microworkers platform, they have been redirected to our crowdsourcing server and welcomed by our framework. They have been informed of the scope of the experiment and the methodology. We asked them also to accept that we collect their answers but also anonymous information regarding the machine and behavior. After acceptance, participants have been asked to fill the demographic questionnaire.

Successively we showed the main interface. Before starting the test we provided graphical instructions indicating how to use the interface. Then we let participants express the social context that best fits each portrait, by dragging and dropping each image in box representing a category. After rating all portraits, we asked participants to indicate a second choice for the social context of each portrait. It was not possible to express twice the same context as both first and second choice: we disabled this behavior by preventing it in the user interface¹¹. Implicitly, we then obtained their third choice too. This approach allowed us to have a ranking of the three contexts for each image.

After rating pictures, we thanked participants and gave him the codes needed to be paid by Microworkers. At the same time we provided participants the possibility to leave a comment, in order to have useful feedback to monitor our experiment.

¹¹An error message would appear and ask to repeat the choice.

Outliers analysis

Despite these numbers, many participants withdrew the experiment before the end. Many withdrew just after the demographic questionnaire, and only 440 actually participated to the experiment. Another small part of participants withdrew during the experiment itself. We cannot say whether they considered their participation not worth the price or whether they experienced technical problems. Investigating experimental timings with server logs, we found that many users in remote regions, notably Eastern Asia, reloaded the test web page more than once in the first crowdsourcing campaign. This fact might indicate poor network connections, and as such, being unable to rapidly load images, even if scaled. However, we collected more than 17000 subjective evaluations (best fit + second choice). Comments provided at the end of the test were dominantly positive; users noted that the experiment was clear and asked to be informed of future tests.

Evaluations were checked for reliability using hidden honeypots. In particular, even if the second honeypot was quite explicit - provided instructions clearly stated to label non-portrait pictures as «non faces», many participants instead labeled them as portraits. While there might be multiple causes for such errors (e.g., misunderstanding the instructions, poor English comprehension or poor attention) we considered participants falling in one of the honeypots as outliers. Around 20% of participants failed to properly indicate the non-portrait images.

Again, around 20% of the participants failed to provide a reasonable birth year, our first honeypot. Many participants left the birth year unmodified (reporting 1880 as birth year), many others modified only last two digits instead (i.e. reporting 1870, 1880, 1890 ...), only very few of them just typed it wrong (i.e. 194). Figure 5.4.4 shows answers histogram. Participants who took care about this detail and those who didn't are two clearly separated groups, as can be seen in figure 5.4.4. We considered them equally outliers. However, it is interesting to investigate if those participants who left the birth year unmodified and those who did not notice the faulty century value behaves differently in the test. For this purpose we retrieved the evaluations of these groups and checked if the two distributions over the three contexts are statistically different. A Fisher exact test underlined that indeed the two groups behave differently ($p = 4.8315e - 005$). This finding suggest that maybe not all of them behave as outliers; however, lacking of further details to make a distinction between these two participants' groups, we considered them all equally outliers.

Very few participants (around 1%) have been marked as outliers considering the third honeypot, related to the answer timings. We can safely affirm that to this outcome greatly contributed the displayed warning message previously described. The two groups of outliers - spotted by first and second honeypot - partially overlap: some participants failed both honeypots. In total, less than 40%

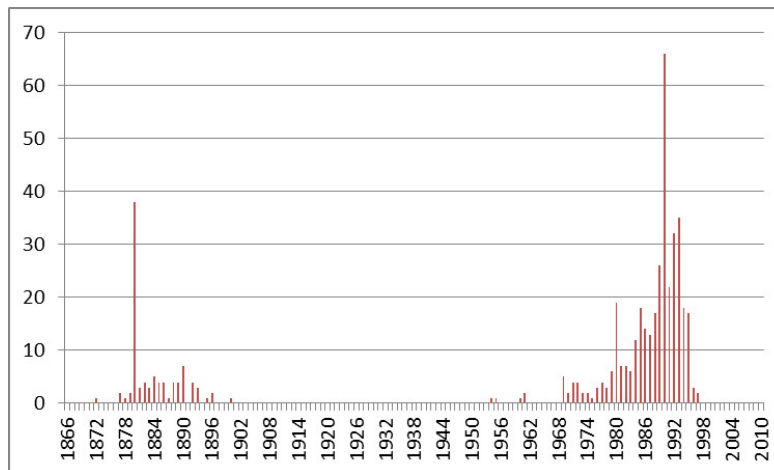


Figure 5.4.4: Histogram of reported age from participants. Two groups are clearly visible, on the left those who did not take care about birth year century, and on the right those who did. The peak on the left group corresponds to participants who did not change at all the default value.

of participants were outliers. The corresponding submitted jobs were discarded and participants were notified. Outliers were not uniformly distributed on all experiment sessions and some images have then few evaluations less than others. The analysis of collected evaluations is detailed in following section 5.5.

Discussion

Crowdsourcing technique has proven to be very powerful: we quickly collected many subjective evaluations, from all over the world. A similar test with the same number of subjects in laboratory would have required around eleven days. We gathered the same amount of participants in half the time, limiting by purpose the acceptance rate (to limit demographic unbalance, ref. 3.3). Implemented interface, providing a labeling approach to portrait context, appeared to be clear for participants and correctly adopted. In the end, we gathered enough valid subjective social context evaluations for our portrait subset.

However particular attention is required as many provided results are not satisfactory. Proposed honeypots demonstrated to be effective. Both the first one and the second one underlined that some participants did not pay enough attention to the test or did not understand the instructions. The third honeypot instead underlined few outliers; however many visual warnings have been raised to participants. It seems that a proper behaviour is shown when participants are explicitly

told that they will be excluded if they do not provide evaluations correctly.

Providing the possibility to leave a comment was useful too. While many participants just inputted messages like 'Thanks', others underlined technical problems encountered and allowed us to quickly respond. In some cases participants just send an empty message instead. This fact gave us the idea for a new honeypot: as the comment is sent only when the 'send' button is clicked, users then just clicked on that without reading the instructions¹².

5.5 Analysis of gathered social context evaluations

This section is dedicated to the analysis we carried out on social context evaluations. Sections cover different research questions that we considered important for our work. Here we consider all outliers removed.

5.5.1 General results

In the end, we obtained more than 6000 valid context rankings for further analysis and all images had a sufficient number of evaluations (>20). Figure 5.5.1 summarizes participants demographics information after outliers removal. In particular, most participants have been found to be young male subjects from Eastern Asia. Mostly they have a Bachelor degree or have at least completed the high school. More than half of them already participated similar tests previously.

Through described experiment, each portrait of adopted subset has been associated with a context that fits it the most. As it is a subjective test, we have a preference distribution between the three proposed contexts. So for example, we can have a portrait that has been evaluated 80% related to work, 10% for friends and 10% for dating purposes. The same holds for the second and third possible contexts fit that we asked. An excerpt of results is given in table 5.1. Figure 5.5.2 shows portrait distribution in our social context space.

As explained before, participants expressed a first and second - different - choice for portrait contexts. We investigated the «evolution» of the second choice respect to the first one, that is to say if there is a more probable second choice once a first choice context is chosen. We analyzed this point computing the average probability of transitioning from context A to context B on our data set, for each possible transition (i.e. first choice «friends», second «for work», ...). Figure 5.5.3 shows the resulting probabilities. Interestingly, the most probable second choice for a portrait labeled as for work or dating (as first choice) is much more likely to become for friends than other. A possible explanation for this result can be that

¹²We discard here the possibility of a technical problem impairing the correct message dispatch

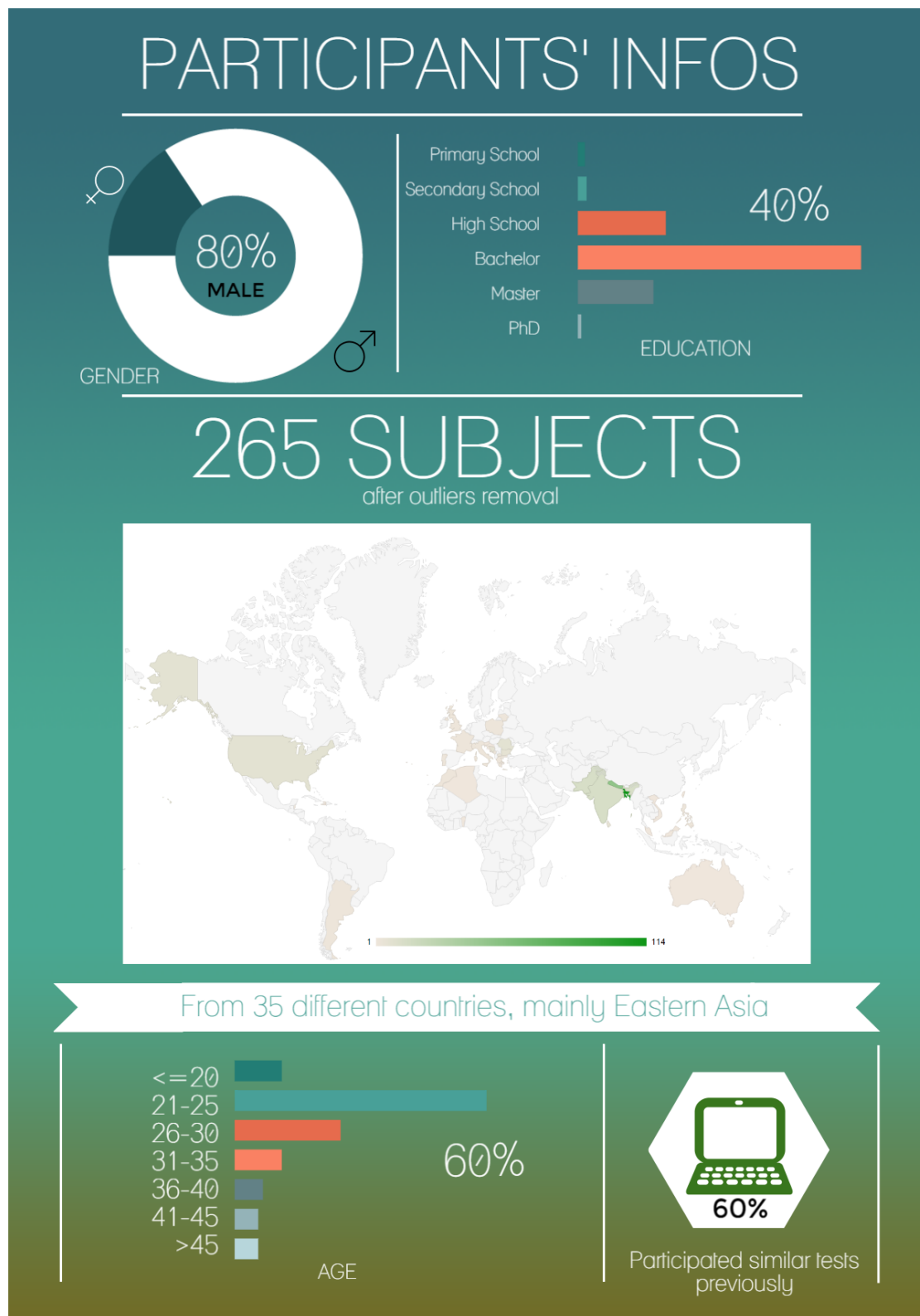


Figure 5.5.1: Participants demographic information for conducted study.

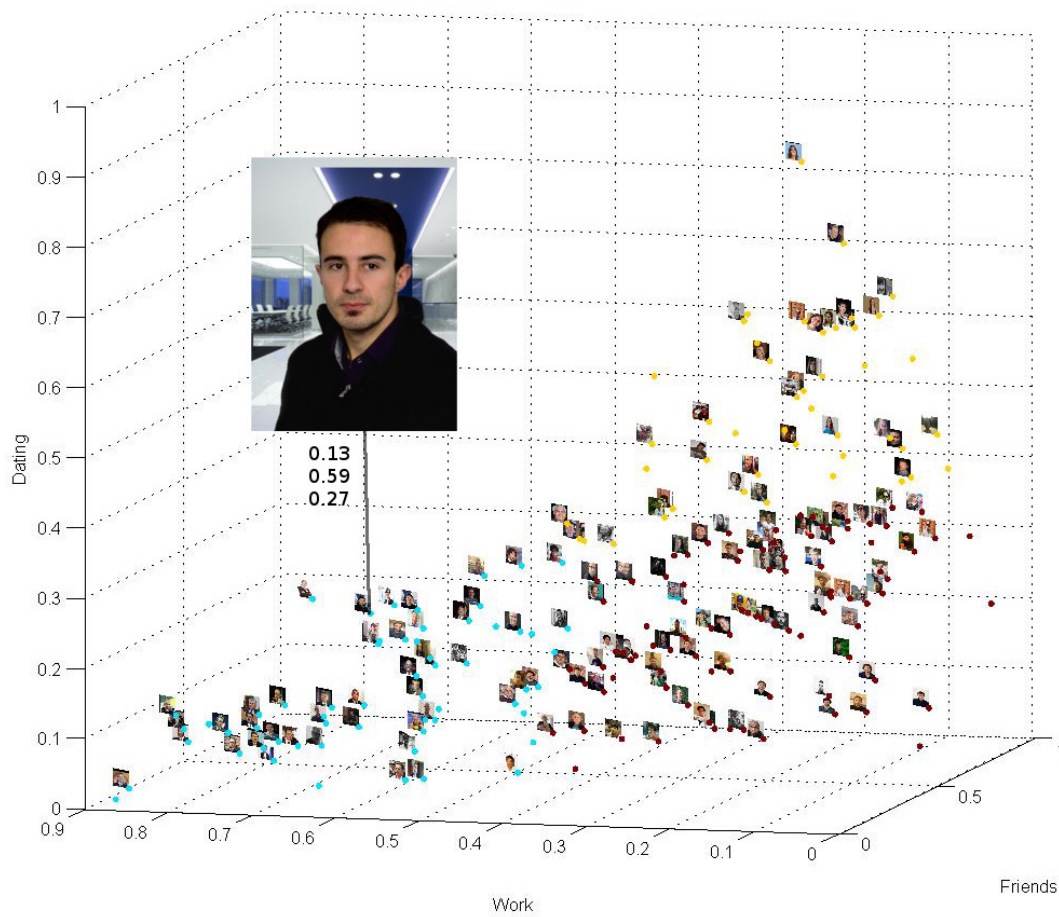


Figure 5.5.2: Representation of a part of portraits labeling in a 3D space, with Friend, Work and Dating labels as dimensions. Each point correspond to a portrait; zoom over a portrait shows choices as percentages.

PORTRAIT	Friends	Work	Dating
img1	0.19	0.14	0.66
img2	0.33	0.57	0.09
img3	0.38	0.28	0.33
...
img216	0.45	0.22	0.32

Table 5.1: An extract of subjective preferences expressed for social context of portraits. First context chosen is shown, preferences are expressed in percentage. Similar results have been gathered for the context of second and third choice.

portraits for work or dating context are hardly interchangeable between the two contexts. It has to be noted that, while for each evaluation the first and second choice are different, averaging on all evaluations it can happen that a portrait is mostly labeled for one purpose on both first and second choice¹³.

As we detail in successive chapter, in the analysis considering portrait features we adopt discrete labels for our portraits. For this reason we threshold these probabilities adopting a majority rule, a common strategy to obtain ground truth from multiple assessors [Welinder 10]; in our case we then have for the the social context $sCtx$:

$$sCtx_i = \arg \max(\overrightarrow{sCtx_i})$$

where i indexes the portraits, and $\overrightarrow{sCtx_i}$ is the vector of expressed preferences. Adopting this method, each portrait is labeled with the most selected social context. So, if a picture was reported to fit a work purpose by 90% of the participants, it is considered as such. Adopting this threshold, our data set presents 50% of the portraits labeled as friends, 32% for work and 18% for dating. Figure 5.5.4 shows some example portraits. While this unbalance can bias successive analysis, we will consider also other labeling strategies as later explained.

5.5.2 Do observers agree on social contexts of portraits?

It is interesting to see if expressed choices present then a statistically significant relevance and if so, for which images. To this extent we can see if choices distributions are significantly different from fate, for each portrait. In practice, this is

¹³I.e. a portrait with probability distribution [.5 .25 .25] within the three contexts (if first choice evaluations spread on the two .25 bins converge in the second choice for the mostly chosen context as first choice, we would end up having again .5 on that same context).

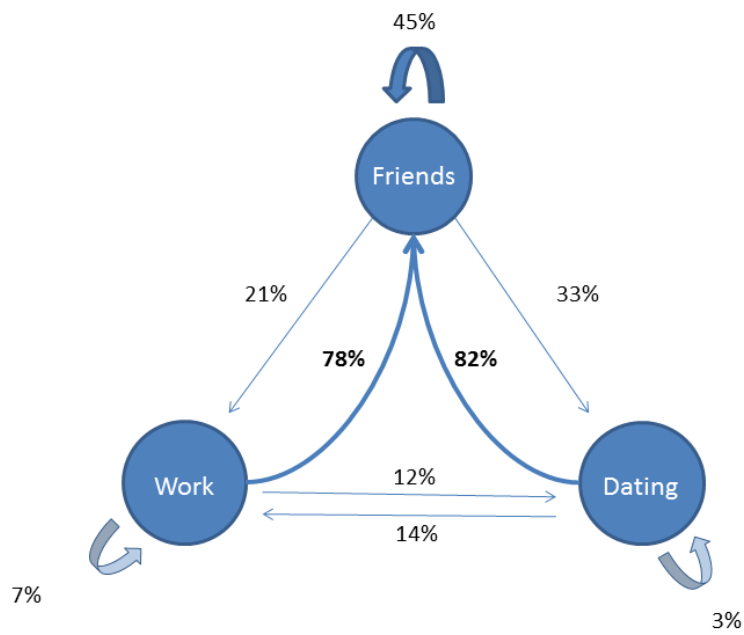


Figure 5.5.3: Diagram showing the transition probability from first to second choice for each possible context transition.

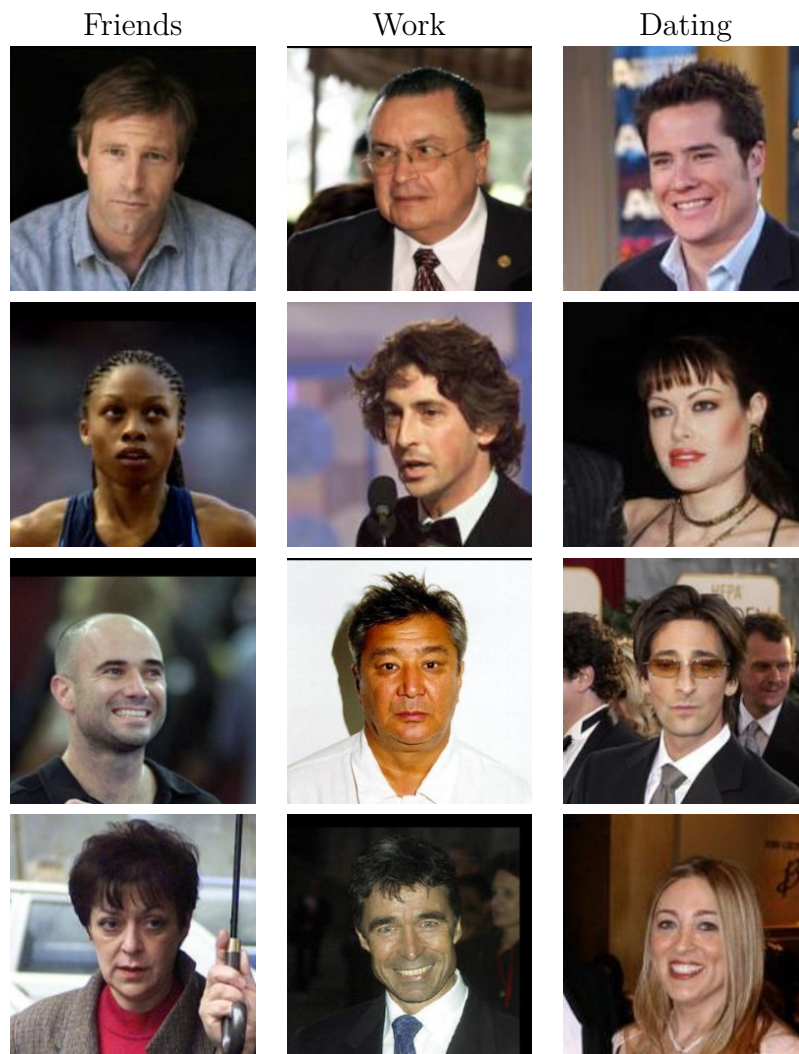


Figure 5.5.4: Some portraits of our data set, labeled respectively as for friends, work or dating purposes after applying a majority rule. Images come from the LFW subset.

equal to check if the a posteriori probability of the most frequent chosen context is significantly different from $1/3$, as three are the possible contexts in our case. A binomial test has been adopted: this is an exact test to check statistical significance of deviations from expected distributions. As in our experiment three are the possible cases, we could check statistical significance also through a multinomial distribution. However we simplify here the problem considering only the first most selected context.

The analysis underlined that at least 75% of portraits present a statistically significant context choice¹⁴. These portraits are then clearly perceived as belonging to one of the social contexts. However the fact that many portraits do not present a statistically relevant choice does not mean that they won't be useful for our analysis. We can still use them to run statistical analysis to investigate which factors influence choices, as we will explain in next chapter. Figure 5.5.2 shows some examples of portraits with their p-value for the binomial test.

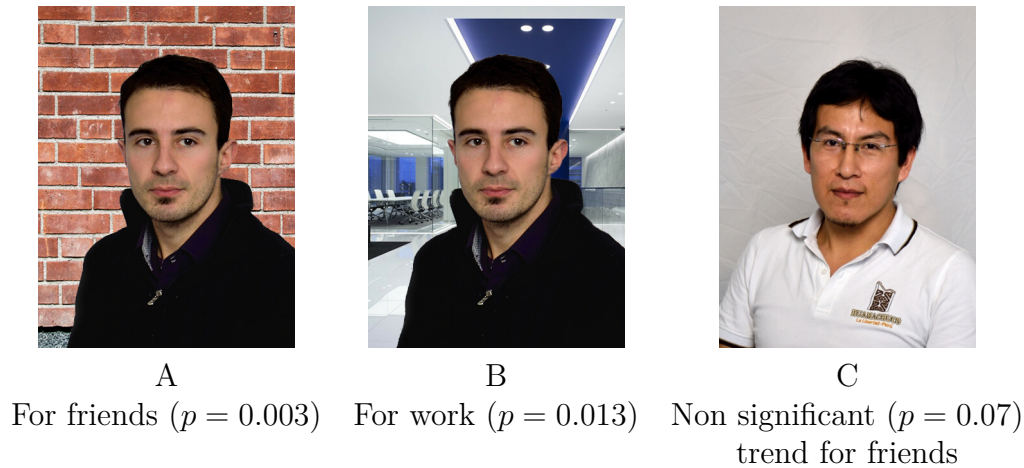


Figure 5.5.5: Some portrait images of selected subset, with most selected first context and choice significance. Original images: made in laboratory for experiment in previous chapter. Background has been modified for A and B.

During this analysis we came up with a finding that pointed out a feature to evaluate. As we can see for pictures A and B, initially made for previous pilot study and then modified, simply changing background shifts perceived context from working purposes (picture B shows as background what can be described as an office or a meeting room) to friend interactions (picture A has a wall of bricks as background). While for this particular portrait (subject in A and B)

¹⁴Analysis considered only first context chosen. Significance threshold $\alpha = 0.05$.

the background is a factor of influence, we cannot say that background scene is a global influential factor. This element suggests anyway that background scene modify somehow the cognition of the portrait itself and that this element should be further investigated. We will include scene background interpretation within the high level features to evaluate, explained in next chapter.

5.5.3 Are chosen social contexts too strict? Hierarchical clustering approach to visualize social context mixtures

As we've discussed in previous section, not all portraits are clearly belonging to one of the three contexts that we proposed. Many of them have been placed by participants in between two contexts. This fact let us consider that maybe proposed contexts may be limiting the expression of subjective opinion as actually it would have been plausible to add other options such as «portrait for either friends or dating». To evaluate the data partitioning that underlies gathered evaluations we adopted a hierarchical clustering approach. This technique allows us to check the presence of portraits that are a «mixture» of different social contexts and which groups and subgroups of portraits have been outlined by our labeling strategy.

Hierarchical clustering organizes data set in hierarchical groups, either joining data points (agglomerative approach) or dividing groups (divisive approach) at each step. This algorithm does not require an a priori number of clusters K as other clustering algorithms, i.e. K -means; indeed it has the advantage of outputting a measure of dissimilarity between created sets, useful to find natural clusters in data set. Only a distance metric between data points and a linkage criteria to split or merge groups are needed. This kind of approach for image data sets has been adopted especially in the field of Content Based Image Retrieval [Krishnamachari 98, Pandey 13] to improve search performances.

We run hierarchical aggregative clustering algorithm different times adopting different metrics and linkage functions; obtained partitions have been evaluated with correlating original distances between data points and resulting distances from the three, measuring how well pairwise distances between original and clustered data points are preserved. These have been computed adopting cophenetic distances [Sokal 62]. Empirically we've obtained best results using correlation as distance metric between data points and a simple unweighted average distance as linkage function between clusters ($cophenet = 0.78$). Figure 5.5.6 shows dendrogram representation of obtained hierarchical clustering. Leaves of dendrogram represent our data points, the portrait images. Graph pies have been added to indicate labeling classes in percentages for portraits within a cluster.

We can see from the dendrogram - from top to bottom - that either 3, 4 or 5

clusters are needed to obtain high consistency within clusters¹⁵. Dividing the data in three parts, the three social context we provided are well represented: a cluster mainly for dating, working or friends purposes. The clustering algorithm suggests to split the first branch, the dating context, in two parts; in particular it puts aside in a new cluster few portraits that in average do present an even probability of being chosen as one of the three social contexts. In order to consider 5 clusters instead, the algorithm suggests to split the friend context in two: one cluster mainly for friends+work and another for friends+dating. This finding reinforces our idea that some portraits for friends may indeed be adopted for the other contexts as alternative. However, due to the fact that the number of images in our data set is limited, we prefer to start our analysis with three clusters.

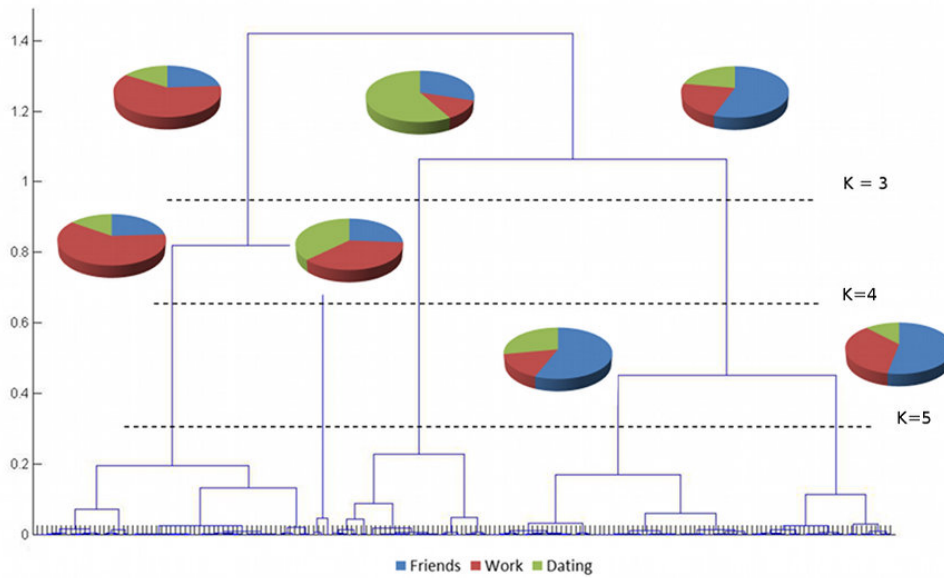


Figure 5.5.6: Dendrogram illustrating hierarchical clustering of our data set. Only 180 stimuli belonging to main clusters are represented for clarity.

5.5.4 Demographic issues with social context evaluations

We discuss here a demographic issue that has been remarked during analysis of high level features - subject of next chapter - but that we prefer to detail here

¹⁵Groups can be defined as inconsistent when distant data points start to be merged together. Split of these groups (showing as links below the split in the dendrogram) are said to be more consistent

as it is related to the bias of test participants gender on portraits social context labeling.

