Modeling user factors in multimedia preferences

Guntuku, Sharath Chandra

2017

Guntuku, S. C. (2017). Modeling user factors in multimedia preferences. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/72810>

<https://doi.org/10.32657/10356/72810>

MODELING USER FACTORS IN MULTIMEDIA PREFERENCES

Sharath Chandra Guntuku

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement
for the degree of Doctor of Philosophy

2017

Acknowledgments

In the course of pursuing a PhD, I got the opportunity to meet, interact and work with some wonderful people. I would like to acknowledge and thank those who helped me along the way:

- To Prof. Lin Weisi - for giving me the opportunity to pursue PhD. You have been extremely supportive, encouraging and have always been very open to what I wanted to work on. Your down-to-earth nature and positive outlook have been a great inspiration for me. Thank you for creating a warm and comfortable environment in the research group, for giving me feedback on my papers upon such short notice and for always welcoming new ideas (and suggestions) on research topics (no matter how tangential they might appear in the first go). I feel fortunate to have a supervisor like you.
- To Prof. George Ghinea - for encouraging and supporting the perception study with videos. I enjoy working with you (and also travelling with you!). Thanks for providing the necessary confidence to push through the ideas and work when I most needed it!
- To Dr. Sujoy Roy - for teaching me how to be a researcher and guiding me through it. I had a very skewed idea of what research is and you very patiently helped me to unlearn my misconceptions and learn the right way to do research - starting from how to read papers with a critical eye, how to identify 'novel' problems, how to perform the right experiments, how to analyse the results and how to write 'technical' papers! Thank you for spending the time and effort so that I get the fundamentals right.
- To Prof. Ng Wee Keong - for your inputs on my projects and your support.
- To Prof. Lyle Ungar - for giving me the opportunity to work with an amazing team of experts in Social Science, Psychology and Health at the World Well-Being Project enabling me to expand my understanding of the problems and their application in these domains.

- To my teacher - for transforming my life to live it based on sound principles, backed by the right purpose and for giving me the right motivation to pursue a PhD.
- To Koteswar - for being a healthy critic of my work (and for giving me the confidence that I can be a good coder!), and also for being very friendly and supportive roommate; Joey - for helping me formulate nice machine learning solutions to the problems I am working on; Michael - for helping me with statistical analysis; and Daniel - for leading by example how one can efficiently work on multiple projects at once; Vikram - for helping me with proofreading the thesis; Ankit - for having time for discussions on work or otherwise, whenever needed.
- To our group members: Manoj, Vinit, Issac, Yang Huan, Yuan Yuan, Yabin, Wentao, Li Qiaohang, Tra Trung Anh, Xiaolong Bai, Xu Long, Yuming, Manish, Jinjian, Gu and Dong Lu - for asking me thought-provoking questions during group meetings and for maintaining a very friendly atmosphere in the group.
- To Mr. Chiew Lerk Chern, Ms. Chua Poo Hua and Ms. Ng-Siom Siew Ling - for your support on all the issues I came to you regarding the GPU cluster and other lab issues.
- To Wan Ji, Shirin, Doyen, Vikram, Dev, Suman, Kalindu, Murat, Chinmay - for helping me smoothly adapt to the new environment when I came to NTU.
- To the School of Computer Science and Engineering, NTU - for its support in pursuing and completing this Thesis.
- To my Thesis Examiners - for your insightful and useful comments on the subject matter as well as my writing.
- To Abha - for being such a source of steadfast support under all circumstances, and for introducing me to seaborne. I could not have asked for anything more!
- To my parents and mother-in-law - for all your love and support in making the right choices.

Contents

Acknowledgments	i
List of Figures	v
List of Tables	ix
Abstract	xii
1 Introduction	1
1.1 Motivation	2
1.2 Contributions	4
1.3 Organisation	7
2 Background	8
2.1 On Multimedia Perception: Modeling Affect and Enjoyment	9
2.2 On Automatic Personality Modeling	14
3 Influence of Individual User Factors on Multimedia Perception	18
3.1 Research Questions	19
3.2 Data	20
3.3 Methodology	24
3.4 Results and Discussion	27
3.5 Conclusion	34
4 Modeling Personality based on Posted Content	36
4.1 Personality from Selfies	36
4.1.1 Research Questions	37
4.1.2 Related Work	38
4.1.3 Data	39

4.1.4	Methodology	39
4.1.5	Results and Discussion	41
4.1.6	Conclusion	45
4.2	Personality from Image Posts	46
4.2.1	Research Questions	46
4.2.2	Related Work	46
4.2.3	Data	47
4.2.4	Methodology	48
4.2.5	Results and Discussion	51
4.2.6	Conclusion	55
5	Deep Feature Representations to Model Users' Likes	56
5.1	Research Questions	58
5.2	Related Work	59
5.3	Data	61
5.4	Methodology	61
5.5	Results and Discussion	69
5.6	Conclusion	79
6	Modeling Personality based on Image Likes	81
6.1	Research Questions	82
6.2	Related Work	84
6.3	Data	85
6.4	Methodology	87
6.5	Results and Discussion	93
6.6	Conclusion	101
7	Conclusion and Future Work	103
7.1	Main Contributions	103
7.2	Future Work	105
	Publications	108
	References	110

List of Figures

1.1	Objective Tasks vs. Subjective Perception	2
1.2	Multimedia perception is dependent on both the nature of content and also user's cultural and psychophysical frameworks.	3
1.3	The framework of this thesis	4
3.1	Distribution of positive affect in the dataset. Mean is represented by the dotted line.	22
3.2	Distribution of negative affect in the dataset. Mean is represented by the dotted line.	22
3.3	Distribution of enjoyment in the dataset. Mean is represented by the dotted line.	23
3.4	Distribution of ratings on movie clips with high enjoyment and high -ve affect.	33
3.5	Predictive models' accuracy when employing different features.	33
4.1	Selfies are a medium of self-expression: Users can control what is in the selfie. Models can be trained to predict mid-level cues (See Table 4.1) in images which can predict users' personality. Cartoon source: [1]	37
4.2	Prediction Accuracy of Mid-Level cues: Low-level visual features (listed in Section 4.1.4) are used to train models which detect mid-level cues in selfies. All models are combined (decision-fusion) for final prediction.	41
4.3	Performance (RMSE) of various features at predicting personality scores. Mid-level detectors outperform other features at predicting most traits	42

5.1	Graph showing the Normalized Feature counts present in the images ‘liked’ by different users in the dataset. Many users have overlapping feature counts, thereby making it difficult to distinguish their visual preferences. However, the first few features i.e. the high-level features (left) that we consider in this work have greater discriminative power when compared to low level features (right).	57
5.2	Random images ‘liked’ by 5 users on Flickr. Users’ 1-4 have certain visual patterns in the way they seem to like images (people vs. no people; b/w vs. color; scenary vs. cluttered, etc.). However, in User 5’s case, the images seem to be very diverse w.r.t the visual content (car, squirrel, bird, people, railway station, etc.), the tags convey that the user was probably looking at vacation images in London. In modeling such users’ likes where context is important, tags augment the information conveyed by visual features.	59
5.3	Framework for modeling user likes. There are two separate parts in this framework. (a) Identifying features which influence a user in liking the image. (b) Learning the deep bimodal representation of images.	65
5.4	Bi-Modal Deep Feature Representation - Multi-Layer Homogenous Feature Learning and Cross-Layer Heterogenous Feature Mapping.	68
5.5	High-Level features require lesser amount of training data to capture users’ preferences when compared to low-level features.	70
5.6	Image retrieval performance to users is higher with high-level features than by using low-level features.	71
5.7	Top 30 test images for 4 random users based on the score given by the models discussed in Section 5.5.1. Images outlined are recommendations, from the test set, offered to each user by our model.	72
5.8	Features weighted for three users and a sample of images they tagged as favorite. Users’ preferences can be represented using high-level visual features .	74

5.9	Comparison of bimodal deep representation with individual modalities.	
	Each graph shows the recall of predicting the user who tagged test images as favorite in the top-k attempts. The baseline performance (with features used in [2]) is mentioned for each top-k setting. Learning multi-layer bimodal representation gives better performance than using shallow individual modalities. Layer 0 indicates the original features and Layers 1 & 2 the deep features for each modalities. 'Combined' indicates the performance of the bimodal representation at each layer.	75
5.10	WordClouds representations of Activating and Inhibiting Features for 2 neurons in the network.	
	Each of the words represents a feature we used as input to train the network. For 2 randomly chosen neurons, activating features are represented by WordClouds with a white background and inhibiting features are represented by those with a black background.	76
5.11	Matrix showing the features learnt by neurons at layer 1.	
	A white dash represents features activating the neurons and a black dash represents features inhibiting them. Weights have been thresholded to clearly visualize the strongest activating/inhibiting features. While some neurons seem to be activated by almost all the features, others are activated only for a particular set of features. Such neurons specifically might help in distinguishing the users' likes	77
5.12	Specific example to describe the visualizations in Figures 5.11.	
	The image in the example has the feature counts as shown by the barplot below it. This image activates 6 Neurons in layer 1. The boxplot of the features learnt by those 6 Neurons is shown to the right of the image.	78
5.13	Top 10 images activating the strongest and weakest neurons with KL-Divergence as the distance metric.	
	Images activating the strongest neuron are have more intra image diversity wrt color and other image properties whereas the images activating the weakest neuron are more uniform.	78

6.1	Proposed pipeline of Features to Answers and Answers to Personality: a-b) User likes images and answers self-assessed BFI-10 questionnaire; c) Assessors look at images liked by the user and answer ‘others-assessed’ BFI-10; d) Features are extracted from images liked by the user; e-f) F2A model (proposed FS-OR) is used to predict answers to BFI-10 questions and BFI-10 scoring scheme (A2P) is used to calculate scores on five personality traits. g) F2P model is used to predict personality directly from image features. F2P approach is used by current computational models in Personality Computing. We compare its performance with our proposed F2A+A2P (FS-OR) approach. .	83
6.2	Need for including Semantic Features. Left: High score on Conscientiousness; Center: Low score on Extraversion; Right: High score on Neuroticism.	87
6.3	Sample of the images ‘liked’ by users with different personality profiles. Left: High score on Openness; Center: High score on Extraversion; Right side: High score on Agreeableness.	88
6.4	Top 10 Images retrieved by F2A+A2P for a sample of users based on their Personality Profile. The images with green border are ‘liked’ by the user. Retrieved images not belonging to the user’s liked set can be seen to have lot of similarity w.r.t cues associated with the corresponding personality traits. Explained in Section 6.5.	98
6.5	Mean Recall (for all users) using top k images retrieved by F2P and F2A+A2P models. Some images that were highly ranked by the model but were not counted towards recall might have been liked by the user, had the user viewed them.	99

List of Tables

2.1	Datasets for Affective Modeling of Videos: Most of them implicitly assume that, given a video, the affect experienced by different users will be the same. MM - Mainstream Movies; CC M - Creative Commons Movies; MV - Music Videos; #V - No. of Videos; #A - No. of Annotators	12
2.1	Datasets for Affective Modeling of Videos: Most of them implicitly assume that, given a video, the affect experienced by different users will be the same. MM - Mainstream Movies; CC M - Creative Commons Movies; MV - Music Videos; #V - No. of Videos; #A - No. of Annotators	13
2.2	BFI-10 questionnaire and associated Personality Traits: Each question is associated to a Likert scale. Questions are taken from [3].	16
3.1	Marginal means of perceived responses (affect and enjoyment) on clips, after fixing the co-variates. The different emotion categories from the Differential Emotions Scale were rated on a Likert Scale of 1-5. Enjoyment was rated on a 5-point Guttman-type scales. The emotions joy, warmth, love, calm, and so on were clubbed as positive affect, and anger, fear, anxiety, sadness, etc. were grouped as negative affect and their aggregate scores were calculated.	21
3.2	Sample Descriptives for the user (five personality-related and six culture-related) factors that were considered in this study in addition to the three system factors (framerate, bitrate and resolution).	24
3.3	List of Features	27
3.4	Baseline fixed-effect multilevel linear regression model	28
3.5	Extended fixed-effect multilevel linear regression model	29
3.6	Optimistic mixed-effect multilevel linear regression model	29

3.7	Paired t-test showing the comparison of models for all three responses (w.r.t MSR)	30
3.8	Significant correlations between enjoyment and experience of affect ($p < 0.05$)	32
4.1	Mid-Level Cues detected from Selfies	40
4.2	Significant Correlations ($p < 0.01$) b/w Self-Assessed Personality Traits and Mid-Level Cues	43
4.3	Significant Correlations ($p < 0.01$) b/w Others-Assessed Personality Traits and Mid-Level Cues	44
4.4	$\mathcal{D}_2$: Pearson correlations between color features extracted from tweet images and user personality. Correlations for personality traits are controlled for age and gender. Positive correlations are highlighted with green ($p < .01$, two-tailed t-test) and negative correlations with red ($p < .01$, two-tailed t-test). . . .	50
4.5	$\mathcal{D}_2$: Pearson correlations between image tag clusters extracted from images embedded in tweets and personality (maximum 8 per personality trait). All correlations are significant at $p < .01$, two-tailed t-test. Results for every personality trait are controlled for age and gender and other four corresponding personality traits. Tags are sorted by occurrence in our data set within a tag cluster.	53
4.6	$\mathcal{D}_1$: Prediction results for personality traits with all features. Performance is measured using Pearson correlation in 10-fold cross-validation.	54
4.7	$\mathcal{D}_2$: Prediction results for personality traits with all features. Performance is measured using Pearson correlation in 10-fold cross-validation.	54
5.1	List of Features. High-Level features are described in Section 6.4 and Low-Level features are taken from [4].	62
6.1	Krippendorff's α measuring reliability of others-assessment, taken from [4] . .	86
6.2	Spearman correlations (along with p-values) between self-assessed and others-assessed personality scores.	86
6.3	Notations used in FS-OR formulation	91

6.4	Automatic personality recognition performance comparison between our approach and [4] in terms of Spearman Correlation Coefficient between actual and predicted trait. All correlations for our approach are found to be significant ($p < 0.05$).	95
6.5	Results with low-level features of predicting self-assessed personality-profiles for both F2A+A2P (FS-OR) and F2P (LASSO and sparse SVR) approaches, represented in terms of (RMSE, correlation). Significant correlations ($p < 0.05$) are marked with *	95
6.6	Results with combined features of predicting self-assessed personality-profiles for both F2A+A2P (FS-OR) and F2P (LASSO and sparse SVR) approaches, represented in terms of (RMSE, correlation). Significant correlations ($p < 0.05$) are marked with *	96
6.7	Measuring actual deviations in prediction of personality scores (which are converted into High, Medium, Low categories based on thresholding)	96
6.8	Weights allocated to features by the trained model, indicating their influence on personality traits along with Pearson correlation ($p < 0.05$) in <i>italics</i> of the feature values with the self-assessed traits.	97
6.9	Given a set of images, performance of using combined features at predicting the right user	99
6.10	Confusion matrix for mapping images to users' ethnic profiles.	100
6.11	Confusion matrix for mapping images to gender	100
6.12	Confusion matrix for mapping images to Personality Profile Clusters	101

Abstract

With the increasing proliferation of data production technologies, such as cameras, and consumption avenues, such as social media, multimedia has become an interaction channel among users today. Consequently, automatically modeling users' perception (i.e. liking an image, or finding a video funny etc.) is an important and challenging problem. This challenge goes beyond carrying out a content- or genre-based multimedia analysis because the same content can arouse varying (and possibly opposite) perceptions, depending on the users' cultural and psychophysical frameworks.

However, existing multimedia systems largely ignore the subjective nature of such perception where individual user factors (such as personality and culture) play a crucial role. While individual differences are extraneous for objective tasks like identifying objects or colors present in the content, neglecting them would be inappropriate in subjective perception tasks such as modeling experience of affect or concepts users 'like' etc. In this thesis, we carry out a series of studies to answer two questions pertaining to individual user factors: 1) If and to what extent individual factors play a role in influencing users' multimedia perception, specifically perception of affect & enjoyment in videos?, and 2) How can user-level personality prediction models be built based on social media data?

The challenge in considering individual user factors in practical systems is that they are not freely available, and need to be collected using survey questionnaires. Use of social media has become widespread due to ubiquitous Internet access and multimedia enabled devices, and this has led to the possibility of using multimedia as a source for user-modeling. Users' innate preferences and tastes are conveyed based on media they consume and share, thereby contributing to the field of personalized recommender systems and content generation systems. A key factor to consider in personality recognition is the source used for acquiring users' personality, i.e. the behavior we base the assessment upon. Studies in Psychology show that to get strong

personality cues, users should be given the necessary freedom of control and motivation to express themselves through that behavior. However, in the computational domain, which focuses mostly on curated behavior of participants involved, these two criteria are given amiss. In this thesis, we aim to address this by exploring how non-intrusive data sources – such as posted and liked content on social media can be leveraged to automatically predict users’ personality profiles.

In modeling users personality based on posted content, we crawl selfies from Sina Weibo – a popular Chinese microblogging site, propose several mid-level cues (such as presence of duck faces, pressed lips etc.), and detect them computationally using low-level visual features (like LBP, BoVW etc). These mid-level cues are then used as predictors to model users’ personality measured using the Big Five model, and are shown to outperform low-level visual features. In a second study with image posts from Twitter, we identify the way in and extent to which personality differences are related to online image sharing in general. We extract interpretable semantic image features (color aesthetics and objects) using large-scale content analysis in the posted images. We then give insights into the personality differences among people posting them, and build predictive models of users’ personality.