Next chapter deals with the social context evaluations that we gathered to evaluate which features impact context perception. Data analysis underlines that for predicting the appropriateness of a portrait for dating purposes, also the gender of the portrayed person is a statistically significant contributor. In particular, we discovered that portraits depicting female persons are reported as more appropriate for dating purposes. This result is somehow surprising and it raises the question if it is not due to a social bias. In this respect, it is interesting to check if to this outcome contributes also the gender of the rater. To this extent we adopted permutation test as in our previous pilot test (ref. 4.4) to check if the gender of participants in our experiment is a global influential factor for social context evaluations. The factor investigated was the gender of the participant as obtained from the demographic questionnaire, while the dependent variable was the context evaluation: we computed a 2x3 contingency matrix for each portrait, being the rows male or female participants evaluations. Chi-square test has been employed to check if a statistical significant difference between rows is present in contingencies matrices. We run 1000 permutations, randomly changing reported gender to half of participants each time. Permutation tests underlined that a global effect is present ($p < 0.05$): male and female evaluations are statistically different.

To analyze if the effect is present especially on one context evaluation, i.e. belief of some portraits more for dating, we repeated tests considering one context (i.e. work) against the other two (i.e. friends+dating): we run other three permutation test, for the three possible combinations respectively. Gender was found statistically influential only for the appropriateness of portraits for dating purposes, underlining that portraits were evaluated differently between men and women only in the context of dating. Figure 5.5.7 shows significant outcomes distributions from the permutation tests.

This finding pointed out that gender related issues are present and that this effect should be more explicitly considered in future research. A possible strategy would be for example to show participants targeted stimuli considering participant gender (i.e. stimuli in which the gender of the portrayed person and the gender of the participant are the same). Alternatively, analysis that consider multiple factors at the same time must be adopted. We discuss these methodologies in next chapter.

To conclude, the same approach can be followed for the other demographic information that we asked in the questionnaire, except for the reported age that has been previously shown affected by outliers (section 5.4). However in this work we prioritized the data analysis that we considered important, such as the one carried out in this section. Still it is interesting to underline, regarding the reported

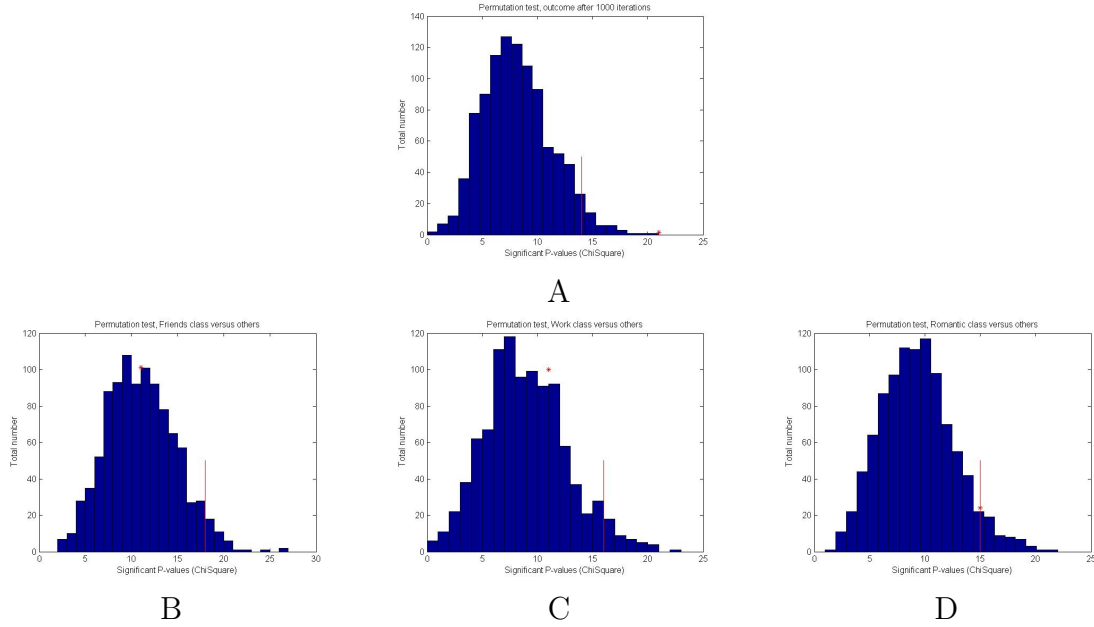


Figure 5.5.7: Histograms of statistically significant outcomes of permutation tests, studying the effect of participants gender against social context labeling. Number of significant outcomes in x-axis, occurrences in y-axis. A star (*) indicates the original outcome in the experiment previous to permutations: if it lies over the 95% percentile of the distribution (red threshold), then it means that is very unlikely to have the same result by fate. Global test (A) underlines the presence of a significant difference between evaluations done by male or female participants. Tests in B, C and D repeat permutations considering social contexts singularly to underline where the difference is present: friend (B), work (C) or dating context (D) respectively. Apparently, men and women differ evaluating dating context.

resource, once taken really care about experiment design and participants reliability. Our framework showed to be flexible and easy to customize. The reuse of previously developed module (i.e. demographic questionnaire) helped us greatly.

To make crowdsourcing effective we had to forecast outlying behaviors during portraits labeling and deal with them during data analysis. We designed and implemented two simple yet effective honeypots. Our experiment underlined that almost 40% of participants was not completely reliable. While some honeypots failure may be due to simple lack of attention more than willingness to cheat, we preferred to exclude outliers from analysis. Most portraits' context evaluations that we gathered were correctly provided and enough to continue our analysis of influential factors.

Statistical analysis of preferences significance underlined that only a minority of portraits do not show a statistically significant social context choice. We adopted a hierarchical clustering approach to analyze the natural clusters of portraits given our labels. Analysis underlined especially that the friend context could be split in two to better model the fact that some portraits are perceived for other purposes than friend relationships as well. A qualitative analysis of results let us find that background scene may be an important factor: while this discovery came from a single stimuli, so we cannot say anything about statistical relevance, it pointed out another picture feature to consider.

Our previous experiment underlined also that demographic considerations must be taken into account carefully when dealing with researches that can be affected by socio-cultural differences. Permutation tests underlined the joint effect of participants and portrait depicted subject gender over the choice of a portrait to fit dating purposes. While this can be seen as a bias effect, we have to consider that participants may have (maybe also cultural dependent) opinions that are not necessarily biased or outliers behaviors but must be considered to understand which factors influence the perception of a portrait.

Gathered context evaluations represent a subjective perception our portraits. To understand which elements within the pictures actually brought to these assessments, we conducted statistical analysis taking into account image features as explained in next chapter.

Keypoints

Context

- ❑ A data set of social context labeled portraits does not exist in scientific research.
- ❑ Crowdsourcing has been positively adopted for generic picture labeling.
- ❑ Honeypots have been shown to be valuable for picture labeling purposes too.

Contributions

- ❑ SoA of crowdsourcing for labeling purposes analysis.
- ❑ We setup our framework to allow portrait labeling in crowdsourcing, with an easy to employ drag and drop interface.
- ❑ We proposed effective and easy to setup honeypots for picture labeling.
- ❑ We built a portrait database labeled on social context, either for friends, work or dating purposes respectively.
- ❑ We underlined a demographic issue related to the gender of test participants while evaluating portraits for dating purposes.

“When in doubt, do the simplest thing that could possibly work.”

(Ward Cunningham)

**UNDERSTANDING AND
MODELING PORTRAIT
FEATURES INFLUENCE IN
PERCEIVED SOCIAL CONTEXT**

Chapter 6

Understanding the importance of image features in portraits social context classification

Previous chapter described how we collected perceived fit of various social context categories for a set of portrait images. Categories were selected based on current typologies in social networks, mainly for friends, work or dating purposes. In this chapter we describe how we exploited gathered evaluations in order to model category fit based on images features. For this purpose we used linear models and machine learning and adopted both low and high level portrait features. While the first ones focus on pixel intensities, the latter are focusing on complex features related to image content interpretation. In order to extract these high level features we relied again on crowdsourcing, since computer vision algorithms are not yet sufficiently accurate for the features we needed. Our results underline especially the importance of some high level content features, e.g. the dress of the portrayed person and scene setting, in categorizing portrait social context.

6.1 Introduction

Our research focuses on understanding which are the elements in a portrait image that influence the perception of the portrait itself. As we said in previous introductory chapters, different images of the same person may give very different messages. We restricted the concept of portrait perception to a specific one: the perception of which social context best fits a particular portrait. As an example a portrait may be perceived subjectively more for work than for dating purposes. With this premise in mind we collected real online portraits from different platforms, as explained in Annex F.

However we had to collect also portrait's subjective evaluations for the social context, as collected images reflect only the perception that the author has for them (i.e. he may think that his LinkedIn picture is professional and fits well that purpose) but this opinion is subjective. We then collected subjective evaluations with crowdsourcing as explained in previous chapter. These evaluations correspond for us to the ground truth on which carry our experiments to understand which are the characteristics that influence perception. For this purpose, we need to extract portrait features and mathematically evaluate the influence that each feature has on social context perception. This is the topic of the current chapter.

We referred to literature in image analysis to select and extract classical image features based on pixel values. However we believe that these features alone will not be enough to address the problem. As Fedorovskaya pointed out in her recent review on «human centered multidisciplinary studies related to imaging» [Fedorovskaya 13], research slowly started to include more and more «higher level psychological» elements in image related studies. We then considered also some research works within the social sciences domain, as we thought that these can suggest elements to analyse within portraits. We will refer to these last as high level features, generalizing the concept expressed by Ke in [Ke]. In contrast, the former kind of features will be named here as low level features.

To evaluate which features influence portrait perception, we then extracted both low and high level features from these portraits, as explained in next section.

6.2 Portrait image features

This section is dedicated to the evaluation of portrait features that we believe to be important in the assessment of social context. These features, chosen after research literature survey, can be divided in two main groups: low and high level features. The difference stands in the fact that the latter are not simple measures based on pixel values but deal with a deeper cognition of the whole scene in the picture. The former ones were directly computed from pixel intensities (e.g., contrast), whereas the latter - related to content interpretation (e.g. dress typology of portrayed person) - were assessed subjectively. To avoid at first complex computer vision approaches to high level features and having at the same time enough subjective evaluations, we opted again for a large scale subjective campaign via crowdsourcing. Automatic feature assessment algorithms are not subject of this section, but they will be outlined in next chapter.

We review literature involving image features analysis in order to have a broader view of what has been done. We briefly discuss their findings and construct a set of features to consider in our statistical analysis. We want to stress the fact that many other features could have been considered in our analysis; however we preferred to

limit them and focus also on the overall methodology at first.

6.2.1 Low level features

The designation of «low level» for certain image features comes from the link that these features have with «low level biological visual processing» hardwired in our brain; this is the case for example for edges detection [Folsom 90]. With time, this term has been used to indicate features that are directly related to image pixel values [Szummer 98]. Generic low-level features come without «explicit heuristic meaning related to the features values» [Xue 13]. In last decades these kind of features has been widely adopted in researches dealing with image assessment and classification [Luo 01, Datta 06]. Usually, these features are easily computed through mathematical expressions on pixels; most typical examples are image luminance and contrast (see fig. 6.2.1 and 6.2.2). Also the color is included in this definition: other examples of low-level features are color saturation and histogram.

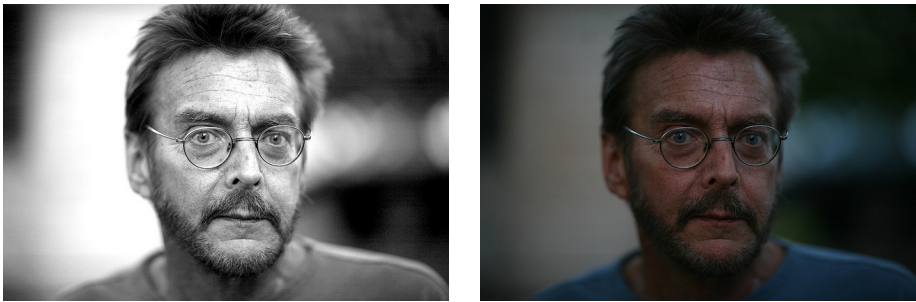


Figure 6.2.1: Two versions of the same portrait, both online and modified by the author. The effect they give is very different. SOURCE: Flickr, Alan Turkus in Richard. CC License¹

Interpretation of color in an image, sometimes called colorfulness, has been adopted in aesthetic research [Amati 14]. Even more complex features as GIST [Xue 13], SIFT and HOG are labeled as low-level features [Totti 14]. It is not straightforward to conduct a complete review of low level features as in some cases the same feature is called differently between different works, but they refer to the same concept (i.e. sharpness or blurriness). Table 6.1 summarizes the most important features we found in reviewed literature. We want to stress the fact that we will use low level features just as a tool in our research, and that it is not purpose of this work to dig into the subject. We want instead to focus on high

¹<https://www.flickr.com/photos/aturkus/2434132519>

level features directly related to semantics, as explained in next section. We will then rely mainly on existing work that review low level features.



Figure 6.2.2: Two versions of the same portrait, with different contrast levels. The right one has been modified to simulate a hard contrast. Many more details are visible in the original image than in the second one. SOURCE: Flickr, F.Stender in Namics, CC license²

A good list of visual features is given by Datta in [Datta 06], who explains also their mathematical computation. In this work the author focuses on the challenge of automatically inferring perceived aesthetic quality of pictures adopting a «computational approach». Positive results have been obtained adopting chosen low-level features in discriminating pictures' aesthetics between two levels (high and low). Pictures were subjectively evaluated online, on the popular network Photo.net. Still regarding image aesthetic assessment, a large amount of research using these features is reported in literature and reviewed by [Khan 12]. Many of the proposed features are also used to assess human portraits, considering regional statistics. Between them, Khan underlines how the composition of light and dark areas are «important factors for aesthetic appeal». In this respect, we added also contrast in our low level features, considered for image aesthetics assessment by [Wong 09]. Wong stresses also the importance of image sharpness, computing

²<https://www.flickr.com/photos/namics/7163386518>

different measures. We also believe that this feature is important (i.e. fig. 6.2.3), but we preferred to exploit crowdsourcing - adopted for high level features - asking sharpness subjective perception instead of computing textures on salient regions.



Figure 6.2.3: Two versions of the same portrait, made by us in laboratory conditions. The right one has been downsampled to simulate a low resolution camera. Assessors will likely agree on the fact that the first one is more professional than the second one.

In many cases a priori information on image content has been considered while computing low-level features to have more powerful features. These are mainly related to image composition as done by Dhar in [Dhar 11], or with the rule of thirds by Datta in [Datta 06] or in [Tong 05] by Tong adopting a saliency map. All these features somehow consider image semantics: cognition of image content is required to evaluate the composition. Even if sharp edges in the image can be used to automatically assess image composition, they may not reflect the intention of the author, that may be put the focus on other areas of the image. In that case, composition would not consider the real focus of the image. Ideas coming from professional photography techniques pushed authors of [Luo 08] to consider the distinction between background and main subject. Their automatic assessment is based on the principle that most attention in a picture should be on the subject, removing other objects. This effect in a picture is obtained for example with a short depth of field. Still, a priori knowledge of the main subject is required: just measuring the difference in terms of blurriness between foreground

and background won't be enough, as many examples of mistaken shots are available. These features in our opinion lie in the middle between low-level features, as directly computed from pixel values, and high level features as considering semantics. The description «mid-level feature» to indicate features to bridge the gap between low-level descriptors and semantic concepts has been already adopted in literature; in [Totti 14] Totti indicates a bag of feature scheme - aggregations of different computations on pixel values considering the whole data set - as a way to include complex representations. While these features are interesting and can be valuable, we prefer not to include them in our set of automatically computed features and focus on high level features. Mid-level features could be addressed later on considering subjective assessments (i.e. to have more reliable ground truth regions of interest) once other factors will be mastered.

As we believe that aesthetic quality can be also a factor influencing portrait perception, we partly followed the work of Datta and Totti for selecting our low level features and implementing computer vision algorithms. In particular, we adopted HSV space statistics (mean and standard deviation) of whole image, aspect ratio and resolution. Image resolution also has been underlined to be an important low-level feature in [Chu 13]. Authors show that image resolution and also physical dimension influence subjective aesthetic perception in a complex way. Being these features very easy to include, we added them to our feature set. We added aspect ratio in our subset especially because we found different picture formats and orientation. Low level features adopted in this research are summarized in tables 6.3. These have been computed using Matlab (R).

6.2.2 High level features

Higher level cognition of picture content is related to scene awareness and image semantics. The human brain elaborates information from an image producing a much richer idea of what is happening in the scene. Every detail in a picture can add a meaning to the image and bias the evaluation, and it is not possible to consider these effects adopting only low level features. Then, we consider also high level features that for us are those related to the cognition of the scene. Research focusing on image assessment already addressed many higher level features, considering image content analysis.

Psychology studies suggest that brain information processing is even more important when people are depicted in the picture: a lot of studies have been carried on personality perception of depicted subjects, without a priori information [Todorov 11]: cognitive biases can influence the perception. We believe that socio-psychological researches can underline interesting factors to consider; we then investigated not only literature dealing with image assessments but also some works involving social sciences. However, we want to stress that our concept of content

LOW LEVEL FEATURE	REFERENCE
Resolution	[Datta 06, Chu 13, Totti 14]
Aspect ratio	[Datta 06, Totti 14]
Luminance	[Totti 14, Tang 13, Datta 06]
Contrast	[Tong 05, Datta 06, Totti 14, Redi 15]
JPEG Quality	[Li 02, Redi 15]
Noise	[Li 02, Redi 15]
Colors	
- basic and dominant color	[Totti 14]
- colorfulness	[Ke , Tang 13, Li 10b, Totti 14, Aydin 14, Amati 14]
- HSV statistics (mean, ...)	[Datta 06, Xue 13, Totti 14, Redi 15]
Focus	
- region of focus (DOF)	[Datta 06, Totti 14]
- centrality	[Totti 14]
- density	[Totti 14]
Sharpness	[Li 02, Aydin 14, Redi 15, Xue 13, Tang 13]
Background percentage	[Totti 14]
Composition	
- rule of thirds	[Datta 06, Totti 14, Redi 15, Xue 13]
- complexity	[Totti 14, Tang 13]
- symmetry	[Redi 15, Li 10b]
Texture	[Datta 06, Totti 14]

Table 6.1: A non-exhaustive list of low level features found in recent literature on image-related researches, that helped us with our feature subset choice.

awareness and cognition discussed here comes without any psychological claim and we just refer to findings in social sciences research.

A very large number of high level features can be thought but, as underlined by [Isola 11a], choosing a large set of features leads to redundancy and it becomes then crucial to make a selection. We review here the most important high level features adopted in literature or underlined by psychology studies, dividing them between those focusing on the scene and on the portrait subject.

Features regarding the scene

Authors of [Totti 14] investigate many semantic features, mainly related to the setting of the scene. They consider in particular the setting, considering location and event, i.e. citylife, partylife, homelife, indoor/outdoor. They focus on the scope and view of the picture, considering the background scene (i.e. showing a city, a forest, a desert or rural, ...), weather and moment of the day. High level describable attributes are used in [Dhar 11], even if their focus is on attributes of non-portrait images. They investigate predictors for aesthetic purposes such as compositional attributes and content attributes as well as more precise ones as sky-illumination characteristics. Portrait specific feature have been added also by Redi in [Redi 15]; her work focus instead on factors of influence in portraits' aesthetics. In her work scene semantics and portrayed subject informations are added to low level features. A large number of manually labeled high level features has been adopted by Isola to understand the memorability of images [Isola 11a]. In this work crowdsourcing has been adopted to label images in order to describe both the scene and the people in the picture, when present. They considered in particular the space depicted in the picture, if for example the scene is showing an open space or a closed one like a cluttered room, and the dynamics of the action depicted, describing if the scene is static or otherwise which action is going on or about to. This last detail is in our opinion quite interesting as is really related to the cognition; however at the same time it can be very subjective.

With these premises, our subset of high level features regarding the scene considers then scene setting (indoor outdoor) and location (city, office, ...) features. Regarding the scene background, we also asked to note if it was blurred or not respect to the foreground, like with Depth of Field in [Datta 06].

Features regarding the subject

Previously cited [Totti 14] considers also the subject of the picture, when present, including subject age, gender and relationships between subjects - if more than one. Particular focus regarding features of humans is given also in [Isola 11a],

³<https://www.flickr.com/photos/thestylepa/6041278814>



Figure 6.2.4: Two similar close ups of the same subject, differing for smile presence and opened eyes. Both elements have been underlined as important by psychological studies. Viewers will probably agree that even if close, effect is different. SOURCE: Flickr.com, Jenny in London Retro Glasses, CC license³

where visibility, clothing and appearance are considered. Still, as said before, many psychology studies dealing with personality perception are useful to spot important high level features.

Common sense suggest us that the eyes are an important element in social interactions. This finding is supported by psychology research too. In [Chen 13] researchers underline that *gaze* between listeners and a speaker positively influences persuasiveness ratings, even if a prolonged gaze can be detrimental. We believed that gaze could have been important (i.e. fig. 6.2.4) and we finally added it in our subset. Deeper investigations considered even the eye color as an element biasing trustworthiness, discarding this hypothesis [Kleisner 13]. In [Forster 13] Foster et al. discuss how *glasses* may affect the perception of people personality, without having any prior information. Their studies evidence that different effects are elicited on intelligence, trustworthiness and attractiveness ratings by different types of glasses. We then included this feature too - the presence of glasses - between our high level features, especially because we consider working and dating social contexts, that can actually be related to intelligence and attractiveness respectively. Another element that we investigated in psychology literature is the influence of the *dress* on personality perception. Again, research confirms what common sense suggests, that clothing has a great impact on how we are perceived in both working and personal life. Multiple messages can be transmitted through dress, as Damhorst underlined in [Damhorst 90] after reviewing more than 100 precedent studies on the topic. Johnson and Lennon summarize studies findings explaining how the «content of the information communicated by dress was competence, power, or intelligence» and that in the majority of studies transmitted

messages were about «character, sociability, and mood» [Johnson 15]. Research confirms this element even when evaluating pictures: in [Behling 91] results indicate that «perception of intelligence and academic achievement are influenced by dress». For these reasons we added the evaluation of portrayed subject clothing in our analysis. We asked high level features assessors to evaluate the category of the upper body garment, giving four non overlapping choices as in table 6.4⁴. Some interesting findings in psychology are instead surprising even for the common sense. In [Conesa 95] results show that «statistically significant differences were found between the incidence of half-left and half-right *profiles*». Authors underline that these differences are consistent with developed models for attentional bias and perception of emotion that consider right vs left brain hemisphere activations. We then added portrayed subject face direction (frontal or profile left / right) as high level feature. Still focusing on the depicted subject, we included its *gender* and if he/she was *smiling*. While smile effect can be subtle in attractiveness perception [Whitehill 08], we didn't add emotion as a feature because we considered that smiling was a sufficient predictor to coarsely discriminate between main emotions as in [Xue 13].

It is important to remark that while many psychological studies considered facial traits too, in this research we do not consider them because we are not interested in influencing factors related to physical appearance of portrayed subject. We underline that for our purpose we are more interested in elements that can be modified other than the face⁵. This approach can be useful for example in order to look for the best picture respect to a given social context within a personal collection.

High level features adopted in this research are summarized in table 6.4. Many of our features are categorical variables, as head tilt and orientation (left,center,right) or scene setting (don't know/indoor/outdoor), and some of these are ordinal as subject size (from small to big).

6.2.3 Assessing high level portrait features in crowdsourcing

We discuss here about high level feature evaluation for each picture in our subset of portraits.

High level features are partly related to low level ones, however this is a very complicated task; how to express high level factors adopting appropriate low level

⁴We focus only on upper body considering that our portraits do not focus on full body poses (ref. portrait definition in chap 1)

⁵While technically it would be possible to photo retouch face traits, we do not consider these cases as they won't represent the original subject anymore. However, this possibility depends on the practical application of this research.

⁶<https://www.flickr.com/photos/x1brett/13942900580/in/album-72157644568241053/>



Figure 6.2.5: Two similar portraits of the same subject. The main difference is the profile side, left or right, and a small face inclination. Psychology findings suggest that different brain areas are activated, possibly eliciting different perceptions. SOURCE: Flickr.com, Brett Jordan, CC license⁶

LOW LEVEL		FEATURE	VALUE
	A	Aspect ratio	continuous value
	B	Resolution	
	C	Hue mean	
	D	Hue standard deviation	
	E	Saturation mean	
	F	Saturation standard deviation	
	G	Value mean	
	H	Value standard deviation	
	I	Image contrast	

Table 6.3: Low level features adopted.

HIGH LEVEL (I)		FEATURE	VALUE
	J	face size	More than chest Chest Close-Up Only the face (or so)
	K	face orientation	Profile Partially Frontal
	L	scene setting	Indoor Outdoor Can't say
	M	beard	Yes, long Only mustache and/or goatee Yes, short Just a shadow Not at all
	N	dress typology	Business suit / Formal dress Normal shirt T-shirt / Not formal dress Can't see
	O	glasses	Yes, sunglasses Yes, normal eye glasses No Can't say
	P	smile	Yes, showing teeth Yes, a little No
	Q	gaze	At the left At the right At the camera
	R	background type	A neutral background (white,black,...) City (a street,...) An office A wall / a room Nature (park, natural landscape,...) Don't know

Table 6.4: High level features adopted, with possible values aside.

HIGH LEVEL (II)	S	subject gender	Male Female
	T	face tilt	Yes, toward left Yes, toward right No
	U	background blurring	Yes, a lot Yes, a little No
	V	eyes opened	Yes, widely Yes, a little No

Table 6.5: High level features adopted (continued).

features can be very difficult [Tong 05]. Some authors even remark that «low-level contents cannot always describe the high level semantic concepts in the user’s mind» [Zhou 00]. Some high level features can actually be assessed automatically (i.e. background blur), but computer vision based high level features extraction can be time consuming and error prone. As first step, we then preferred to adopt a different methodology for all of them, in order to focus on features’ importance: we adopted crowdsourcing to manually label them. The advantage of this approach is twofold: (1) it greatly speeds up the process and avoids errors, and (2) it offers a subjective opinion for some cognitive features that might be perceived differently between people (e.g., where a portrait has been shot). For this purpose, we followed the same crowdsourcing strategy adopted in previous chapter, changing only the task for the participants. They were provided with a web interface to evaluate each feature for each picture: figure 6.2.6 shows the upper part of the interface as seen by participants. As done by [Lintott 08], we added visual icons as guidelines near answer options to simplify the understanding of each proposed value for some features in our labeling (e.g., subject profile side). We also gave the possibility to answer ‘don’t know’ for some features when in doubt. Within a week, 745 participants from all over the world participated to the experiment, each one evaluating 25 portraits.

The mutual exclusion of some high level features values were used as honeypots to detect outliers. In particular we adopted the depicted person’s gender and beard presence for obvious reasons. Around 18% of participants failed in this honeypot and have been excluded. A second honeypot we designed, the mutual exclusion of scene setting between outdoor and office/room as background, outlined around 40 % of participants as outliers. This result is in line with our previous crowdsourcing experiment. Jointly, the two honeypots outlined around half of participants to be

outliers⁷. However, all portraits present a sufficient number of evaluations for the high level features in order to continue the analysis.

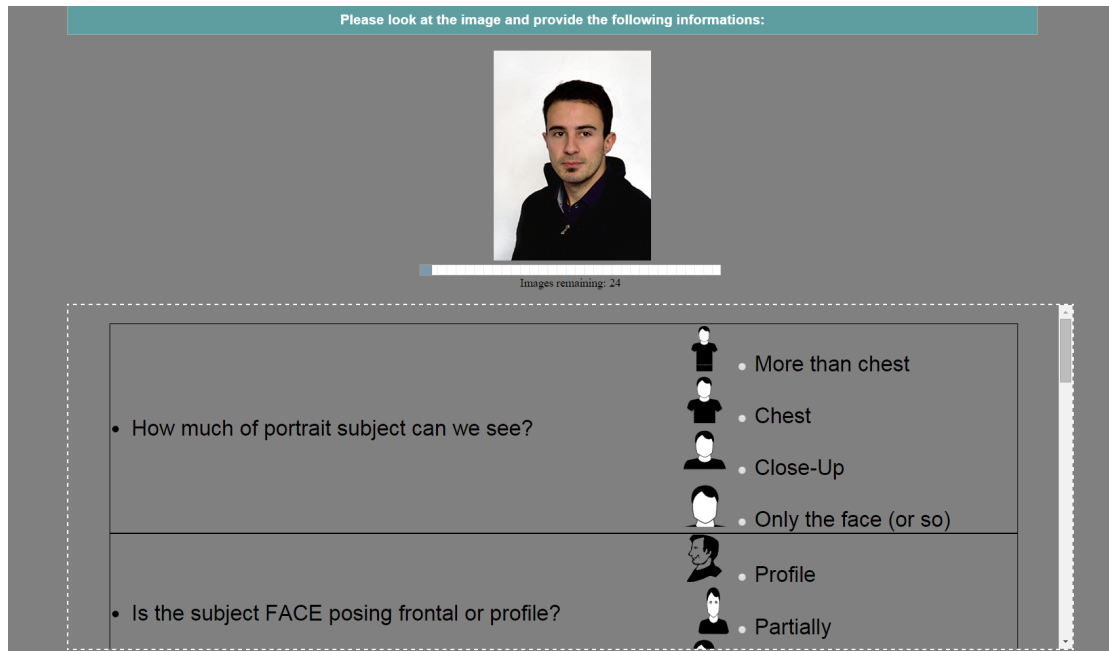


Figure 6.2.6: First page of the interface adopted in crowdsourcing for evaluating high level features of portraits.