In modeling users personality based on liked content, we use images tagged as ‘Favorite’ on Flickr. Here, there exist two challenges - a) how best to represent users’ likes on images, and b) how to build effective models to predict users’ personality using images they liked. To address the first challenge we represent images using semantic features which are interpretable - for instance, is the image colored or black & white, is the image rendered by graphics or a natural image, are there people in the image and what their characteristics are and so on. We further propose a deep bi-modal representation exploiting the visual content along with the tags associated with images to better the state-of-the-art on user ‘likes’ by 15-20%. To address the second challenge, we use the semantic interpretable features that we identified previously and a novel Feature Selection based Ordinal Regression for personality prediction. Extensive experiments show that our proposed method significantly improves personality prediction performance.

Chapter 1

Introduction

Perception of warmth and competence shapes our judgement of people and organizations, and when perceived together they cause active behavioral responses from viewers. For instance, every day we witness several organizations put forward their missions in the form of advertisement campaigns – while most of these ads fail to attract our attention, some of them leave a lasting impression in our minds. Successful ad campaigns tend to have compelling storytelling components with emotion evoking communication, structured to stimulate action. Ad campaigns are but one specific scenario which illustrate the importance and challenge of modeling multimedia perception.

In today's technology landscape, production and diffusion of multimedia have become a ubiquitous part of people's lives. While the characteristics of multimedia systems that deliver content influence users' multimedia experience, the experience itself is individual. And individual user factors such as personal preferences, cultural and social backgrounds, personality, gender, age along with physiological characteristics play a key role in ascertaining users' multimedia experience.

For instance, in Figure 1.1, given the image of a humanoid and butterfly, two sets of questions could be posed. The first set, as on the left – 'What is in the image?', 'What is happening in the image?' – concerns with objective facts as the answers indicate. However, the second set on the right – 'Do you like this image?', 'What emotion do you perceive from the image?' – is subjective and the response depends upon the person to whom the question is posed. The problem of modeling these responses is not just limited to content- or genre-based analysis of multimedia because the same image might arouse varied (and possibly different) responses in different people and consequently, such perception is dependent on both the content and also

Figure 1.1: Objective Tasks vs. Subjective Perception

user's psychocultural frameworks (Figure 1.2). Accurate modeling of users' perception has a lot of impact on industrial downstream applications such as targeted advertising and system personalization.

1.1 Motivation

In this thesis, we carry out studies to answer two questions pertaining to the role of individual factors. The following is a description of the questions and motivation behind studying them.

- **Do individual user factors play a significant role in influencing the users' perception of multimedia affect and enjoyment?**

User factors are already being considered in today's multimedia landscape, albeit still from an average/demographic perspective. For instance, in advertising, the same product is advertised differently in different markets according to their demographics, and may also be branded differently to take into account local sensibilities or preferences [5]. Such a 'personalized' approach, considering individual user factors, is being considered also in education, energy conservation, and health, to name but a few domains.

However, most traditional multimedia experiences have been optimized by using responses measured through Mean Opinion Scores [6], ignoring the individual nature of

Figure 1.2: Multimedia perception is dependent on both the nature of content and also user's cultural and psychophysical frameworks.

the experiences involved. While it might be acceptable to ignore individual differences for objective tasks like identifying objects or colors present in the content, doing so might be inaccurate in subjective perception tasks such as modeling experience of affect (emotion) or concepts the users like, etc. Therefore, we examine the role played by individual user factors in the domain of multimedia perception. We study the intensity and differences in the affective responses evoked by multimedia contents depending on an individual's personality and culture.

- **How to automatically build predictive models for users' personality based on social media data?**

The challenge in considering individual factors (such as personality) in practical systems is that such user profiles are not freely available, and often need to be collected using survey questionnaires and other intrusive techniques. Therefore, we explore how non-intrusive data sources such as social media can be leveraged to automatically predict users' personality. Prior research [7] suggests that personality is strongly expressed on a platform which offers users sufficient self-expression and freedom of control. Use of social media has become widespread due to ubiquitous Internet access and multimedia enabled devices, and users online leave behind traces of their psychological traits. There-

fore we use both posted content and liked content from social media to build predictive models of users' personality.

For posted content, we use selfies users post on Sina Weibo, a popular Chinese microblogging website, and image posts on Twitter. For liked content, we use images the users tag as 'Favorite' (termed as 'like' in this thesis) on Flickr. When a collection of images associated with a user is available, we can computationally find out the factors that contribute towards the users' online behavior. These factors can be the emotional associated with the image, the concepts present in the image (viz. the objects and their perspectives/poses, etc.), the aesthetic and artistic aspects of the image, a relatable associated context, and any combination of these factors. Precisely, the subquestions this which open up are: how best to quantitatively represent some of these semantic characteristics in images, and how to estimate the contribution of these factors in understanding the differences in the users' image posting/liking behavior? By computationally modeling this relationship, we would be able to automate users' personality recognition.

1.2 Contributions

Towards answering the questions outlined before, this thesis presents the following contributions:

Figure 1.3: The framework of this thesis

- **Influence of Personality on Users' Multimedia Response**

Towards understanding the role of individual factors on users' multimedia perception, we first build a suitable dataset of 144 video clips rated by 114 participants with their individual user factors and ratings, since most of the available datasets collect ground truth on users' perception (be it of emotional response or enjoyment) by averaging or taking the majority vote from the users. The implicit assumption in previous works averaging the ratings is that users with different psychocultural backgrounds give similar responses. To address this assumption, we explore the influence that two primordial dimensions of user factors - personality and culture - in conjunction with system factors - frame rate, resolution, and bit rate - have on user affect and enjoyment of multimedia presentations.

To this end, a two-site, cross-cultural study was undertaken, the results of which produced three predictive models. Personality and culture traits were shown statistically to represent 5.6% of the variance in positive affect, 13.6% in negative affect and 9.3% in enjoyment. The correlation between affect and enjoyment was significant. Predictive modeling incorporating user factors showed about 8%, 7% and 9% improvement in predicting positive affect, negative affect and enjoyment respectively when compared to models trained only on system factors. Results and analysis indicate that individual factors play a significant role in influencing users' experience while watching multimedia.

- **Automatic Personality Assessment from Social Media Data**

Towards building automatic user profiling models, we use both i) posted and ii) liked content on social media. For posted content, we use users' selfies (profile images) from Sina Weibo and image posts from Twitter, and for liked content, we use images tagged as 'Favorite' on Flickr. We extract several interpretable features from images to gain insights into the expression of personality differences on both the content posted and liked on social media.

– Personality from Posts

Selfies We propose several mid-level cues specific to selfies (such as presence of duckface, if the user is alone, emotional positivity, etc.), and analyse how they influence both recognition and perception of personality profiles. Low-level visual features (like LBP, BoVW, etc.) are extracted from selfies of 123 users from a popular microblogging site, Sina Weibo, and mid-level cue detectors are trained based on them. These are then used to predict the users' personality (based on Five Factor Model). They are seen to outperform state-of-the-art features for most traits. Using the trained computational models, we then present several insights on how selfies reflect their owners' personality and how the users' are judged by others based on their selfies.

Image Posts We identify the way in and extent to which personality differences, measured using the Big Five model, are related to online image sharing based on a topic-based semantic analysis of image posts from Twitter. In two experiments, the larger one consisting of $\sim 700,000$ Twitter images posted by ~ 4000 users (an order of magnitude larger than previous work), analyze personality differences, and we build personality prediction models based on the interpretable features.

– Personality from Likes

Modeling users personality based on images they like on Flickr is more challenging than modeling users' personality through explicit cues such as facial expressions, which has been widely investigated. In addressing this problem, there exist two parts - a) how best to represent users' likes on images? and b) how to predict the users' personality based on image they like?

Learning Feature Representation Automatically understanding and discriminating between different users' liking for an image is a challenging problem. This is because the relationship between image features (even semantic ones extracted by existing tools, viz. faces, objects, etc.) and the users' likes is non-linear, influenced by several subtle factors. We present a deep bi-modal knowledge representation of images based on their visual content and associated tags (text). A mapping step between the different levels of visual and textual representations allows for the

transfer of semantic knowledge between the two modalities. Feature selection is applied before learning deep representation to identify the important features for a user to like an image. The proposed representation is shown to be effective in discriminating between the users based on images they like and also in recommending images that a given user 'likes', outperforming state-of-the-art feature representations by around 15-20%. Beyond this test-set performance, an attempt is made to qualitatively understand the representations learned by the deep architecture used to model user likes.

Modeling Personality We model personality traits of users based on a collection of images they like on Flickr. We use the set of semantic features previously proposed for representing different concepts in images to model the users personality. A novel machine learning approach is developed to model the users' personality based on the image features, resulting in upto 15% improvement over the state-of-the-art. Then, the efficacy of the semantic features and the modeling approach is shown in personality modeling-based recommendation and, the factors that might influence the users with different personality traits to like an image are explained.

1.3 Organisation

The organization of this thesis is shown in Figure 1.3. In the next chapter, we give an overview of works on the two connected parts - a./ multimedia perception (surveying how previous works mostly ignore the signal from user factors, and how they have been shown to be useful in several domains), and b./ automatic user factors modeling (restricted to personality, owing to lack of data on culture). Then in individual chapters, we present a more detailed review of the related works specific to the sources we consider to model users personality since it would be easy for the reader to grasp the motivation for the approaches we take in individual chapters. In Chapter 3, we present our study on the significance of personality and culture on users' perception of affect and enjoyment. In Chapter 4, we describe how posted images, namely image tweets and selfies, can be used to infer personality differences. In modeling personality based on liked content, we discuss how effective feature representations can be learnt on image likes in Chapter 5, and how personality can be modelled based on image likes in Chapter 6. In Chapter 7, we conclude this thesis, and present the future directions that it opens.

Chapter 2

Background

In this chapter we discuss prior works in the field of multimedia perception (specifically modeling affect and enjoyment) in Section 2.1. Most of the studies that model affective response implicitly assume that, given a video, the affect experienced by different users will be more or less the same. This is equally the case with affective video datasets, highlighting the need to consider individual user factors in such subjective perception modeling.

We then present the literature on automatic user personality modeling in Section 2.2. We explore the works that model users personality based on different sources and later in independent chapters, present the state-of-the-art on automatic personality modeling using each source we use (namely posted and liked content on social media). Most of the works in automatic personality modeling work use textual features, and the few that use images do so with features which are not easily interpretable, and consequently lacking insights.

Definitions: The following as the definitions of some of the keywords as used in this thesis.

- **Affect** is defined as a psychological response of a user, including but not limited to emotions and mood, in a given context [8]. In other words, it can be understood as the ‘experience of an emotion’. In this thesis, affect is used as a noun as defined in Dictionary of Psychology by the American Psychological Association [9].
- **Enjoyment**, in this thesis, is defined as the state of taking pleasure in an experience.
- **Personality** is defined as a series of “internal properties” that relate to overt behaviors [10]. The Five Factor Model (FFM) [11] is one of the most used theories to examine the predictive utility of personality. This consists of: openness to experience, conscientiousness, extroversion, agreeableness and neuroticism.

- **Culture** is defined by traits that are associated with the local environment in which one lives and grows, which shape one's perception and cognition through shared conceptions and collective norms [12]. Hofstede's Six Factor Model (HM) [13] studies these cultural differences. The model includes six cultural traits: power distance; individualism; uncertainty avoidance index; masculinity; pragmatism; and indulgence.

2.1 On Multimedia Perception: Modeling Affect and Enjoyment

Theories of Affect: To quantify affect, two major methods are found in the literature - the dimensional approach and the categorical approach. The dimensional approach [14] maps emotions into a 3-dimensional space consisting of arousal, valence and dominance axes. On the other hand, the categorical approach [15] groups emotions into different categories such as happiness, surprise, sadness, anger, fear, disgust and so on.

While there does not exist one grand unified theory of emotion, several studies [15, 16, 17] show that measuring and classifying emotions into different discrete classes can be correlated with specific action tendencies, physiological responses, motor expressions and feeling states. A lack of agreed standards on subjective ratings of emotions could make absolute continuous scores fuzzy. While studying the subjective nature of emotions, it would be interesting to not just study the variance in intensity and valence (in a broad sense), but also assess the variance in different categories on emotions that viewers manifest.

Automatic Affect Modeling: There exist two broad categories of research in automatic affect modeling: one based on facial expressions and physiological sensors, and the second which is a direct content based approach.

An implicit way of analysing emotional impact of multimedia is to model the viewers - their facial and physiological responses. Several techniques such as facial action unit modeling, [18], active appearance model [19] and optical flow-based approaches [20] were adopted and developed by researchers working on facial expression based affect modeling [21, 22, 23, 24, 25] (see [26, 27] for a thorough review). Similarly, a range of complementary sensors (e.g., heart rate, EEG) were used to measure the evoked emotions [28, 29, 30]. However, the extent to which physiological responses capture the subjective intensity of affect (which varies as a

consequence of the users' innate psychology) is unclear. Consequently, we focus on works which model affect directly based on the content.

Modeling affect directly based on video content dates back to 2005, where video frames specific features were extracted and mathematically mapped to arousal and valence space. While lot of works [31, 32, 33] on developing classical feature descriptors involved representing the semantics of the content such as color, texture and shape abstractions, there is an increased interest in understanding users' response to the viewed content – more concretely how content could be represented based on the affect it induces on the viewers.

This led to research on designing features that can be used to model content based on the response they would generate in the users. A comparative study between Google and Bing in [34] showed that only 8% of the users' preference for one search engine over another was due to 'topical relevance', whereas more than 60% were influenced by images' 'attractiveness'. On similar lines, authors in [35] identify cues that can be used to rank images based on 'interestingness'. There were also works on Color based bag-of-emotions [36], Visual sentiment Ontology and Detectors using Adjective Noun Pairs [37] and studies on image memorability [38, 39], popularity [40] and so on which used cues from both Computer Vision and Social Media (to a limited extent) to develop affective representations of images. While most of these were in the last decade, various research attempts were made starting early 21st century to address the problem of aesthetics, mood and emotion inference in images [41, 42] to further the understanding of how users perceive the affect in different images.

Most of these proposed features are based on experiments in psychology, art theory and photography. For instance, Valdez and Mehrabian from their psychological experiments [43] show that the HSV color space can be directly related to pleasure, arousal and dominance (according to the dimensional approach to emotions [44]). There are also studies which show that users response to images depend on composition, colorfulness, spatial organization, emphasis, motion, depth or presence of humans [45, 46, 47].

Also there exist principles in photography like rule of thirds, close-up shots with high dynamic ranges (HDR) and use of complementary colors which distinguish professional shots from amateur ones. Rule of thirds principle states that if an image is divided horizontally and vertically into three parts, 'well-shot' photos have the main elements lying at the four intersections. Another technique used by photographers is the large aperture setting or the use

of macro lenses to reduce the depth of field while shooting single objects. Other observations like using complementary colors to bring out clear distinction between subject and background, proper choice of camera parameters and appropriate lighting conditions to create a surrealistic feel and so on are identified and computationally represented in [41, 42, 48]. Similarly in art, for the paintings which do not have colors, brushstrokes provide signature for artists' styles and various works like [49] tried to build stochastic models based on this feature.

Though the discussion until now was w.r.t image-based features, a parallel can be drawn for videos as well, processing them frame-by-frame. Starting from tasks like video shot detection [50, 51] to inferring the affect induced by videos [52, 53, 54], a lot of research has gone into finding suitable representations for each of the video understanding tasks. One significant difference w.r.t videos and images is that apart from the visual channel, videos also have the auditory and sometimes the lexical (in the form of subtitles) channels. The added temporal dimension, along with the huge amount of raw data and little prior structure makes it more challenging to find feature representations for videos. The pipeline for analysing videos begins with structure analysis (shot boundary detection, key frame extraction and scene segmentation) followed by Feature extraction (like static features from key frames, object features and motion features from segmented shots).

Previous works dealing with modelling affect focus on distilling the influence of specific cinematographic theories [54], types of segment and shot [55], the use of colour [56]. Apart from the works mentioned above, there has been research focused on modeling the different audio-visual features to predict emotions [57, 58, 59, 60, 61, 62]. The features used in this thesis are inspired by those used in the literature, along with certain content-based descriptors which have been shown to perform well in several content understanding tasks [63, 64].

Putting users back in the loop: Personalised Affect: Works described so far aim at a generic understanding of the content, independent of users who view the content. The assumption in these works is that affective response is homogeneous across users, irrespective of their varied individual and cultural backgrounds.

A consequence of this is that such studies implicitly assume that, given a video, the affect experienced by different users will be more or less the same. This is equally the case with affective video datasets (as seen in Table 2.1). However prior research shows that individual differences can lead to varied experiences [65]. To illustrate this, evidence reveals a complex

relationship between affective video tagging and physiological signals [53, 66, 67]. As such, it is important to consider the subjective nature of affective perception.

It is seen that emotions are a consequence of the subjective evaluation of a given content (appraisal theory [68]), implying that given the same content there can be varied (and possibly opposite responses) depending on the users' psychocultural natures. Moreover, based on users' personality, the intensity of emotion perceived could possibly change as well. Consequently, there has been an increased attention towards user-specific personalised affect analysis of content with the promise of improved performance and usability of applications.