6.2.3.1 Challenges with high level features and analysis of uncertainty

We underlined that high level features assessment is a complicated subject. As said, high level features assessment is a challenging computer vision task. Moreover, the problem lies in the inherent subjectivity of scene cognition: some high level features may be subjective even for people. This problem has been already noticed in research related to Content Based Image Retrieval: keywords can be adopted to map toward high-level semantics, but the mapping can be tricky. As stated by Zhou in [Zhou 00], «people use the same word for different meanings in different context, or use different words for similar or even the same concepts».

To check how much the subjectivity influences our high level features assessment, we measured the level of agreement between assessments given by participants. In this analysis we do not take into account outliers, that we consider removed at this point. For this purpose we computed the percentage of participants that agree on evaluations for each feature, for each portrait. For example, a feature evaluated

⁷Some participants failed both honeypots, and the two groups partially overlap.

as '1' by 10 participants, '2' by 5 and '3' by other 5 participants, will show an agreement of 10/20. While we admit that other more rigorous statistical test could have been carried out, i.e. Cohen's kappa or a Fisher test between actual distributions and distributions by chance, we underline that we just wanted a simple measure that can underline features that can be doubtful, either in general or for some portraits only. Moreover this measure can also be introduced in our crowdsourcing framework for real time checks. The average level of agreement underlines that while some subjectivity is present, participants usually agree on all of our features. Every feature presents an agreement of more than 60% of our participants (figure 6.2.7); clear features, i.e. related to gender, show a very high agreement as expected.

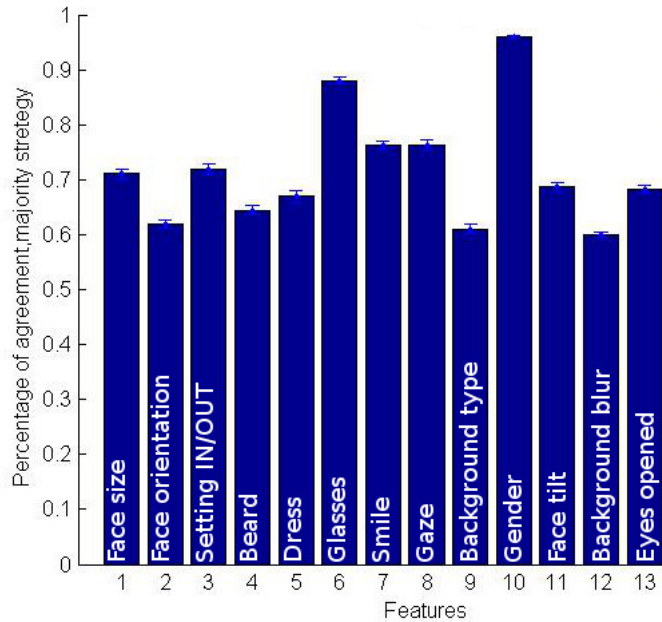


Figure 6.2.7: Graph bar of observers agreement for high level features, in percent.

Participants opinions indeed diverge assessing some features for few portraits. In figure 6.2.8 we show some evaluated portraits that underline different opinions between observers. In particular, these underline how some features are inherently subjective, and slightly different opinions between participants are normal. For example, an office in the background may not be perceived as such by some observers, that may think it is a normal room. This fact will probably change the perception of the social context for the portrait itself. With the approach that we adopted we cannot check this fact directly, unfortunately: people that evaluated

the high level features are not the same that evaluated portraits social context.⁸

To sum up, we want to stress that the interface design and the instructions given for the crowdsourcing assessment have a great impact on the expression of features subjectivity. In fact, precise instructions with examples and design interface with icons near the answers will force participants to comply with high level feature examples - reflecting only our subjective opinion. If this is what we want, depends on the task.

It must be noted that a different opinion on high level features, that is to say a different understanding of depicted scene in a portrait, can give a different opinion of social context. To this extent, we investigated statistically if the uncertainty of social context assessments is related to the uncertainty on high level features labeling. The uncertainty - for both quantities - has been measured as the standard deviation of the assessments; a bigger deviation corresponds to a lower agreement between observers. We then computed, for each image on our data set, the uncertainty for both the social context and each high level feature. To find if a consistent relation exists, we fitted a linear⁹ model with the social context uncertainty as dependent variable. This model has been chosen to find if a relation exists and also the contribution of each feature¹⁰. Obtained model shows a good fit on data (deviance of fit $\sigma = 1.06$), underlining a dependency between the uncertainties. Fitted model regression coefficients are shown in figure 6.2.9. Associated F statistics underlined some features as statistically significant in the model ($p - values < 0.05$), as portrait depicted subject face orientation, head tilt and beard level (red crosses on figure 6.2.9). While this finding does not imply a cause¹¹, it underlines that empirically a relation between these quantities exists; reading our results, i.e. a bigger uncertainty on portrait background (features L - U) leads to a bigger uncertainty on social context.

6.3 Analysis of influential features in portraits social context perception

This section deals with the mathematical approaches to infer the contribution of each portrait feature on social context perception. Once again, subjective perception has been collected as described in previous chapter, while features have been described in previous section.

⁸Otherwise, we could have run the analysis considering the participant factor.

⁹Scatter plots between each feature and social context did not underline non-linear correlations.

¹⁰Model assumptions have been checked; small deviations from normality are present. However, they have minor consequences on these models [Ramsey 02].

¹¹Our study is an observational study, with not every possible factor under control.

	<p>Beard: shaved or just a shadow? Opinions diverge: 34% vs 42% (remaining 24%: outliers).</p> <p>Simon Bostock, in Selfies, www.flickr.com/photos/bfchirpy/10560642264/in/photostream/</p>
	<p>Face: frontal or profile? Opinions split: 28% frontal, 38% half profile, 34% full profile.</p> <p>Surya Teja, in NaG, www.flickr.com/photos/suryateja/14045039949/</p>
	<p>Background interpretation: 40% of participants reported it to be nature, 19% to be a street. Around 25% of participants skipped this question (remaining 16%: different answers).</p> <p>Vivian Farinazzo, in Thales (Sonic Dash), www.flickr.com/photos/lifeissimpleinthemoonlight/12055876983/</p>
	<p>Background interpretation: 38% of participants reported it to be an office, 23% reported to be a wall. Around 33% skipped this question (6% outliers).</p> <p>Forgemind ArchiMedia, in Webuse 0001, www.flickr.com/photos/eager/14249857795/in/photostream/</p>

Figure 6.2.8: Examples of high level features subjectivity in our data set. Outliers have been removed when not mentioned. Images from Flickr, Creative Commons licenses.

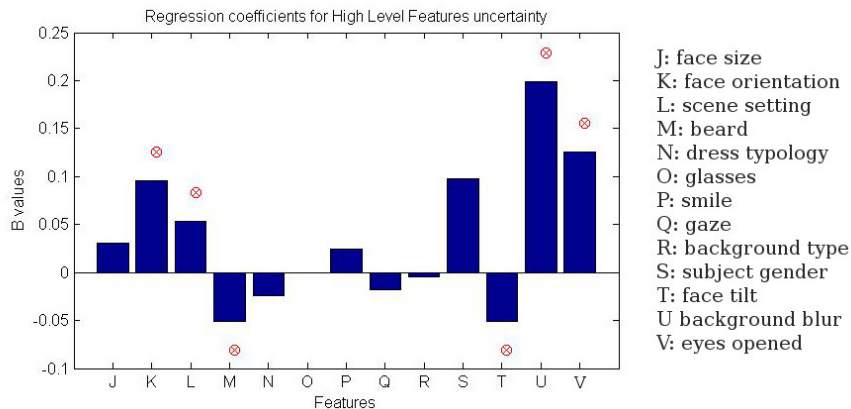


Figure 6.2.9: Bar plot of fitted model coefficients, related to high level features uncertainties. Letters follow tables 6.4 and 6.5. Red crosses indicate statistically significant regressors.

Different approaches may be adopted to learn a model based on input features from subjective assessments. We can roughly divide the approaches in two big groups: white and black box. Black box ones are probably the most frequently adopted in current research dealing with image analysis; widely adopted examples are Neural Networks (NNs) and Support Vector Machines (SVMs). With black box methods the mathematical functions linking inputs / outputs are not easily explainable as very complex interactions usually occurs. For example, it would be extremely hard to explain mathematically how a regressor influences a particular result in a fully connected NN, due to the large number of links between nodes. The same holds for SVMs, as data is projected on higher dimensional spaces. However these methods are able to model complex non linear interactions between features and so they are considered more powerful than simpler linear methods. For this reason they usually provide better results (i.e. higher classification accuracy). With the term white box approaches we refer instead to much simpler methods such as General Linear Models or Linear and Logistic Regression. In these cases from the model we obtain a regression function that directly links dependent and independent variables. It is important to underline that many of our features are categorical variables - as head tilt and orientation (left,center,right) or scene setting (don't know/indoor/outdoor) - and some of these are ordinal - as subject size (from small to big). This point limits the kind of possible analysis as not all independent variables are normally distributed.

In our research, we adopted both kinds of approach. We are interested in black box ones to see which results can we obtain adopting more powerful models, and we are interested in white ones to better explain obtained models. These ap-

proaches are described in dedicated sections below. Looking at current literature, we adopted Neural Networks, SVMs, Logistic Regression and Decision Trees. These models can also be adopted to classify data samples. We then measured classification accuracy as a quantitative measure of model fit, splitting our dataset in 75% for training and the remaining for testing. For every classifier we adopted a leave-one-out (LOO) cross validation, due to the small size of our dataset, providing mean classification accuracy and confidence intervals. For confidence intervals, being our model predictions discrete¹², we did similarly to [Jiang 08] considering a Binomial distribution but adopting a Wald method.

The ground truth for the analysis is given by collected portrait social context subjective assessments. We can use this data in multiple ways, as each portrait has been ranked between the three possible contexts. As said in 5.5.1, a common method is adopting a majority rule on choices; we adopted this strategy too to obtain discrete labels. However, as we explained in 5.5, many portraits do not present a clear context choice, i.e. when a portrait has been evaluated 50% of times for working purposes and 50% for friends. To take into account this fact we added other three classes to the ground truth - social context «mixtures»: Friends/Work, Work/Romantic and Romantic/Friends - to better model our data set. We then put in these classes portraits labeled in between two contexts and that do not present a statistically significant choice as explained in 5.5.3. We refer to the two different class labeling strategies - considering or not context mixtures - as standard and detailed strategies respectively.

We used a majority strategy also to define the value of the high level features for each portrait picture. While high level features subjectivity may be valuable for our study, we prefer to avoid this additional uncertainty at the moment, and leave raw feature evaluations for future works. Still, methods to reduce uncertainty improving the a majority strategy exist (i.e. [Peng 13]). However these cannot be adopted in our case, as they rely either on evaluators assessment (i.e. based on his previous job) or on stimuli ground truth; our experiment did not include any of these elements. To conclude data pre-processing, all features were normalized to the same mean and range, between 0 and 1, so that analysis results (i.e. computed coefficients in models) could reflect actual relevance weights. For this purpose we used the common formula:

$$Y = \frac{X - \min(X)}{\max(X) - \min(X)};$$

where Y is the normalized data, X is the original data vector.

¹²We consider here the prediction of the social context of a portrait, that can either be correct or not. This outcome does not follow a Normal Distribution.

6.3.1 Black box approaches: Neural Networks and Support Vector Machines

We discuss here the adoption of black-box approaches, more complex non-linear models. Based on current research, previously reviewed, we opted for Neural Networks and Support Vector Machines.

6.3.1.1 Neural Network

We decide to adopt a simple Neural Networks to address the problem, due to their large adoption in literature for classification problems. We then approach the problem as a portrait classification between the three context classes. Our inputs are images features, normalized between $[-1; 1]$. We tested different network sizes, obtaining the highest classification accuracy of 65% (c.i. 6%) with one hidden layer of 150 neurons. Results are higher than chance (three classes imply a 33% accuracy of random choice), however this approach does not allow an easy interpretation of each feature contribution. While different feature selection strategies have been proposed in literature [Leray 98], a first approach to assess features contribution by selectively remove them from inputs and check variations of classification accuracy. We adopted this strategy with proposed NN; even if some differences between features are present, no feature per se has been underlined as critical in the classification. Results are shown in figure 6.3.1. Neural Networks approach has been adopted following also detailed strategy, considering mixture contexts. Different network sizes have been tested also in this case, maintaining same network topology. A classification accuracy better than chance, 41% (c.i. 4.4%), has been obtained adopting double the quantity of neurons (300 instead of 150); this result seems to underline that the detailed strategy adds complexity instead of simplifying the classification task.

6.3.1.2 Support Vector Machines

Another very popular approach within black box like machine learning approaches are Support Vector Machines (SVMs). As underlined in [Mohammadi 12], logistic regression might result into low accuracies and the experiment can be completed by using a SVM. Basically, SVMs are binary classifiers that separate data points after projecting them in a higher dimensional space through a kernel function. Again many positive examples in image analysis literature are present, as in previously cited Datta's work[Datta 06] or in [Dhar 11]. In our case however data belongs to more than one class, as a portrait belongs either to Friends, Work or Dating purposes. We then adopted a multiclassSVM, an extension of SVM reducing the multiclass problem to multiple binary problems. A C -Support Vector Classifica-

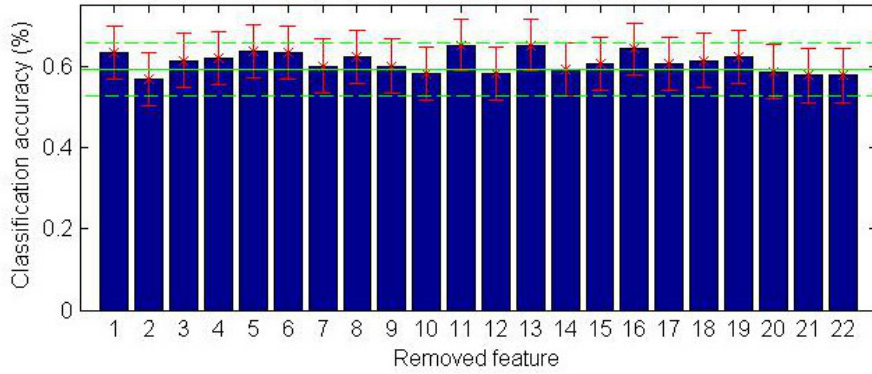


Figure 6.3.1: Neural network classification accuracy (μ and c.i.), removing one feature at a time. In green accuracy obtained considering all features.

tion model has been adopted using the popular library LIBSVM [Chang 11]. We started the analysis following the standard strategy (ref. previous section) and tested different SVM parameters; the best classification accuracy (70% , c.i. 6%) has been obtained adopting a radial basis kernel and a regularization parameter C of 1. We also tested a linear kernel for our SVM to check if radial kernel nonlinear approach adopted is really an advantage. Classification accuracy dropped slightly (3%), within computed confidence intervals. A Barnard test on classification outcomes of LOO cross-validation between the linear and the radial kernel fails to reject the null hypothesis of difference ($p = 0.27$). Regarding the detailed strategy for our ground truth, we obtained an accuracy of only 43.5% (c.i. 6.6%) adopting a sigmoid kernel. We did not dig further into feature selection for this result.

Anyway, these results come with no explanation of features individual contributions. In order to investigate this point, we run a feature selection algorithm based on [Kohavi 97]. The algorithm creates a relevant subset of features by sequentially add them to an initial selection and learn a different model each step¹³. Based on the classification error on test data, validated with a 10 folds cross-validation, the algorithm selects features to add in order to produce best results. Every fold picked randomly 10 times (MonteCarlo simulation). Not considering different initial conditions that slightly modify performances, final features included by algorithm were four: image resolution, subject dress, presence of glasses and background blur. With only these features, we achieve around 74% percent of accuracy (c.i. 6% with LOO). In order to understand which features are discriminant for the two classes Work and Dating, we repeated the same analysis adopting a one

¹³It is also possible to proceed backwards, removing features, but it's unfeasible to add and remove in the same optimization: to consider all the 2^n subsets of n features would be unfeasible.

VS all strategy, merging classes together: once Friends+Dating vs Work pictures, once Friends+Work vs Dating pictures. For these two cases, we then repeated the same feature selection algorithm as before. For work purposes the results underlines that the «dress» is an important feature . For dating purposes, instead the algorithm underlines dress and glass presence as important features.

6.3.1.3 Discussion

Proposed approaches provide satisfactory results - remarkably higher than chance - even if accuracy can be improved. This outcome can be due to multiple causes, as for example high level features uncertainty as well as the lack of other explanatory variables - that we did not take into account in this work. Surprisingly the detailed strategy, that we believed to better model the ground truth, performed worse in both adopted approaches. The need of more neurons in our Neural Network for this strategy underlines its higher complexity. This result may be due to the fact that examples are not enough and evenly distributed between the six classes to fit the models: many portraits present no significant clear choice and then mixtures classes have more samples than the others. Anyway we take these results as a reference for successive analysis related to white box approaches, topic of next section.

6.3.2 White box approaches

In this section we discuss white box methods adopted: the Logistic Regression and the Decision Tree. Obtained results (in terms of classification accuracy) are slightly lower compared to black box approaches, however they are useful to explain regressors contribution.

6.3.2.1 Logistic Regression

For the analysis of influential factors, we adopted a Logistic Regression, using our features as regressors and the context ranks as observations to fit. We want to underline that we cannot use a simple linear regression, as done before for uncertainties analysis (ref. 6.2.3.1), because the dependent variable is not a ratio variable (i.e. a real continuous value) but a nominal variable instead (a social context class either 1,2 or 3). Since the dependent variable had three possible discrete outcomes, we used a Multivariate Logistic Regression. The model, in case of a single observation can be written as:

$$y_n = \beta_0 + \sum_{k=1}^K x_{nk}\beta_k + \epsilon_n \quad (6.3.1)$$

where y is the dependent variable - the context rank for a particular category - x are our predictor values, β are the coefficients to be estimated and ϵ indicates the error term. Even if we expect our model to be more complex than linear, we adopted this method to have interpretable outputs for our features. We fitted three independent models, one for each context, adopting the ranks of that context for all portrait pictures as responses. The results of the logistic regression fit are shown in Figure 6.3.2.1. The Friends context was selected as reference in our model. This means that each computed coefficient expresses the expected influence of a feature on the relative chance that a portrait picture is perceived in another context (i.e., Work or Dating) than the reference, where this chance is expressed in log odds. So for example, an increment of one unit in the feature "Dress" increases β_k times the relative log odds of a portrait picture being perceived as Work context, where β_k is the coefficient related to the feature "Dress" in the model of the Work context - assuming everything else being equal¹⁴.

With these coefficients we can compute the probability of a portrait to belong to a context, given its features. This probability can be used to classify new data. Our model achieved a classification accuracy of 66% (c.i. 6.3%) on our data set following the standard strategy previously mentioned. For the sake of clarity, we underline that we should compare this result with pure chance, picking one context over three possible (33%). Adopting the detailed strategy with our ground truth, model regression did not fully converge. This is due to the low number of samples for the classes in the detailed strategy. As result, its accuracy did not exceed 40% (c.i. 6.5%). We will not analyze further this case in successive steps.

Logistic Regression also provides a measure of the statistical influence of each feature in the model via their p-value. These p-values are shown in table 6.6. They illustrate that the prediction model for a portrait picture having a Work context is significantly affected by the dress (N) that the person in the picture wears, as well as by the portrait setting (L) and by the low level feature of mean saturation (E) in the picture. As expected, a formal dress increases the appropriateness of the portrait for working purposes, while instead an increase in saturation decreases it. For predicting the appropriateness of a portrait for dating purposes instead the gender of the portrayed person, and his/her face orientation are statistically significant contributors. In particular portraits depicting female persons are reported as more appropriate for dating purposes.

Logistic Regression allows to express these results as a formula, linking addressed features with the relative probability of a portrait being for work or dating more than for friends. Considering only most important features, we have:

$$\ln\left(\frac{P(Work)}{P(Friends)}\right) \approx \beta_{0W} - 1.8X_{\mu Sat} + 1.1X_{Scene IN/OUT} + 2.6X_{Dress}$$

¹⁴Assuming that this would be possible.

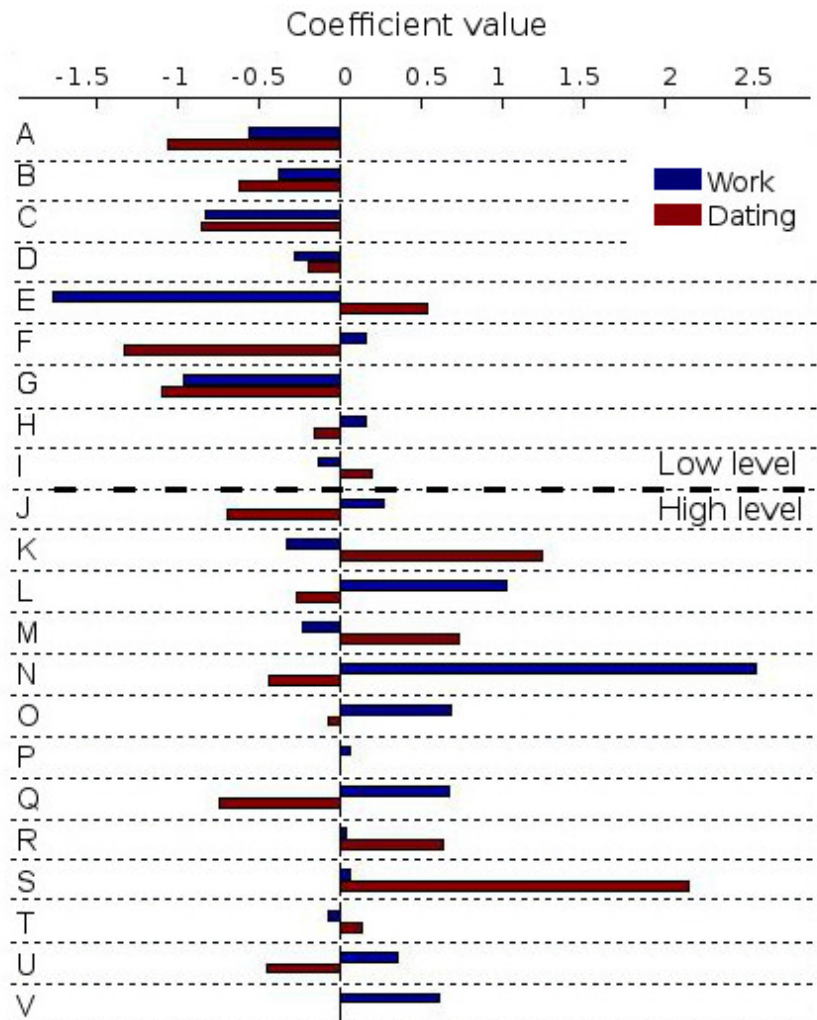


Figure 6.3.2: Features regression coefficients for first choice. Friend context has been taken as reference. Features, here in capital letters, are encoded following table 6.6.

$$\ln\left(\frac{P(Dating)}{P(Friends)}\right) \approx \beta_{0D} + 1.4X_{Face\ Orientation} + 2.1X_{Gender}$$

In the equations β_{0W} and β_{0D} are the constant values from our model, for Work and Dating respectively, while X are actual features values. However, it is worth to underline that it is unwise to consider these models as *exact* regression equations describing the phenomenon [Ramsey 02]; there is always some uncertainty, many models are possible and ours are the ones expressing expected variations of dependent variables considering *only* the independent variables we included.

As we can see in the second equation, β coefficient related to depicted portrait subject gender is relatively important. Moreover, as underlined previous chapter (section 5.5.4), in respect to this finding we analyzed if this result is influenced also by the gender of the evaluators. That analysis underlined that indeed this factor has a significant influence in portrait perception for dating purposes. This finding strongly underline that different models should be built for men and women. However, we cannot made this distinction in this research as too few female subjects actually participated the test (ref. 5.5.1). This effect should be taken into account while designing future works.

6.3.2.2 Decision Tree

Another possible approach is to adopt a decision tree modeling. This approach has already been employed in image aesthetics assessments, obtaining positive results adopting low level image features[Datta 06]. The model builds a tree of decisions to classify the input data. Leaves are classification outcomes (i.e. responses to inputs), and each input corresponds to a path on the tree, starting from the root. Each internal node is labeled with a feature, on which successive splits are made (i.e. decision). Originated branches differentiate by the value on this feature. The learning algorithm then tries to infer the best features and thresholds to construct the tree. At each step, the algorithm examines all possible data splits for every predictor variable, applying the one that maximizes a certain criterion. In our case we adopted binary splits as in mentioned reference [Datta 06] . Our stopping criterion is instead the requirement that all leaves must correspond to a class observation. We fitted two decision trees, one for each class labeling strategy. For the majority strategy this approach gives on our data set a 58% classification accuracy (c.i. 6.5%). Obtained decision tree is visualized in figure 6.3.3. Dress and gender are outlined between the first discriminative features (top roots in the obtained tree). For the detailed strategy accuracy significantly decreases as for logistic regression, achieving only 34% (c.i. 6%).

Feature	Work	Dating
A: Aspect ratio		
B: Resolution		
C: Hue μ		
D: Hue σ		
E: Saturation μ	*	
F: Saturation σ		
G: Value μ		
H: Value σ		
I: Contrast		
J: Face size		
K: Face orientation		**
L: Indoor/outdoor	**	
M: Beard		
N: Dress	***	
O: Glasses		
P: Smile		
Q: Sight		
R: Background		
S: Gender		***
T: Head tilt		
U: Background blur		
V: Eyes opened		

Table 6.6: Features significance with Logistic Regression.*= $p < 0.05$, **= $p < 0.01$, ***= $p < 10^{-3}$

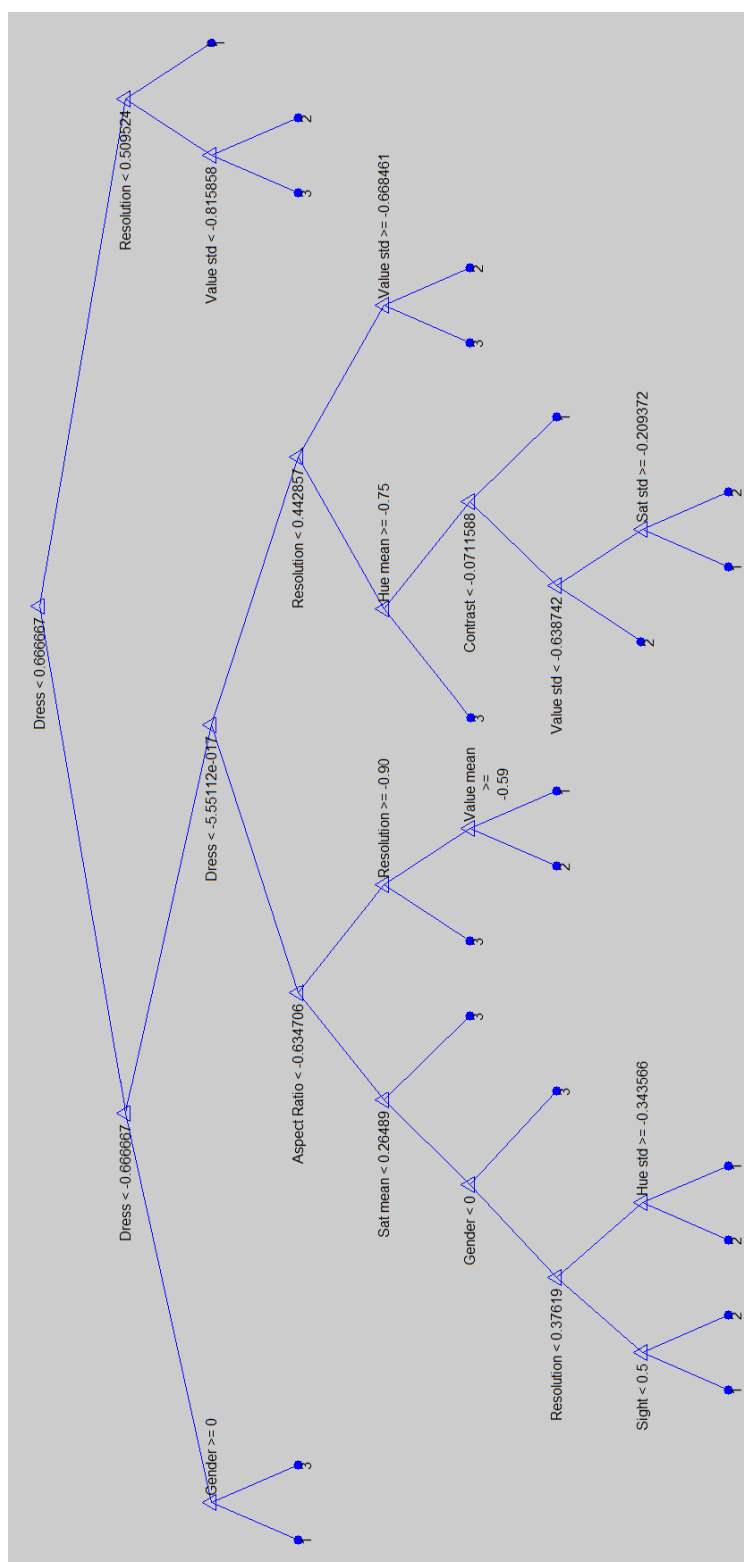


Figure 6.3.3: Trained decision tree. It can be seen that 'Dress' and 'Gender' features are important features influencing portraits classification. Social context are encoded as 1,2 and 3 being respectively Friends, Work and Dating.

6.3.3 Discussion

Results indicate that the clothes of portrait subject significantly affect the perception of the picture to fit working purposes. This result is perfectly in line with expectations coming from empirical experience and supports the validity of our analysis: a formal dress increases the appropriateness of a portrait for working purposes. This finding is confirmed by black approaches results too. Also the portrait setting (L) and a lower mean value for saturation (E) in the picture contribute to increase this fit. Viceversa, an higher saturation increases slightly the chances of a portrait to be perceived for dating purposes more than for friends. Empirically, we can probably explain this outcome with a «warmer» perception of the picture. More important, portrayed subject face orientation and gender are statistically significant contributors for this fit, not considering for now the evaluator gender bias as explained in previous chapter 5.5.4. We are indeed surprised that background scene perception has not been strongly underlined as important, especially after qualitative assessment shown in 5.5.2, but only background interpretation indoor / outdoor marginally influences context perception. We are not able to say if test power is insufficient to outline an effect (i.e. not enough examples for each scene setting) or other causes are influencing the result (i.e. noise in evaluations).

Regarding proposed methods, they seem to be equally valuable, but they all have pros and cons. In the end, only the logistic regression gave numeric results considering all the features at once. This model offers many advantages: first of all results are interpretable, as we link directly each feature contribution on context probabilities. Moreover coefficients give us a quantitative measure of each contribution. Secondly, we can also compute a p-value for each feature in the model, indicating the statistical influence of each regressor. Lastly, the model is relatively easy and computationally inexpensive. Decision Tree modeling too provides interpretable results and is computationally light, but it does not provide direct statistical measures of each feature importance. In the end, this lower complexity comes also with the price of considering only linear relationships, in both models.

We expected black boxes models to be sensibly more accurate, but no big difference in results has been underlined in terms of classification accuracy with trained models. However we tested different models and parameters. This finding may point out that our set of independent variables model relatively well our problem and that class are linearly separable in this feature space.

Method	Standard Strategy	Detailed Strategy
Neural Network	$65 \pm 6\%$	$41 \pm 4.4\%$
SVM	$70 \pm 6\%$	$43 \pm 6.6\%$
Logistic Regression	$66 \pm 6.3\%$	$40 \pm 6.5\%$
Decision Tree	$58 \pm 6.5\%$	$34 \pm 6\%$

Figure 6.3.4: Table summarizing classification accuracy for the different methods.

6.4 Conclusion

In this chapter we addressed the problem of determining influential factors from gathered perceived context of portrait pictures. Our analysis considered some classical low level features as well as higher level features related to image interpretation. To chose a subset of features to assess we addressed current literature both in our field and in social sciences, as social context assessment is greatly influenced by cognitive biases. It is clear that considering all possible factors of influence in only one analysis is a challenging task. However, isolating them to determine each feature's separate impact is at least as hard and time consuming. We approached the problem considering features we supposed mostly important for our purpose, basing our choice on current literature. We also showed how multiple features can be addressed simultaneously: to this extent the large amount of crowdsourcing social context evaluations, gathered as said in previous chapter, was very useful. We adopted crowdsourcing not only to determine the appropriateness of a portrait picture for a given context, but also to evaluate the high level features. In fact, extracting these only with computer vision tools can be error prone. To this extent we adapted our crowdsourcing framework and run a crowdsourcing campaign to label high level features of all portraits. Low level features, much easier to be addressed, have been computed instead with computer vision. We used a white box approaches as well as a black box ones for the statistical analysis. Even if we expected the former less accurate than the latter, they allows easier interpretation of the results. In particular, the Logistic Regression model offers three important advantages. First of all results are interpretable, as we link directly each feature contribution on context probabilities. Secondly, computed coefficients give us a quantitative measure of each contribution. Lastly, the model is relatively easy and computationally inexpensive.

While the proposed analysis is not able to explain all variability in the subjective assessments, some statistically relevant features have been underlined and greatly helped to discriminate between portrait contexts. Qualitative results are the same for all methods: cloth of the portrayed person is shown to be discriminative for

the likelihood of a portrait to be perceived as work related. Also the gender of the portrayed person appeared to be influential, more particularly for the likelihood of a portrait to be perceived as dating related. The latter, however, was also dependent on the gender of the participant as said in previous chapter; this element should be taken into account in future works. Our first results are then consistent with expectations from empirical experience. Interestingly the background interpretation was not as influential as we expected: background interpretation only contributed very little to our model. With these results in mind, we will focus on computer vision efforts on features underlined as influential.

To conclude, many improvements can be considered. First of all a broader scale analysis should be carried out, both in terms of stimuli number and variety, to carry out more accurate statistical analysis. Secondly, other features that we underlined in our literature review could be included in the analysis, still assessing them in crowdsourcing. However more evaluations will probably be needed to conduct such an analysis with many factors. Lastly, it would be interesting to consider different models for different demographic groups, in particular for different cultures in the world (i.e. one model for Europe and USA, another for ASIA).

Keypoints

Context

- ❑ While subjective opinions have already been correlated to image features for different research purposes (i.e. for quality assessment, for aesthetics or memorability) this has never been done for linking portrait features to social context perception.
- ❑ Literature underlined that only low level features are not enough to model high level concepts in multimedia.
- ❑ High level features have been proposed; some of them are related to social sciences findings.

Contributions

- ❑ We reviewed low and high level features adopted; we selected a subset of features underlined as important considering findings in image analysis research as well as in psychology.
- ❑ We underlined statistically the importance of some image features, especially high level ones (i.e. dress, portrait setting, gender), in predicting the social context; results are consistent with empirical experience.
- ❑ We statistically underlined: (i) that the uncertainty of portrait social context is also partially dependent on the uncertainty of some high level features; (ii) that the importance of portrayed subject gender is also dependent on evaluators gender.

“The miracle isn’t that I finished. The miracle is that I had the courage to start.”

(John Bingham)

Chapter 7

Conclusions and perspectives

In this section we outline our latest ongoing work and we recall the main objectives, briefly summarizing our contributions. In the end, we discuss limitations and possible improvements.

7.1 Ongoing work with Computer Vision tools to automate portrait analysis

The objective of ongoing work is the automatic extraction of high level features, focusing on those underlined as significant by our analysis (see chapter 6, sec. 6.3). The purpose of this evaluation is double; first, it greatly reduces the amount of work needed to label portraits¹. The second purpose is practical, as an automatic extraction would allow to build an automatic portrait evaluation system, based on obtained model. Such a system could be a software in which users upload their portrait and get the suggested social context for it. As said in previous chapter, to evaluate reliably high level features with computer vision is a challenging task. Next paragraph outlines the SoA on this topic, while in sec. 7.1.2 we describe their implementation in practice. With these automatically evaluated features, and our model, we have been able to build an automatic portrait evaluation system in the form of a web page. This system is still under development and it is not publicly accessible.

¹High level features have been manually evaluated for each portrait, and a larger amount of samples would allow to build a better model.

7.1.1 High level features assessment with computer vision

Research on computer vision addressed the evaluation of some high level features that we underlined as important. Face presence, position and size respect to the picture are nowadays considered reliable measures. These are mainly provided from Viola Jones algorithms and its improvements [Viola]. Mentioned features have already been adopted for aesthetic assessments, as in [Li 10a], where authors also estimate pose and mouth expression, applying a particular filtering. Interestingly, Viola Jones algorithms can be used to detect any object, like i.e. glasses. Instead, more complex methods are applied to detect facial landmarks; authors of [Sun 13] use deep convolutional neural networks cascades, obtaining good results also with pictures showing face profiles. Gender and age have been addressed too, and results are quite reliable. Regarding gender, a well known technique is machine learning with Principal Component Analysis ([Valentin 97]), that produces eigenfaces. This allows to classify male and female faces with high accuracy ($> 95\%$, [Cheng 08]). Age predictors are instead approached with many different strategies, and recent research achieved prediction errors of about 5 years ([Fu 10]). Clothing parsing is a more complex task. Current approaches use a large number of features (SURF, HOG, PHOW, LBP), usually extracted from regions of interest, and construct bags of features descriptors. These then feed black-box machine learning algorithms as SVMs and Random Forests as done in [Di 13] and [Bossard 13]. While classification accuracy varies sensibly with the number of considered classes, results are encouraging: the average accuracy is around 35% with 15 classes. Conditional Random Fields are used instead by [Yamaguchi 12] and [Chen 12]. Research focused too on image features considering the scene. Again, large number of SIFT features have been adopted to recognize scene type and attempt outdoor/indoor classification, obtaining interesting results ([Dhar 11], [Lazebnik 06]).

Some software frameworks have been developed for the assessment of face-related high level features. OpenCV² implements Viola Jones algorithm and eigenfaces classification; however reliability greatly vary from data set to data set. This is especially true when portraits are taken 'in the wild' and not in controlled conditions. Commercial tools, collecting multiple features' evaluation at once, have been developed and used in scientific research. This is the case of Face++³, BetaFace⁴ and SkyBiometry⁵ (ref. [Lienhard 15a]).

²Retrievable at <http://opencv.org/>, Sept 2015

³<http://www.faceplusplus.com/>

⁴Betaface Advanced face recognition, www.betaface.com

⁵SkyBiometry, Cloud-based Face Detection and Recognition API, www.skybiometry.com/

7.1.2 Practical implementation

To implement high level features detection, we looked for ready to use algorithms implementations. However, this process still takes a considerable amount of time, as to adapt them to our purposes (i.e. adopted coding, image format, size, ...) is not straightforward. Moreover, often the practical details are usually omitted in research publications when describing the algorithms (i.e. machine learning tuning and image pre-processing details).

To evaluate face related features we opted for Face++ software, due to its good performances and because is free of charge for non-commercial use. This software evaluates face presence, gender, race, smile and detects glasses. The algorithm outputs also face roll, pitch and yaw angles; this information can be used for the face orientation feature. However, we cannot use these latter directly as our model considers discrete feature values (i.e. right or left profile, instead of angles). An example of results is provided in figure 7.1.1. Regarding the clothing category,

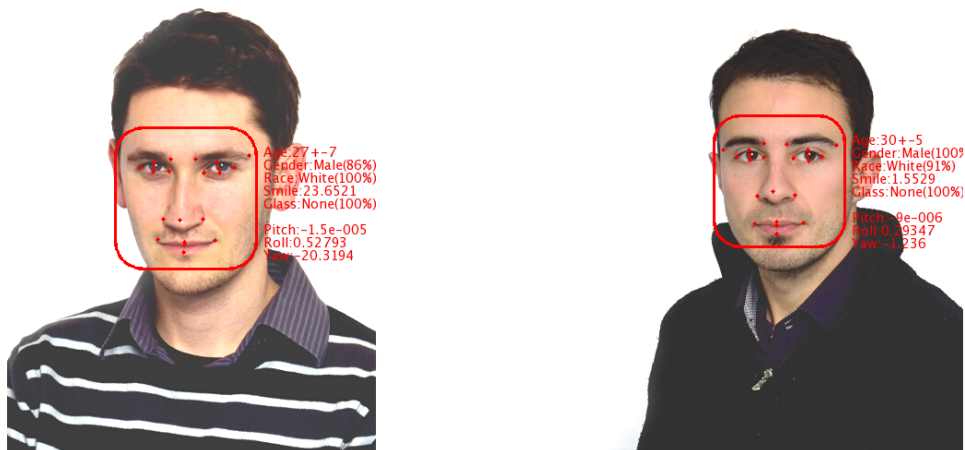


Figure 7.1.1: Two portraits from our first data set, automatically tagged by the BetaFace software.

we did not find any publicly released ready-to-use algorithm focused on upper garment clothing (as we considered in our analysis, see chap. 6). While method of [Yamaguchi 12] is available, it works only on full body pictures. For this reason, we implemented features extraction as in [Bossard 13], but adding skin detection. Results have been tested on the smaller data set provided by [Chen 12], providing

7 cloth categories⁶. A classification accuracy similar to the original research work has been obtained (75%, 10-folds cross validation).

Face++ is accessible as a web service too, allowing to evaluate portraits through the web. This feature allows us to easily deploy the prediction model as a web page. We then implemented a simple interface on our server to upload a portrait and evaluate its social context, following our model. Low level features are evaluated with the same scripts as in previous chapter. Unfortunately, we are still missing some important features to reliably output the predicted social context: clothing and scene type underlined in section 6.3.3. Moreover, Face++ smile feature and face angles are given as continuous variables⁷. In order to use these features we have to discretize obtained values; while we can make this decision based on the mean values, this choice adds other uncertainty in the evaluation. We remind that these inputs are discrete variables in our model (profile or frontal-face, smile presence).

Described work is still a draft and not yet publicly released; model runs on a local server. An example of the result is shown in figure 7.1.2. In particular, at the moment we are working on maximizing the reliability of features assessments. In fact, features evaluation is not yet accurate enough to make satisfactory the automatic portrait assessment. We cannot make a comparison with manually labeled images to evaluate our system due to the fact that many features are still missing an automatic evaluation. Even if they are not statistically influential per se, the lack of their joint contribution in the model is strongly biasing social context predictions.

7.2 Summary and Contributions

In this thesis we present our study focusing on the social context perception of portrait pictures. The work done is essentially composed of three parts. In the first one we study novel methodologies currently used for multimedia assessments experiments, related to the field of Quality of Experience in particular. Investigated methodologies are two: electrophysiology and crowdsourcing. In the second part, we address the practical adoption of crowdsourcing for collecting portraits and their social context evaluations. In the third part, first we describe the selection and the evaluation of a set of low and high level image features. Then, we evaluate through mathematical analysis the impact of each feature on social context perception.

⁶Shirt, sweater, t-shirt, outerwear, suit, tank top, dress.

⁷Smile feature is computed as a score measuring the strength of this expression.

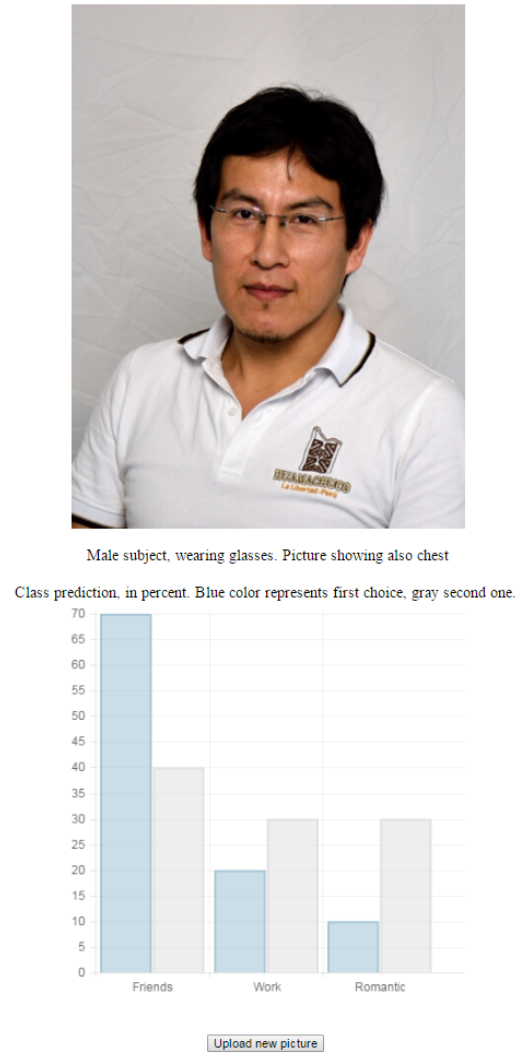


Figure 7.1.2: An example of implemented portrait evaluation web page. Results are computed with off-line model, due to the lack of automatic evaluation on all features.

Find a suitable alternative methodology

Our work first goal was to find a suitable methodology to assess the subjective social context perception. In order to accomplish this task, we investigated current research that is closely related to our study. In this panorama, we focused on Quality of Experience, a research field that encompasses a broader concept of multimedia evaluation; the focus is not only on technical multimedia quality but incorporates also factors related to the context and especially the user. Electrophysiology emotional assessment has been tested at first, as it has been shown to be useful to measure user reactions to stimuli. Some preliminary positive examples are present in research, for some specific cases. To better investigate its usefulness in our case, we preferred to run two pilot studies. These underlined different problems that impact methodology reliability and usefulness (i.e. lack of standard reference procedures, large inter & intra subject variability, ...). More than continuing digging into these problems, we preferred to investigate another methodology adopted in QoE research: crowdsourcing. While useful for different purposes, in scientific research this technique can be seen as an internet outsourcing of normal in-laboratory assessments. Through crowdsourcing, a large amount of cheap subjective assessments can be gathered fast. In particular, compared to normal in-lab studies, it provides a broader audience, richer in terms of demography. However, reliability problems arise, due to multiple cons (i.e. different experiment conditions, instructions misunderstanding, outliers). Nevertheless crowdsourcing has been largely adopted and related problems mitigated. This fact convinced us to adopt it for our research.

Run crowdsourcing in practice

After the decision of adopting crowdsourcing, our goal was to run this technique in practice. Previous SoA review underlined that three main elements are needed: a platform to gather participants, a software framework to run the experiment online and a data set. This last point is detailed in next paragraph. The platform gathers crowdsourcing participants and manage all the needs (i.e. advertise the experiment, pay the participants,...); many commercial services are available for this purpose. We decided to adopt one of the most popular in Europe and within the QoE community, Microworkers. The second point to run the experiment online we need to host a software framework on our servers. Few frameworks have been developed and even less are freely available. These are however mostly focused at simple tasks (i.e. image quality evaluation) and not easily customizable. For these reasons we preferred to deploy a custom solution developing a personal framework. We developed a modular system based on the SoA as well based on our needs. To test chosen crowdsourcing platform and our solution, we then run a pilot study.

We compared the effect of self shot portraits against professional ones on the perception of a job candidate, simulating a resume selection case. Adopted stimuli have been gathered as described in next paragraph. The study underlined that i) adopted methodology is effective and efficient also for our research and that ii) an effect of the portrait is present and statistically measurable.

Find a suitable portrait data set

An additional implicit objective has been considered: our work required also a suitable portrait data set to run tests with. At first we clearly defined to which images we are interested in, as many different types of portraits are possible: we posed our working definition at the beginning of the thesis (ref. 1.3.1). Our contribution brought to the analysis of possible face image sources, based on literature. We considered pros and cons of four main available sources. These are some image databases containing face images adopted in scientific research, personal photo collections and online communities of photo enthusiasts; the fourth alternative is to shot portraits by ourselves. After a deep analysis, we concluded that when few stimuli are needed it is easier to shot portraits in controlled conditions, as we have full control over all the aspects of portraits. Instead, when more portraits are needed and less constraints are imposed on portrait conditions, it is better to use online communities, that offer a rich and varied collection of many portraits and social contexts. However legal constraints must be considered adopting this option as licenses, retrieval methodology and privacy issues may pose problems. With these considerations in mind, we prepared two data sets for our first pilot study and our main experiment; we adopted self shot portraits and online sources respectively.

Gather social context subjective evaluations

Between the main goals of the thesis was to collect a social context subjective evaluations. These constitute for us a ground truth to run the analysis. We achieved this goal adopting crowdsourcing: participants' task was to label portraits stating the best purpose for each image. Possible choices were three social contexts, based on current trends on nowadays social networks: "for friends", "for work" and "for dating". While different and more contexts could be imagined, these represent a good trade-off to start with. Crowdsourcing allowed to easily obtain enough evaluations for our main portrait data set. The majority of our portraits do not present a clear fit for a purpose and opinions are more likely to diverge. Nevertheless, in the following image feature analysis, we found that this uncertainty is partially due to image features uncertainty; this underlines that the scene in portrait may

be perceived differently by different people and this reflects on perceived portrait purpose.

Understanding and modeling image features on social context perception

To fulfill the main goal of this thesis, we looked for portrait features to consider. We looked for i) low level image features adopted in multimedia research related literature but also for ii) high level features related to the semantic of the image. A part of our contribution resides in the review of current literature in the multimedia domain as well as - for the high level ones - of some remarkable results within the psychology domain. We then selected a subset that took into account various low level features, some features related to the subject in the picture that can elicit social biases (i.e. clothing typology, gaze direction, ...) as well as some scene-related features (i.e. environment). We then evaluated both low and high level features for our data set, through computer vision algorithms and manually in crowdsourcing respectively. The uncertainty on these last is overall small. Our main contribution is in the adoption of different mathematical approaches to model social context perception based on features. White box ones have been very useful to explicit features influence. Our results are in line with empirical experience and psychology findings related to chosen high level features.

7.3 Discussion: limitations and improvements

Our research investigated which image features influence the perception of a portrait to be more suitable for a purpose more than another. This perception, already addressed in psychology and sociology, has never been investigated in light of engineering and informatics at the best of our knowledge. Still, we found many similarities with closely related fields, as Quality of Experience and Image Aesthetics, that helped us with the methodology. Still, we found many problems, mainly practical, that imposed us to make choices limiting our work.

First, reviewing literature to select features, we found that *many are the factors* that can influence social context perception. It is impossible to consider them all for obvious feasibility reasons. We then had to make a choice, discarding many factors. For example, it would have been interesting to consider the make up in female portraits, the haircut style, the clothing color just to name a few. These factors can be added in future research, repeating the same procedure for high level features assessment in crowdsourcing, but with the new features.

Secondly, we found that the *perception of some high level features can be subjective*. This is the case of clothing for example: opinions split for some portraits

reporting depicted subjects to have a formal or informal dress. While this element can underline outliers in many cases, sometimes this can be due to a cultural bias. It would be great to consider this bias in our analysis, however many more evaluations, differentiated in terms of participants demography, should be gathered.

Still, crowdsourcing greatly helped our research as it provided many evaluations, needed for a study that considers many factors at the same time. However, crowdsourcing took a remarkable amount of time in order to run it in practice, especially considering reliability controls. This amount of time was not considered at the beginning and slowed down the research. Future experiments and data analysis will definitely profit of this in terms of time.

Cultural bias plays another important role in our analysis. We underlined in fact that different models for male and female subjects should be done. However, in our research this cannot be done due to the lack of a sufficient number of female subjects. This limitation can be overcome either adding more female participants or eliminating this factor showing only portraits depicting subjects of the same gender of the participants.

Another limitation is the *number of evaluated portraits*. More stimuli would make our model more precise and allow a deeper analysis. However, while the Web is full of images, it must be considered that evaluating them both in terms of social context and high level features has a cost. Our data set represented for us a good trade off between analysis feasibility and economical costs. Forms of free-of-charge crowdsourcing can be imagined, but pros and cons must be considered as well; for example preparing a gamification strategy takes time and efforts, while preparing a volunteer based experiments for university students on campus limits the demography. In this respect, it is interesting to mention the great value of *social networks*, that could be a huge resource in terms of portraits and evaluations. These offer different faces as well as different social contexts. However, two main points must be considered: privacy limitations and how to successfully “exploit” the crowd of social network users. Privacy is a concern that is clearly important, considering how we will use the images (we already discussed this point in 4.3, more info in Annex F). The second point needs to be addressed carefully, as discussed while talking about portrait sources (ref. Annex F). Still, social networks could be an important source of social context evaluations, if a proper strategy to address evaluations and motivate participants is designed. In fact, we underline once again that we cannot easily exploit implicit information in the form of user interactions (i.e. assuming the adoption of a particular picture as a subjective opinion), as these are not statistically relevant due to the low number of evaluators. In this direction, it would be really interesting to design a social network application where profile owners can upload their profile picture and let the community rate it (in terms of social context). At the same time, people would like to participate as

they will know what people think of their picture while we will gather evaluations. To conclude, the different limitations constitute possible improvements, leaving margin for interesting further works in this research.

Annexes

Appendix A

Qualinet COST Action

In this annex I give some details regarding the scientific community called Qualinet, in which I took part during while working at my thesis. This participation boosted substantially the research work, offering many exchanges with experts in crowdsourcing as well in the close field of multimedia assessment.

This community is formally an Action of the COST framework, a European framework supporting and funding the cooperation among researchers on technology related programs¹. Actions are the actual research driven programs, launched and funded by COST. They have clear defined objectives, goals and deliverables to be pursued within a four year time span. Qualinet is the COST Action IC 1003, focused on “Quality of Experience in Multimedia Systems and Services”².

It’s focused on the concept of Quality of Experience (QoE), that encompasses in a much broader way the concept of overall quality. The attention in QoE research is then focused on the overall experience and not only on the media itself or the sole user - as would it be i.e. in classic quality assessments. As said in chapter 1, research underlined that inherent media quality is not everything and there are other aspects to consider in the evaluation. Qualinet conducted a remarkable research effort on the subject, formally defining the QoE and its components considering the experience of the overall scientific community taking part in the Action. For Qualinet the QoE is “the degree of delight or annoyance of the user of an application or service” and underlines that three are the main characteristics that influence QoE (influential factors): the user, the system and the context [Le Callet 12], that must be considered to “better express everything involved in a [...] service”. The work done by this Action is not only theoretical but practical too, as it conducted multiple experiments to push forward QoE research.

¹More details on official website, http://www.cost.eu/about_cost, retrieved Aug 2015.

²http://www.qualinet.eu/index.php?option=com_content&view=article&id=2, retrieved August 2015.

This has been done under different aspects and along different axes: five are the working groups that structure Qualinet. The first one is focused on contexts and applications to which QoE is related to. The second one focuses on the mechanisms and models of human perception, fundamental to be understood for a research considering also the user. The third and fourth focus on technical elements of research, addressing the study of reliable quality metrics and the construction of multimedia databases respectively. The last one is focused on the standardization and dissemination of results. Every working group is internally divided in Task Forces, specific to a subtopic.

Within this community, part this thesis work has been a cooperation and a contribution to the second working group, that takes into account the user. Contributions are the work related to electrophysiology and to crowdsourcing. In fact this working group investigates multiple levels of features - involving the user - that impact the perception, such as cognitive, emotional and social aspects. First output, related to the work done with the Emotions Task Force, is the work described in 2.4.2 and published in [De Moor 14]. This contribution was supported by a Short Term Scientific Mission at VTT Institute of Technology in Finland (Oulu).

The second output is related to the Crowdsourcing Task Force: this task force has the main objective of identifying strengths and scientific challenges for QoE assessment via CS as well as to derive a methodology for such a task³. As an output from the task force, we recently released a white paper to publish best practices and recommendations for crowdsourced QoE, based on practical experience ([Hoßfeld 14]). This white paper includes our recommendations described in chapter 4. In particular, it outlines practical implementation problems - found during our experiments too - and it summarizes lessons learned through the different experiences and exchanges related to crowdsourcing.

³Official page on [#qualinet_wg2_taskforce_crowdsourcing](https://www3.informatik.uni-wuerzburg.de/qaewiki/qualinet:crowd), retrieved June 3rd, 2015.