Table 2.1: Datasets for Affective Modeling of Videos: Most of them implicitly assume that, given a video, the affect experienced by different users will be the same. MM - Mainstream Movies; CC M - Creative Commons Movies; MV - Music Videos; #V - No. of Videos; #A - No. of Annotators

Dataset	Category	#V	#A	Ann. Type	Users
Mutual Information-Based Emotion Recognition [69]	2 MM	655	-	One value each (0,1) for valence and arousal	No
Predicting Emotions in User-Generated Videos [70]	User Generated Videos	1101	-	Plutchik's emotions used as search keywords on Youtube and Flickr	No
A Connotative Space for Supporting Movie Affective Recommendation [71]	25 MM	25	240	Warmth of the scene atmosphere, Dynamic pace of the scene, Energetic impact	No
Music Video Affective Understanding Using Feature Importance Analysis [72]	Multilingual MV	250	11	4 point valence, arousal	No
Utilizing Affective Analysis for Efficient Movie Browsing [73]	13 MM	4000	-	One value each (0,1) for valence and arousal	No
Affective visualization and retrieval for music video [74]	MV	552	27	One value each (0,1) for valence and arousal	Ind. Rating
Affective level video segmentation by utilizing the pleasure-arousal-dominance information [75]	13 MM	43	14	Ekman's 6 Emotions (1-10)	No
Emotional identity of movies [76]	87 MM	87	-	First 2 Genres from IMDB	No

Table 2.1: Datasets for Affective Modeling of Videos: Most of them implicitly assume that, given a video, the affect experienced by different users will be the same. MM - Mainstream Movies; CC M - Creative Commons Movies; MV - Music Videos; #V - No. of Videos; #A - No. of Annotators

Dataset	Category	#V	#A	Ann. Type	Users
Affective audio-visual words and latent topic driving model for realizing movie affective scene classification [77]	24 MM	206	16	Plutchik's 8 emotions (1-7)	No
Determination of emotional content of video clips by low-level audiovisual features [62]	24 MM	346	16	Pleasure Arousal Dominance values (1-7)	No
LIRIS-ACCEDE [78]	160 CC M	9800	1517	Rating-by-Comparison on Valence-Arousal	No
FilmStim [79]	70 MM	70	364	24 emotional classification criteria	No
MediaEval [80]	Travelogue series	126	-	Popularity/Boredom	No
Content-based prediction of movie style, aesthetics and affect: Data set and baseline experiment [81]	14 MM	14	73	ValenceArousal	No
DEAP [29]	MV	40	32	PhysiologicalSignals, FaceVideos, Valence, Arousal	No
MAHNOB HCI [30]	MM	20	27	ValenceArousal, 6 emotion categories	No
CP-QAE-I (Our Dataset)	14 MM	144	114	16 sets of emotion-related adjectives from DES	Yes

There are recent works on incorporating users' feedback, preference and profile in music video retrieval [74], use of connotative space as mid-level representations to capture emotional properties of movies [71] and implicit video tagging [30]. However, as will be discussed in the following subsection, modeling affective response incorporating users' innate psychological nature is still an open problem, holding promise in several applications involving cold-start situations. Therefore, we focus on investigating the influence of personality and culture on affective response in this thesis. It is to be noted that we do not aim at creating a large-scale video dataset for affective modeling, rather our aim is to understand the influence of users' individual

traits on their perception of affect and develop a dataset towards this goal. Consequently, we create the dataset CP-QAE-I and study the influence of individual user traits. Table 2.1 shows a review of other affective video analysis datasets.

2.2 On Automatic Personality Modeling

Individual differences were a subject of scientific inquiry since the bygone ages starting from the Greek Philosophers in 4th century BC [82]. Modern psychologists ascertain these differences to several factors, one of which is the ‘personality’ of every individual. Personality studies aim “to distinguish internal properties of the person from overt behaviours, and to investigate the causal relationships between them” [11]. Patterns of thought, emotion and behavior can be predicted (with a reasonable accuracy) using the current personality constructs, as will be explained in this Section. Though there are many variants (or theories), we refer to trait models [10], which “provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences” [83].

Five Factor Model (FFM) [3] is a widely used personality traits model. Users’ personality, according to FFM, consists of: conscientiousness (C), openness to experience (O), agreeableness (A), extroversion (E) and neuroticism (N). For collecting the Personality information, subjects are asked to answer a psychology questionnaire (typically BFI-10 [84], NEO-FFI [85] or NEO-PI-R [86]). There are studies [87, 88, 89] which show that BFI-10 has good correlation with other independent indices measured simultaneously with the personality test [11]. For example, it was seen that a person who had a high score on extraversion had more social contacts [90] and the correlation between different personality traits and speech traits was studied in [91, 92].

According to Brunswick Lens model [93], personality is externalized through ‘distal cues’ which is any behavior that others can perceive - like choice of words in language [94], intonation of voice while speaking [95], kind of photos one likes [4], type of people one befriends [96] and so on. Our focus here is to understand the state-of-the-art in user modeling (in other words Automatic Personality Recognition - “inferring self-assed personalities from machine detectable distal cues” [97]).

Sources for modeling users In literature, both explicit and implicit cues were used to model users personality. This can be summarized as follows:

- **Text:** from essays & blogs that users write or comments they post on social media
- **Audio:** Prosody statistics in user generated contents like songs, speech, etc.
- **Visual:** Facial expressions, bodily actions, posture, images, etc.
- **Usage Logs of mobile devices & egocentric networks:** Centrality measure, usage statistics, etc.

Textual cues were extracted using a psychologically oriented tool - Liguistic Inquiry and Word Count (LIWC)[98] from which word categories were adopted and used as independent variables and the Big Five traits as the dependent variables. Source of data included essays that students wrote [99] and blogs that users maintained [100, 101]. The task was performed using ordinal logistical models (as in [101]) or Naive Bayes classifiers ((with 2,3 or 5 classes) in [100]) or N-gram models (in [102]). In these works, the motivations driving different kinds of bloggers were inferred (for instance that extraverts blogged to talk about their life while neurotics did so to release their tensions) and in the other work, students could be categorised as belonging to upper/lower half of each of the traits from the essays they wrote.

Audio-Visual cues like statistics of pitch, intensity, speaking rate, features from openS-MILE [103], and other prosody measurements (for the audio channel), features from LIWC, Medical Research Council (MRC) Psycholinguistics Database (for the lexical channel) and posture, motion activity, walking velocity, interpersonal distances, eye gaze direction and face/hand/head movements (for the visual channel) were used in [104, 105, 106]. The data was in the form of verbal conversations [104], self presentations over Skype [107], video clips of meetings [108] to cite a few. All of these works used either existing classification or regression models to predict the personality of the subjects whose data was being analysed in the above-mentioned modalities.

From the social media (where works dealt with data from Facebook in [96], Twitter in [109, 110], RenRen in [111], FriendFeed in [112], and blogs in [113]), textual features associated with the profile information, egocentric networks, number of followers, usage statistics, emotional states of the posts, LIWC categories were extracted to find the correlation between

Table 2.2: BFI-10 questionnaire and associated Personality Traits: Each question is associated to a Likert scale. Questions are taken from [3].

Question	Trait
I have few artistic interests	O
I have an active imagination	O
I tend to find fault with others	A
I am generally trusting	A
I get nervous easily	N
I am relaxed, handle stress well	N
I am reserved	E
I am outgoing, sociable	E
I tend to be lazy	C
I do a thorough job	C

the typical behaviors of users and their dominant personality traits. Recently, there has been a surge in the number of works using images to model users' personality [114, 115, 116, 117, 118], however the features they work with are not interpretable thereby lacking the ability to derive insights into user behavior.

From mobile and wearable devices features like demographic information, statistics on mobile phone usage [119, 120], behavioral evidence like speaking activity, movement, proximity with others, number of face-to-face interactions [121], centrality, betweenness and other metrics associated with social network [122] were extracted to infer the personality traits of the users.

Summary There is an increasing necessity for multimedia services to bring user into the loop by designing content and delivery mechanisms in a personalised manner to be able to sustain the current customer demands in the times of information deluge. However, most research on modeling multimedia responses, as reviewed in this chapter, carry forward an implicit assumption that content perception is homogenous across users, irrespective of users' varied individual and cultural backgrounds. This motivates us to understand the different factors that can be used to represent user differences, and assess if and to what extent they influence responses to multimedia.

Research findings show that the psychological concept of personality accounts for individual differences in several manifestations – emotional, interpersonal, and motivational [85].

Several studies showed how personality factors have significant correlations with users behavior in a variety of fields. However, personality acquisition was possible only through questionnaires, and there is an increased interest to develop technologies to unobtrusively assess users' personality, from data sources which are available widely and freely – such as social media. While text offers the barrier of language constraints for training models, images are universal in expression. Further while dealing with images, lack of interpretable features limits the scope and insights based on results obtained. Consequently, we turn to mine large amount of digital traces left by people on social media especially in the form of images they post and like, using interpretable features.

Deriving meaningful insights into users' behavior can have huge impact on multiple domains like recommender systems, personalized advertising, tuning retrieval technologies and so on.

Chapter 3

Influence of Individual User Factors on Multimedia Perception

Multimedia content produces diverse affective (emotional) responses in humans. Warmth and competence shape our judgements of people and organizations, and when perceived together they cause active behavioral responses from the viewers [123]. Daily we witness several organizations put forward their missions in the form of ad campaigns. While most of these ads fail to attract our attention, some of them leave a lasting impression in our minds. Take the example of the campaign by Volvo, which was listed as one of most unforgettable ad campaigns of 2013 [124] or Singapore's Ministry of Education 'Teach' campaign. The huge success of such ad campaigns is attributed to how story-telling components are shaped into emotion-evoking communication, structured to stimulate action.

Ad campaigns are but one specific scenario which illustrate the importance and challenge of modeling multimedia-evoked emotions. Publicity campaigns, movies, sports, educational material, games, to name a few, all require research into investigating a user's Quality of Experience (QoE) [125, 126], of which affect is an important dimension. Experience of affect is defined as the positivity of emotions that viewers feel while watching videos. The problem is not just limited to content- or genre-based analysis of multimedia. This is because a video which arouses a positive emotion in one person might arouse a negative emotion in the other (depending on the nature of content and users' cultural and psychophysical frameworks which influence their perception) [65]. Whilst this is understood, how system parameters impact on the affective experience of those viewing multimedia content remains largely unexplored. What is also relatively unexplored is whether, and if so, to what degree, individual user factors also impact upon affective responses. These are the two main issues which we address in

this paper - does multimedia content and system quality parameters with which it is presented evoke different affective responses depending on an individual's personality and culture?

Answering this question involves understanding the subjective nature of emotions and how crucial a role individual user factors play in modeling experience of affect (emotion), thereby addressing users' needs for emotion-sensitive video retrieval [66]. In this work, we attempt to understand how personality [11] and culture [13] influence users' experience of affect and enjoyment in multimedia.

3.1 Research Questions

Specifically, the following research questions are posed:

- RQ 1. Can a model based on multimedia system characteristics (Bit-Rate, Frame-Rate and framesize) and individual user factors (i.e. personality and culture) predict the intensity of affect (both positive and negative) and enjoyment?
- RQ 2. Which system characteristics and individual user factors influence the experience of affect and enjoyment the most?
- RQ 3. How do positive & negative affective experiences correlate with enjoyment across stimuli?
- RQ 4. How do predictive models perform on the task of automatic assessment of experience of affect and enjoyment of videos?

By investigating how different dimensions of these individual user factors modulate users' experience of affect and enjoyment, and specifically by understanding the correlation between enjoyment and perception of affect, we intend to provide initial findings for multimedia content creators and service providers to achieve maximal user satisfaction with respect to the content they create and deliver to diverse users. Also, service providers can exploit the correlations between system factors and the different personality traits to personalise content delivery mechanisms by considering the perceived quality of users [127].

3.2 Data

To address the concern of modeling the influence of individual differences on the experience of affect and enjoyment, we build a dataset using videos which are annotated by users with diverse personality and cultural traits. This section describes the videos, the procedure to used to collect annotations and the descriptive statistics.

3.2.1 Video Dataset

This study uses the CP-QAE-I dataset [128] (<http://1drv.ms/1M1bnwU>), which consists of 12 purposively selected short clips from popular movies to cover different affective categories [79]. Clips from a wide range of valence but low variance on arousal [79] have been adopted to reduce content-based biases, and the clips, their description and the means of positive and negative affect are given in Table 3.1 for the reader’s reference. The content parameter also varies in cinematographic schemes utilized in the original movie production. There are three video coding parameters, namely bitrate – with the settings 384kbps and 768kbps, resolution (frame-size) – with the settings 480p and 720p, and framerate – with the settings 5fps, 15fps and 25fps, and these result in twelve coding-quality combinations. As a result, the dataset contains 144 (12*12) video sequences.

From the analysis with G*Power 3 [129] using the F -statistic as well as repeated measures, the minimum required sample size is 64, utilizing the conventional error probabilities ($\alpha = 0.05, \beta = 0.2$) and with the assumption of existing medium effects ($f = 0.39$) with $r = 0.8$ correlation.

3.2.2 Procedure

The participants for data collection were 57 college students from each of the NTU, Singapore and Brunel Univeristy, London (so totally 114 participants): 43 from Britain, 22 from India, 16 from China, 15 from Singapore and 18 from other nationalities. 28.9% of the participants were female and 23.9 years was the mean age, $\sigma = 3.68$. The corresponding cultural and personality traits are given in Table 3.2.

We applied a lab-based subjective testing approach. A set of videos were located locally on servers at the authors’ universities. As we asked users to view the videos streamed over

Table 3.1: Marginal means of perceived responses (affect and enjoyment) on clips, after fixing the co-variates. The different emotion categories from the Differential Emotions Scale were rated on a Likert Scale of 1-5. Enjoyment was rated on a 5-point Guttman-type scales. The emotions joy, warmth, love, calm, and so on were clubbed as positive affect, and anger, fear, anxiety, sadness, etc. were grouped as negative affect and their aggregate scores were calculated.

MovieClip (Duration in Mins:Secs)	Description from [79]	+ve Affect	-ve Affect	Enjoyment
A_FISH_CALLED_WANDA (2:56)	<i>One of the characters is found naked by the owners of the house</i>	0.184	-0.536	-0.037
AMERICAN_HISTORY_X (1:06)	<i>A neo-Nazi kills an African-American man, smashing his head on the curb</i>	-0.397	0.756	-0.607
CHILDS_PLAY_II (1:07)	<i>Chucky beats Andy's teacher with a ruler</i>	-0.231	0.698	-0.158
COPYCAT (1:04)	<i>One of the characters gets caught by a murderer in a toilet</i>	-0.33	0.418	-0.315
DEAD_POETS_SOCIETY_1 (2:34)	<i>A schoolboy commits suicide</i>	-0.331	0.341	-0.504
DEAD_POETS_SOCIETY_2 (2:23)	<i>All the students climb on their desks to express their solidarity with Mr Keating, who has just been fired</i>	1.053	-0.553	0.725
FOREST_GUMP (1:47)	<i>Father and son are reunited</i>	0.992	-0.523	0.656
SE7EN_1 (1:39)	<i>By the end of the movie, Kevin Spacey tells Brad Pitt that he beheaded his pregnant wife</i>	-0.346	0.248	0.42
SE7EN_3 (0:24)	<i>Policemen find the body of a man tied to a table</i>	-0.431	0.03	-0.306
SOMETHING_ABOUT_MARY (2:00)	<i>Mary takes sperm from Ted's hair mistaking it for hair gel</i>	0.468	-0.72	0.471
THE_PROFESSIONAL (2:44)	<i>The two main characters are separated forever</i>	-0.194	0.216	0.254
TRAINSPOTTING (0:40)	<i>The main character dives into a filthy toilet</i>	-0.477	-0.389	-0.654

Based on estimated marginal means of a mixed-effects regression model. Covariates in the model are evaluated at the following values: AGREEABLENESS = 7.45; EXTRAVERSION = 5.42; CONSCIENTIOUSNESS = 6.59; OPENNESS = 6.77; NEUROTICISM = 5.67; POWER DISTANCE = -34.29; MASCULINITY = -6.73; INDIVIDUALISM = 22.44; UNCERTAINTY AVOIDANCE = 40.83; INDULGENCE = -11.60 PRAGMATISM = 22.82;.

a network, we attempted to remove any issues caused by file format (w.r.t compatibility with the browser etc.). This was the motivation for keeping the file format and network protocol constant across both universities. Users answered an online questionnaire from the local server (to avoid any latency issues over the Internet). Each user saw all 12 clips (in a random order), with different system characteristics and rated the experience of affect and their enjoyment of each sequence by completing questions immediately after viewing each. Informed consent and anonymity were assured at every stage of the study.

Since individual user factors are studied, we aim to maximize ecological validity in recording the users' viewing behavior, so there was no limit on time to finish the survey. However, owing to the nature of such studies, some participants dropped out.

Participants started the survey by answering the BFI-10 [130] and the VSM-2013 [13] for assessment of personality and cultural traits. Afterwards, they were shown 12 videos under

test, and were expected to give their ratings on all sequences. 73.7% of the 114 participants did so. However, all participants rated a minimum of 3 videos, with the average being 10.8 ($\sigma = 2.56$). Over all, 1,232 ratings were collected (90% of the maximum possible).

Figure 3.1: Distribution of positive affect in the dataset. Mean is represented by the dotted line.

Figure 3.2: Distribution of negative affect in the dataset. Mean is represented by the dotted line.

3.2.3 Measuring Users Response

Positive and Negative Affect was measured using Differential Emotions Scale [131]. This includes 16 sets of emotion-related adjectives. We refer the reader to [131] for the list of

Figure 3.3: Distribution of enjoyment in the dataset. Mean is represented by the dotted line.

all sets. Each such set is linked to one of the 5 Likert scales and each participant rated the intensity of felt emotion. The emotions joy, warmth, love, calm, and so on were clubbed as positive affect, and anger, fear, anxiety, sadness, etc. were grouped as negative affect and their aggregate scores were calculated. Descriptive statistics on ratings are shown in Figures 3.1 and 3.2.

Enjoyment This was measured using one of the 5-point Guttman-type scales, and each participant indicated how much he/she enjoyed a video sequence. A value of 1 represents “no” enjoyment and a value of 5 denotes “high” enjoyment. Descriptive statistics on ratings are shown in Figure 3.3.

3.2.4 Measuring Users Traits

Culture It was measured using the VSM-2013 questionnaire [132] according to the following aspects: individualism (IDV), power distance (PDI), uncertainty avoidance (UAI), pragmatism (PRG), masculinity (MAS), and indulgence (IVR). We refer the reader to [133] for the list of all questions and how they relate to different cultural traits. Most of these questions are on a scale of 1-5.