Appendix B

Theories of Emotions: external stimuli and elicited emotions

In this annex we detail the different theories that have been made regarding the generation of emotions. Its objective is to complete the scope given in chapter 3, underlining how complicated is the subject even in the neuro-psychology field, without going too much in detail. Firstly, a clarification is needed as different terms have been defined in psychology to define the different reactions related to affective behaviors. An *emotion* is defined by social science as “an episode of interrelated, synchronized changes in the states of [...] organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism” [Scherer 05]. In practice is a “state of arousal associated with varying degrees of physiological activation” [Basavanna 00]. In fact, many theories about the relation nervous response-emotion have been proposed (as detailed in annex B). A *feeling* is defined as the “conscious subjective experience of an emotion” [APA 07], or the “pleasure and pain dimension of emotion” [Basavanna 00]. It is then a “component of the emotion itself” [Scherer 05]. *Mood* is the “affective state of relatively long duration; usually less intense than typical emotional reactions” [Basavanna 00]. In practice, is the mental state that we are experiencing, product of all the interactions both affective and of other nature; usually is not directed toward something in particular but to the overall environment around us [Frijda 09]. Here we focus either on emotions, considering the effect of the stimuli, or on the overall user’s mood, considering the whole process linked to the content fruition.

Even if science is still researching how the emotions generate in human brain, some areas seem to be related to their generation [MACLEAN 52]. The limbic system in particular has been pointed between the main actors; this part of the brain is deep inside the brain itself and linked to the spinal cord that irradiates all over

the body. As emotions are so entangled in nerves, it is not surprising to find that physiology reactions too are closely linked to emotions. This dual interconnection opens then the question to how these two - emotions and physiological reactions - work together and which one influences the other. Different theories have been formulated on the subject; four of them are the most discussed [Satu 14]. A first one, called James-Lange and formulated independently by these two psychologists at the end of 19th century, indicates emotions only as «a perception of changes in the body» [Cannon 27]: the cause of emotions is then a physical reaction to an outside stimuli, almost as an higher level interpretation of it. The opinion of Dr. Lange differ however from James' one, only from where the physiological reaction comes from, as Lange narrows down the source only to the circulatory system. This theory raised some critics especially from physiologist Cannon and his doctoral student Bard. In particular they argued that many physiological conditions having similar body reactions do not produce the same emotion (i.e. fever and asphyxia). Moreover they argued that «visceral changes are too slow to be a source of emotional feeling». They developed then a new theory, called Cannon-Bard, that states that physiological reactions and experiencing emotions occur together but are separated and independent form each other. This fact then rejects that one is causing the other. Researchers pointed the thalamic brain region as source of emotions, after conducting surgical experiments on animals. Studies pointed the thalamic brain region as source of emotions. Decades later the two components - physiological reaction and emotions - have been linked together again by two psychologists, Schachter and Singer. Their studies brought to the development of another theory, stating that emotions are based both on body arousal perception and cognitive processing. This theory is so called two factor theory of emotion and justify also why it is possible to experience very different emotions in different situations, expressing similar physiological reactions [Satu 14]. We do not detail the the forth theory, called Opponent-Process theory and developed by Richard Solomon and John Corbit, as it does not consider physiological reactions but focuses on opposite emotions balance.

Appendix C

Laser Doppler Perfusion Monitoring

Laser Doppler Perfusion Monitoring (LDPM) is a technique used in medicine for studying the perfusion of blood in microcirculation of tissues. It is widely adopted for analyzing tissues damaged by heat as in burn assessment or in necrotic pathologies. In general, laser doppler velocimetry adopts the shift in the frequency of a low power laser's reflected light as a measure of the quantity and velocity of particles in fluids [Durst 76]. This is why it is also known as laser doppler flowmetry or velocimetry. The principle is the following: a laser light beam of a fixed wavelength is emitted toward the fluid, so that the incident light is reflected by particles in it. A receiver measures the light reflected at a fixed angle. Frequency of reflected light is shifted hitting moving particles due to the Doppler Effect. As part of emitted light is absorbed and diffracted, the measure takes into account light reflected by an area defined by the characteristics of the laser frequency and of the material under test. In our device's probe, a 780 nm laser light is emitted 0.25 mm far from the receiver; considering the average composition of surface skin, we are able to measure blood perfusion 1 mm beneath skin. Blood perfusion is directly related to both the amount and speed of particles in the capillaries, and it's influenced by many different factors, like position, age, temperature, health, heart rate and blood pressure. As these last physiological factors are ruled by the autonomous nervous system, we are investigating if this measurement links with users psychological reactions. LDPM shows fluctuations in flow's speed. This allows to measure homeostatic reactions like vasoconstriction and also heart beats; in the first case signal shows a decay due to lowering of flow, while in the second faster oscillations will be summed to the signal baseline. With more accuracy and sampling, also heart valves opening/closing can be seen. A deeper analysis of laser Doppler perfusion signal frequencies measurable from human skin is present in [Kvandal 06], also during thermal tests [Maniewski 99]. This measurement gives an absolute value - in Perfusion Units (PU) - of the irroration of tissue due to

microcirculation. As said, perfusion is related to reflected light; the reflected light is related both to the velocity and quantity of particles. Two different measures are then available, called respectively Velocity and Concentration of Moving Blood Cells (CMBC). The perfusion is given by their product. At the moment, to the best of our knowledge, LDPM has been used in affective research with multimedia stimuli only once as a complementary measure, to check the presence of vasoconstriction due to a stimuli [Kistler 98]. As said in chapter 3, skin blood flow has been monitored with LDPM in affective research for other kind of stimuli, notably tastes and odors of water [Haese], obtaining good results. In this work the main measurement adopted was infrared thermography on fingertips; however also blood pressure and perfusion measurements were taken.

Appendix D

Details regarding first electrophysiology pilot study: Investigating Electrophysiology for Measuring Emotions Triggered by Audio Stimuli

Stimuli selection

As said in Chapter III, we've adopted the International Affective Digitized Sounds (IADS) database, provided under request by the NIMH Center for Emotion and Attention (CSEA) at the University of Florida [Bradley 07]. This database has been employed in other affective researches [Mühl 11, Viinikainen 12] . This database consists of many different pure sounds - without any speech - of different nature, evaluated subjectively for their emotional impact. The latest version of this database take into account 167 different sounds of 6 seconds length in average, rated each one from at least 100 participants. Sounds' emotional impact has been assessed with the Self Assessment Manikin (SAM), providing a rate in the PAD space.

We selected a subset of this database to limit experiment duration, as longer times are more likely to cause stress and/or boredom to the user, impairing affective assessment. Subset has been constructed choosing sounds the more possibly spaced on the PAD space. Clustering and selection of representants has been adopted to restrict sound number, as done in [Viinikainen 12]. Before clustering the space we restricted the number of samples selecting the ones with lower standard deviation in assessments. For clustering we adopted the KMean algorithm.

At first we empirically choose 5 different regions in the PAD space looking at the sound distribution and selected the three sounds closest to cluster centroid to represent it; ANOVA analysis on the three separate dimensions has been run to avoid taking outliers of a cluster. As we do not have raw self assessment data - not provided with the IADS database - we adopted the procedure described in [Cohen 02] using directly mean value and standard deviation of sounds evaluations. The whole procedure brought to underline two big group of sounds, that we indicate as A and B respectively, as shown in figure 1, of lower or higher affective impact.

For each test run we randomized cluster order and sound order in each cluster, as we did not want bias introduced by presentation order or by any cumulative effect on emotion that can arise.

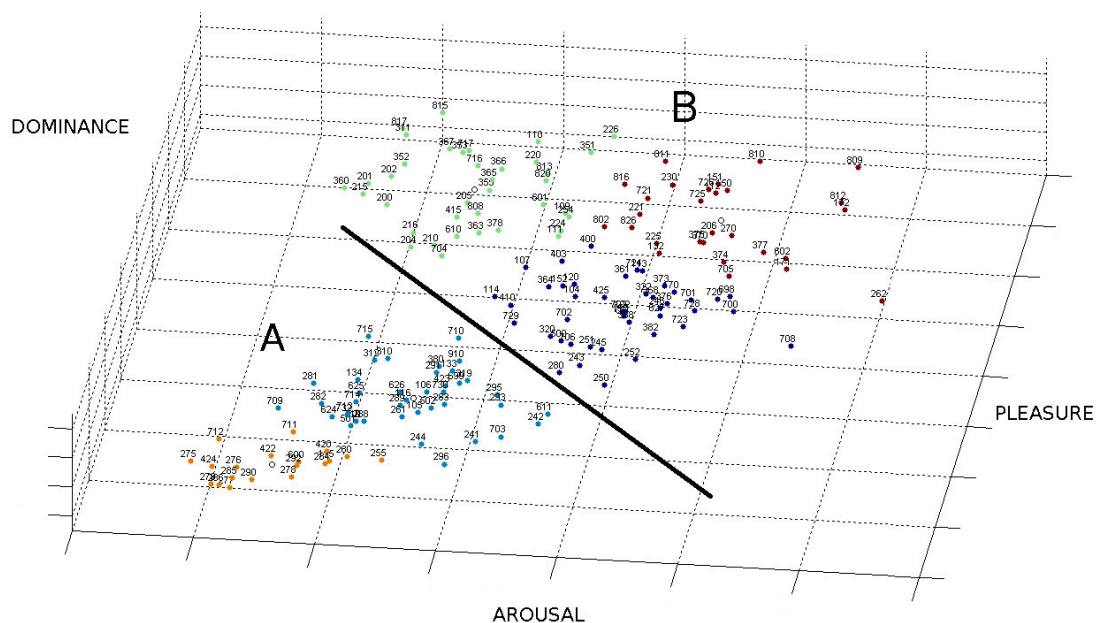


Figure D.1: IADS sound clustering; two main regions underlined

Each cluster, made of three representatives, has been presented once to each subject. A fixed amount of time has been waited before providing the next stimuli burst. The purpose of this pause, is to relax the user before the next stimuli set. During pre-tests, we empirically determined that a pause of 10 seconds was in average enough to let the user relax and restore a baseline in physiological signals to start from. It has to be noticed that this baseline can differ from the one at the beginning, as the overall state of the user can change, although the randomization at the beginning mitigate possible bias or accumulation effects in the user. We

noticed also that more seconds are not likely to restore this baseline easier, as fluctuations arise again. Our hypothesis is that a longer time can cause the user to focus on something else and this can provide measurable reactions on the user that can bias the experiment. For example, a longer period may induce the user to think about something wrong with the experiment or on how much time has passed and think on what he has to do next, causing a reaction dependent on his state of mind.

Data Analysis

We started analysis adopting techniques already used in different electrophysiology measurements, notably as with our EEG experiment previously cited, as there is no literature regarding LDPM in detail for affective research. We focused our attention directly on the perfusion signal instead on only Velocity or CMBC signals as the first is related to both of them and we don't know which one can be more representative of an emotion. Signals have been inspected manually in order to check for clear errors or impairments, as due to data communication errors or probe shifts, ending in removing a user from the dataset as probe had a discontinuous contact with user's finger. Other smaller impairments have been successively removed manually. Signals have been aligned cutting unneeded seconds from beginning and end of recordings, belonging to setup and post-test moments. This visual inspection showed also a large variation between signals, showing a different degree of reaction between users; only in few cases very small or no variations from baseline were present. It is not possible to say if this is only due to their personal variation in perfusion or to an higher reaction to stimuli. In some cases a slow 'fall-rise' pattern is observed soon after a stimuli. However delay and amplitudes are strongly subjective; in any case no stimuli elicited a fall-rise pattern longer than 3 seconds until now. However, considering pattern features as mean, standard deviation or derivate, the analysis carried out did not underline a strong correlation stimuli/observed pattern. A first simple analysis to carry is to find if with only the LDPM we can detect the presence of a stimuli, that is to say if the stimuli perception provoked a variation on signal's baseline.

To analyze mathematically perfusion signal's evolution after a sound stimuli we the filtered heart rate components, visible on perfusion signal, and higher frequencies. Low frequency oscillations are confirmed on Laser Doppler Flowmetry of blood [Kvandal 06] and our analysis shows 99% of the power distributed below 4 Hz. Heart rate pulses have been removed with a notch filter as here we are not interested in heart pulse measure to our affective evaluation. Signals have then been low pass filtered, decimated and z-scored. Starting from previous observations we selected signal features to use on a machine learning algorithm. Our aim here is

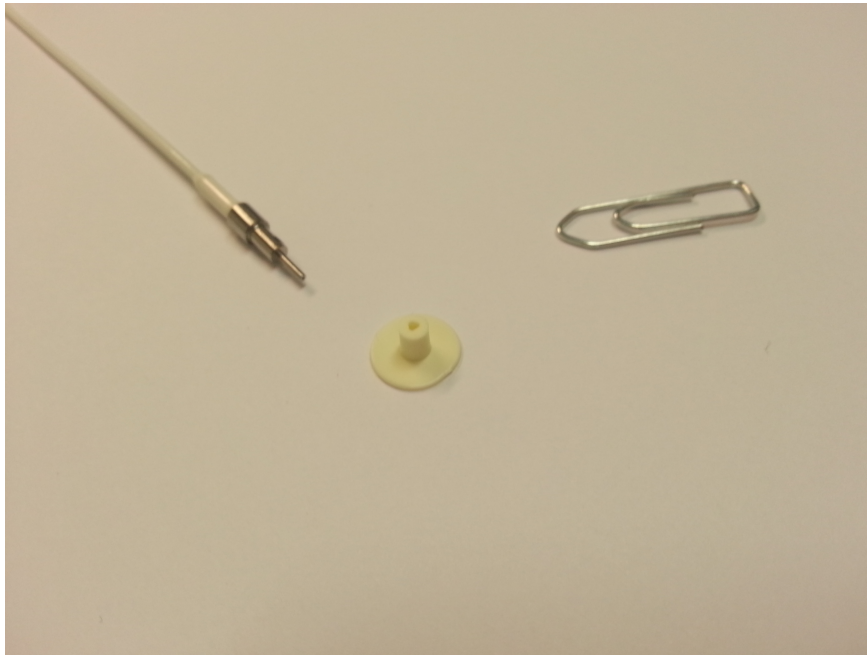


Figure D.2: The LDPM probe tip, with a regular clip aside for comparison.

to check if we can discriminate between user state while listening to a sound belonging to cluster A or cluster B previously described. Signal has been windowed in order to separate reactions belonging to different sounds. Power of different frequencies have been extracted from perfusion signals based on proposed features in literature, related to other physiological responses, as in [Bos 06, Koelstra 12]. We adopted in detail a multilayer perceptron, feedforward back-propagation, with one hidden layer, tuning it accordingly; 15% of data has been used for the validation stage. Classification shows accuracy better than chance, as presented in confusion matrix in figure~\ref{fig:confusion}. Results are impaired mostly from erroneous classification of second cluster, relative to higher values in PAD space, as a low value one. The other one instead, although presents some misclassification, is better recognized. These results are less performing compared with the ones of our previously cited EEG experiment, from which we adopted methodology and data analysis. In that experiment, adopting as for LDPM two classes for low and high impact stimuli, we achieved an accuracy in correct classification of 82% for the first and 76% for the second class.

Appendix E

Details regarding pilot study 2: Chamber QoE – A Multi-instrumental Approach to Explore Affective Aspects in relation to Quality of Experience

This annex gives details regarding the second pilot study described in chapter 2.

Experiment took place in VTT institute of Finland, in an experiment room. Room light was optimized, in order to allow a clear vision of the screen as well as a clear illumination of participants faces, needed for Emotracker. Luminosity was between 21-25 Lux from the back of the screen. Videos were reproduced on a 17" TFT monitor, with a 1280x1024 resolution. Sound volume, reproduced in by normal speakers, was fixed on an audible fixed level for all users and set to approximately 60 dB. Particular attention has been given to avoid any unwanted sudden external sound or remove corresponding measurements, in order not to consider unwanted user reactions. It has not been possible instead to remove any electromagnetic signal present (i.e. WiFi connections) for obvious reasons; these caused rare interference for the EEG wireless connection from time to time. Corresponding signal excerpts - of the order of around 1 second each - have been discarded.

Their physiological reactions were recorded through an EEG headset.

Questionnaires have been prepared extending and tailoring questionnaires used in previous studies by Dr. K. De Moor from NTNU. Questions were aimed mostly at discovering the affective impact of the content; for this purpose we adopted both pictorial scales like Self-Assessment Manikin (SAM) and Pick A Mood (PAM).

*Appendix E Details regarding pilot study 2: Chamber QoE – A
Multi-instrumental Approach to Explore Affective Aspects in relation to Quality
of Experience*

These provided an easy to use complimentary ground truth measure for affective state. How to use them has been explained to the user at the beginning of the experiment. Additionally, a 10 points Absolute Category Rating for rating quality of videos. This data however is not adopted here as meant for a different research.

Raw face recordings as well extrapolated gaze data and calculated facial expressions are available from the Emotracker device. After data analysis, calculated expressions showed very few emotional activations for the majority of participants. This measure, due to the very low number of effective measures, did not provide any useful information. The cause of this outcome has been underlined to be due to two factors: first, users slightly moved from time to time from optimal position while looking at the screen. With the current technology, it is not possible to check immediately if face recordings can provide useful results as the elaboration is not possible in real time; this means that if a user slightly move from optimal position results facial expressions may not be accurately computed. Secondly, adopted stimuli are not strongly affective ones, then reactions are far less strong than those the Emotracker can detect. Another possible cause is that users reacted less as annoyed by the experiment environment: many users underlined in fact the impact of the Emotracker camera in front of them. However this is only an hypothesis and further study are needed to confirm it.

EEG signal has been filtered with a low-band to the frequency of 64Hz, the maximum allowed by our sampling rate of 128Hz, following Nyquist theorem. Band power has been extracted for alpha (8-12 Hz) and beta (13-26 Hz), adopting a sliding temporal window of 2 seconds length, similarly as done in [Bos 06]. First window from every recording (one recording per video per participant) has been removed from analysis to avoid considering participants reactions to the beginning of video reproduction (i.e. surprise to see the content). Videos begin and EEG recordings have been synchronized through a common clock adopting a common local NTP server connected to both the user display and the computer controlling the EEG. Time error has been measured being in the order of ten ms before the begin of the experiment.

Appendix F

Portrait sources for research purposes

In this annex we explicit our review of different image sources found in literature, summarized in sec. 4.3 of chapter 4. In particular, we differentiated adopted image data sets in four classes: public image databases, personal collections, online portraits and shooting a portrait set in laboratory. These groups are discussed in next four sections.

F.0.1 Public image databases

Many portrait databases have been built and published in literature. They differ between them in many aspects; first of all for the content, as some have a large spectrum of images while others are focused on a small number of subjects. Secondly, they greatly differ for the number of images: some databases are focused on a particular image typology and feature a small number of samples while broader collections can contain thousands of images. We will focus here on databases containing face pictures, mentioning remarkable works using them.

Many data sets have been developed for research on face detection or recognition, emotion or pose estimation. Gur et Al. of University of Pennsylvania's Brain Behavior Laboratory proposed different data sets. Their data sets are made for face memory studies [Gur 01] and for research on facial emotion recognition [Erwin 92]. These black and white databases focus on the sole face, on black background. Successive studies from the same lab increased the number of stimuli and proposed color stimuli, but the general characteristics are the same.

Another available database within the field is the *FEI Face Database*, made by

¹Webpage University of Pennsylvania's Brain Behavior Laboratory, retrieved on June 2015; <http://www.med.upenn.edu/bbl/downloads/2Dfaces/>



Figure F.1: Example stimuli from Gur et Al. studies. SOURCE: University of Pennsylvania's Brain Behavior Laboratory website¹

the FEI University in Brasil [Leo 05]. This database features face pictures from 200 individuals, each taken at 11 different yaw angles, showing then different face profiles. Subjects have neutral expression in all shots, except in one taken with a smile. One underexposed shot has also been taken for each subject. Half of subjects are men and half women, aged between 19 and 40 years old. Pictures are taken on a white background and show also subjects' shoulders. The database also features manual landmarks annotations for frontal images. The *Yale Face*



Figure F.2: Example stimuli from FEI Face Database. SOURCE: database webpage on FEI University website²

Database and its extended version [Yale Univ 01] are well known too. The original one contains 165 grayscale images of 15 people, proposing 11 images each subject, one per different facial expression or configuration (i.e. different light direction, expression, with and without glasses, ..). However images shows almost only the face and neck. The *CMU Pose, Illumination, and Expression (PIE) database*

²<http://fei.edu.br/~cet/facedatabase.html>

[Sim 02] is also focused on different images of the same subject. This data set features over 40 thousands images from 68 subjects; this huge amount of images comes from the combinations of the different factors taken into account. For memorability purposes, a huge database has been proposed by W. Bainbridge, within the team of A. Oliva [Bainbridge 13b]. The database is called *The 10k US Adult Face Database*, and features more than 10 thousands portraits of around 2 thousands people. Pictures have been taken from online sources and are however cropped around the face.

All mentioned data sets are however useless for our purposes, as they do not convey information regarding the social context, focusing only on the face. More interesting is the well known *Labeled Faces in the Wild (LFW)* [Huang 07], collecting portraits of public figures taken in real situations (fig. F.0.1). Images present an high variability of both face pose, lighting and expression as well of the background and context. This data set is labeled and mainly dedicated to face detection-recognition. Compared to other databases LFW is more interesting for our purpose as images partially show background and cloth of the person, elements related to social context. However many images depict famous people and we must pay attention in our research as this can be a possible biasing factor. Moreover, another possible bias can come from pictures having too low resolution. These may be interesting for face recognition purposes but may pose problems: a too low resolution can impair people from understanding correctly the scene content. Very similar to LFW is the *PubFig* database from Columbia University [Kumar 09]. While still offering online retrieved portraits of public figures, this data set offers many more shots of less people: it then offers more contexts for the same person. However, being them public figures, the contexts are actually many similar between them.

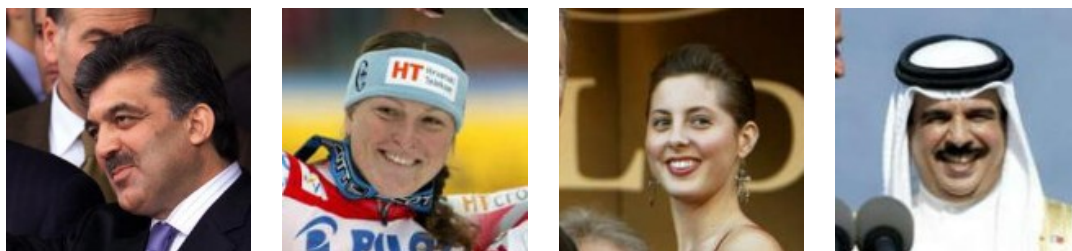


Figure F.3: Example stimuli from Labeled Faces in the Wild data set. SOURCE: LFW website³

California Institute of Technology proposed instead a database composed of labeled images (Caltech DB), belonging to many different categories. Mainly it

³<http://vis-www.cs.umass.edu/lfw/>

is aimed to object recognition purposes. The face category contains 450 shots of 27 subjects in different conditions [Fei-Fei 07]; however shots are scaled to low resolution, and this can pose problems as said before. Recently, the *FAVA database* has been proposed to study aesthetics of facial images [Lienhard 15b]. While it is not yet public, this data set is composed by 300 images subset of AVA data set, a much larger image database created for aesthetic purposes by [Murray 12]. AVA contains more than 250 thousands images taken from the DPChallenge online network, later on described in this section. The same research group proposed also the *Human Face Scores* data set [Lienhard 14], made with images from LFW and Caltech databases as well as shots from a private collection. 250 images have been taken and their aesthetic value has been subjectively evaluated in laboratory.

A very important resource is provided by *ImageNet*, a huge collection of labeled pictures, developed for object recognition purposes [Deng 09]. Many computer vision challenges have been launched adopting the database and still active online ⁴. ImageNet has been used as image resource many times in research, especially for image recognition approaches adopting deep neural networks [Krizhevsky 12, Sermanet 13]. Images are related to a large number English words and are categorized in a hierarchical structure. Images have been collected from the Internet through querying different search engines and then cleaned relying on human labor. At present, the database of ImageNet accounts 15 millions of images, related to 21 thousands concepts⁵ and it is the biggest online and available at the best of our knowledge. Between its categories, there is of course also the “face” category, accounting at present around 1500 images. While many images in the category may be useful in our research, there is a huge variability between shots. Many shots are blurred, manipulated via software, contain children, show only partially the face or contain multiple faces; some of them are also Copyrighted⁶ and therefore cannot be used for many purposes. Manual feedback on images is possible through ImageNet website interface - to clean the data set. As many images from this data set have been taken from Flickr - being these indexed in adopted search engines - we preferred to look for these pictures directly in Flickr, as said later. Many face images are also proposed by *FaceTracer* database [Kumar 08], that provides more than 15000 labeled aligned face images, exploited by authors for face verification and image search; however, their complete data set (not yet labeled or released) is made of around 3.1 million pictures [Kumar 11]. While it offers a good variety in terms of subjects (both on age and demographics) and face statistics are reported (exact location and rotation angles), images are cutted near the face, removing a lot of context. As images have been retrieved online, original

⁴I.e. <http://www.image-net.org/challenges/LSVRC/2014/>

⁵ImageNet statistics, webpage <http://image-net.org/>

⁶At least in the original source; images retrieved June 17th, 2015.

URLs are reported; still, many of them are protected by copyright. Tables F.1-2.3 summarize databases found.

F.0.2 Personal collections

As alternative to available databases it is worth to mention the use of personal photo collections as photo resources in research. These data sets may be very interesting for us as they provide different portrait version of the same subject in different contexts. This variability in portrait elements (other than the face) may be useful to conduct statistical analysis. Personal collections have been sometimes merged with other data sets in order to increase the number or stimuli. Authors of [Lienhard 14] added personal images to FLW and Caltech Face Dataset to have enough portraits to address their image segmentation study. No particular reason for this choice is given. Undisclosed personal portrait collections have been adopted also by the team that developed Google Picasa’s face movie transition, called photobios [Kemelmacher-Shlizerman 11]. In this case the adoption of such collections is straightforward, considering that Picasa software organizes personal collections of pictures and that their algorithm was meant to aim this tool. In [Redi 13a] most images have been taken from a personal collection of a photographer. While the reason is not explained, we believe that authors preferred this solution as the study focuses on pictures aesthetic assessment and involves many image descriptors, some being also uncommon as “simplicity”. More than 100 personal shots taken in different conditions have been exploited. Personal collections have also been adopted by themselves, as in [Pigeau 10] where personal photo collection organization is the target of the study. In this work a large number of geographical tagged photos is needed in order to test accuracy of a multidimensional clustering algorithm. The use of this kind of pictures is justified by the objective of the work as well as the need of a relatively small spread in terms of time-space between pictures tags. Different image collections have been constructed, accounting between 700 and 1700 pictures each. In [Ferré 07] authors exploited a very large personal collection (5000 pictures) to test an organizing and browsing system, mainly based on Concept Analysis of metadata. The choice of a personal collection was due to the aim of organizing the collection itself, which ground truth was probably known especially by authors.

An interesting “phenomenon” is today appearing on the web: shooting a selfie per day for a whole year and post the series online, sometimes in form of an animation (ex. figure F.0.2). This trend is providing a large amount of portraits of the same subject, in a wide variety of conditions. However these are usually unavailable to be downloaded as such but only in the form of animations made by authors themselves.

DATABASE	AUTHOR	TYPE	# IMAGES	PUBLICATION	ORIGINAL PURPOSE	NOTES
Black and white face memory stimuli	University of Pennsylvania's Brain Behavior Laboratory	Frontal grayscale neutral faces		[Gur 01]	Face memorability	Black background
Black and white emotional face memory stimuli	University of Pennsylvania's Brain Behavior Laboratory	Frontal grayscale faces; happy, sad, and neutral expressions		[Erwin 92]	Emotion recognition	
FEI Face Database	FEI University in Brasil	Face profiles of 200 individuals, normal shot and underexposed version. Men and women, aged between 19 and 40 years. Neutral expressions + smile.	~2800	[Leo 05]	Face recognition	White background. Pictures show also subjects' shoulders. The database also features manual landmarks annotations for frontal images.
Yale Face Database	Yale University	Grayscale images of 15 people, different facial expression or configuration (i.e. changing light direction, expression, glasses, ...)	165	[Yale Univ 01]	Face detection / recognition	Images show almost only the face and neck

Table F.1: Summary of reviewed publicly available databases

DATABASE	AUTHOR	TYPE	# IMAGES	PUBLICATION	ORIGINAL PURPOSE	NOTES
FAVA	GIPSA-Lab	Color face images	300	[Lienhard 15b]	Image aesthetics	Taken from AVA database [Murray 12]
Human Face Scores (HFS)	GIPSA-Lab	Set of 7 different color images of 20 persons, and 110 additional images of different persons.	250	[Lienhard 14]	Image aesthetics	Taken from AVA and Caltech data sets, plus private shots. Subjectively evaluated for aesthetics
Caltech Face Dataset	Computational Vision Lab, Caltech University	Frontal faces, 27 subjects, unclear mix of conditions	450	[Fei-Fei 07]	Image classification	Scaled images
CMU Pose, Illumination, and Expression (PIE)	The Robotics Institute, Carnegie Mellon University	Color portraits of 68 subjects; different combinations of factors (as lights, expression, talk, glasses)	>40000	[Sim 02]	Face detection / recognition	
The 10k US Adult Face Database	MIT	Face photographs of around 2000 people	>8500	[Bainbridge 13b]	Image memorability	Pictures taken from online sources; cropped around the face

Table F.2: Summary of reviewed databases - continued from table 2.1.