Personality This was measured using the BFI-10 [130] questionnaire, according to the FFM [3], measuring conscientiousness (Con), openness (Ope), Extroversion (Ext), Neuroticism

Table 3.2: Sample Descriptives for the user (five personality-related and six culture-related) factors that were considered in this study in addition to the three system factors (framerate, bitrate and resolution).

User Factors	Min	Max	$\bar{x}(NTU)$	$\bar{x}(BUL)$	$\bar{x}$	σ
Personality factors						
Openess	4	10	6.60	6.91	6.75	1.424
Conscientiousness	2	10	6.40	6.70	6.55	1.523
Extroversion	2	9	5.61	5.46	5.54	1.689
Neuroticism	2	10	5.56	5.68	5.62	1.716
Agreeableness	3	10	7.33	7.31	7.22	1.533
Culture factors						
Individualism	-140	140	25.79	11.67	18.73	50.619
Power Distance	-155	140	-35.61	-36.32	-35.96	53.219
Masculinity	-140	105	3.68	-6.14	-1.23	53.483
Pragmatism	-130	155	16.14	17.54	16.84	58.090
Uncertainty Avoidance	-120	130	52.54	36.67	44.61	47.182
Indulgence	-220	185	-22.63	-11.32	-16.97	65.522

(Neu), and Agreeableness (Agr). The questions of BFI-10 along with the corresponding traits is shown in Table 2.2. Most of these questions are on a scale of 1-5.

3.3 Methodology

In this section, we introduce different statistical methods used, the features extracted to build the predictive models and the evaluations.

3.3.1 Statistical Analysis

The analysis has been conducted in PASW 20.0, which is Statistical Analysis tool [134]. Linear mixed-effects modeling has been adopted for repeated measures, with model parameters determined with the restricted maximum-likelihood method.

We build three computational models (namely baseline, extended and optimistic) to investigate the influence of system factors – framerate, bitrate and resolution, individual user factors – five personality factors and six culture factors, on the experience of affect and enjoyment. Each of them along with corresponding findings will be described in this section. Afterwards a comparison between the three models will be presented to address the four questions we pose in this paper.

Baseline Model This model considers only system factors. For the CP-QAE-I video dataset, there are 12 variations of the system factors – framerate, bitrate, and resolution/framesize. Factors such as format of the files and network protocol were held constant. Due to the expected interactions between these conditions (e.g., an attempt to minimise Bit-Rate while maximising Frame-Rate and framesize would likely create artefacts), they have been modelled as factorial interactions. In addition, the movie clip itself is included as a parameter to reflect differences in cinematographic schemes used to create the movies, along with the nature of the content. This was modelled as a main effect.

Extended Model The extended model adds additional fixed parameters to the baseline model. These were cultural traits: individualism, power distance, masculinity, pragmatism, uncertainty avoidance, and indulgence. Additionally, personality traits were also added: extroversion, agreeableness, conscientiousness, neuroticism, and openness. These were incorporated into the model as covariates with direct effects.

Optimistic Model While a model aims at predicting a dependent variable as precisely as possible, not all of the residual variance can be solely attributed to individual user factors. A non-trivial proportion of the residual variance can also, to name but a few, be attributed to random error, measurement error, and the limitations of the modelling technique (in this case, generalised linear regression). As such, an optimistic model can be used to estimate the part of the residual variance that might possibly be attributed to individual user factors in general and, to a small extent, because of the factors such as context and limitations in experimental control. This is achieved by modelling every participant as a random effect. In other words measurements have been repeated to obtain a different intercept during the regression for every participant.

3.3.2 Predictive Modeling

While statistical analysis provides an understanding of the relationships of different dependent and independent variables in the data, predictive models help to forecast users' responses on new data samples.

We propose a prediction framework which takes input features in video content, system characteristics, personality and cultural traits, predicting experience of both positive and negative affect, and enjoyment on the video clips using L_1 regularised L_2 -loss sparse SVM with a

linear kernel. We chose a linear kernel to avoid the problem of overfitting (as seen in literature [63, 135]). We use the libsvm framework [136] to carry out these experiments.

We use 4 categories of features to represent visual-based, audio-based, system, and individual user factors, in an attempt to describe various features which might influence the experience of affect and enjoyment. They are described as follows.

Visual features To represent the different concepts in the video, we used HybridCNN [64], Adjective-Noun Pairs [63], Color histograms, Bag-of-Visual-Words [137], Aesthetic features [55, 138] and LBP features [139]. Representation for every video is gotten by mean-pooling on features obtained from all frames.

Color Histogram: Users' perception is greatly influenced by the color hues in the videos. Therefore, color histograms are chosen to represent users' color inclination.

Aesthetic Features: We used 2 sets of features to represent the aesthetic characteristics in videos: a.) art and psychology based features [138] to describe photographic styles (rule-of-thirds, vanishing points, etc.). b.) psycho-visual characteristics at cell, frame and shot levels proposed by [55].

LBP: LBP was used to encode users' perception of texture in the videos. As LBP represents facial information well and many of the videos have people, we use LBP features.

Bag-of-Visual-Words [137]: A 1500 dimension feature based on vector quantization of key-point descriptors is generated for every frame. A bag of visual words representation is used by mapping the keypoints to visual words.

HybridPlacesCNN : CNN features from fc7-ReLu layer from ImageNet [140] and Places dataset [64] are used to represent objects and places in videos.

Adjective-Noun Pairs (ANP): Emotion based attributes are detected (2089 ANPs) [63] using the sentibank classifiers. 8 emotion categories [141] are used to define the adjectives, and objects and scenes are used to define the nouns.

Audio features Affective characteristics associated with the audio content in videos are extracted (OpenSmile [142]) as musical chroma features [143], prosodic features [144] and low-level descriptors such as MFCC features, intensity, loudness, pitch, pitch envelope, probability of voicing, line spectral frequencies, and zero-crossing rate. The visual affective content in the videos was expected to be represented by the aesthetic and ANP features described above.

Table 3.3: List of Features

Category	Feature	Description
Visual features	Color Histogram	based on RGB values of each frame
	Visual Aesthetics	art and psychology based features to describe photographic styles (rule-of-thirds, vanishing points, etc.); psycho-visual characteristics at cell, frame and shot levels
	LBP	users' perception of texture in the videos
	BoVW	1500-dimension vector based on quantization of keypoint descriptors
	HybridPlaces CNN	features from the ReLu layer following fc7 layer of a CNN trained on 1.3 million images from ImageNet and 2.5 million images on the Places dataset
Audio features	Sentibank	measuring visual sentiment by detecting of 2089 Adjective Noun Pairs (ANPs) based on emotion related concepts
	Musical chroma	Opensmile
	Prosody	
System-factors	Low-level descriptors	intensity, MFCC, loudness, pitch, pitch envelope, probability of voicing, zero-crossing rate and line spectral frequencies
	Bit Rate, Frame Rate, Resolution	
User-factors	Perceptual Characteristics	Quality of the distorted image is expressed as a simple distance metric between the model statistics and those of the distorted image
	Personality	openness, conscientiousness, extraversion, agreeableness, neuroticism
	Cultural traits	power distance, masculinity, individualism, uncertainty avoidance, indulgence, pragmatism
	Demography	gender, age and nationality

System-factors Bitrate, Framerate, resolution and perceptual characteristics [145] are used to represent quality characteristics of videos. Perceptual characteristics describe the no-reference quality metric [146]. Perceptual characteristics were represented by temporal distortions in the video, spatial domain natural scene statistics, statistical DCT features motion coherence feature describing the coherence in strength and direction of local motion due to temporal distortions, and reflecting the perceptual difference between pristine and distorted videos.

User-factors The five personality factors, six cultural traits, gender, age and nationality of the users are used to represent individual user factors.

3.4 Results and Discussion

Results target at answering the four research questions raised at the outset of the paper. Sections 3.4 and 3.4 deal with RQ1 and RQ2, Section 3.4 deals with RQ3, and Section 3.4 answers RQ4.

Table 3.4: Baseline fixed-effect multilevel linear regression model

Parameter	df_{num}	Positive Affect			Negative Affect			Enjoyment		
		df_{den}	F	p	df_{den}	F	p	df_{den}	F	p
Movie Clip	11	156.009	25.315	0.00	144.643	33.932	0.00	177.09	40.14	0.00
Frame Rate (FR)	2	803.739	0.32	0.73	710.192	0.056	0.95	1131.23	5.173	0.006
Frame Size (FS)	1	809.889	0.006	0.94	729.398	3.298	0.07	1146.39	2.846	0.092
Bit-Rate (BR)	1	816.675	1.724	0.19	714.909	0.30	0.58	1139.69	0.474	0.491

Interactions of System Factors namely $FR \times FS$, $FS \times BR$, $FR \times BR$, $FR \times FS \times BR$ were found to be insignificant predictors and hence not included in the above table.

3.4.1 Statistical Analysis

3.4.1.1 Baseline Model

Table 3.4 shows the results. The impact that each factor had on the outcomes is measured by the F -statistic in the table, and the significance of the interaction between a factor and outcome is measured by a p -value ($p < 0.05$ indicated that the factor is a significant predictor for the outcome). The movie clip itself had the largest impact on experience of affect and enjoyment (indicated by the highest value for the F -statistic). However, an interesting observation is that only Frame Rate had a statistically significant effect on enjoyment. This shows that system factors alone do not make a huge impact on how the content is perceived. That is, given two videos of different natures at different bitrate, resolution/framesize and framerate, the nature of the content alone is more likely to influence how it is perceived than the system settings at which it is delivered. Our findings can be corroborated by similar observations in QoE [147, 148].

3.4.1.2 Extended Model

Table 3.5 shows the results of the extended model. Many personality and cultural traits seem to be significant predictors of experience of affect and enjoyment. Among personality traits, extraversion and conscientiousness are significant predictors for positive affect, and agreeableness, neuroticism and conscientiousness are significant for negative affect. Conscientiousness and openness are significant predictors for enjoyment [149, 150]. Among cultural traits, masculinity and indulgence are significant predictor for positive affect, indulgence alone for negative affect and uncertainty avoidance for enjoyment. None of the system factors (except Frame-Rate for enjoyment) and their interactions are significant predictors.

Table 3.5: Extended fixed-effect multilevel linear regression model

Parameter	df_{num}	Positive Affect			Negative Affect			Enjoyment		
		df_{den}	F	p	df_{den}	F	p	df_{den}	F	p
Movie Clip	11	193.163	35.925	0.00	206.260	39.739	0.00	171.956	39.733	0
Frame Rate (FR)	2	1071.695	0.18	0.84	1045.660	0.48	0.62	1136.577	4.695	0.009
Frame Size (FS)	1	1074.152	0.54	0.46	1061.874	2.10	0.15	1151.402	3.336	0.068
Bit-Rate (BR)	1	1083.535	2.334	0.13	1044.851	0.06	0.807	1145.171	0.257	0.612
Extraversion	1	1074.324	4.559	0.033	1059.767	0.08	0.78	1150.401	0.024	0.877
Agreeableness	1	1072.223	1.876	0.17	1059.481	24.314	0.00	1152.475	2.001	0.157
Conscientiousness	1	1077.950	9.474	0.002	1041.655	3.964	0.047	1141.249	5.271	0.022
Neuroticism	1	1084.026	0.02	0.888	1050.845	25.227	0.00	1146.479	0.05	0.823
Openness	1	1074.213	2.670	0.103	1058.628	2.110	0.147	1145.365	4.344	0.037
Power Distance	1	1073.888	4.676	0.031	1055.500	0.00	0.985	1152.465	9.138	0.003
Individualism	1	1070.708	2.148	0.143	1052.462	2.486	0.115	1150.026	0.674	0.412
Masculinity	1	1074.304	4.874	0.027	1043.258	1.061	0.303	1141.312	3.312	0.069
Uncertainty Avoidance	1	1077.284	0.534	0.465	1044.360	0.306	0.580	1144.106	5.751	0.017
Pragmatism	1	1069.661	0.886	0.347	1064.578	0.175	0.676	1160.7	0.604	0.437
Indulgence	1	1070.162	5.863	0.016	1051.545	4.863	0.028	1149.178	2.206	0.138

Interactions of System Factors namely $FR \times FS$, $FS \times BR$, $FR \times BR$, $FR \times FS \times BR$ were found to be insignificant predictors and hence not included in the above table.

Table 3.6: Optimistic mixed-effect multilevel linear regression model

Parameter	df_{num}	Positive Affect			Negative Affect			Enjoyment		
		df_{den}	F	p	df_{den}	F	p	df_{den}	F	p
Movie Clip	11	178.713	42.312	0.00	152.624	55.782	0.00	179.877	46.99	0.00
Frame Rate	2	701.036	1.788	0.168	945.140	1.392	0.249	1116.89	8.025	0.00
Frame Size	1	695.825	0.002	0.965	969.366	5.764	0.017	1120.818	3.13	0.077
Bit-Rate	1	715.664	1.159	0.282	972.050	1.457	0.228	1121.96	0.054	0.816

Interactions of System Factors namely $FR \times FS$, $FS \times BR$, $FR \times BR$, $FR \times FS \times BR$ were found to be insignificant predictors and hence not included in the above table.

This suggests that multimedia system characteristics have little to no influence on the intensity of the affect that viewers experience. Additionally, there appears to be a different set of predictors for affect compared to overall enjoyment. F -statistic is generally quite small for most of the predictors. However, the predictors of agreeableness and neuroticism, for negative affect, are notably much larger. This suggests that a considerable amount of the variance in negative affect can be explained by these parameters.

3.4.1.3 Optimistic Model

Table 3.6 shows the results of the optimistic models. The model is quite similar to the baseline model with the exception of larger F -statistic, indicating that larger proportion of variance is explained because of considering random effects. Additionally, interaction between framesize and experience of affect is now significant.

3.4.1.4 Model Comparison

Paired t-tests on Mean Squared Residuals (MSR) are used to compare the models as shown in Table 3.7. Proportional reduction in overall MSE is shown (see [151]). We used the MSR to calculate the percentage of variance explained. The three models are then compared to provide attributions. User factors were drawn from: (Optimistic - Baseline). The extended model, however, examined the relative contribution of personality and culture as a proportion of the overall user factors: (Extended - Baseline / Optimistic - Baseline).

The results show that individual user factors, namely personality and culture, play a crucial role in modeling the experience of affect and enjoyment, indicating that content production and delivery mechanisms should not just take into account multimedia system factors but also individual user factors to achieve maximal user satisfaction.

Models for Positive Emotion From the baseline to optimistic model, the MSR reduced from 0.6304 ($\sigma = 1.050$) to 0.4051 ($\sigma = 0.886$; $p < 0.005$), representing a predicted variance of 55.3%. A part of this is contributed by culture and personality. 5.6% of variance attributable to individual user factors is predicted by the extended model, reducing the baseline MSR to 0.6177 ($\sigma = 1.005$; $p = 0.021$).

Models for Negative Emotion From the baseline to optimistic model, the MSR reduced from 0.6514 ($\sigma = 0.889$) to 0.3615 ($\sigma = 0.536$; $p < 0.00$), representing a predicted variance of 58.1%. 13.6% of variance attributable to individual user factors, reducing the baseline MSR to 0.6118 ($\sigma = 0.8278$; $p < 0.00$).

Table 3.7: Paired t-test showing the comparison of models for all three responses (w.r.t MSR)

Models	Positive Affect				Negative Affect				Enjoyment			
	$\Delta\bar{x}$	σ	t	p	$\Delta\bar{x}$	σ	t	p	$\Delta\bar{x}$	σ	t	p
<i>Baseline</i> $\rightarrow$ <i>Extended</i>	0.013	0.193	2.311	0.021	0.039	0.277	5.008	0.00	0.039	0.430	3.219	0.001
<i>Baseline</i> $\rightarrow$ <i>Optimistic</i>	0.2253	0.924	8.552	0.00	0.2898	0.726	14.014	0.00	0.4199	1.129	13.069	0.00

Models for Enjoyment From the baseline to optimistic model, the MSR reduced from 1.3684 ($\sigma = 1.63$) to 0.9481 ($\sigma = 1.22$; $p < 0.00$), which makes up 23.0% of the overall variance predicted. 9.3% of variance due to individual user factors is predicted by the extended model, which decreases the baseline MSR to 1.3290 ($\sigma = 1.58$; $p < 0.001$).

3.4.1.5 Correlation between Affect and Enjoyment

As introduced at the beginning of the article, there is a very close and significant relationship between what users enjoy and the emotion it evokes (results from correlation analysis are shown in Table 3.8). In all clips, enjoyment is significantly correlated with interest, joy, satisfaction and the latent factor, positive emotion. This means that for a user to enjoy a video the content has to definitely draw his/her interest, but must also have moments of happiness and deliver something which satisfies the viewer [152].

There are also very few instances of negative emotions (sad, fearful, guilty, and ashamed) giving enjoyment to the users. For instance, enjoyment was seen to have a significant positive correlation with emotions Ashamed and Guilty for both the clips AMERICAN_HISTORY_X (in which the protagonist is seen brutally torturing someone) and DEAD_POETS_SOCIETY (in which one of the main characters commits suicide by shooting himself) and with emotion Ashamed for SOMETHING_ABOUT_MARY (in which there is an obscenity involved and yet sounds joyful/funny) and Guilty for TRAINSPOTTING (in which a person is seen to get inside a dirty toilet bowl, and yet the music is of a totally different contrast). The predominant emotion in these clips are not widely enjoyable. However, these might be associated with how certain the users (possibly with high scores on neuroticism) perceive certain contents [153, 154]

Apart from that, even the nature of the content itself can arouse contradictory emotions. For example, enjoyment is observed to be positively correlated with sadness in the movie clip FOREST_GUMP which has a defining excerpt in which the leading character encounters his son for the very first time. This is a scene with bitter-sweet connotations for viewers due to the fact that the protagonist, quite belatedly in his life, is faced with the news that not only has he fathered a son, but also that the son is doing well in school and is a fine student. So, such occurrences are due to the interaction of individual user factors and nature of the content both.

It is interesting to note that while most of the users might associate enjoyment with positivity, there are certain users who need to experience negative emotions to connect to the content's

Table 3.8: Significant correlations between enjoyment and experience of affect ($p < 0.05$)

Clip	Interest	Joy	Sad	Fearful	Disgust	Surprise	Warm	Loving	Guilty	Moved	Satisfied	Calm	Ashamed	+ve Affect
C-I	0.579	0.699	-	-	-0.332	0.430	0.626	0.419	-	-	0.456	0.451	-	0.610
C-II	0.505	0.304	-	-	-0.332	0.380	0.350	0.383	0.258	-	0.526	0.404	0.291	0.483
C-III	0.596	0.485	-	-	-0.314	0.250	0.298	0.247	-	-	0.432	0.286	-	0.426
C-IV	0.444	0.424	-	-	-	-	0.287	0.243	-	-	0.255	-	-	0.304
C-V	0.469	0.263	-	-	-	-	0.368	0.271	0.244	0.219	0.288	0.232	0.250	0.298
C-VI	0.514	0.385	-	-0.239	-	-	0.353	0.325	-	-	0.493	-	-0.215	0.456
C-VII	0.549	0.643	0.248	-	-	0.293	0.479	0.513	-	0.508	0.510	0.407	-	0.560
C-VIII	0.550	0.408	-	-	-	-	-	-	-	0.346	0.445	0.319	-	0.421
C-IX	0.340	0.394	-	-	-	-	0.323	0.251	-	-	0.219	0.244	-	0.292
C-X	0.512	0.658	-	-	-	0.258	0.301	-	-	-	0.267	0.354	0.290	0.419
C-XI	0.541	0.266	-	-	-	0.232	-	-	-	0.277	0.358	0.325	-	0.347
C-XII	0.590	0.688	-	-	-0.401	0.317	0.434	0.419	0.244	0.295	0.488	0.353	-	0.542

Movie Clips:- C-I: A_FISH_CALLED_WANDA; C-II: AMERICAN_HISTORY_X; C-III: CHILDS_PLAY_II; C-IV: COPYCAT; C-V: DEAD_POETS_SOCIETY_1; C-VI: DEAD_POETS_SOCIETY_2; C-VII: FOREST_GUMP; C-VIII: SE7EN_1; C-IX: SE7EN_3; C-X: SOMETHING_ABOUT_MARY; C-XI: THE_PROFESSIONAL; C-XII: TRAINSPOTTING. No significant correlations were observed between Enjoyment and Anxious, Angry, Disdain & -ve Affect and thus are not shown in the table. Entry $E_{i,j}$ represents the correlation between enjoyment and affect for movie clip i and emotion category j

message. This insight gives content creators a better understanding of how to influence users with different personality and cultural traits to establish an emotional connection with them, which is very important to drive behavioral action (especially in scenarios involving ad campaign design, etc.).

To investigate this further, we selected the records with high enjoyment (i.e. 4 and 5) and with high negative affect (above mean) and low positive affect (below mean). The filtered subset (about 5.4% of the total ratings) was investigated to understand the nature of content which likely influences such rating behavior. Figure 3.4 shows distribution of clips in the filtered subset.

3.4.2 Predictive Models

Experiments are run on leave-one-video-out setting by binarising perceptual quality and enjoyment scores. Figure 3.5 depicts the accuracy with which the trained models predict user affect and enjoyment. User, system, content and emotion factors (the latter based on audio affect) were employed, and features extracted to represent the same.

As far as positive and negative affect are concerned, models which gave the best performance were those trained on emotion and content factors (in this order). This is intuitive, as

Figure 3.4: Distribution of ratings on movie clips with high enjoyment and high -ve affect.

Figure 3.5: Predictive models' accuracy when employing different features.

the nature of content plays the most important role in influencing viewers' experience of affect. It was also seen in our statistical analysis (in the previous section).

It was then explored if adding individual user factors to those pertaining to system, content and emotion would improve the predictive modeling performance.

In terms of positive affect, the combination of individual user and content, followed by that of individual user and emotion factors yielded the best results. It must be noted that the features representing content factors also include ANPs, which are especially designed for visual sentiment prediction [63]. This explains their almost equivalent performance when compared to emotion factors. Similar performance was seen when three factors (namely content, individual user and emotion) are combined to train the model. Similar observations are made for negative affect, however the performance was lower than that of predicting positive affect. This

is similar to the observations made in other works [155, 156, 157, 158, 159], possibly due to the intrinsic challenge of modeling the nature of negative emotions.

As far as enjoyment is concerned, models trained on individual user, emotion, and content factors performed better than others. Our statistical analysis showed evidence of this, as content was found to be significantly correlated with enjoyment. The impact of factors associated with emotion has been explored in other studies [160, 161, 162].

3.5 Conclusion

Experience of affect and enjoyment in multimedia is influenced by an intricate interplay between characteristics of stimuli, individuals, and systems of perception. Returning to the research questions posed at the outset of the paper, we can now state that:

- RQ1 For positive affect, negative affect and enjoyment, personality and culture respectively represented 5.6%, 13.6% and 9.3% of variance. Notwithstanding the fact that these constitute sizeable proportions, follow up studies need to explore other potential contributing factors, such as sensory impairments/acuity, user cognitive style, and domain expertise [163].
- RQ2 Traits of extraversion, conscientiousness, masculinity and indulgence are significant predictors for positive affect, and agreeableness, neuroticism, conscientiousness and indulgence were important predictors for negative affect. Conscientiousness, openness and uncertainty avoidance were significant predictors for enjoyment.
- RQ3 The majority of the movie clips which were enjoyed were also rated high on positive affect, with a small exception of clips having high correlation between negative affect and enjoyment. Such behavior is possibly due to the interchange that potentially takes place between individual user factors (e.g., neuroticism) and media content.
- RQ4 Predictive models trained with a mixture of individual user factors and content, emotion and emotion factors yielded the highest achievement for positive affect, negative affect and enjoyment respectively with an accuracy of 79%, 77% and 76% respectively.

It is important to know the impact of individual user factors on user enjoyment, as this allows one to optimise this latter parameter especially in conditions in which other more traditional forms of adaptation (such as layered adaptation) are difficult/impractical to perform. As highlighted above, results obtained in our study showcase the important part that individual user factors have on two important aspects of user QoE, namely affect and enjoyment. Thus, integration of individual user factors in the optimistic model was shown to have significantly improved modeling performance; however, extended models based on personality and culture did not have impacts of the same magnitude. This means that several other individual user factors, apart from those taken into consideration in this study such as user behavior in a particular domain of study (e.g., movies rated or images liked) or other psychological constructs like mood, etc. [164, 165, 166] can be explored. Nonetheless, results show that individual user factors, namely personality and culture, exert an important influence in modeling the experience of affect and enjoyment, indicating that content production and delivery mechanisms should not just take into account multimedia system factors but also individual user factors, in order to achieve maximal user satisfaction.

Chapter 4

Modeling Personality based on Posted Content

In this chapter, we present the work on modeling personality based on content posted by users on social media. In two independent studies, first with selfies users post on Sina Weibo (presented in Section 4.1, and second with image posts on Twitter Section 4.2, we address the problem of automatic personality modeling. While the first study deals with profile photos of users, the second deals with more generic posts from users' timeline.

4.1 Personality from Selfies

'Selfie' refers to a self-portrait taken by a person. Selfies have become so popular that 2013 was dubbed as 'the year of the selfie'. Social networking sites saw millions of selfies being uploaded (beginning from famous celebrities at important events like the Academy Awards, astronauts in space to anyone who has a smart-phone) [167]. Emergence of selfies as a new medium for self-expression and self-representation opens up the possibility of systematically assessing users' personality.

There are two dimensions to automatic personality prediction [97]: a) *personality recognition*, which deals with predicting personality of users which they externalize through distal cues, i.e. any form of observable behavior as detailed in Section 2.2, and b) *personality perception*, which deals with what others perceive about the user through proximal cues. While automatic personality recognition is about inferring self-assessed personalities from machine detectable distal cues, automatic personality perception is about inferring the personality that observers attribute to a given user from proximal cues.

Figure 4.1: **Selfies are a medium of self-expression:** Users can control what is in the selfie. Models can be trained to predict mid-level cues (See Table 4.1) in images which can predict users' personality. Cartoon source: [1]

A key factor to address the problem of automatic personality prediction, be it recognition or perception, is the source used for acquiring users' personality, i.e. the behavior we base the assessment upon. Studies in Psychology show that to get strong personality cues, users should be given the necessary freedom of control and motivation to express themselves through that behavior [7]. Many experimental studies have shown that standardized photos contain valid personality-related cues [168, 169, 170]. Similar work was conducted on profile photos and photos posted on social networks which provided evidence of personality expression in photos [171, 172].

However, when compared to other types of photos, selfies give individuals more freedom of controlling their face visibility, emotional expression, and camera position, and are often posted on social media platforms used for self-presentation [173, 174]. Social Psychology research shows that there are several picture-coding cues (or mid-level cues) that are appropriate for coding selfies [170, 171, 172, 175, 176]. We consider some of them as listed in Table 4.1 and as shown in Figure 4.1. Using these, we attempt to computationally model users' personality based on their selfies.

4.1.1 Research Questions

In this work, with selfies as the basis, we examine the following questions:

RQ 1. If and to what extent computer models can predict users’ self-assessed personality (personality recognition), and how others assess users’ personality (personality perception)?

RQ 2. Which cues aid in personality recognition and perception? What behavioral insights can we gather based on personality differences?

Towards the above-mentioned goals, we build a selfie image dataset of 123 users from a popular microblogging website (Sina Weibo). The users also provide their self-assessed personality scores (using BFI-44 questionnaire [177]). We recruited eight research assistants from the Department of Psychology to provide their assessment of selfie subjects’ personality using BFI-44. We also recruited two volunteers to annotate these images with mid-level cues.

Contributions: 1) We train models to detect mid-level cues (which are relevant to personality prediction based on psychology literature [170, 171]) in selfies, using low-level visual features. 2) The trained mid-level cue detectors are used to automatically predict users’ personality, outperforming state-of-the-art features for most traits. 3) We present several insights on which mid-level cues contribute to personality recognition and personality perception.

4.1.2 Related Work

Personality of users can be quantitatively measured and is found to be consistent w.r.t users behavior [11]. According to Brunswick Lens model [93], personality is externalized through ‘distal cues’ (i.e. any behavior that others can perceive), enabling others to form impression about the users’ personality. With images becoming a wide-spread channel for communication and expression, there has been a recent trend to use facial images from Facebook [115], random portraits on the web [178] and existing face recognition datasets [179] to model users’ personality.

However, face recognition datasets [179] and random portraits on the web [178] would not have provided subjects in the images an appropriate platform for self-expression and freedom of control [171], which are necessary cues for personality to be strongly expressed [7].

Therefore we use selfies of users and propose a novel approach to build mid-level cue detectors (which are especially relevant for selfies and facial images [170, 171]) using visual

features. These detectors are then used to model users' personality. To the best of our knowledge, no previous work has targeted the task of building such mid-level cue detectors for facial images (though similar approaches are shown to be successful in domains such as sentiment analysis [63]).

4.1.3 Data

To participate in our study, we recruited 612 users from Sina Weibo (a popular microblogging website similar to twitter in China). Previous studies show that profile photos may have more impact on personality judgements than textual self-disclosures [115]. Therefore, profile pictures of all participants were downloaded. Two independent raters identified 123 pictures (out of 612) to be selfie portraits, which are used in our study to model personality.

Each of the 123 users (89 females and 34 males) answered the BFI-44 [177] and gave us their profile user names, usage frequency and other basic information (namely gender, age, country of residence and ethnicity). The scores on the five factors were calculated using BFI scheme (ranging between 1 and 5).

In this work, we aim to build mid-level detectors to automatically predict both self-assessed and others-assessed personality scores of users. To get the ground truth on others-assessed personality of users, eight research assistants from the Department of Psychology viewed each user's selfie and rated their impression of the selfie subjects personality using the same BFI questionnaire (See [97] for more details on formats). To get the ground truth for learning the mid-level cue detectors, two independent raters annotated the selfies with the mid-level cues listed in Table 4.1. The two raters achieved a 90.81% inter-rater agreeability ($p < 0.001$) in their coding.

4.1.4 Methodology

The relationship between users' personality and image features is highly non-linear with many intermediate latent spaces. Therefore, mapping between the feature descriptors and the user response would benefit from representations for the latent spaces (obtained either by learning as in the case of deep learning object detection pipelines [180], or those designed based on domain knowledge in the case of this study). These 'mid-level representations' might possibly give better understanding of users' personality. Towards this goal, we attempt to learn mid-level

Table 4.1: Mid-Level Cues detected from Selfies

Mid-Level Cue	Label Description
Face Visibility	Not Visible/Partial/Complete
Photoshop Editing	Present/Absent
Public Location	Yes/No
Private Location	Yes/No
Duckface	Present/Absent
Pressed Lips	Present/Absent
Emotional Positivity	Negative/Neutral/Positive
Alone	Yes/No
Amount of Body	Face Only/Shoulder-up/Waist-up
Eyes Looking at Cam.	Yes/No
Camera Height	Below/At/Above Head-level
Camera In Front	Yes/No
Gender	Male/Female
Age	Below 18/18-20/21-25/Above 26

cues such as the different characteristics of a selfie (entire list in Table 4.1) using traditional image features. Then we use the mid-level cues to predict personality.

4.1.4.1 Feature Extraction

We extracted 10 different visual features motivated by their relevance to capturing different aspects of the images' characteristics.

Color Histograms: are evaluated in the RGB space. These are supposed to represent users' inclination to different colors.

Aesthetic Features [138]: are used to characterise photographic styles (e.g., rule-of-thirds, vanishing points, etc.) based on art theory and psychological studies. These are supposed to give an idea about users' aesthetic preferences.

GIST: Spectral information and coarse localization are used to represent global structure of the scene in an image. These convey information about the scene in the selfies.

LBP: LBP is used to encode visual texture perception information. As most of the area in selfies is occupied by users' faces, LBP features represent facial information as well.

Attribute Features [181]: are useful for characterizing abstract sentiments associated with the image by automatically designing attributes encoded by a compact category-attribute matrix.

Bag-of-Visual-Words [137]: A vocabulary is generated by vector quantization of keypoint descriptors, which in this case is taken as 1500 dimension.

Figure 4.2: **Prediction Accuracy of Mid-Level cues:** Low-level visual features (listed in Section 4.1.4) are used to train models which detect mid-level cues in selfies. All models are combined (decision-fusion) for final prediction.

Fisher encodings of SIFT, SURF, HOG and MSER: The Fisher Vector encoding extends the Bag-of-Visual-Words by going beyond 0-order statistics and by encoding the second order statistics about each visual word’s local descriptors’ distribution.

4.1.4.2 Training Mid-Level Cue Detectors

After extracting the features from images, we used LibSVM with RBF kernel to train mid-level cue detectors. Parameter tuning was performed on a 5-fold cross validation and results are reported on 25 % test data split. We use the output of mid-level detectors as features to predict the scores on personality (both self-assessed and others-assessed) and we compare the results with those achieved using visual features. Since personality scores were continuous values between 1 and 5, we employed LASSO formulation to model them.

4.1.5 Results and Discussion

4.1.5.1 Mid-Level Cue Detectors for Selfies

The accuracy of detecting mid-level cues using visual features (late-fusion) is shown in Figure 4.2. Each of the visual features were used to train models to detect mid-level cues and then all the models were combined using a late-fusion technique. Cues such as alone and

Figure 4.3: Performance (RMSE) of various features at predicting personality scores. Mid-level detectors outperform other features at predicting most traits

public_location are detected with >75% accuracy with almost all the features. However, other cues need specific features for good detection performance. For example pressed lips is detected the best by LBP (followed by BOW), which is known to capture facial features well. Similarly emotional positivity is best detected by Fisher_SIFT and Aesthetic Features (which are designed specifically for emotion detection in natural images).

Age and camera height are the least performing detectors, probably because of the complex nature of age estimation and the absence of specific features to detect pose respectively. The final mid-level cue detectors we use for personality modeling are built using a late-fusion scheme of all features. It should be noted that mid-level detectors are not cent percent accurate (that is there exists a confidence level to their predictions) and consequent errors would propagate to personality predictions as well.

4.1.5.2 Predicting Personality

The results of predicting users' personality using mid-level cue detectors are shown in Figure 4.3. In predicting self-assessed personality (recognition) we find that mid-level cue detectors outperformed other features in agreeableness, conscientiousness and neuroticism. And in predicting others-assessed personality (perception), they outperformed other features in all traits except extraversion. In both case, they outperform the features LBP and SIFT, HOG (Fisher encodings), which are used by previous work [115, 178] on predicting personality based on facial images.

The lower performance in personality recognition, when compared to personality perception, might be due to users' manipulation of their self-presentation to create a socially desirable

Table 4.2: Significant Correlations ($p < 0.01$) b/w Self-Assessed Personality Traits and Mid-Level Cues

Self-Assessed trait	Sig. Cue	Corr.
Extraversion	-	-
Agreeableness	Emotional Positivity	0.18
	Camera Height	-0.20
Conscientiousness	Private Location	-0.20
Neuroticism	Duck Face	0.21
Openness	Emotional Positivity	0.22

self-image [173]. This implies that there is a possibility of users' exaggerating their expressions (and other mid-level cues), thereby leading to higher errors in personality recognition. However, this exaggeration might be visible to a person who is perceiving the users' personality, leading to a stronger association between mid-level cues and personality perception. There also exists the biases present in selfies posted on social media - most of the users seem to be smiling and having a positive look, thereby reducing the diversity of the self-presentations. This might be another reason for performance on personality perception to be higher.