DATABASE	AUTHOR	TYPE	# IMAGES	PUBLICATION	ORIGINAL PURPOSE	NOTES
Labeled Faces in the Wild (LFW)	Computer Vision Laboratory, University of Massachusetts	Portraits of ~5700 world famous public figures taken in real situations	>13000	[Huang 07]	Face recognition	High variability of face pose, lighting and expression as well of the background and context.
PubFig	Columbia University	Color portraits of 200 public figures	~58000	[Kumar 09]	Face verification	Large variation as in LFW
Data set of FaceTracer	Columbia University	Color portraits, large variety of subjects in terms of age and demographics	~15000 aligned labeled (~3 100 000 in total, not yet public)	[Kumar 08]	Image retrieval	Face statistics are reported: exact location and rotation angles. Images are cutted near the face, removing a lot of context
ImageNet	Princeton University	Online hierarchical image database, manually tagged	~15 000 000 (~1500 in 'face' category)	[Deng 09]	Image classification	Images clustered in 21 thousands concepts

Table F.3: Summary of reviewed databases - continued from table 2.2.

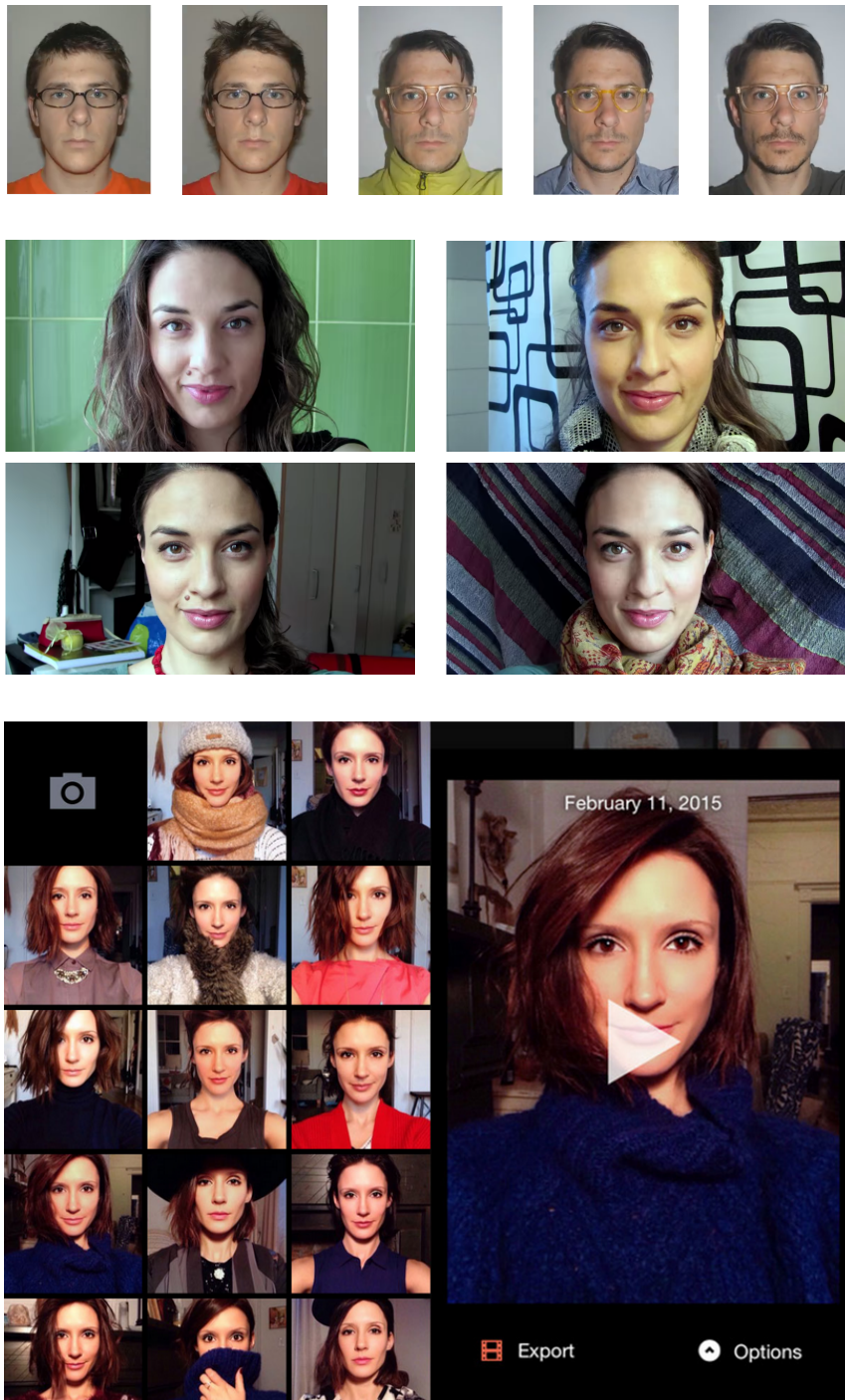


Figure F.4: Top and middle: online projects of daily selfies: Living My Life Faster, 2014, JK Keller; One photo a day in the worst year of my life, 2012, B92. Bottom: screenshot of Everyday-app, a mobile app for this purpose. ⁷

After this review, at the best of our knowledge no personal collection used for research has been made publicly available and none of them is focused on portraits, except those merged into previously mentioned databases.

F.0.3 Online portraits

Nowadays a huge image resource is the web. This is even more true for portraits, as online communities and social networks are constantly filled with this kind of pictures. Moreover these portraits are usually taken in real use cases and this characteristic makes them even more interesting for our research on social context. Online sources have been used in research multiple times; we review here the most important online communities used with remarkable works using them. However “online” does not imply that images are freely available to be downloaded and adopted; we will discuss this point in next subsection.

Social Networks

A first resource is given by social networks (SNs), that today are extremely popular. *Facebook* (FB) is probably the most famous social network in Europe and US⁸, counting more than 700 millions of active users daily [Sedghi 14]. Aside the social and playful aspect, its popularity and huge amount of data made it really interesting for research too; this network has been largely adopted to conduct network analysis of large communities and adopted for image related studies: around 300 millions images are uploaded and shared daily [Meeker 14]. Portrait typology in this and other networks present a huge variety (i.e. fig. F.0.3). In [Wood 12] where researchers focused on exploiting semantic information aside images to provide better search results and entertainment to users, providing more pertinent results; their approach positively exploited FB images metadata and human labor through the public API of this SN (i.e. user tags). Also in [Hum 11] researchers positively exploited FB data but they instead focused only on portrait images in order to investigate pictures variability and how contents differ by users’ gender. It is important that researches exploiting this data, as the cited ones, had to ask permission to selected participants before. Another popular social network is *Twitter*⁹, where users’ interactions are more focused on broadcasting and resharing information: around 100 million users connect to Twitter daily to post or retrieve “tweets”. This community has been primarily adopted for studies regarding user experience, sentiment and content analysis [Kivran-Swaine 14, André 12] or users

⁷<http://everyday-app.com/>

⁸While present even in other parts of the world, in some countries this network is blocked and alternatives are present (i.e. in China).

⁹<https://twitter.com/>

personality studies [Quercia 11]. Indeed, no specific research on profile images carried on Twitter has been found. A small survey we conducted to evaluate the usefulness of this source outlined that many users do not adopt a personal picture but a logo or a group picture (see appendix ...)? Other SNs exist for different specific purposes: *LinkedIn* is probably the most popular for social connections related to work while *Meetic* is very popular for dating purposes. These last SNs are even more rich in portrait pictures as it is particularly important for users to focus only on their pictures, for obvious reasons due to the nature of these SNs. However, while these social networks are filled with portrait images, many limitations are enforced for privacy and security issues, as discussed in next subsection. This is why the use of LinkedIn and Meetic in research is quite limited and research focused only on LinkedIn social connections analysis, as in [Skeels 09] and [Gloor 07].

Photo sharing communities

Photo sharing communities became important within the last decade. *Flickr*¹⁰ is a huge resource for images; it is a popular image and video hosting website opened around ten years ago. Its network is very big and active; a recent analysis done with its public API revealed that an average of around 60 millions of public pictures have been uploaded monthly in the last two years [Michel 15]. People adopt Flickr as platform to save and share their images, both publicly and privately. This network has been adopted multiple times in research. Mainly it has been adopted for aesthetic assessment purposes, as in [Li 10b] and in [Pogačnik 12]. Flickr has been adopted also for researches related to face pictures. In [Males 13] authors collected almost 400 images from Flickr for aesthetic assessments of headshots. More recently [Lienhard 15a] investigated the instantaneous feeling of a facial picture adopting stimuli from different sources, Flickr included.

Due to the easiness of use, the positive results shown in literature and the variety of images, we preferred to collect mainly from Flickr the pictures for our work related to social context perception, as explained in F.0.6.2. Other popular photography communities exist online. *Photo.net*¹¹ is quite popular between photography amateurs and has been adopted multiple times in research. Between the most known works adopting them Datta's research on aesthetics [Datta 06], that proved this resource to be useful for images but not necessarily for evaluations, as they are biased by the presence of professional photographers. More recently the Photo.net has been adopted in [Amirshahi 14] for evaluating photography rules. *DPChallenge*¹² is another known community in the field, featuring digital photography contests on different themes. It has been adopted to construct previously

¹⁰ *Flickr*, <https://www.flickr.com/>

¹¹ *Photo.Net*, <http://www.photo.net>.

¹² *DPChallenge*, <http://www.dpchallenge.com/>



Figure F.5: Portraits similar to many others found online in social networks. Many of them show playful moments, depending on the targeted social network. As shown, sometimes funny profile pictures may not even show the face, and a careful selection must be done. SOURCE: personal collection; from top to bottom, left to right: K. De Moor, L. Krasula, V. Skodras, M. Masoura & I. Hupont, S. Tavakoli.

mentioned AVA dataset as well as on other studies on aesthetics in [Pogačnik 12] and in [Joshi 11]. Both Photo.net and DPChallenge feature peer-reviewed pictures, rated by the community itself; they slightly differ in the rating system as described in [Datta 08]. While useful for consumer pictures, it has however been remarked that these two communities are mostly focused on professional shots¹³[Li 10b].

Online remote work marketplaces

Many online *marketplaces for remote work* are also useful communities providing profile images. Between these, we can mention *Elance*, *Upwork*, *Worknhire.com* or *Freelancer.com*¹⁴. These are platforms gathering people willing to work remotely, providing resumes to employers (managing similar to LinkedIn), managing job posting, contracts and salaries. These communities too are really interesting for portrait images as workers are interested in putting their portraits; in some cases this is also compulsory: Elance for example forbids to users the adoption of non professional or group pictures, as well as too small or unidentifiable portraits [Elance 15]. However no study has been conducted adopting pictures posted on these sites, but only studies regarding remote work misbehavior [Clarke 13, Motoyama 11].

Search engines image request

Normal web *search engines* have been found to be useful too. These, as Google and Yahoo, often provide a dedicated search for images (i.e. Google Images¹⁵). In [Shah 12] authors focusing on photo enhancement retrieved 14k face images from the web adopting Google. This approach easily allowed authors to retrieve both positive and negative samples with simple queries. With time search engines have been improved, providing other tools than simple query by text; some engines provide also the ability to perform queries with images, to find visually similar ones. Recently [Nieuwenhuysen 14] investigated the accuracy of such systems, finding that in some cases is also possible to retrieve semantically related images. Other tools have been put in place by search engines that can help in the purpose, as face presence in images or color matching. In [Van De Weijer 07] instead Google is adopted to avoid test subjects to collect a data set of images related to different color palettes. However the quality of results querying face pictures has been questioned and machine learning improvements have been proposed, obtaining promising results [Kumar 11]. Better results have been achieved inputting names

¹³We refer here to picture quality, not to the social context.

¹⁴URLs are homonyms with their names.

¹⁵<https://images.google.com/>

as search queries, as done in [Bainbridge 13a] to construct previously mentioned *The 10k US Adult Face Database*.

While these sources would be able to provide a huge amount of portraits for our research, communities pose a lot of limitations regarding posted material, pictures included. These issues, topic of next paragraph, must be considered carefully.

F.0.4 Issues in adopting online resources

While a huge number of portraits can be collected online, many problems arise, since many online resources do not allow - or strongly limit - pictures' disclosure. It is really important to underline that while technically in most cases it is possible to retrieve displayed images, this may infringe the law. This is even more important when images must be published, as when doing public subjective assessments (i.e. in crowdsourcing, ref. next chapter). Three points must be considered while retrieving images: licenses, privacy issues and automated retrieval ban.

Licensing

When accessing images, particular attention must be taken on the *license* attached to the image and the task to accomplish with the image itself: an image is not freely available just because appears online. Considering this aspect, Flickr is in our opinion between the most interesting tool for online image retrieval, as uploaded images are associated with a specific license. In particular, users can choose when uploading an image if this must be protected by copyright or otherwise it is a Creative Commons work. This last is a category of licenses that usually allow content redistribution under certain conditions, as citing the source¹⁶. Limitations are clearly stated near pictures when they are displayed; search tools on the platform allow to filter for a particular license. ImageNet features many creative commons images as well as copyrighted ones. For these last - retrieved through search engines - the website gives only the link and display them as found, citing the source. However some original images have been removed from the web and this fact opens another problem: it has to be remarked that licenses for some contents may change ¹⁷. This may happen for example if the platform hosting contents changes its policies or - when the law allows it - if the content owner changes his mind. For the same reason, owners can remove the content. This can be a problem when using such images, as tracking licence changes and reacting to it is really hard, especially considering that Internet contents may be stored and

¹⁶Creative Commons, <https://creativecommons.org/>

¹⁷However, it should be noted that CC licenses are not revokable [wiki.creativecommons.org/wiki/Frequently_Asked_Questions].

successively mirrored multiple times. Search engines too put in place some filters for this purpose, to show only Creative Commons only contents. However at the best of our knowledge and considering our practical experiences far less results are provided with these filters in place, as many elements online do not explicit any license and then are not reported between results. Regarding our work, we adopted data sets composed of private images as well as CC attribution licensed images from Flickr, as later explained in this chapter.

Privacy

In addition to images' licenses, *privacy limitations* must be considered. In particular, social networks enforce strong privacy limitations in their Terms of Service to protect their users. For example, Facebook and Linkedin forbid using posted material (included pictures) without explicit permission of profile owners: images, even if open and accessible, cannot be used legally. While asking per se won't be a problem, to build a big data set can become easily time consuming. For this reason portraits from these databases weren't used in our experiment. In Twitter instead images and other information related to the profile are, according to Twitter privacy policy, public on users page [Twitter Inc. 15].

Retrieval

Lastly, problems may arise even grabbing images from the source, even when license and privacy allow. Many websites in fact forbid the use of automated systems to retrieve data on their pages, except those tools authorized especially for the scope (i.e. Application Programming Interfaces - APIs). This is the case of Twitter ¹⁸ and Flickr ¹⁹ for example, featuring image URL retrieval with APIs, making download possible. Web crawlers, small software to scrape web pages and extract information, are a valid alternative when allowed. While they are sometimes forbidden, they are a solution for search engines image retrieval: specific tools have been developed and are available online (i.e. image downloaders in Google Chrome extensions). To conclude, where no API is available and crawlers are forbidden, the only solution is to grab images manually (i.e. for some remote work platforms).

Considering metadata

For the sake of clarity, it is useful to make some additional remarks. At the beginning we thought that taking online portraits would have been useful also

¹⁸<https://dev.twitter.com/rest/reference/get/users/lookup>, retrieved on June, 2015

¹⁹<https://www.flickr.com/services/api/misc.urls.html>, retrieved on June, 2015

because we can retrieve more information regarding images, usually missing in normal databases. First, we can retrieve pictures meta data (i.e. EXIF data containing camera model, ...), but in practice we found that fetch this data is possible and quite easy only on few social networks; many instead strip away this information. Moreover, EXIF data present a huge variety of possible values. As complementary information we can also extract picture descriptions, posted by users. This task too is quite easy, but to process this information revealed to be a very complex task: descriptions are in the form of free text. In the end we desisted due to the inherent complexity of this task, to focus on our primary task. Another information is the source of the portrait itself (i.e. LinkedIn), as it can be an indicator of associated social context. However, this information reflects only the opinion of the picture owner, that put it in that particular social network, and his opinion may be different from the public one. Then we cannot use this information as a ground truth reflecting the overall public perception of social context.

F.0.5 Shooting a portrait set in laboratory

A possible alternative is to build a personal specific data set for the ongoing study. This solution leaves much more freedom as we completely control the portrait creation. In this case we can shoot multiple pictures of the same subject in different conditions, trying to cover all possible contexts in which we are interested. However, it demands more time and exhibits higher costs, especially if a large number of stimuli is needed. Moreover, it is sometimes really hard to have some particular conditions, i.e. specific backgrounds (without considering for now software postprocessing). Data sets constructed in laboratory have been adopted for example to create some databases that we previously mentioned in subsection F.0.1, as the CMU PIE or the Yale DB. This strategy is instead a good solution if few stimuli are needed. In [Bashir 14] authors are interested in a very specific case for a psychology study: the influence of red color on the persuasion of a message. To this extent they took pictures of a communicator with different clothes colors, everything else being equal, and showed the different versions to different subjects. Finding online such images would have been very difficult.

F.0.6 Our data sets

We give here more details about the two data sets adopted in this thesis, described in chapters 4 and 5 respectively.

F.0.6.1 First data set: professional shots versus selfies

In chapter 4 we set up a simulated hiring process for an invented company, for which we prepared fake CVs with a portrait attached. We focused on two picture versions of the same subject, one being a professional shot and one a selfie. As said, our purpose is very specific and we we opted for making portraits in laboratory.



Figure F.6: The photo booth and lightings adopted to take professional shots of subjects for our first data set.

Two different portrait versions - a selfie and a professional shot - have been realized for 6 fake candidates, for a total of twelve portrait images. These have been adopted as resume pictures. Pictures has been taken in our laboratory. Subjects come from different countries and have different face traits. They are aged between 23 and 40. Female portraits have not been taken due to the lack of sufficient different models. Subjects are or have been working in our laboratory, participated voluntarily for free and did sign a disclaimer to allow us to use pictures anonymously.

For the high quality shots, these have been taken in controlled conditions. We adopted a professional photo booth with proper photographic lamps offering diffused lighting (figure F.6). Shots have been taken with a mid-high range Nikon DSLR. Many shots have been taken for each subject, with different combinations of lighting, facial expression and posing. Best shots have been taken, selecting were possible more natural shots. Known photography rules have been followed

for subjects pose and lighting.

For the second portrait version we collected instead selfies. These pictures have been taken with a mobile phone. Mobile device was a Samsung modern smartphone, from which we adopted the frontal camera, offering 2MP shots. We did not use a state-of-the-art mobile offering a more powerful camera as we did want pictures representing “normal” selfies, that do not offer high quality shots but more common ones found in social networks. Subjects were guided by us in taking the shot in order to make all portraits similar (face distance, angle, comparable sharpness, ...), but they have been left free as much as possible to have real selfies. To be sure that images would have not resizing on participants’ browsers, we resized them to a size of 400 x 533 px, considering internet users screen resolutions as done in [Gardlo 12a]. We also shrank images size converting pictures to JPEG file format, setting quality to 85%. This quality setting was found as the best trade off allowing small file size and minimal almost unnoticeable compression artifacts. Figure F.7 shows the data set. As previously mentioned, we manually retouched pictures via software, using proprietary Nikon photo editing and GIMP²⁰, to improve them. The objective has been to reduce imperfections and make portraits more professional. Some of the pictures in this subset have been modified successively to try to change perceived social context and added in the second data set, as explained in next subsection. Realized portraits are shown in figure F.7.

F.0.6.2 Second data set: real online portraits for social context evaluation

The second data set has been prepared to fulfill the main aim of our research: to evaluate social context influential factors. For this purpose we are interested in face pictures that convey more complementary information. This is the case for example of shots taken in real conditions, that contain a broader range of details regarding the context i.e. regarding the background or the cloth. Hence, we looked for amateur or semi-professional pictures, reflecting less formal and “posed” portraits - a characteristic that we suppose influence subjective context perception. For those reasons, we preferred to avoid available databases, with the exception of LFW database as said later in this section. As previously explained (ref. section F.0.4), we took care about collecting only publicly accessible images and considering licenses’ restrictions. Based on previous considerations on image sources, we looked for portraits in Flickr, where we found portraits that we believe fit the categories friends and dating purposes. To this extent, we adopted Flickr API and searched for terms like “face”, “portrait” and more specific terms like “businessman”. Copyleft images or CC attribution licensed images have been taken, as done in [Males 13], provided as found and citing sources.

²⁰<http://www.gimp.org/>

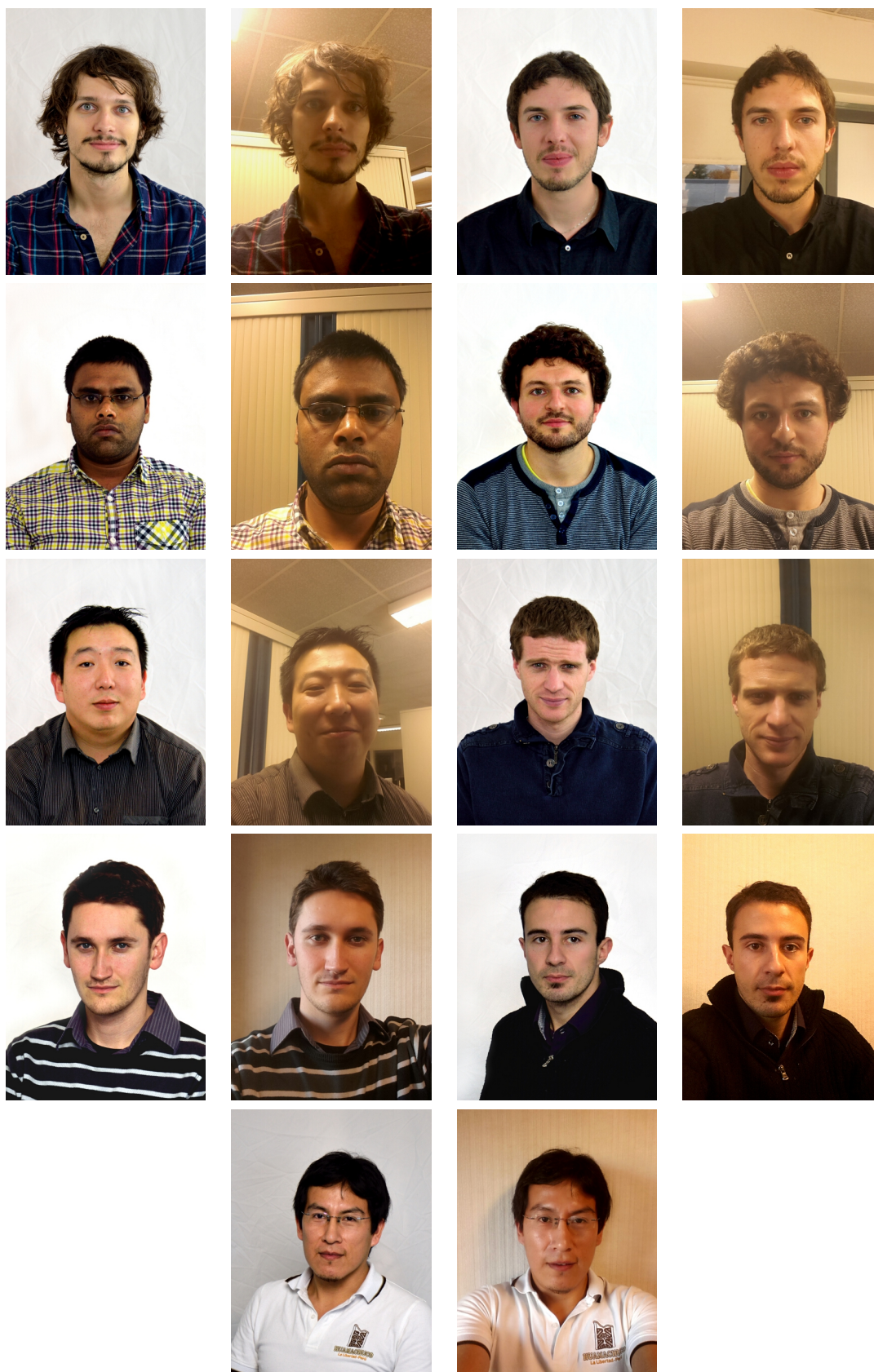


Figure F.7: First dataset created for our pilot study. Subjects are from our research team and voluntarily participated in the shooting.

A careful image selection was required. We manually checked content discarding all non portraits, non-adult subjects or inappropriate content. Extreme close-ups or too small pictures were also not selected, as content information was actually lacking. On the other hand, also pictures showing the context too clearly (e.g., showing signing a contract in a working environment) were discarded. In the end, Flickr did not provide in our opinion many examples for working purposes. To gather work related portraits we then considered previous mentioned work related social networks (i.e. for freelancers). We based this choice on their nature and on the fact that Terms of Service frequently ask for “professional portraits” within profiles. Where no API was available and crawlers were forbidden, we gathered images manually. Still to have more examples possibly related to the working category, in addition we adopted 35 pictures from previously mentioned LFW data set. Even if we preferred to avoid available databases to favor less posed portraits, we found many shots suitable for work category as features shots of public events - i.e. a public speech and real condition shots. We also found some alluring portraits, for both subject pose and expression, maybe useful for dating context. We avoided portraits of very famous people (e.g., world known politicians), since knowing their profession might influence too much the assessment of the fit to a category. Where possible we took different versions of the same subject²¹, both for LFW and Flickr portraits. To add more high quality portraits and selfies we also used the best portraits we created for our previous data set; eleven portraits have been added to the set.

Finally, we also digitally modified the background of some of these images. The purpose is double; first, to have more portraits showing multiple versions of the same subject. While other modifications would have been possible (i.e. changing clothes), this has been empirically found to be the easiest to implement manually for us and the less noticeable. Secondly, it has the advantage of offering an insight into the importance of background in social context bias. Background was changed proposing two different scene setting: a warm brick wall and a cold modern office, taken from open sources previously mentioned. Successive analysis (ref. 5.5) underlined that this element is statistically influential for these shots, even if more tests are needed to confirm the hypothesis in general. Obtained model from all shots will underline overall influential elements.

While it would have been possible to have a much bigger number of portraits, we limited the actual number for the experiment since all pictures had to be assessed also on their high level features, as we will explain in chapter 6. This process has been done manually. Thus, as a compromise in terms of accuracy and reasonable time/cost for the experiment, we decided to use 216 collected pictures. When

²¹These will be shown to different participants in successive tests, in order to avoid potential biases.

adopting this data set, no information about depicted subject was given in order to preserve anonymity, except for the link to original resource when the picture license imposed it.