To further understand which specific mid-level cue detectors were significant in predicting self-assessed traits, we performed correlation analysis (shown in Tables 4.2). The following are some observations:

1. Contradictory to a previous psychology finding where extraversion was related to positive emotional expression [169], it was not significantly correlated with any cue in our study. One reason could be that users, irrespective of their degree of extraversion, tend to show positive emotion in selfies to maintain positive self-impression [173].
2. Agreeableness was negatively correlated with camera height indicating that agreeable individuals more likely take selfies from below.
3. Conscientiousness was negatively correlated with private location indicating that conscientious people do not like to expose their private space in the background as corroborated by a psychology finding [182].
4. Neuroticism positively correlated with duckface indicating that neurotic individuals make duckfaces in their selfies.
5. Openness positively correlated with emotional positivity, which was not observed in previous findings.

Table 4.3: Significant Correlations ($p < 0.01$) b/w Others-Assessed Personality Traits and Mid-Level Cues

Others-Assessed trait	Sig. Cue	Corr.
Extraversion	Pressed Lips	-0.19
	Emotional Positivity	0.29
Agreeableness	Emotional Positivity	0.50
	Eyes Look at Camera	0.24
Conscientiousness	Duck Face	-0.31
	Emotional Positivity	0.25
	Location Information	0.30
	Public Location	0.25
	Photoshop Editing	-0.20
Neuroticism	Duck Face	0.25
	Emotional Positivity	-0.40
	Face Visibility	-0.21
	Amount of Body	-0.22
	Alone	0.22
Openness	Location Information	-0.19
	Emotional Positivity	-0.22
	Eyes Look at Camera	0.21
	Face Visibility	-0.26

A similar analysis on others-assessed traits, to identify cues that observers used when judging personality, revealed the following:

1. Rating on extraversion was negatively correlated with pressed lips, possibly because it's being seen as a symptom of shyness. Extraversion was positively correlated with emotional positivity, supporting the relationship between perception of extraversion and smiling [169].
2. Agreeableness ratings were positively correlated with eyes looking at camera, supporting that users who had eye contact are seen as more agreeable [183].
3. Rating on neuroticism was positively correlated with duckface and negatively correlated with face visibility, indicating that showing a full duckface implies that the user is moody.

4.1.6 Conclusion

In this work, we built mid-level cue detectors for selfie portrait images by exploiting several visual features. These mid-level detectors were then used to model personality of users in the selfies. To answer the research questions we posed in this work:

RQ1 In predicting self-assessed personality (recognition) we find that mid-level cue detectors outperformed other features in agreeableness, conscientiousness and neuroticism. And in predicting others-assessed personality (perception), they outperformed other features in all traits except extraversion. The lower performance in personality recognition, when compared to personality perception, might be due to users' manipulation of their self-presentation to create a socially desirable self-image [173].

RQ2 We performed correlation analysis on the different mid-level cues and the personality traits to understand selfie-taking behaviors of users with different personalities and how others would perceive the users based on their selfies. As mentioned in the previous section, some of the findings were corroborated by previous findings in Psychology literature, while others were novel.

It would also be interesting to explore the degree of exaggeration that users exhibit w.r.t different mid-level cues (for example, if the user is posing in a duckface to an extent that it looks awkward and uncomfortable vs. otherwise).

4.2 Personality from Image Posts

People are increasingly using social media platforms in order to keep records of their daily activities, and the selection of content they post reflects how they see the world around them including their activities and preferences. This interaction allows researchers to study how people differ in content sharing preferences. With the ubiquitous nature of camera-enabled mobile devices and Internet in many regions of the world, the complexity of content shared has increased in the recent years from simple text updates to multimedia such as images focused around platforms such as Twitter, Instagram or Snapchat [184]. Current computer vision algorithms allow us to automatically analyse large scale data sets and extract shallow color features as well as semantic content such as objects or scenes.

The aim of this work is to study if personality can be predicted from posted image content. For instance, extraversion is in part an orientation toward engaging with the external, social world, which may result in users high in this trait to post images of multiple people. Users high in openness to experience have an appreciation for art, which could lead them to post sketches or pictures of musical instruments.

Computational models that predict user traits based on their online footprints have several applications in targeted online marketing, increasing acceptance of HCI systems, personalized search and recommendations and exploring social science hypotheses based on large-scale social media data.

4.2.1 Research Questions

The questions we investigate in this work are:

RQ 1. How do personality differences manifest in users' image posting preferences on social media?

RQ 2. If and how well can we computationally predict personality traits from image posts?

4.2.2 Related Work

Recent research has examined the interplay between users' personality traits – measured using the Big Five model [185, 186] – and their social media data [187]. Facebook 'Likes' and

status updates were used to accurately infer users' personal information [188]. Users choice to disclose particular sections of their social media profile was used to study their personality traits [117]. While text posted by users has found great success for behavior analysis [189], images on social media are now increasingly being used due to their increased production and exchange in the recent years [184].

Previous work in personality analysis through images are either focused on profile images [114, 115, 116, 190] which could be a matter of self-presentation [171, 172] or use small data sets and shallow image features, thus limiting the scope and interpretation of results [117, 118]. Following the research in Psychology [7] highlighting the importance of freedom of control and self-expression, we use users' posts which offer users the opportunity to reveal more about themselves through diverse and more amount of data than they can in profile images.

4.2.3 Data

We use two Twitter data sets in our experiments which differ in size and the set of available user traits. Both studies received approval from the Institutional Review Board (IRB) of our institution.

First, we use a data set ($\mathcal{D}_1$) containing 436 Twitter users whose Big Five personality scores were computed based on the International Personality Item Pool proxy for the NEO Personality Inventory Revised (NEO-PI-R) [185]. These users also self-reported their age and gender (binary male–female). For each user we collected up to 3200 of their most recent tweets and downloaded all images that were embedded in these posts. In total, we downloaded 579,929 tweets which contained 34,875 embedded photos across 232 users. A total of 161 users posted at least 10 photos and we only include these in our experiments.

A second data set ($\mathcal{D}_2$) is an order of magnitude larger and consists of 4132 Twitter users who self-reported their age and gender while performing various studies (unrelated to personality). We do not have personality computed via surveys for this data set, as these are very costly and time-consuming to administer. Hence, similar to previous work on personality analysis from profile images [116], we use an automatic text-regression method to assign each user scores for the Big Five personality traits [189]. The model was trained on a sample of over 70,000 Facebook users, using tokens and topics extracted from status updates as features. In

their original validation, the model achieved a Pearson correlation of $r \sim .35$ predictive performance on average for all five traits [189, 191], which is considered a high correlation in psychology, especially when measuring internal states [192]. We downloaded the user tweets leading to a data set of 5,547,510 tweets, out of which 700,630 contained a photo across 3498 users. We exclude the texts of the tweets which contain images when predicting personality to limit any potential confound. Our image analysis is performed on the 2606 users who posted at least 20 photos.

4.2.4 Methodology

In order to uncover personality differences in image sharing, we extract multiple types of image features:

4.2.4.1 Feature Extraction

Colors: The colors of an image represent the most notable features to a user. Research has shown that colors can invoke emotions [193] or psychological traits [194]. Colors can be expressed in various color spaces. Here, we use the HSV (Hue–Saturation–Value) as this provides a more intuitive representation of colors for users [195]. A pixel in the HSV space is characterized by three numbers: (1) *Hue* – the color type ranging between 0 and 360 degrees e.g., 0 is red, 60 is yellow; (2) *Saturation* – the intensity of the color ranging from 0 to 1 e.g., 0 represents no color and is a shade of gray; (3) *Value* – the brightness of the color ranging from 0 to 1 e.g., 0 represents black. Using the images converted to the HSV space, we first determine whether an image is **grayscale** using the method from [48], which we ignore (4.99% of post images) from subsequent color analysis. We compute **saturation** and **brightness** as the average saturation and value respectively of the pixels in the image. An experimental study of colors established a linear relationship between saturation and brightness and the dimensional model of affect containing three factors: Pleasure, Arousal and Dominance [196, 197]. We also compute the standard deviation of the HSV values, which we will use for prediction only.

Using the hues of an image, we compute the **hue count** of a photo as a measure of its simplicity using the method described in [48]. Professional photos usually have a lower number of unique hues – although each color may be rich in tones – which leads to a simpler composition. In addition, we compute a 6-bin histogram which splits pixels into the primary and

secondary colors as well as a 12-bin histogram containing also the tertiary colors. Finally, we compute the ratio of pixels with warm colors. These features are used in prediction only. A subset of these features were used in a smaller experiment on social media image content and personality [118].

Image Content: Images can have very diverse content beyond faces, especially in the case of images embedded within tweets on a user’s Twitter feed. With the goal of aiding our analysis section, we use automatic content analysis techniques to generate tags for images.

Imagga: First, we use the Imagga tagging API¹ as our content analysis engine, which was previously used with good results [198]. We labeled all images with the Imagga Tagging API and generated for each image a *bag-of-tags* out of the top-10 predicted tags, based on the developers’ recommendations. We removed all tags that occurred less than 200 times in our large data set, leaving us with 1299 distinct tags.

Preliminary analysis showed multiple tags which were very similar and usually co-occurred together (e.g., glass – drink – beverage, waves – shoreline – seascape – seaside). Hence, we reduce the feature space in order to increase interpretability and decrease sparsity and the number of features used in the prediction experiments. To achieve this, we learn tag clusters that contain frequently co-occurring tags. We use a procedure that was originally applied to words in a tweet, which produced very accurate results for prediction and analysis [199]. First, we compute the Normalised Pointwise Mutual Information (NPMI) between all pairs of tags [200]. This measures the degree to which two tags are likely co-occur in the same context – here, image – and takes a maximum value of 1 if two tags always co-occur and a value of 0 if they occur as according to chance. By replacing negative values with 0, we can use this metric as a similarity measure and compute a tag $\times$ tag similarity matrix. We feed this matrix to the spectral clustering algorithm, a hard-clustering method appropriate for generating non-convex clusters [201, 202, 203], which performs a graph partitioning on the Laplacian of the similarity matrix. The number of clusters needs to be specified in advance, and, after preliminary experiments, we decided to use 400 clusters for the rest of the study. Once these clusters of semantically similar tags are created, we represent each image as a vector containing the normalized number of times each tag cluster is detected. For each user, we derive a feature vector

¹<http://docs.imagga.com/\#auto-tagging>

Table 4.4: $\mathcal{D}_2$: Pearson correlations between color features extracted from tweet images and user personality. Correlations for personality traits are controlled for age and gender. Positive correlations are highlighted with green ($p < .01$, two-tailed t-test) and negative correlations with red ($p < .01$, two-tailed t-test).

Feature	Ope.	Con.	Ext.	Agr.	Neu.
Colors					
Grayscale	.039	-.130	-.128	-.152	.262
Brightness	-.108	.040	.124	.027	-.020
Saturation	-.017	.023	.102	.076	-.077
Pleasure	-.017	.032	-.079	.037	-.024
Arousal	-.007	.005	.119	.048	-.054
Dominance	.005	-.013	.113	.010	-.021
Hue Count	-.094	.040	.118	.085	-.103
Content					
% Posts with People	-.106	.109	.116	.082	-.059
# Posted Images	.068	-.025	.094	-.141	.093

of *image content topics* as the normalized number of times each topic cluster is present in the tweet-embedded images.

VGG-Net: Secondly, we use a pre-trained version of the 19-layer VGG-Net image classifier based on convolutional networks [204], trained on the ImageNet data set to predict a class probability for 1000 objects in the ImageNet tagset. However, in preliminary analysis, we found that this tagset is very limited and does not contain categories of core interest for our analysis (e.g., faces, building interiors, sports), and hence we use it for personality prediction experiments only, and not in the correlation analysis. We also extracted features from the final fully-connected 4096 dimension layer of the convolutional network, but did not find significant results in prediction and thus omit from presenting these for brevity.

4.2.4.2 Correlation Analysis and Prediction

We perform univariate correlation tests between each feature and personality trait to uncover the associated image features. We control for age and gender effects using partial correlation so that our analysis is not skewed by demographic biases. Additionally, in content analysis, as some personality traits are inter-correlated, we control for all other four traits in order to isolate the peculiarities of each personality dimension.

We study the predictive power of image features in the task of user personality profiling. We treat this as a regression problem to which we apply a linear regression algorithm with Elastic Net regularization [205]. Since VGG and Imagga Clusters are high-dimensional features, we employed PCA for dimensionality reduction by setting the explained variance ratio threshold as 95% for $\mathcal{D}_2$, and choosing 51 principal components for $\mathcal{D}_2$ (owing to the small size of the dataset). We have also experimented with non-linear methods (Support Vector Machines with RBF kernel), but found that the results were not statistically significant better and hence we omit them. To evaluate our results, we split our data into 10 folds and performed cross-validation on one held-out fold at a time. For all our methods, we tune the parameters of all our models on a separate validation fold. The overall performance is assessed using Pearson correlation of the predicted value to the self-reported score.

4.2.5 Results and Discussion

4.2.5.1 Correlation Analysis

Results for our analysis using image colors and image content on the $\mathcal{D}_2$ data set are presented in Tables 4.4 and 4.5 respectively. The same set of experiments on personality correlations was conducted on the $\mathcal{D}_1$ data set, but the correlations are not robust to randomness. This shows the need for this behavior to be studied using larger samples. In Table 4.4 we included an additional feature coding the percentage of posts that contained people, identified through the presence of a curated list of Imagga Clusters in the top 10 predicted ones (46.4% of the images), and another coding the raw number of posted images.

Openness to Experience: Color analysis shows openness is associated with posting images in grayscale and having low hue count, which indicate artistic and well-composed images. Users high in this trait also prefer less bright images and are more likely to post a greater number of images. Image content analysis (Table 4.5) reveals two major themes for the users high in openness to experience. First, posting images containing drawings (*art, drawing, decoration* – clusters denoted by their first tag) indicates engagement with aesthetic domains, a known characteristic of openness to experience [206]. Secondly, posting content associated to travel (*ancient, building*) suggests a general interest in novel places and cultures. On the negative side, the users low in openness prefer posting images of sports or games (*player, football,*

game), which are conventional interests in the US, as well as images of people, especially children (*child*).

Conscientiousness: Conscientiousness is a trait associated with orderliness and preference for planned behaviour. Users high in this trait do not prefer grayscale images and prefer photos of people rather than objects, suggesting a preference for generally normative and conventional posting behavior. Content analysis reveals images containing either landscapes (*sky, building*) or formal, office environments (*classroom, office*). The latter corresponds to the fact that conscientious users have an orientation towards work. On the other hand, low conscientiousness is portrayed by images of drawings (*art, drawing*) or tech (*computer*).

Extraversion: Extraversion is associated with images that are bright, saturated and more chromatically complex as revealed through a high hue-count. This trait is characterized by engagement with other people and, in accordance to theory, users who score high are specifically interested in posting images of people (*people, face, women, body*). In addition, extraverts are perceived as being energetic and this is reflected by images portraying active interests (*disco*). Introversion is portrayed through drawings (*art, drawing*) or abstract symbols (*design, casual*) and by posting fewer images in general.

Agreeableness: Agreeableness shows similar patterns to extraversion in color associations, which is expected based on their common positive emotion core. Users high in agreeableness are characterized by kindness, generosity and optimism, and their images show a breadth of aspects of the good life: friends (*happiness, women*), pets (*cat*), flowers (*flower*) or music (*music*). On the other hand, low agreeableness is associated with images of office environments (*office, people*), text-related content (*press, newspaper*) and fewer photos of people overall.

Neuroticism : Neurotic users are associated most strongly with posting grayscale images. Their images are less saturated, contain fewer hues and tend not to contain people. Perhaps surprisingly, neurotic users post more images, although they prefer pictures of objects rather than people. Through content analysis, we observe that high neuroticism is disclosed through posting images containing documents (*paper*) or animals (*cat, animal*). Although they are less likely to post pictures of people overall, they favor a type of images containing people, perhaps attractive or sexy ones (*face, cute*). Users low in neuroticism are characterized by posting images of art and abstract symbols (*art, design*) and tech objects (*casual, device*).