Bibliography

- [Alonso 09] Omar Alonso & Stefano Mizzaro. *Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment*. Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, pages 15–16, 2009. [www](#)
- [Amati 14] Cristina Amati, Niloy J. Mitra & Tim Weyrich. *A study of image colourfulness*. Proceedings of the Workshop on Computational Aesthetics - CAe '14, pages 23–31, 2014. [www](#)
- [Amirshahi 14] Seyed Ali Amirshahi, Christoph Redies, Joachim Denzler & Gregor Uwe Hayn-Leichsenring. *Evaluating the Rule of Thirds in Photographs and Paintings*. Art & Perception, vol. 2, no. 1-2, pages 163–182, jan 2014. [www](#)
- [André 12] Paul André, Michael S Bernstein & Kurt Luther. *Who Gives A Tweet ? Evaluating Microblog Content Value*. CSCW '12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pages 471–474, 2012.
- [APA 07] American Psychological Association APA. Dictionary of Psychology. American Psychological Association, 2007.
- [Arndt 11] Sebastian Arndt, Jan-Niklas Antons, Robert Schleicher, Sebastian Moller, Simon Scholler & Gabriel Curio. *A Physiological Approach to Determine Video Quality*. 2011 IEEE International Symposium on Multimedia, pages 518–523, dec 2011. [www](#)

- [Aydin 14] Tunc Aydin, Aljoscha Smolic & Markus Gross. *Automated Aesthetic Analysis of Photographic Images*. IEEE Transactions on Visualization and Computer Graphics, vol. 2626, no. c, pages 1–1, 2014. [www](#)
- [Bainbridge 13a] Wilma Bainbridge, Phillip Isola, Idan Blank & Aude Oliva. *Establishing a Database for Studying Human Face Photograph Memory*. Conference of the Cognitive Science Society, no. 2012, pages 1–6, 2013.
- [Bainbridge 13b] Wilma a Bainbridge, Phillip Isola & Aude Oliva. *The intrinsic memorability of face photographs*. Journal of experimental psychology. General, vol. 142, no. 4, pages 1323–34, nov 2013. [www](#)
- [Barnard 47] George Alfred Barnard. *Significance Tests for 2x2 Tables*. Biometrika, vol. 34, no. 1/2, pages 123–138, 1947.
- [Barnett 01] E Barnett & M Casper. *A definition of "social environment"*. American journal of public health, vol. 91, no. 3, page 465, 2001.
- [Basavanna 00] M. Basavanna. Dictionary Of Psychology. 2000.
- [Bashir 14] Nadia Y. Bashir & Nicholas O. Rule. *Shopping under the Influence: Nonverbal Appearance-Based Communicator Cues Affect Consumer Judgments*. Psychology & Marketing, vol. 31, no. 7, pages 539–548, 2014. [www](#)
- [Behling 91] D. U. Behling & E. A. Williams. *Influence of Dress on Perception of Intelligence and Expectations of Scholastic Achievement*. Clothing and Textiles Research Journal, vol. 9, no. 4, pages 1–7, 1991.
- [Borod 00] Joan C. Borod. The Neuropsychology of Emotion. Oxford University Press, 2000.
- [Borsboom 12] Barry Borsboom. *Guess Who?: A game to crowd-source the labeling of affective facial expressions*

- is comparable to expert ratings.* mediatechnology.leiden.edu, no. june, 2012. [www](#)
- [Bos 06] Danny Oude Bos. *EEG-based emotion recognition: The influence of visual and auditory stimuli*. Emotion, 2006.
- [Bossard 13] Lukas Bossard, Matthias Dantone & Christian Leistner. *Apparel classification with style*. Computer Vision–ACCV ..., 2013. [www](#)
- [Bradley 07] M M Bradley & P J Lang. *The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual*. ... FL, Tech. Rep ..., 2007.
- [Bragg 13] Jonathan Bragg, Mausam & Daniel S. Weld. *Crowdsourcing Multi-Label Classification for Taxonomy Creation*. Hcomp, pages 25–33, 2013. [www](#)
- [Braun 99] Christoph Braun, Martin Grundl, Claus Marberger & Christoph Scherber. *BeautyCheck*. Rapport technique 1, apr 1999. [www](#)
- [Brew 10] Anthony Brew, Derek Greene & Pádraig Cunningham. *Using crowdsourcing and active learning to track sentiment in online media*. Frontiers in Artificial Intelligence and Applications, vol. 215, pages 145–150, 2010.
- [Bunn 14] Geoffrey C Bunn. *The Truth Machine: A Social History of the Lie Detector*. Technology and Culture, vol. 55, no. 3, pages 752–754, 2014. [www](#)
- [Cannon 27] Walter B. Cannon. *The James-Lange theory of emotions: a critical examination and an alternative theory*. The American journal of psychology, vol. 39, no. 1-4, pages 106–124, 1927. [www](#)
- [Carney 05] Dana R. Carney, Judith a. Hall & Lavonia Smith LeBeau. *Beliefs about the nonverbal expression of social power*. Journal of Nonverbal Behavior, vol. 29, no. 2, pages 105–123, 2005.

- [Castillo 06] Oriana Yuridia Gonzalez Castillo. *Survey About Facial Image Quality*. Rapport technique, Fraunhofer Institute for Computer Graphics Research, 2006.
- [Catellier 14] Andrew A Catellier & Luke Connors. *Web-Enabled Subjective Test (WEST) Research Tools Manual*. Rapport technique, 2014.
- [Chandler 13] Damon M. Chandler. *Seven Challenges in Image Quality Assessment: Past, Present, and Future Research*. ISRN Signal Processing, vol. 2013, pages 1–53, 2013. [www](#)
- [Chanel 11] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt & Thierry Pun. *Emotion assessment from physiological signals for adaptation of game difficulty*. Systems, Man and ... , vol. 41, no. 6, pages 1052–1063, 2011. [www](#)
- [Chang 11] Chih-chung Chang & Chih-jen Lin. *LIBSVM : A Library for Support Vector Machines*. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, pages 1–39, 2011.
- [Chen 09] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang & Chin-Laung Lei. *A crowdsorceable QoE evaluation framework for multimedia content*. Proceedings of the seventeen ACM international conference on Multimedia - MM '09, page 491, 2009. [www](#)
- [Chen 10] Kuan-Ta Chen, Chi-Jui Chang, Chen-Chi Wu, Yu-Chun Chang & Chin-Laung Lei. *Quadrant of euphoria: a crowdsourcing platform for QoE assessment*. IEEE Network, vol. 24, no. 2, pages 28–35, mar 2010. [www](#)
- [Chen 12] Huizhong Chen, Andrew Gallagher & Bernd Girod. *Describing clothing by semantic attributes*. In European Conference on Computer Vision (ECCV), 2012.
- [Chen 13] Frances S Chen, Julia a Minson, Maren Schöne & Markus Heinrichs. *In the eye of the beholder: Eye*

- contact increases resistance to persuasion.* Psychological Science, vol. 24, no. 11, pages 2254–2261, 2013. [www](#)
- [Cheng 08] Quanhua Cheng, Zunxiong Liu & Guoqiang Di. *Facial Gender Classification with Eigenfaces and Least Squares Support Vector Machine.* Journal of Artificial Intelligence, vol. 1, no. 1, pages 28–33, jan 2008. [www](#)
- [Chu 13] Wei-Ta Chu, Yu-Kuang Chen & Kuan-Ta Chen. *Size Does Matter: How Image Size Affects Aesthetic Perception?* vol. 1, 2013. [www](#)
- [Clarke 13] Robert Clarke & Thomas Lancaster. *Commercial aspects of contract cheating.* In Proceedings of the 18th ACM conference on Innovation and technology in computer science education - ITiCSE '13, page 219, New York, New York, USA, 2013. ACM Press. [www](#)
- [Cohen 02] B H Cohen. *Calculating a factorial ANOVA from means and standard deviations.* ... Statistics: Statistical Issues in Psychology, Education, ..., 2002. [www](#)
- [Conesa 95] J Conesa, C Brunold-Conesa & M Miron. *Incidence of the half-left profile pose in single-subject portraits.* Perceptual and motor skills, vol. 81, no. 3 Pt 1, pages 920–2, dec 1995. [www](#)
- [Cosmides 00] Leda Cosmides & John Tooby. *Evolutionary Psychology and the Emotions.* In M. Lewis & J. M. Haviland-Jones, editors, Handbook of Emotions 2nd Edition. 2000. [www](#)
- [Damasio 94] Antonio R Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain.* Putnam Publishing, 1994.
- [Damhorst 90] M. L. Damhorst. *In Search of a Common Thread: Classification of Information Communicated Through Dress.* Clothing and Textiles Research Journal, vol. 8, no. 2, pages 1–12, 1990.

- [Datta 06] Ritendra Datta, Dhiraj Joshi, Jia Li & James Z Wang. *Studying aesthetics in photographic images using a computational approach*. Computer Vision-ECCV 2006, 2006. [www](#)
- [Datta 08] Ritendra Datta, J Li & JZ Z Wang. *Algorithmic inferencing of aesthetics and emotion in natural images: An exposition*. ICIP 2008, pages 8–11, 2008. [www](#)
- [De Moor 14] Katrien De Moor, Filippo Mazza, Isabelle Hupont, Miguel Ríos Quintero, Toni Mäki & Martín Varela. *Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience*. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas & Huib de Ridder, editors, Proc. SPIE, Human Vision and Electronic Imaging, page 90140U, feb 2014. [www](#)
- [Demartini 10] Gianluca Demartini, Malik Muhammad Saad Misen, Roi Blanco & Hugo Zaragoza. *Entity summarization of news articles*. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10, page 795, New York, New York, USA, 2010. ACM Press. [www](#)
- [Deng 09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li & Li Fei-Fei. *ImageNet: A large-scale hierarchical image database*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 2–9, 2009.
- [Desmet 03] Pieter M A Desmet. *Measuring emotions; development and application of an instrument to measure emotional responses to products*. In P C Wright M.A. Blythe, K. Overbeeke, A.F. Monk, editeur, Funology: from Usability to Enjoyment, pages pp. 111–123. Kluwer Academic Publishers, 2003.
- [Desmet 12] P. M. a. Desmet, M.H. Vastenburg, D Van Bel & N. Romero. *Pick-A-Mood; development and application of a pictorial mood-reporting instrument*. In

- Proceedings of the 8th International Design and Emotion Conference, numéro September, pages 11–14, 2012.
- [Dhar 11] Sagnik Dhar, Tamara L. Berg, Stony Brook, Vicente Ordonez & Tamara L. Berg. *High level describable attributes for predicting aesthetics and interestingness*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1657–1664, 2011.
- [Di 13] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu & Neel Sundaresan. *Style Finder: Fine-Grained Clothing Style Detection and Retrieval*. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 8–13, jun 2013. [www](#)
- [Donmez 09] Pinar Donmez, Jaime G Carbonell & Jeff Schneider. *Efficiently learning the accuracy of labeling sources for selective sampling*. Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (2009), page 259, 2009. [www](#)
- [Durst 76] F. Durst, A. Melling & J.H. Whitelaw. Principles and practice of laser-Doppler anemometry. 1976.
- [Eaton 10] Kit Eaton. *MECHANICAL TURK’S UNSAVORY SIDE EFFECT: MASSIVE SPAM GENERATION*, 2010. <http://bit.ly/1jdtmJ5>
- [Edler 01] R J Edler. *Background considerations to facial aesthetics*. Journal of orthodontics, vol. 28, no. 2, pages 159–68, jun 2001. [www](#)
- [Eisenthal 06] Yael Eisenthal, Gideon Dror & Eytan Ruppin. *Facial attractiveness: beauty and the machine*. Neural computation, vol. 18, no. 1, pages 119–42, jan 2006. [www](#)
- [Ekman 87] Paul Ekman & W V Friesen. *Universals and cultural differences in the judgments of facial expres-*

- sions of emotion. Journal of personality and social psychology, vol. 5, no. 4, pages 712–717, 1987. [www](#)
- [Ekman 99] Paul Ekman. *Handbook of Cognition and Emotion, Basic Emotions*. Numeéro 1992, chapitre 3, page 13. 1999.
- [Elance 15] Elance. *Elance Site Usage Policy*, 2015. <http://www.nli.ie/en/site-usage-policy.aspx>
- [Erwin 92] R J Erwin, R C Gur, R E Gur, B Skolnick, M Mawhinney-Hee & J Smailis. *Facial emotion discrimination: I. Task construction and behavioral findings in normal subjects*. Psychiatry research, vol. 42, no. 3, pages 231–240, 1992.
- [Etcoff 11] Nancy L Etcoff, Shannon Stock, Lauren E Haley, Sarah a Vickery & David M House. *Cosmetics as a feature of the extended human phenotype: modulation of the perception of biologically important facial signals*. PloS one, vol. 6, no. 10, page e25656, jan 2011. [www](#)
- [Fedorovskaya 13] Elena a. Fedorovskaya & Huib De Ridder. *Subjective matters: from image quality to image psychology*. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas & Huib de Ridder, editeurs, SPIE Proceedings Vol. 8651, volume 8651, pages 86510O–86510O–11, mar 2013. [www](#)
- [Fei-Fei 07] Li Fei-Fei, Rob Fergus & Pietro Perona. *Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories*. Computer Vision and Image Understanding, vol. 106, no. 1, pages 59–70, 2007.
- [Ferré 07] Sébastien Ferré. *CAMELIS: Organizing and browsing a personal photo collection with a logical information system*. CEUR Workshop Proceedings, vol. 331, pages 108–119, 2007.

- [Figuerola Salas 13] Óscar Figuerola Salas, Velibor Adzic, Akash Shah & Hari Kalva. *Assessing internet video quality using crowdsourcing*. Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia - CrowdMM '13, pages 23–28, 2013. [www](#)
- [Folsom 90] Tyler C. Folsom. *A modular hierarchical neural network for machine vision*. International Joint Conference on Neural Networks, 1990.
- [Forgas 87] J P Forgas & G H Bower. *Mood effects on person-perception judgments*. Journal of personality and social psychology, vol. 53, no. 1, pages 53–60, jul 1987. [www](#)
- [Forster 13] Michael Forster & Gernot Gerger. *The Glasses Stereotype, revisited*. The Jury Expert, vol. 25, no. 2, pages 1–9, 2013.
- [Frijda 09] N. H. Frijda. *Emotion Experience and its Varieties*. Emotion Review, vol. 1, no. 3, pages 264–271, jun 2009. [www](#)
- [Fu 10] Yun Fu, Guodong Guo & Thomas S. Huang. *Age synthesis and estimation via faces: A survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 11, pages 1955–1976, 2010.
- [Gadiraju 15] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze & Gianluca Demartini. *Understanding Malicious Behavior in Crowdsourcing Platforms*. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15, pages 1631–1640, New York, New York, USA, 2015. ACM Press. [www](#)
- [Gardlo 12a] Bruno Gardlo. *Quality of Experience Evaluation Methodology via Crowdsourcing*. Doctoral thesis, University of Zilina, 2012.
- [Gardlo 12b] Bruno Gardlo, Michal Ries, Raimund Schatz, Tobias Hoßfeld & Raimund Schatz. *Microworkers vs. facebook: The impact of crowdsourcing platform*

- choice on experimental results.* In 2012 4th International Workshop on Quality of Multimedia Experience, QoMEX 2012, pages 35–36, 2012.
- [Gloor 07] Pa Gloor, Pierre Dorsaz & Hauke Fuehres. *Analyzing Success of Startup Entrepreneurs by Measuring their Social Network Distance to a Business Networking Hub*. Ickn.Org, 2007. [www](#)
- [Gosling 07] Samuel D Gosling, Sam Gaddis & Simine Vazire. *Personality Impressions Based on Facebook Profiles*. ICWSM, pages 1–4, 2007. [www](#)
- [Grady 10] Catherine Grady & Matthew Lease. *Crowdsourcing document relevance assessment with Mechanical Turk*. In CSLDAMT '10 Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, numéro June, pages 172–179, 2010. [www](#)
- [Greenleaf 92] E A Greenleaf. *Improving rating scale measures by detecting and correcting bias components in some response styles*. Journal of Marketing Research, vol. 29, no. 2, pages 176–188, 1992. [www](#)
- [Gunes 06] H. Gunes & M. Piccardi. *A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior*. In 18th International Conference on Pattern Recognition (ICPR'06), pages 1148–1153. IEEE, 2006. [www](#)
- [Gur 01] Ruben C. Gur, J. Daniel Ragland, Paul J. Moberg, Travis H. Turner, Warren B. Bilker, Christian Kohler, Steven J. Siegel & Raquel E. Gur. *Computerized neurocognitive scanning: I. Methodology and validation in healthy people*. Neuropsychopharmacology, vol. 25, no. 5, pages 766–776, 2001.
- [Haese] Gwenaëlle Haese, Philippe Humeau, Fabrice D E Oliveira, Patrick Le Callet & Pierre L E Cloirec. *Tastes and odors of water - Quantifying objective analyses : a review (In Press)*. Critical Reviews in

- Environmental Science and Technology, pages 1–66.
- [Hatfield 93] Elaine Hatfield, John T Cacioppo & Richard L Rapson. *EMOTIONAL CONTAGION*. Current Directions in Psychological Science, vol. 2, pages 96–99, 1993.
- [Hirth 10] Matthias Hirth, Tobias Hoßfeld & Phuoc Tran-gia. *Cheat-Detection Mechanisms for Crowdsourcing*. Rapport technique August, 2010.
- [Hirth 11a] Matthias Hirth, Tobias Hoßfeld & P Tran-Gia. *Anatomy of a crowdsourcing platform-using the example of microworkers. com*. Innovative Mobile and Internet . . . , pages 1–8, 2011. [www](#)
- [Hirth 11b] Matthias Hirth, Tobias Hoßfeld & P Tran-Gia. *Human Cloud as Emerging Internet Application-Anatomy of the Microworkers Crowdsourcing Platform*. Rapport technique February, 2011. [www](#)
- [Horch 11] Clemens Horch, Christian Keimel & Klaus Diepold. *QualityCrowd. Crowdsourcing for Subjective Video Quality Tests*. Rapport technique, Technische Universität München, 2011.
- [Hoßfeld 11a] Tobias Hoßfeld. *Modeling YouTube QoE based on Crowdsourcing and Laboratory User Studies*. Rapport technique, COST QUALINET STSM report, 2011. [www](#)
- [Hoßfeld 11b] Tobias Hoßfeld, Matthias Hirth & Phuoc Tran-Gia. *Modeling of crowdsourcing platforms and granularity of work organization in Future Internet*. In 2011 23rd International Teletraffic Congress (ITC), pages 142–149, 2011.
- [Hoßfeld 11c] Tobias Hoßfeld, Raimund Schatz, Sebastian Biedermann, Alexander Platzer, Sebastian Egger & Markus Fiedler. *The Memory Effect and Its Implications on Web QoE Modeling*. In 23rd International Teletraffic Congress ITC 2011, 2011.

- [Hoßfeld 14] Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger & Christian Keimel. *Best Practices and Recommendations for Crowdsourced QoE*. Rapport technique, Qualinet White Paper, 2014. [www](#)
- [Hosseini 14] Mahmood Hosseini, Keith Phalp, Jacqui Taylor & Raian Ali. *The four pillars of crowdsourcing: A reference model*. In 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS), pages 1–12, 2014. [www](#)
- [Howe 06] Jeff Howe. *The Rise of Crowdsourcing*. Wired Magazine, no. 14, pages 1–7, 2006. [www](#)
- [Huang 07] Gary B. Huang, Manu Ramesh, Tamara Berg & Erik Learned-Miller. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Rapport technique, University of Massachusetts, 2007. [www](#)
- [Hum 11] Noelle J. Hum, Perrin E. Chamberlin, Brittany L. Hambright, Anne C. Portwood, Amanda C. Schat & Jennifer L. Bevan. *A picture is worth a thousand words: A content analysis of Facebook profile photographs*. Computers in Human Behavior, vol. 27, no. 5, pages 1828–1833, sep 2011. [www](#)
- [Isola 11a] Phillip Isola, Devi Parikh, Antonio Torralba & A. Oliva. *Understanding the Intrinsic Memorability of Images*. NIPS, pages 1–9, 2011. [www](#)
- [Isola 11b] Phillip Isola, Jianxiong Xiao, Antonio Torralba & Aude Oliva. *What makes an image memorable?* Cvpr 2011, no. c, pages 145–152, jun 2011. [www](#)
- [Jain 04] Ramesh Jain. *Quality of experience*. IEEE Multimedia, vol. 11, no. 1, pages 96–95, jan 2004. [www](#)
- [Jain 13] Ankit K. Jain, Can Bal & Truong Q. Nguyen. *Tally: A web-based subjective testing tool*. In Proceedings of Fifth Workshop on Quality of Multimedia Experience (QoMEX2013), pages 128–129, 2013. [www](#)

- [Jang 09] SooCheong (Shawn) Jang & Young Namkung. *Perceived quality, emotions, and behavioral intentions: Application of an extended Mehrabian–Russell model to restaurants*. Journal of Business Research, vol. 62, no. 4, pages 451–460, apr 2009. [www](#)
- [Janssen 09] Joris H. Janssen, Egon L. van den Broek & Joyce H. D. M. Westerink. *Personalized affective music player*. 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pages 1–6, sep 2009. [www](#)
- [Jarque 87] Carlos M Jarque & Anil K Bera. *A Test for Normality of Observations and Regression Residuals*. Science (New York, N.Y.), vol. 55, no. 2, pages 163–172, may 1987. [www](#)
- [Jiang 08] Wenyu Jiang, Sudhir Varma & Richard Simon. *Calculating confidence intervals for prediction error in microarray classification using resampling*. Statistical applications in genetics and molecular biology, vol. 7, no. 1, page Article8, 2008.
- [Johnson 15] Authors Kim K P Johnson & Sharon Lennon. *The Social Psychology of Dress - Berg Fashion Library*, 2015. <http://www.bergfashionlibrary.com/page/The\protect\T1\textdollar0020Social\protect\T1\textdollar0020Psychology\protect\T1\textdollar0020of\protect\T1\textdollar0020Dress/the-social-psychology-of-dress>
- [Joshi 11] Dhiraj Joshi & Ritendra Datta. *Aesthetics and emotions in images*. IEEE Signal Processing Magazine, vol. 28-5, no. SEPTEMBER 2011, pages 94–115, 2011. [www](#)
- [Karger 13] David R. Karger, Sewoong Oh & Devavrat Shah. *Efficient crowdsourcing for multi-class labeling*. ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 1, page 81, 2013. [www](#)

- [Kawrykow 12] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette & Jérôme Waldispühl. *Phylo: A citizen science approach for improving multiple sequence alignment*. PLoS ONE, vol. 7, no. 3, 2012. [www](#)
- [Kazai 12] Gabriella Kazai, Jaap Kamps & Natasa Milic-Frayling. *An analysis of human factors and label accuracy in crowdsourcing relevance judgments*. Information Retrieval, vol. 16, no. 2, pages 138–178, jul 2012. [www](#)
- [Ke] Yan Ke, Xiaoou Tang & Feng Jing. *The Design of High-Level Features for Photo Quality Assessment*. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06), vol. 1, pages 419–426. [www](#)
- [Keimel 12] Christian Keimel, Julian Habigt & Klaus Diepold. *Challenges in crowd-based video quality assessment*. Quality of Multimedia ..., pages 13–18, 2012. [www](#)
- [Kemelmacher-Shlizerman 11] Ira Kemelmacher-Shlizerman, Eli Shechtman, Rahul Garg & Steven M. Seitz. *Exploring photo-bios*. ACM Transactions on Graphics, vol. 30, no. 4, page 1, 2011.
- [Khan 12] Shehroz Khan & Daniel Vogel. *Evaluating visual aesthetics in photographic portraiture*. Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging (CAe '12), pages 55–62, 2012. [www](#)
- [Khosla 12] Aditya Khosla, Jianxiong Xiao, Antonio Torralba & Aude Oliva. *Memorability of Image Regions*. Nips, no. 1, pages 1–9, 2012. [www](#)
- [Khosla 13] Aditya Khosla, Wilma a Bainbridge, Antonio Torralba & Aude Oliva. *Modifying the Memorability of Face Photographs*. In International Conference on Computer Vision (ICCV), 2013. [www](#)

- [Kistler 98] a Kistler, C Mariauzouls & K von Berlepsch. *Fingertip temperature as an indicator for sympathetic responses*. International journal of psychophysiology : official journal of the International Organization of Psychophysiology, vol. 29, no. 1, pages 35–41, jun 1998. [www](#)
- [Kittur 08] A Kittur, E H Chi & B Suh. *Crowdsourcing user studies with Mechanical Turk*. Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pages 453–456, 2008. [www](#)
- [Kivran-Swaine 14] F Kivran-Swaine, Jeremy Ting & JR R Brubaker. *Understanding Loneliness in Social Awareness Streams: Expressions and Responses*. In 8TH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2014. [www](#)
- [Kleisner 13] Karel Kleisner, Lenka Priplatova, Peter Frost & Jaroslav Flegr. *Trustworthy-looking face meets brown eyes*. PloS one, vol. 8, no. 1, page e53285, jan 2013. [www](#)
- [Koelstra 12] Sander Koelstra, C Muhl, Mohammad Soleymani, Jong-seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt & Ioannis Patras. *DEAP: A Database for Emotion Analysis using Physiological Signals*. Affective Computing, ..., pages 1–15, 2012. [www](#)
- [Kohavi 97] R Kohavi & R Kohavi. *Wrappers for feature subset selection*. Artificial Intelligence, vol. 97, no. 1-2, pages 273–324, 1997. [www](#)
- [Kondo 09] Narihiko Kondo, NAS A S Taylor & M Shibasaki. *Thermoregulatory adaptation in humans and its modifying factors*. vol. 13, pages 35–41, 2009. [www](#)
- [Kreibig 07] Sylvia D Kreibig, Frank H Wilhelm, Walton T Roth & James J Gross. *Cardiovascular, electrodermal, and respiratory response patterns to fear-*

- and sadness-inducing films.* Psychophysiology, vol. 44, no. 5, pages 787–806, sep 2007. [www](#)
- [Krishnamachari 98] Santhana Krishnamachari. *Hierarchical clustering algorithm for fast image retrieval.* Proceedings of SPIE, no. 13, pages 427–435, 1998. [www](#)
- [Krizhevsky 12] Alex Krizhevsky, I Sutskever & GE Hinton. *Imagenet classification with deep convolutional neural networks.* Advances in neural ... , pages 1–9, 2012. [www](#)
- [Kuikkaniemi 11] Kai Kuikkaniemi. *White paper : Crowdsourcing in Media Industry - Quality and Reward Mechanisms.* Rapport technique, 2011.
- [Kumar 08] Neeraj Kumar, Peter Belhumeur & Shree Nayar. *FaceTracer: A Search Engine for Large Collections of Images with Faces.* In Computer Vision – ECCV 2008, pages 340–353, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. [www](#)
- [Kumar 09] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur & Shree K Nayar. *Attribute and simile classifiers for face verification.* 2009 IEEE 12th International Conference on Computer Vision, pages 365–372, sep 2009. [www](#)
- [Kumar 11] Neeraj Kumar, A Berg, P. N. Belhumeur & S. Nayar. *Describable visual attributes for face verification and image search.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 10, pages 1962–1977, oct 2011. [www](#)
- [Kvandal 06] Per Kvandal, Svein Aslak Landsverk, Alan Bernjak, Aneta Stefanovska, Hebe Désirée Kvernmo & Knut Arvid Kirkebø en. *Low-frequency oscillations of the laser Doppler perfusion signal in human skin.* Microvascular research, vol. 72, no. 3, pages 120–127, nov 2006. [www](#)
- [Kvernmo 98] H D Kvernmo, A Stefanovska, M Bracic, K A Kirkebø en & K Kvernebo. *Spectral analysis of the*

- laser Doppler perfusion signal in human skin before and after exercise*. Microvascular research, vol. 56, no. 3, pages 173–82, nov 1998. [www](#)
- [Lang 80] PJ Lang. *Behavioral treatment and bio-behavioral assessment: computer applications*. In Technology in mental health care delivery systems, pages 119–137. 1980. [www](#)
- [Lang 97] PJ Lang, MM Bradley & B.N. Cuthbert. *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*, 1997.
- [Lassalle 12] Julie Lassalle, Laetitia Gros, Thierry Morineau & Gilles Coppin. *Impact of the content on subjective evaluation of audiovisual quality: What dimensions influence our perception?* In IEEE international Symposium on Broadband Multimedia Systems and Broadcasting, pages 1–6. IEEE, jun 2012. [www](#)
- [Lazebnik 06] Svetlana Lazebnik, Cordelia Schmid & Jean Ponce. *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pages 2169–2178, 2006.
- [Le Callet 12] Patrick Le Callet, Sebastian Möller & Andrew Perakis. Qualinet White Paper on Definitions of Quality of Experience, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), version 1.2. 2012.
- [Leo 05] Leonel De Oliveira Junior Leo. *Captura e Alinhamento de Imagens : Um Banco de Faces Brasileiro*. Rapport technique, Centro Universitario da FEI, 2005. [www](#)
- [Leray 98] Philippe Leray & Patrick Gallinari. *Feature Selection with Neural Networks Feature Selection with Neural Networks*. Behaviormetrika, vol. 26, pages 16—6, 1998. [www](#)

- [Li 02] Xin Li Xin Li. *Blind image quality assessment*. In Proceedings of International Conference on Image Processing (ICIP), volume 1, pages 449–452, 2002.
- [Li 09] Mu Li & Bao-Liang Lu. *Emotion classification based on gamma-band EEG*. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, vol. 2009, pages 1323–6, jan 2009. [www](#)
- [Li 10a] Congcong Li & Andrew Gallagher. *Aesthetic quality assessment of consumer photos with faces*. Proceedings of IEEE ... , pages 3–6, 2010. [www](#)
- [Li 10b] Congcong Li, Andrew Gallagher, Alexander C. Loui & Tsuhan Chen. *Aesthetic quality assessment of consumer photos with faces*. In 2010 IEEE International Conference on Image Processing, pages 3221–3224. IEEE, sep 2010. [www](#)
- [Li 13] Jing Li, Marcus Barkowsky & Patrick Le Callet. *Subjective assessment methodology for preference of experience in 3DTV*. In IVMSP Workshop, 2013 IEEE ... , 2013. [www](#)
- [Lienhard 14] Arnaud Lienhard, Marion Reinhard, Alice Caplier & Patricia Ladret. *Photo Rating of Facial Pictures based on Image Segmentation*. In 9th Int. Conf. on computer Vision Theory and Applications, VISAPP, pages 329–336, Lisbonne, 2014. [www](#)
- [Lienhard 15a] Arnaud Lienhard, Patricia Ladret & Alice Caplier. *How to predict the global instantaneous feeling induced by a facial picture?* Signal Processing: Image Communication, pages 1–14, 2015. [www](#)
- [Lienhard 15b] Arnaud Lienhard, Patricia Ladret & Alice Caplier. *Low Level Features for Quality Assessment of Facial Images*. In Computer Vision Theory and Applications (VISAPP), 2015.