Table 4.5: D_2 : Pearson correlations between image tag clusters extracted from **images embedded in tweets** and personality (maximum 8 per personality trait). All correlations are significant at $p < .01$, two-tailed t-test. Results for every personality trait are controlled for age and gender and other four corresponding personality traits. Tags are sorted by occurrence in our data set within a tag cluster.

r	Image Tag Clusters	r	Image Tag Clusters
Openness (+)		Openness (-)	
0.155	senior, old, mature, elderly, husband, grandma, retired, grandfather, retirement	0.207	ball, sports_equipment, game_equipment, basketball, basketball_equipment..
0.148	art, cartoon, clipart	0.195	player, athlete, baseball, contestant, base, baseball_equipment, ballplayer...
0.137	drawing, representation, diagram	0.154	football, helmet, back, football_helmet, crash_helmet
0.120	ancient, palace, castle, historic	0.119	game, puzzle, jigsaw_puzzle
0.119	decoration, sketch, tattoo, graffiti	0.117	clothing, garment, shirt, consumer_goods, jersey, sleeve
0.108	office, businessman, professional, corporate, executive, suit, handsome, men	0.109	tennis, racket
0.102	artwork	0.105	runner, diversion, track
0.093	building, architecture, city	0.099	child, boy, kid, little, childhood, baby, children
Conscientiousness (+)		Conscientiousness (-)	
0.153	classroom, whole	0.125	face, pretty, hair, model, fashion, sexy, lady, brunette, glamour
0.132	office, businessman, professional, corporate, executive, suit, handsome, men	0.111	art, cartoon, clipart
0.114	building, architecture, city	0.105	drawing, representation, diagram
0.104	structure, fountain, gas	0.101	expression, looking, beard, head
0.101	business	0.101	cute, eyes
0.099	paper, document, writing, menu, pen, fare, book	0.097	artwork
0.094	furniture, table	0.072	computer, equipment, technology, display,screen, monitor
0.092	sky, landscape, travel	0.068	people, person, adult, caucasian, man, male, happy, portrait, attractive, smile...
Extraversion (+)		Extraversion (-)	
0.321	people, person, adult, caucasian, man, male, happy, portrait, attractive, smile...	0.253	art, cartoon, clipart
0.243	face, pretty, hair, model, fashion, sexy, lady, brunette, glamour	0.213	drawing, representation, diagram
0.182	women, group, friends, friend, girls, friendship, buddy	0.165	design, sign, icon, graphic, symbol, internet, set, web, button, icons
0.175	lifestyle	0.149	casual, silhouette, sport, laptop, covering, businessperson, element, light...
0.171	happiness, couple, together, love, fun, two, family	0.128	black, african, picture, dark
0.136	disco, cabaret, spot, ballroom	0.127	artwork
0.127	body, swimsuit, slim, bikini, maillot, tights	0.125	text, 3d
0.120	expression, looking, beard, head	0.125	decoration, sketch, tattoo, graffiti
Agreeableness (+)		Agreeableness (-)	
0.144	happiness, couple, together, love, fun, two, family	0.211	office, businessman, professional, corporate, executive, suit, handsome, men
0.108	performance, concert, stage, platform, musician	0.177	press, print_media
0.098	flower, floral, garden, flowers, petal, blossom, flora, bloom	0.132	business
0.094	cat, feline, pet, domestic, fur, kitten, kitty, pets, domestic_cat, whiskers, furry...	0.129	newspaper, money, currency, finance, cash, bank, banking, financial, bill
0.094	trees, season	0.101	work, success, manager, confident, successful
0.093	women, group, friends, friend, girls, friendship, buddy	0.099	signboard, billboard, scoreboard, street_sign
0.089	decoration, sketch, tattoo, graffiti	0.092	people, person, adult, caucasian, man, male, happy, portrait, attractive, smile...
0.081	music, guitar, stringed_instrument	0.081	publication, magazine, book_jacket, comic_book, jacket, wrapping
Neuroticism (+)		Neuroticism (-)	
0.107	paper, document, writing, menu, pen, fare, book	0.108	casual, silhouette, sport, laptop, covering, businessperson, element, light...
0.104	happiness, couple, together, love, fun, two, family	0.096	computer, equipment, technology, display, screen, monitor
0.104	cute, eyes	0.094	device, machine, slot_machine, slot, vending_machine
0.099	face, pretty, hair, model, fashion, sexy, lady, brunette, glamour	0.064	design, sign, icon, graphic, symbol, internet, set, web...
0.092	animal, dog, domestic_animal, canine	0.063	art, cartoon, clipart
0.082	cat, feline, pet, domestic, fur, kitten, kitty, pets, domestic_cat, whiskers, furry...	0.060	communication, telephone, phone, mobile, call, cellular_telephone
0.079	people, person, adult, caucasian, man, male, happy, portrait, attractive, smile...	0.057	space, digital, fractal, laser, optical_device, glow, render
0.072	women, group, friends, friend, girls, friendship, buddy		

4.2.5.2 Prediction

Results from the prediction experiments are presented in Tables 4.6 and 4.7. Feature combination is performed using a linear ensemble over the individual predictions of each feature set. The same patterns hold when evaluating the results with Root Mean Squared Error for regression and we omit them for clarity.

On the $\mathcal{D}_1$ data set, colors and imagga clusters seem to outperform VGG-Net classes possibly because of the lack of diversity in the Imagenet classes that VGG-Net is trained upon. It is to be noted that the small size of the data set might be acting as an artifact. On the $\mathcal{D}_2$ data set, image content features extracted from either VGG-Net or Imagga performed better than color features for openness, conscientiousness and extraversion, while for agreeableness and neuroticism, colors outperform the other two feature sets. Combining all features using an ensemble works better in predicting all traits, showing colors and content are complimentary to some extent. Prediction performance using images is overall high, especially in the case of openness and neuroticism.

Table 4.6: $\mathcal{D}_1$: Prediction results for personality traits with all features. Performance is measured using Pearson correlation in 10-fold cross-validation.

Feature	# Feat.	Ope	Con	Ext	Agr	Neu
Colors	33	.093	.147	.022	.229	.085
Imagga Clusters	400	.044	.077	.183	.085	.404
VGG-net Classes	1000	.056	.051	.061	.037	.222
All (Image)	3	.094	.177	.187	.229	.416

Table 4.7: $\mathcal{D}_2$: Prediction results for personality traits with all features. Performance is measured using Pearson correlation in 10-fold cross-validation.

Feature	# Feat.	Ope	Con	Ext	Agr	Neu
Colors	33	.251	.343	.272	.289	.417
Imagga Clusters	400	.263	.381	.317	.221	.383
VGG-net Classes	1000	.410	.383	.319	.198	.398
All (Image)	3	.448	.479	.369	.336	.503

4.2.6 Conclusion

In this work, we analyze image posting preferences using interpretable aesthetic and semantic features using an order of magnitude larger data set than previous work, with the aim of identifying the way and extent to which personality differences measured using the Big Five model are related to online image sharing. We created tag clusters in order to analyze the content of images shared online and identified idiosyncrasies of each personality trait, while controlling for all other four traits in order to isolate the peculiarities of each personality dimension. Answering the questions we posed in this work:

RQ1 Findings from color analysis show that openness is associated with posting images in grayscale and having low hue count, which indicate artistic, well-composed, less brighter images and are more likely to post more images; conscientious users do not prefer unconventional grayscale images and prefer photos of people rather than objects; extraversion and agreeableness is associated with images that are bright, saturated and more chromatically complex as revealed through a high hue-count; neurotic users post more images, with a preference for objects than people and are associated most strongly with posting grayscale, less saturated and images containing fewer hues. Image content analysis shows that users high in openness to experience post images containing drawings and those associated to travel, and users low in openness prefer posting images of sports or games as well as images of people, especially children; conscientious users post about landscapes or formal, office environments, and users with low in conscientiousness post images of drawings or tech; extraverted users specifically post images of people and portray active interests, and introversion is portrayed through drawings or abstract symbols and by posting fewer images in general; agreeable users's posts show a breadth of aspects of the good life such as friends, pets, flowers or music, and low agreeableness is associated with images of office environments, text-related content and fewer photos of people overall; and neuroticism is disclosed through images containing documents or animals and a type of images containing people whereas emotionally stable users are characterized by posting images of art and abstract symbols and tech objects.

RQ2 Predictive models of personality traits from image posts show reliable accuracy on held-out data using image features.

Chapter 5

Deep Feature Representations to Model Users' Likes

This chapter presents an investigative study into learning feature representations which are effective in discriminating users based on images they like, and consequently which can capture the differences in semantics of images that different users like. The notion of like is very subjective and subtle, and hence hard to describe by methods or processes.

In a computational manner, it is challenging to ascribe reasons or find out the factors that led to a user liking a single image. However, if we know a set of images that a user has liked (which is quite a common place with social media sites where users openly express their preferences), we can computationally find out the factors that contribute towards the user liking the set of images. These factors can be affective in nature (viz. the emotional message in the image), the concepts present in the image (viz. the objects and their perspectives/poses ,etc.), the aesthetic and artistic aspects of the image, a relatable associated context, and any combination of these factors; all of which induce the user to have some connection with the image.

We aim to address the problem of user likes in two ways: a. discriminating users based on the semantic characteristics of images that they liked and consequently b. using these differences to recommend images that a given user likes. Precisely, the question this work investigates is: given that we have a way to represent quantitatively and model some of these semantic characteristics (using features mentioned in Table 5.1), can we estimate how these factors influence the differences in different users' image liking behavior?

Figure 5.1 shows how different features vary in the images 'liked' by different users. Users' patterns often overlap, thereby making the problem of distinguishing their preferences difficult.

is often captured by the tags associated with an image [208]. While tags primarily convey the concepts in an image which visual detectors can be trained to detect (for example, the objects in the image, etc.), in many cases they convey semantic concepts that go beyond what traditional visual detectors can be trained to detect (See Figure 5.2).

We aim to learn layered deep representations for images based on a broader collection of content specific, contextual, aesthetic, and affective factors that induce different users to like these images, a process we term as modeling user likes throughout the text. We use the notation ‘high-level features’ to denote them (see 5.1). The quantitative values for the factors are measured from the visual and textual information (tags) associated with the image. At every layer of the representation there is knowledge ‘translation’ between the textual and visual domain. We concatenate the visual and tag features at every layer and use this as the representation for every image. The efficacy of the proposed approach (usage of high-level features and the learning process of deep bimodal representations) is evaluated in an recommender scenario. Along with this test-set performance of the deep bi-modal representations, an attempt is made to qualitatively understand the representations learned. Throughout the text, ‘high-level features’ refers to the features mentioned in Section 5.4 and ‘low-level features’ refers to the features mentioned in [2]. The proposed deep representation is denoted as ‘deep bimodal feature representation’.

5.1 Research Questions

In this work, we investigate the following specific questions:

- RQ 1. How do high-level features compare with low-level features in representing user likes measured by $recall@K$ performance (as we have positive only feedback for training)? Recall@K is recall evaluated at a given cut-off rank K, considering only the topmost results returned by the system.
- RQ 2. How well does the proposed system detect features that influence a user the most in ‘liking’ an image?
- RQ 3. How well do the bimodal deep representations perform in modeling the users’ preferences when compared to individual modalities?

Figure 5.2: **Random images ‘liked’ by 5 users on Flickr.** Users’ 1-4 have certain visual patterns in the way they seem to like images (people vs. no people; b/w vs. color; scenery vs. cluttered, etc.). However, in User 5’s case, the images seem to be very diverse w.r.t the visual content (car, squirrel, bird, people, railway station, etc.), the tags convey that the user was probably looking at vacation images in London. In modeling such users’ likes where context is important, tags augment the information conveyed by visual features.

RQ 4. What do the mid-level feature representations learn and how do we evaluate if they are learning something meaningful?

While the first two questions deal with the efficacy of using the high-level features (without any deep learning involved) in representing user likes, the third question deals with the efficacy of using the learned bimodal deep representations in representing user likes. The fourth question deals with giving an interpretation to what the hidden layer neurons are learning.

5.2 Related Work

Several works need a mention to justify clearly the scope and context of this work. The previous section explains the issues with understanding user likes and our approach to learn deep mid-level representations that help map images to the user response like. The approach of learning deep representations has been very successfully applied to understanding concepts in images [209, 210], scene labeling [211], image retrieval [212, 213], activity classification [214] and

video tracking [215]. However, there are some clear differences that prevent straightforward application of existing ‘concept detection’ tools to the problem at hand.

Difference from Deep Representations for Concepts Detection There is a clear difference between the two objectives – concept detection and user like prediction. In the learning process, the supervisory information available for modeling a ‘concept’ (such as an object, scene, etc.) is explicit, can be visually verified, and is clearly content dependent. However the supervisory information available for modeling user likes is implicit and not visually obvious or content dependent. Thus the mid-level representations learnt for concept detection are also content dependent and in most cases visually verifiable. In contrast, the mid level representations for user likes go beyond the content. They would factor in contextual [209], aesthetic (as modeled by [216] using style attributes) and affective (as applied by [217]) factors and thus making the modeling process even more challenging.

Difference from Multi-Modal Fusion Approaches A popular method named Canonical Correlation Analysis (CCA) [208]) essentially finds linear projections of two random vectors that are maximally correlated. Kernel (KCCA) [218] versions of such linear projection based methods have also been proposed where the time required to train (i.e. to compute the joint representations) scales poorly with training data size. One fundamental idea behind such correlation based approaches to fusing visual and textual information is the assumption that both modalities are *content-based related* with each other, hence the correlation, which is useful for content-based tasks like concept detection, image retrieval [219, 220], etc. However, when we try to model user likes, we are hoping for both modalities to *complement* each other by contributing in terms of affective, aesthetic and contextual information also. In fact the reason for a like may not have anything to do with the prominent concepts in the image. Hence, we need better representations that entail these features at different levels of abstraction.

The quantitative values for the factors are measured from the visual and textual information (tags) associated with the image. At every layer of the representation there is knowledge ‘translation’ between the textual and visual domain. We concatenate the visual and tag features at every layer and use this as the representation for every image. The initial idea of feature translation is inspired from [221] where heterogenous feature mapping was done on textual information from different languages domains. However, our proposed method is different from [221] in two aspects:

- The algorithm in [221] is for heterogeneous transfer learning [219, 222], while we focus on learning multi-modal representations by augmenting features for classification.
- The deep structure in [221] learns feature representation in the same layer across domains, however we propose a new architecture for multi-modal representations as shown in Figure 5.4 based on the intuitions presented in Section 5.4. The translators learnt across layers are expected to aid in capturing more information about user likes, thereby increasing the model performance.

5.3 Data

The data set for our experiments consists of a collection of 4000 images (with their associated tags) crawled from Flickr [223]. 200 images tagged as favourite per user are downloaded along with the associated tags for 20 users. For every image, we extract the features mentioned in Section 5.4 using the default-parameters of the software provided by the cited works. We used manual help to assure that the accuracy of features like gender is 100%. We also extract the low-level features mentioned in [2] for each image. The dataset used in [2] does not contain tags associated with the images and thus is not suitable for our purpose.

5.4 Methodology

5.4.1 Feature Extraction

After motivating the use of high-level features such as people- and content-oriented characteristics for representing users' likes, we describe the features we propose to use in our work.

People Presence of people is a significant characteristic in images which influences several users to like an image. Research shows that human brain is particularly sensitive to faces, due to the presence of specific neural pathways [224] and hence the presence/absence can influence users with different traits (say Extraversion) to like/dislike images.

- **Head and Upperbody recognition** Perceptions of an image can depend significantly on the presence or absence of people in it. And persons' visibility is often partial in images, typically showing only the upper body. To address this, an upper-body detector [225]

was used. And for the cases where only the head was in the field of view, head detector was used [226].

- **Face and Pose recognition** Where people in an image look (towards/away from the camera) also influences users' liking for images. Face detection and pose estimation by [227] was used. Pose angle characterizing profile/frontal face along with relative area of the face was taken.
- **Gender identification** To ascertain the distribution of gender in images, we used IntraFace [228]. We took support of manual annotation to prevent any errors.

Table 5.1: List of Features. High-Level features are described in Section 6.4 and Low-Level features are taken from [4].

Category	Type of Feature	Name of the Feature
High-Level	Content	Saliency - Classification between iconic and spread out image [229] Scene Features: PRICOLBP [230] Visual Clutter [231] Color vs. Black & White Image [232] Object Detectors [233] Graphics Image vs. Natural Image [234]
	People	# Females in the image [228] # Males in the image [228] Face Detection - No. of faces and their orientation [frontal/pose], Relative area of the face wrt image [227] Head Detection [226] Upperbody Detection [225]
Low-Level	Affect	Color Channel Statistics Use of Light - Average Pixel intensity of V channel HSV Statistics - Mean of S channel and standard deviation of S, V channels; <i>Hue angular dispersion</i> in IHLS Emotion Statistics: Pleasure, Arousal and Dominance Colorfulness based on Earth Movers Distance Color Names - Amount of Black, Blue, Brown, Green, Gray, Orange, Pink, Purple, Red, White, Yellow
	Aesthetics	Low Depth of Field - Amount of focus sharpness in the inner part of the image Rule of Thirds - Mean of S,V channels in the inner rectangle of the image Regions - Average size of regions (after mean shift segmentation) Image Parameters (Size) Edge Statistics - Total number of edge points extracted with Canny Level of Detail (Watershed Segmentation)
	Texture	Wavelet Textures - Level of spatial graininess measured with a three-level Daubechies wavelet transform on the HSV channels Tamura Features - Amount of Coarseness, Contrast, Directionality GLCM Statistics - Amount of Contrast, Correlation, Energy, Homogeneity for each HSV channel Image Entropy

Content Users with different personality traits can have varying perceptions of an image based on the content. For instance, those with anxious and tense behavior seem to have different tastes from those who have an easygoing outlook [235, 236]. Also, achieving satisfactory performance in predicting self-assessments depends on 'Availability' [237] which means that the traits must be perceivable to others. Therefore, to represent the high-level semantic concepts from contents in images (which are easier to perceive) based on the motivations presented in the introduction to this Chapter, we extract the following features.

- **Scene Features** Scene classification has been one of the classical problems which researchers have been trying to address. The performance of this task is based on the features that are used. PRICoLBP feature [230] due to its effectiveness in classifying scenes.
- **Graphics rendered images** Attractive computer rendered images often capture the attention of people, and they are exclusively posted and discussed on Flickr [238]. Also, natural images are liked by many. Therefore, to distinguish this cue, we used the feature in [234], to classify images as computer graphics or natural images.
- **Image Saliency** Attention of users can be determined based on image saliency. When images having a scenic backdrop are compared with images with certain objects in the focus, the scale of image can be understood by its saliency. When there is nothing in the image to capture the users' attention, the saliency is usually spread out. This would be the case with highly focused images and those which have a scenic backdrop. But images with clear objects show sharp saliency spikes, which is the usual case with iconic images or those with clear objects. To distinguish between both, saliency was used [229] to compute the mean maximal intensity points in images relative to a threshold.
- **Color vs. black and white image** To check if an image is black and white or a color image, we used the color naming technique proposed in [138]. This is a binary feature where 1 represents a color image and 0 represents a black and white image.
- **Visual Clutter** Busyness of an image can be described by its clutter. As the number of objects in natural scenes is difficult to measure, clutter was measured using feature congestion metric developed in [231].

- **Tags:** We construct a dictionary of tags (with an average of 9.1 tags per image). We removed some stop words like camera names ('kodak', 'fujifilm', 'canon', etc.), lens characteristics ('70mm', 'eos', etc.), and 'generic' words like 'image' and 'photo', just having a single letter ('a-Z') or number ('0-9'). The tag feature matrix is binary and sparse unlike the visual feature matrix. We then apply sparse SVD to convert the sparse tags feature matrix into a compressed representation (refer to [208]).
- **Features used in [2]:** Texture, HSV Statistics, Edge Statistics, Rule of Thirds, etc. (see Table 5.1). These are denoted as low-level features.