- [Lintott 08] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray & Jan Vandenberg. *Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey*. Monthly Notices of the Royal Astronomical Society, vol. 389, no. 3, pages 1179–1189, 2008. [www](#)
- [Loni 13] Babak Loni, Martha Larson, Alessandro Bozzon & Luke Gottlieb. *Crowdsourcing for Social Multimedia at MediaEval 2013 : Challenges, Data set, and Evaluation*. In Proceedings of MediaEval 2013., pages 1–2, 2013.
- [Lövheim 12] Hugo Lövheim. *A new three-dimensional model for emotions and monoamine neurotransmitters*. Medical hypotheses, vol. 78, no. 2, pages 341–8, feb 2012. [www](#)
- [Luo 01] Jiebo Luo Jiebo Luo & a. Savakis. *Indoor vs outdoor classification of consumer photographs using low-level and semantic features*. Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205), vol. 2, pages 745–748, 2001.
- [Luo 08] Yiwen Luo & Xiaoou Tang. *Photo and video quality evaluation: Focusing on the subject*. In Computer Vision–ECCV 2008, pages 386–399, 2008. [www](#)
- [MACLEAN 52] P D MACLEAN. *Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain)*. Electroencephalography and clinical neurophysiology, vol. 4, no. 4, pages 407–418, 1952.
- [Maji 11] Subhransu Maji. *Large Scale Image Annotations on Amazon Mechanical Turk*. 2011.
- [Males 13] M Males, A Hedi & M Grgic. *Aesthetic quality assessment of headshots*. ELMAR, 2013 55th International ..., no. September, pages 25–27, 2013. [www](#)

- [Manant 14] Matthieu Manant, Serge Pajak & Nicolas Soulié. *Do recruiters 'like' it? Online social networks and privacy in hiring: a pseudo-randomized experiment*. Rapport technique 56845, Munich Personal RePEc Archive (MPRA), 2014. [www](#)
- [Mancas 13] Matei Mancas & Olivier Le Meur. *Memorability of natural scenes: The role of attention*. In 2013 IEEE International Conference on Image Processing, pages 196–200. IEEE, sep 2013. [www](#)
- [Maniewski 99] R Maniewski, P Leger, P Lewandowski, A Liebert, P Bendayan, H Boccalon, L Bajorski & K O Möller. *Spectral analysis of laser-Doppler perfusion signal measured during thermal test*. Technology and health care : official journal of the European Society for Engineering and Medicine, vol. 7, no. 2-3, pages 163–9, jan 1999. [www](#)
- [Markic 09] Olga Markic. *Rationality and emotions in decision making*. vol. 7, no. 2, pages 54–64, 2009.
- [Massey 51] Frank J. Jr. Massey. *The Kolmogorov-Smirnov test for goodness of fit*. Journal of the American statistical Association, vol. 46, no. 253, pages 68–78, 1951. [www](#)
- [Mateo 13] Maria Pilar Perla Mateo. *Emotracker, emociones ante la pantalla*. Tercermilenio, 2013.
- [Mauss 09] Iris B Mauss & Michael D Robinson. *Measures of emotion: A review*. Cognition & emotion, vol. 23, no. 2, pages 209–237, feb 2009. [www](#)
- [Mazza 14] Filippo Mazza, Matthieu Perreira Da Silva & Patrick Le Callet. *Would you hire me? Selfie portrait images perception in a recruitment context*. In Bernice E Rogowitz, Thrasyvoulos N Pappas & Huib de Ridder, editors, Proc. SPIE 9014, Human Vision and Electronic Imaging XIX, 90140X, numéro February, pages 2–6, feb 2014. [www](#)
- [Mazza 15] Filippo Mazza, Matthieu Perreira Da Silva, Patrick Le Callet & Ingrid E. J. Heynderickx. *What do you*

- think of my picture? Investigating factors of influence in profile images context perception.* In Proc. SPIE 9394, Human Vision and Electronic Imaging XX, 2015. [www](#)
- [Meeker 14] Mary Meeker. *Internet Trends 2014 – Code Conference KPBC*. Rapport technique, 2014. [www](#)
- [Michel 15] Franck Michel. *Flickr unofficial statistics*, 2015. <https://www.flickr.com/photos/franckmichel/6855169886>
- [Mohammadi 12] Gelareh Mohammadi & Alessandro Vinciarelli. *Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features*. IEEE Transactions on Affective Computing, vol. 3, no. 3, pages 273–284, jul 2012. [www](#)
- [Motoyama 11] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage & Geoffrey M. Voelker. *Dirty Jobs: The Role of Freelance Labor in Web Service Abuse*. In Proceedings of the 20th USENIX Conference on Security, page 14, San Francisco, CA, 2011. Association, USENIX. [www](#)
- [Mühl 11] C Mühl & E van den Broek. *Multi-modal affect induction for affective brain-computer interfaces*. In Affective Computing and Intelligent Interaction, pages 235–245, 2011. [www](#)
- [Murray 12] Naila Murray, Luca Marchesotti & Florent Perronnin. *AVA: A large-scale database for aesthetic visual analysis*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2408–2415, 2012.
- [Nakasone 05] Arturo Nakasone, Mitsuru Ishizuka & Helmut Prendinger. *Emotion recognition from electromyography and skin conductance*. In Proceedings 5th International Workshop on Biosignal Interpretation, pages 219–222, 2005.
- [Nie 11] Dan Nie, XW W Wang, LC C Shi & BL L Lu. *EEG-based emotion recognition during watching movies*.

- Neural Engineering (NER), ..., pages 667–670, 2011. [www](#)
- [Niedenthal 09] Paula M Niedenthal, Piotr Winkielman, Laurie Mondillon & Nicolas Vermeulen. *Embodiment of emotion concepts*. Journal of personality and social psychology, vol. 96, no. 6, pages 1120–36, jun 2009. [www](#)
- [Niemic 02] Christopher P Niemic, Advisor Kirk, Warren Brown & D Ph. *Studies of Emotion: A Theoretical and Emperical Review of Psychophysiological Studies of Emotion*. Journal of Undergraduate Research, pages 15–18, 2002.
- [Nieuwenhuysen 14] Paul Nieuwenhuysen. *Search by image through the Internet : applications and limitations*. In Seventh Shangai International Library Forum, numéro July, pages 145–155, 2014.
- [Nourbakhsh 12] Nargess Nourbakhsh, Yang Wang, Fang Chen & Rafael A Calvo. *Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks*. In OZCHI’12, pages 1–4, Melbourne, Australia, 2012. [www](#)
- [Nowak 10] Stefanie Nowak & Stefan Ruger. *How reliable are annotations via crowdsourcing? a study about inter-annotator agreement for multi-label image annotation*. In The 11th ACM International Conference on Multimedia Information Retrieval (MIR), pages 29–31, Philadelphia, USA, 2010. [www](#)
- [O’Connor 15] Maria Francesca O’Connor & Laurel D Riek. *Detecting Social Context : A Method for Social Event Classification Using Naturalistic Multimodal Data*. In CBAR 2015, 2015. [www](#)
- [O’Hagan 98] Fiann O’Hagan. *No Reason Without Emotion.pdf*, 1998. http://www.informatics.sussex.ac.uk/research/groups/nlp/gazdar/teach/atc/1998/revman/o_hagan.html

- [Oyama 13] Satoshi Oyama, Yukino Baba, Yuko Sakurai & Hisashi Kashima. *Accurate integration of crowd-sourced labels using workers' self-reported confidence scores*. IJCAI International Joint Conference on Artificial Intelligence, pages 2554–2560, 2013.
- [Pandey 13] Shreelekha Pandey. *A Hierarchical Clustering Approach for Image Datasets*. 2013.
- [Pathak 11] Sujata Pathak & Arun Kulkarni. *Recognizing emotions from speech*. In 2011 3rd International Conference on Electronics Computer Technology, pages 107–109. IEEE, apr 2011. [www](#)
- [Peng 13] Jian Peng, Qiang Liu, Alexander Ihler & Bonnie Berger. *Crowdsourcing for structured labeling with applications to protein folding*. ICML Workshop on Machine Learning Meets Crowdsourcing, pages 2008–2012, 2013.
- [Picard 97] Rosalind W. Picard. *Affective Computing*. Numéro 321. Cambridge, 1997.
- [Picard 01] Rosalind W. Picard, Elias Vyzas & Jennifer Healey. *Toward machine emotional intelligence: Analysis of affective physiological state*. ... and Machine Intelligence, ..., vol. 23, no. 10, pages 1175–1191, 2001. [www](#)
- [Picard 10] Rosalind W. Picard. *Affective Computing: From Laughter to IEEE*. IEEE Transactions on Affective Computing, vol. 1, no. 1, pages 11–17, jan 2010. [www](#)
- [Pigeau 10] Antoine Pigeau. *Incremental and hierarchical classification of a personal image collection on mobile devices*. Journal of Multimedia Tools and Applications, vol. 46, no. 2-3, pages 289–306, 2010. [www](#)
- [Plutchik 01] Robert Plutchik. *The Nature of Emotions*. American Scientist, 2001.

- [Pogačnik 12] D Pogačnik, Robert Ravnik, Narvika Bovcon & Franc Solina. *Evaluating photo aesthetics using machine learning*. Data Mining and Data Warehouses, pages 4–7, 2012. [www](#)
- [Povoa 14] Isabel Povoa. *Evaluating the impact of digital filters on the aesthetic appeal of photographs: a crowdsourcing-based approach*. PhD thesis, TUDelft, 2014.
- [Quercia 11] Daniele Quercia, Michal Kosinski, David Stillwell & Jon Crowcroft. *Our twitter profiles, our selves: Predicting personality with twitter*. Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011, pages 180–185, 2011.
- [Rabin 99] M Rabin & J L Schrag. *First impressions matter: A model of confirmatory bias*. The Quarterly Journal of Economics, no. February, 1999. [www](#)
- [Rada 95] H Rada, a Dittmar, G Delhomme, C Collet, R Roure, E Vernet-Maury & a Priez. *Bioelectric and microcirculation cutaneous sensors for the study of vigilance and emotional response during tasks and tests*. Biosensors & bioelectronics, vol. 10, no. 1-2, pages 7–15, jan 1995. [www](#)
- [Rainville 06] Pierre Rainville, Antoine Bechara, Nasir Naqvi & Antonio R Damasio. *Basic emotions are associated with distinct patterns of cardiorespiratory activity*. International journal of psychophysiology, vol. 61, no. 1, pages 5–18, jul 2006. [www](#)
- [Ramsey 02] Fred L. Ramsey & Daniel W. Schafer. *The Statistical Sleuth*. Duxbury Thomson Learning, 2nd editio edition, 2002.
- [Raykar 11] Vikas C. Raykar & Shipeng Yu. *An entropic score to rank annotators for crowdsourced labeling tasks*. In Proceedings - 2011 3rd National Conference on

- Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2011, pages 29–32, 2011.
- [Redi 13a] Judith A. Redi & Isabel Pova. *THE ROLE OF VISUAL ATTENTION IN THE AESTHETIC APPEAL OF CONSUMER IMAGES : A PRELIMINARY STUDY*. In VCIP 2013, 2013.
- [Redi 13b] Miriam Redi. *Novel Methods for Semantic and Aesthetic Multimedia Retrieval*. PhD thesis, 2013.
- [Redi 14] Judith Redi & Isabel Pova. *Crowdsourcing for Rating Image Aesthetic Appeal: Better a Paid or a Volunteer Crowd?* In Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia - CrowdMM '14, pages 25–30, 2014.
[www](#)
- [Redi 15] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal & Alejandro Jaimes. *The Beauty of Capturing Faces: Rating the Quality of Digital Portraits*. In Eleventh IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015), 2015.
- [Ren 13] Peng Ren, Armando Barreto, Ying Gao & Malek Adjouadi. *Affective Assessment by Digital Processing of the Pupil Diameter*. IEEE Transactions on Affective Computing, vol. 4, no. 1, pages 2–14, jan 2013. [www](#)
- [Ribeiro 11] F Ribeiro. *Crowdsourcing subjective image quality evaluation*. 18th Image Processing (ICIP), pages 3158–3161, 2011. [www](#)
- [Riek 11] Laurel D. Riek, Maria F. O'Connor & Peter Robinson. *Guess what? A game for affective annotation of video using crowd sourcing*. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 6974 LNCS, pages 277–285. 2011.

- [Ross 10] Joel Ross & E Al. *Who are the crowdworkers?: shifting demographics in mechanical turk*. In CHI'10 extended abstracts on Human factors in computing systems. ACM, 2010, 2010.
- [Rossi 13] Peter E Rossi, Zvi Gilula & Greg M Allenby. *Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach*. vol. 96, no. 453, pages 20–31, 2013.
- [Rossion 12] Bruno Rossion, Bernard Hanseeuw & Laurence Dricot. *Defining face perception areas in the human brain: a large-scale factorial fMRI face localizer analysis*. Brain and cognition, vol. 79, no. 2, pages 138–57, jul 2012. [www](#)
- [Russell 78] James A. Russell. *Evidence of convergent validity on the dimensions of affect*. Journal of Personality and Social Psychology, vol. 36, no. 10, pages 1152–1168, 1978. [www](#)
- [Russell 08] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy & William T. Freeman. *LabelMe: a database and web-based tool for image annotation*. International Journal of Computer Vision, vol. 77, no. 1-3, pages 157–173, 2008.
- [Satu 14] Toru Satu. *Notes on Four Theories of Emotion*, 2014. <http://webspace.ship.edu/tosato/emotion.htm>
- [Scherer 05] K. R. Scherer. *What are emotions? And how can they be measured?* Social Science Information, vol. 44, no. 4, pages 695–729, dec 2005. [www](#)
- [Sedghi 14] Ami Sedghi. *Facebook: 10 years of social networking, in numbers*, 2014.
- [Sermanet 13] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus & Yann LeCun. *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks*. dec 2013. [www](#)

- [Shah 12] Rajvi Shah & Vivek Kwatra. *All smiles: automatic photo enhancement by facial expression analysis*. Proceedings of the 9th European Conference on . . . , 2012. [www](#)
- [Shamir 14] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M. von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller & John Wallin. *Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls*. The Journal of the Acoustical Society of America, vol. 135, no. 2, pages 953–962, 2014. [www](#)
- [Sheng 08] Victor S Sheng, Foster Provost & Panagiotis G Ipeirotis. *Get another label? improving data quality and data mining using multiple, noisy labelers*. New York, pages 614–622, 2008. [www](#)
- [Shivakumar 12] G Shivakumar & P A Vijaya. *Emotion Recognition Using Finger Tip Temperature: First Step towards an Automatic System*. ijcee.org, vol. 4, no. 3, pages 252–255, 2012. [www](#)
- [Sim 02] Terence Sim, Simon Baker & Maan Bsat. *The CMU Pose, Illumination, and Expression (PIE) database*. In Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, numéro 1, pages 53–58. IEEE, 2002. [www](#)
- [Skeels 09] Meredith M. Skeels & Jonathan Grudin. *When social networks cross boundaries*. In Proceedinfs of the ACM 2009 international conference on Supporting group work - GROUP '09, page 95, New York, New York, USA, 2009. ACM Press. [www](#)
- [Snow 08] Rion Snow, B O'Connor, Daniel Jurafsky & AY Y Ng. *Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. . . . methods in natural language . . .*, 2008. [www](#)
- [Sokal 62] Robert R. Sokal & F. James Rohlf. *The Comparison of Dendrograms by Objective Methods*. Taxon, vol. 11, no. 2, pages 33–40, 1962. [www](#)

- [Soleymani 08] Mohammad Soleymani, Guillaume Chanel, Joep J. M. Kierkels & Thierry Pun. *Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses*. 2008 Tenth IEEE International Symposium on Multimedia, pages 228–235, dec 2008. [www](#)
- [Soleymani 12] M. Soleymani, J. Lichtenauer, T. Pun & M. Pantic. *A Multimodal Database for Affect Recognition and Implicit Tagging*. IEEE Transactions on Affective Computing, vol. 3, no. 1, pages 42–55, jan 2012. [www](#)
- [Stathopoulou] I O Stathopoulou & G A Tsihrintzis. *A NEURAL NETWORK-BASED FACIAL EXPRESSION ANALYSIS SYSTEM*. pages 2–5. [www](#)
- [Steiner 11] Thomas Steiner, Ruben Verborgh, Rik Van De Walle, Michael Hausenblas & Joaquim Gabarró Vallés. *Crowdsourcing Event Detection in YouTube Videos*. In Proceedings of Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), pages 1–10, Bonn, Germany, 2011. [www](#)
- [Stork 99] D.G. Stork. *Character and document research in the Open Mind Initiative*. In Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318), 1999.
- [Sun 13] Yi Sun, Xiaogang Wang & Xiaoou Tang. *Deep convolutional network cascade for facial point detection*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3476–3483, 2013.
- [Szummer 98] Martin Szummer & Rosalind W. Picard. *Indoor-outdoor image classification*. Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, no. 445, 1998.

- [Takahashi 04] K. Takahashi. *Remarks on emotion recognition from multi-modal bio-potential signals*. 2004 IEEE International Conference on Industrial Technology, 2004. IEEE ICIT '04., vol. 3, pages 1138–1143, 2004. [www](#)
- [Tang 13] Xiaou Tang, Wei Luo & Xiaogang Wang. *Content-based photo quality assessment*. IEEE Transactions on Multimedia, vol. 15, no. 8, pages 1930–1943, 2013.
- [Tarasov 10] Alexey Tarasov, Charlie Cullen & S J Delany. *Using crowdsourcing for labelling emotional speech assets*. Science, no. 09, pages 1–5, 2010. [www](#)
- [Times 15] Financial Times. *Crowdsourcing in Financial Times Lexicon*, 2015. <http://lexicon.ft.com/Term?term=crowdsourcing>
- [Todorov 05] Alexander Todorov, Anesu N Mandisodza, Amir Goren & Crystal C Hall. *Inferences of competence from faces predict election outcomes*. Science (New York, N.Y.), vol. 308, no. 5728, pages 1623–6, jun 2005. [www](#)
- [Todorov 11] Alexander Todorov, Christopher P. Said & Sara C. Verosky. *Personality impressions from facial appearance*. In Handbook of ..., pages 1–48. 2011. [www](#)
- [Tong 05] Hanghang Tong, Mingjing Li, Hj Zhang, Jingrui He & C Zhang. *Classification of digital photos taken by photographers or home users*. Advances in Multimedia ..., pages 198–205, 2005. [www](#)
- [Totti 14] Luam Totti, Felipe Costa, Sandra Avila, Eduardo Valle, Wagner Jr. Meira & Virgilio Almeida. *The Impact of Visual Attributes on Online Image Diffusion*. In ACM Web Science Conference (WebSci), 2014. [www](#)
- [Toy 10] Toyama MICT Image Quality Evaluation Database, 2010. <http://mict.eng.u-toyama.ac.jp/mictdb.html>

- [Trzepacz 93] Paula T. Trzepacz & Robert W. Baker. The Psychiatric Mental Status Examination. 1st edition, 1993.
- [Twitter Inc. 15] Twitter Inc. *Twitter Privacy Policy*, 2015. <https://twitter.com/privacy?lang=EN>
- [Valentin 97] Dominique Valentin, Hervé Abdi, Betty Edelman & Alice J. O’Toole. *Principal Component and Neural Network Analyses of Face Images: What Can Be Generalized in Gender Classification?* Journal of Mathematical Psychology, vol. 41, no. 4, pages 398–413, dec 1997. [www](#)
- [Van De Weijer 07] Joost Van De Weijer, Cordelia Schmid & Jakob Verbeek. *Learning color names from real-world images*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
- [Vazire 04] Simine Vazire & Samuel D Gosling. *e-Perceptions: personality impressions based on personal web-sites*. Journal of personality and social psychology, vol. 87, no. 1, pages 123–132, 2004.
- [Viinikainen 12] Mikko Viinikainen, Jari Kätsyri & Mikko Sams. *Representation of perceived sound valence in the human brain*. Human brain mapping, vol. 33, no. 10, pages 2295–305, oct 2012. [www](#)
- [Viola] P. Viola & M. Jones. *Rapid object detection using a boosted cascade of simple features*. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, pages I–511–I–518. IEEE Comput. Soc. [www](#)
- [von Ahn 04] Luis von Ahn & Laura Dabbish. *Labeling images with a computer game*. ACM Conference on Human Factors in Computing Systems, pages 319 – 326, 2004. [www](#)
- [von Ahn 06] L. von Ahn. *Games with a Purpose*. Computer, vol. 39, no. 6, pages 92–94, jun 2006. [www](#)

- [von Ahn 08] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham & Manuel Blum. *reCAPTCHA: human-based character recognition via Web security measures*. Science (New York, N.Y.), vol. 321, no. 5895, pages 1465–1468, 2008.
- [Vugt 10] Henriette C. Van Vugt, Jeremy N. Bailenson, Johan F. Hoorn & Elly a. Konijn. *Effects of facial similarity on user responses to embodied agents*. ACM Transactions on Computer-Human Interaction, vol. 17, no. 2, pages 1–27, may 2010. [www](#)
- [Wang 12] Gang Wang, Christo Wilson, X Zhao & Yibo Zhu. *Serf and turf: crowdturfing for fun and profit*. Proceedings of the 21st international conference on World Wide Web, pages Pages 679–688, 2012. [www](#)
- [Welinder 10] Peter Welinder & Pietro Perona. *Online crowdsourcing: Rating annotators and obtaining cost-effective labels*. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, pages 25–32, 2010.
- [Whitehill 08] Jacob Whitehill & Javier R Movellan. *Personalized facial attractiveness prediction*. 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, pages 1–7, sep 2008. [www](#)
- [Willis 06] Janine Willis & Alexander Todorov. *First impressions: making up your mind after a 100-ms exposure to a face*. Psychological science, vol. 17, no. 7, pages 592–8, jul 2006. [www](#)
- [Wong 09] Lai Kuan Wong & Kok Lim Low. *Saliency-enhanced image aesthetics class prediction*. Proceedings - International Conference on Image Processing, ICIP, pages 997–1000, 2009.
- [Wood 12] Mark D. Wood & Minwoo Park. *Exploring Photos in Facebook*. 2012 IEEE International Symposium on Multimedia, pages 88–91, dec 2012. [www](#)

- [Wu 11] Shaomei Wu & Tao Lin. *Exploring the use of physiology in adaptive game design*. 2011 International Conference on Consumer Electronics, Communications and Networks (CECNet), pages 1280–1283, apr 2011. [www](#)
- [Wu 13] Chen-Chi Wu, Kuan-Ta Chen, Yu-Chun Chang & Chin-Laung Lei. *Crowdsourcing Multimedia QoE Evaluation: A Trusted Framework*. IEEE Transactions on Multimedia, vol. 15, no. 5, pages 1121–1137, aug 2013. [www](#)
- [Xue 13] Shao-fu Xue, Henry Tang, Dan Tretter, Qian Lin & Jan Allebach. *Feature design for aesthetic inference on photos with faces*. In 2013 IEEE International Conference on Image Processing, pages 2689–2693. IEEE, sep 2013. [www](#)
- [Yale Univ 01] . Yale Univ. *The Yale Face Database B*, 2001. <http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/YaleFaceDatabase.htm>
- [Yamaguchi 12] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz & Tamara L. Berg. *Parsing clothing in fashion photographs*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, no. Fig 1, pages 3570–3577, 2012.
- [Yang 12] N Yang, R Muraleedharan & J Kohl. *SPEECH-BASED EMOTION CLASSIFICATION USING MULTICLASS SVM WITH HYBRID KERNEL AND THRESHOLDING FUSION*. ece.rochester.edu, pages 455–460, 2012. [www](#)
- [Zhong 08] Mingjun Zhong, Fabien Lotte, Mark Girolami & A Lécuyer. *Classifying EEG for brain computer interfaces using gaussian processes*. Pattern Recognition Letters, vol. 3, pages 354–359, 2008. [www](#)
- [Zhou 00] Xiang S Zhou. *CBIR: from low-level features to high-level semantics*. Proceedings of SPIE-The International Society for Optical Engineering, vol. 3974, pages 426–431, 2000. [www](#)

Thèse de Doctorat

Filippo MAZZA

Influence of image features on face portraits social context interpretation: experimental methods, crowdsourcing based studies and models

Influence des caractéristiques des images de portrait sur l'interprétation de leur contexte social: méthodologie expérimentale, évaluation par crowdsourcing et modèles

Résumé

Les réseaux sociaux occupent une part croissante de notre vie quotidienne. Sur ces réseaux, les participants constituent un profil virtuel leur permettant d'interagir avec d'autres dans un but précis : rester en contact avec ses amis, entretenir son réseau professionnel, trouver l'amour... Ce profil doit être cohérent avec le contexte dans lequel il est utilisé. Et en particulier la photo de profil, car différentes images de la même personne peuvent transmettre des messages très différents. Dans cette thèse, nous étudions les éléments des photos de portrait pouvant modifier la perception du contexte d'utilisation le plus adapté. Nous définissons ce concept en empruntant la notion de "contexte social" à la psychologie. Pour notre étude, nous prenons en compte des caractéristiques d'image de bas et haut niveau. Les premières sont directement liées aux valeurs des pixels de l'image (ex : contraste). Les secondes sont quant à elles liées à l'interprétation de la scène (ex : influence de l'habillement lors des interactions sociales). L'étiquetage des caractéristiques de haut niveau a été réalisé par crowdsourcing, une technique récente exploitant la puissance du web afin d'externaliser des tâches simples. Nous avons exploité cette même technique afin de recueillir les évaluations du contexte social dans lequel nos images de portrait seraient le plus à même d'être utilisées. Puis nous avons modélisé les liens entre les différentes caractéristiques des images et le contexte social. Il a ainsi été possible de quantifier l'influence de chaque caractéristique, les résultats obtenus étant cohérents avec l'expérience empirique.

Mots clés

Photos de portrait, contexte sociale, crowdsourcing, caractéristiques d'images, apprentissage automatique.

Abstract

Online communities and social networks are more and more present in everyday life. On these networks, people build a virtual profile and interact between them for many different purposes: to be in touch with friends, for business, to make new connections, to find a love partner... Online profiles should be coherent with these tasks, starting with the omnipresent profile picture, as different pictures of the same subject can convey very different messages. This thesis focuses on which elements inside a profile picture modify the perception of the context that best suits the picture itself. We define this concept borrowing the "social context" concept in psychology. Image features considered are both low and high level; while the first are more technical quantities related to the sole pixels values (i.e. brightness, contrast), the latter are related to the understanding of the scene depicted in the picture. These elements are underlined by results of research on psychology (i.e. influence of clothing or gaze direction in social interactions). These features have been evaluated through crowdsourcing, a relatively new technique that exploits the power of the web to outsource simple tasks. We adopted the same technique to gather social context evaluations, being this a subjective perception. Then, through different mathematical approaches, we modelled the social context with image features, to understand and quantify the influence of each feature. Results are in line with empirical experience.

Key Words

Face portraits, social context, crowdsourcing, image features, machine learning.