5.4.2 Learning Deep Bimodal Feature Representation

The framework for modeling user likes is depicted in Figure 5.3. This includes the design steps for bimodal deep knowledge representation and the process of training a model for discriminating users based on images they liked and also to recommend images that a given user liked. There are two separate parts in this framework. (a) Identifying the features in the initial feature collection which influence a user the most in liking the image, which gives us an idea of the features among the initial collection of features that would be considered relevant if we did not have mid-level representations. (b) Learning the deep bimodal representation of images. The feature selection step and learning deep representations are designed for two different purposes. In feature selection step, we aim to identify features which distinguish each user's liking behavior. In learning deep bimodal representation, the goal is to learn, in an unsupervised manner, multi-layer non-linear representations of all visual and tag features, as we believe that non-linear combinations of features from both modalities influence users liking for images. The feature selection step is not carried forward in the pipeline and is completely independent.

We use deep networks to learn multi-layer nonlinear representations of both tag and visual features individually. At every layer, we "translate" the features from the tag domain into the visual domain. We concatenate the visual features and the "translated" tag features at every layer and use this as the representation for every image. The entire process is described formally in Algorithm 1.

Figure 5.3: Framework for modeling user likes. There are two separate parts in this framework. (a) Identifying features which influence a user in liking the image. (b) Learning the deep bimodal representation of images.

5.4.2.1 Multi-layer Homogeneous Feature Learning

Marginalized Stacked Denoising Autoencoders (mSDA) [239] are used to form a deep network wherein layer-wise nonlinear transformations of the Visual and Textual features of an image are learned. We choose mSDA because of its advantages, like faster training time and implementation simplicity, when compared to other forms of Autoencoders. The mid-level representations thus learned by the deep network are found to be very effective features for classification with SVMs [209, 221].

In presenting the formulation, we follow the notations used in [239] for simplicity. We absorb a constant feature into the selected feature vector as $\mathbf{x}_V = [\mathbf{x}_V^\top \ 1]^\top$ or $\mathbf{x}_T = [\mathbf{x}_T^\top \ 1]^\top$, and incorporate a bias term $\mathbf{b}$ within the weight matrix as $\mathbf{W}_V = [\mathbf{W}_V \ \mathbf{b}_V]$. We further denote $\mathbf{X}_V$ as Visual Domain data, and $\mathbf{X}_T$ the Tag Domain data. In this way, we can simply learn the bias term $\mathbf{b}$ for linear representation as follows $\mathbf{x}_V = \mathbf{W}_V \mathbf{x}_V + \mathbf{b}_V = [\mathbf{W}_V \ \mathbf{b}_V][\mathbf{x}_V^\top \ 1]^\top$.

Firstly, for the Visual Domain data, we apply mSDA on $\mathbf{X}_V$ to learn a weight matrix $\mathbf{W}_V \in \mathbb{R}^{(d_V+1) \times (d_V+1)}$ by minimizing the squared reconstruction loss as follows:

$$\sum_{i=1}^m \left\| \mathbf{X}_V - \mathbf{W}_V \mathbf{X}_V^{(i)} \right\|_F^2, \quad (5.1)$$

where $\mathbf{X}_V^{(i)}$ denotes the i -th corrupted version of $\mathbf{X}_V$. The marginalised denoised stacked autoencoder takes a matrix of observations, makes it noisy and finds optimal weights for a linear transformation to reconstruct the original values. To achieve non-linear encoding, output of each layer is filtered through tanh function. Corruption is achieved by binary masking noise. Corruption in stacked denoising autoencoders refers to any change or distortion to the input. This is done so as to learn a stable and robust feature representation wrt the input distribution. In mSDA, a fraction of the features of input features are set to zero. For details of the algorithm, we point to the original work [239]. The solution to (5.1) depends on how the original features are corrupted which can be explicitly expressed as follows,

$$\mathbf{W}_V = \mathbf{P}\mathbf{Q}^{-1} \quad \text{with } \mathbf{Q} = \tilde{\mathbf{X}}_V \tilde{\mathbf{X}}_V^\top \text{ and } \mathbf{P} = \hat{\mathbf{X}}_V \tilde{\mathbf{X}}_V^\top, \quad (5.2)$$

where $\hat{\mathbf{X}}_V = [\mathbf{X}_V \ \mathbf{X}_V \ \cdots \ \mathbf{X}_V]$ denotes the m -times repeated version of $\mathbf{X}_V$, and $\tilde{\mathbf{X}}_V$ is the corrupted version of $\hat{\mathbf{X}}_V$. In general, to alleviate bias in estimation, a large number of m over the training data with random corruptions are required, which is computationally expensive. To address this issue, mSDA introduces a corruption probability p to model infinite corruptions, i.e. $m \rightarrow \infty$. Then a feature vector $\mathbf{q} = [1-p, \dots, 1-p, 1]^\top \in \mathbb{R}^{d_V+1}$ is defined, where $\mathbf{q}_i$ represents the probability of a feature indexed by i “surviving” after the corruption. Thus, we can obtain the expectation of (5.1), and its solution can be written analytically as

$$\mathbf{W}_V = \mathbb{E}[\mathbf{P}]\mathbb{E}[\mathbf{Q}]^{-1}, \quad (5.3)$$

where $\mathbb{E}[\mathbf{P}]_{ij} = \mathbf{S}_{ij}\mathbf{q}_j$, $\mathbf{S} = \mathbf{X}_V \mathbf{X}_V^\top$, and

$$\mathbb{E}[\mathbf{P}]_{ij} = \begin{cases} \mathbf{S}_{ij}\mathbf{q}_i, & \text{if } i \neq j, \\ \mathbf{S}_{ij}\mathbf{q}_j, & \text{otherwise.} \end{cases} \quad (5.4)$$

After $\mathbf{W}_V$ is learned, the nonlinearity of features is injected through the nonlinear encoder function $h(\cdot)$ that is learned together with the reconstruction weights $\mathbf{W}_V$, and mSDA applies a nonlinear squashing-function, e.g., the hyperbolic tangent function $\tanh(\cdot)$, on the outputs of mSDA, $\mathbf{h}_V = \tanh(\mathbf{W}_V \mathbf{X}_V)$, to generate nonlinear features.

5.4.2.2 Cross-layer Heterogeneous Feature Mapping

So far, in a specific layer k of feature learning, we have learned a pair of reconstruction weights $\mathbf{W}_V^k$ and $\mathbf{W}_T^k$, and higher-level feature representations $\mathbf{h}_V^k$ and $\mathbf{h}_T^k$ for the visual and

tags domain data, respectively (architecture shown in Figure 5.4). By denoting $\mathbf{h}_V^k$ and $\mathbf{h}_T^j$ the layer k and layer j feature representations of the cross-domain corresponding instances in the visual and tag domains, respectively, we now introduce how to learn a feature mapping across heterogeneous features $\mathbf{h}_V^k$ and $\mathbf{h}_T^j$.

Algorithm 1 Deep Heterogenous Feature Mapping.

Input: tag domain data $\mathbf{D}_T = \{\mathbf{x}_{T_i}\}_{i=1}^{n_1}$, visual domain data $\mathbf{D}_V = \{(\mathbf{x}_{V_i}, y_{V_i})\}_{i=1}^{n_2}$, a feature corruption probability p in mSDA, a trade-off parameter λ , and the number of layers K .

Initializations: $\mathbf{X}_V, \mathbf{X}_T, \mathbf{h}_V^1 = \mathbf{X}_V, \mathbf{h}_T^1 = \mathbf{X}_T$, and learn $\mathbf{G}_1$ by solving

$$\min_{\mathbf{G}_{1,1}} \|\mathbf{h}_V^1 - \mathbf{G}_{1,1} \mathbf{h}_{T,1}\|_F^2 + \lambda \|\mathbf{G}_{1,1}\|_F^2.$$

Run FGM to select relevant features

for $i = 2, \dots, K$ **do**

1: Apply mSDA on $\mathbf{h}_V^{i-1}$ and $\mathbf{h}_T^{i-1}$:

$$\begin{aligned} \{\mathbf{W}_V^i, \mathbf{h}_V^i\} &= \text{mSDA}(\mathbf{h}_V^{i-1}, p), \\ \{\mathbf{W}_T^i, \mathbf{h}_T^i\} &= \text{mSDA}(\mathbf{h}_T^{i-1}, p). \end{aligned}$$

end for

for $j = 1, \dots, K$ **do**

for $k = j + 1, \dots, K$ **do**

2: Learn heterogeneous feature mapping $\mathbf{G}_{k,j}$:

$$\min_{\mathbf{G}_{k,j}} \|\mathbf{h}_V^k - \mathbf{G}_{k,j} \mathbf{h}_T^j\|_F^2 + \lambda \|\mathbf{G}_{k,j}\|_F^2.$$

end for

end for

Do feature augmentation on visual domain data and mapped tag domain data

$$\mathbf{Z}_V = [\mathbf{h}_V^k \top \cdots \mathbf{G}_{k,j} \mathbf{h}_T^K \top]^\top,$$

and train a classifier f with $\{\mathbf{Z}_V, \mathbf{Y}_V\}$.

Output: $f, \{\mathbf{G}_{k,j}\}_{k,j=1}^K, \{\mathbf{W}_V^k\}_{k=2}^K$, and $\{\mathbf{W}_T^k\}_{k=2}^K$.

Specifically, from layer j of the text domain to layer k of the visual domain, we aim to learn a feature transformation $\mathbf{G}_{k,j} \in \mathbb{R}^{(d_V+1) \times (d_T+1)}$, where a bias term is incorporated by minimizing the following objective:

$$\|\mathbf{h}_V^k - \mathbf{G}_{k,j} \mathbf{h}_T^j\|_F^2 + \lambda \|\mathbf{G}_{k,j}\|_F^2, \quad (5.5)$$

Figure 5.4: Bi-Modal Deep Feature Representation - Multi-Layer Homogenous Feature Learning and Cross-Layer Heterogenous Feature Mapping.

where $\lambda > 0$ is a parameter of the regularization term on $\mathbf{G}_{k,j}$, which controls the tradeoff between the alignment of heterogeneous features and the complexity of $\mathbf{G}_{k,j}$. It can be shown that the optimization problem (5.5) has a closed form solution which can be written as follows:

$$\mathbf{G}_{k,j} = (\mathbf{h}_V^k \mathbf{h}_T^j{}^\top)(\mathbf{h}_T^j \mathbf{h}_T^j{}^\top + \lambda \mathbf{I})^{-1}, \quad (5.6)$$

where $\mathbf{I}$ is the identity matrix of the dimensionality $d_T + 1$.

We note that learning the feature mapping between a layer of tags and all the corresponding higher layers of visual features might probably capture the semantics of the image better than mapping between the same layers. For example, given a picture of a sunset – tag associated with it might be ‘sunset’ whereas the visual features will depict the colors in the image like ‘red’, ‘yellow’ and low clutter in the image and so on. We can see that the tag features in this case are at a higher level of semantics when compared to the visual features, therefore making the mapping between this layer of tags and higher layers of visual features more meaningful.

5.4.3 Interpreting Mid-Level Representations

In order to get the middle level features, we apply the algorithm marginalised Denoising AutoEncoder (mDAE) as follows:

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{m} \sum_{j=1}^m l(\mathbf{x}_i, f_\theta(\tilde{\mathbf{x}}_i^j)), \quad (5.7)$$

where m is the number of corruption versions, $\theta = [\mathbf{W}, \mathbf{b}]$ is the combination of learnt weights and bias on the features, and $f_\theta(\tilde{\mathbf{x}})$ is its reconstruction function of corruption $\tilde{\mathbf{x}}$. Similar to

mSDA, we calculate the expectation of the loss function $l(\mathbf{x}_i, f_\theta(\tilde{\mathbf{x}}_i^j))$ with the first two orders of the Taylor expansion at the mean $\mu_{\mathbf{x}}$ as follows:

$$\min_{\theta} \mathbb{E}[l(\mathbf{x}_i, f_\theta(\tilde{\mathbf{x}}_i^j))] \approx l(\mathbf{x}_i, f_\theta(\tilde{\mathbf{x}}_i^j)) + \frac{1}{2} \text{tr}(\Sigma_{\mathbf{x}} \nabla_{\tilde{\mathbf{x}}}^2 l), \quad (5.8)$$

where $\Sigma_{\mathbf{x}} = \mathbb{E}[(\tilde{\mathbf{x}} - \mu_{\mathbf{x}})(\tilde{\mathbf{x}} - \mu_{\mathbf{x}})^\top]$ and $\nabla_{\tilde{\mathbf{x}}}^2 l$ is the second-order derivative.

After learning θ , the hidden representation $\mathbf{z}$ is computed as $\mathbf{z} = h(\mathbf{W}\tilde{\mathbf{x}} + b)$.

Thereby each neuron in the middle layers (namely layers 1 and 2), can be expressed as a non-linear combination of the features in the earlier layers (namely layers 0 and 1, respectively). The above equation gives us the weights on each of these features. A positive weight on a feature indicates that it is activating the neuron and a negative weight indicates that it is inhibiting the neuron.

With this, for every neuron in the middle layers, we have a set of activators and a set of inhibitors. As the features in layer 0 are semantic features (both low-level and high-level), the activators and inhibitors of neurons at layer 1 can be seen as two sets of semantic features, one set which contribute towards firing the neuron and the other which contributes towards inhibiting it. In the case of vectors, the size represents the average activation from the whole vector since we would like to know how the entire vector (which represents a feature) influences neuron activation. Again, in each of these clusters, depending on the weight assigned to each feature, the degree of each feature's activation/inhibition capacity is calculated. In the case of layer 2, the learnt representations in each neuron can be understood to be a set of learnt representations in layer 1 (i.e. set of sets of semantic features at layer 0).

We attempt to represent these learnt representations as WordClouds where weight on the features denotes the corresponding size in the visualization.

5.5 Results and Discussion

We describe results from experiments conducted to answer the research questions [RQ1, RQ2, RQ3, RQ4] raised after the Introduction of this work. Section 5.5.1 deals with question RQ1, Section 5.5.2 deals with RQ2, Section 5.5.3 deals with RQ3 and RQ4 is answered in Section 5.5.4.

To recap, the first two questions deal with the efficacy of using the high-level features (without any deep learning involved) in representing user likes, the third question deals with the

Figure 5.5: High-Level features require lesser amount of training data to capture users' preferences when compared to low-level features.

efficacy of using the learned bimodal deep representations in representing user likes. The fourth question deals with giving an interpretation to what the hidden layer neurons are learning.

5.5.1 Comparing high-level features with low-level features

We conducted two experiments to verify the efficacy of high-level features. In the first experiment we assessed the performance of predicting the user who 'liked' a given set of images and in the second experiment we assessed the performance of retrieving the images 'liked' by a given user.

Experiment 1: For each user, we divide the data (images) into training and test sets (with three splits as follows: a) 25%/75% b) 50%/50% and c) 75%/25%). The training data is used to learn a one-vs-all classifier (FGM) for discriminating users. Feature selection using the FGM method was used for all training and test data. For each split we examine the recall@k (for $K = 3$ and 5) of the user being correctly predicted for the test images (shown in Figure 5.5).

When low-level features are used, the model's performance varies based on the amount of training data, but when high-level features are used, the model performs well with a lesser amount of training data. The reason for this can be ascertained by the amount of information the model is able to gather with each image based on the features. While low-level features would need a lot of examples to identify the patterns in the users' preferences, the semantic nature of the high-level features ensures a stable modeling of user likes.

Figure 5.6: Image retrieval performance to users is higher with high-level features than by using low-level features.

Experiment 2: We used the above one-vs-all models trained on a 50/50% training/test split for each user to retrieve ‘liked’ images by each user (i.e. 100 images ‘liked’ by the user and a random sample from the images ‘liked’ by the other users as positive and negative samples for training the model). After all user models are trained, a total of 3800 images is provided as test images for each user model. We then ranked the test images for each user based on the score given by the classifiers and we obtained the mean recall curves with high-level features and low-level features (shown in Figure 5.6). Here, data imbalance is likely to affect the modeling process when we take 100% of the images liked by a user as positive samples and all the remaining images (i.e. all the images liked by the other users) as negative samples. However, to avoid this, we randomly choose 50% of the images (100 images) liked by a user as positive samples and 100 images (out of the remaining images) as negative samples. Similar protocol was also followed in [2]. This is a very rudimentary way to model positive-only feedback data – there exist several advanced approaches to handle the same problem [240]. However, in this work, our goal is to understand the importance of feature representations used and how to learn effective representations for the problem at hand and therefore use the protocol used in [2] to arrive at a baseline approach.

For many users, the model also retrieves images in the top-k which were not actually tagged as favorite by the user, but upon visual inspection we found that they had very similar characteristics to that of the images that the user tagged as favorite (See Figure 5.7). These images are

Figure 5.7: Top 30 test images for 4 random users based on the score given by the models discussed in Section 5.5.1. Images outlined are recommendations, from the test set, offered to each user by our model.

not in the set of images that each user liked. As we know only the relevant images (user likes), we use recall as the metric to measure the model's performance. Within the images ranked as top 100, high-level features can retrieve about 30% of the images 'liked' by the user while low-level features can retrieve only 15%. Also high-level features are able to retrieve all the images 'liked' by the user around 15% faster than low-level features, which again confirms the results shown in the previous experiment (in Figure 5.5).

5.5.2 Features Relevant to each User

The features described in Section 5.4 are semantic concepts to represent an image. However it is not clear what influences the differences in the 'liking' behavior of different users, as intuitively only a specific set of features might contribute for each user to like an image. Formally, the selection of most informative features for users can be expressed as solving a sparse SVM formulation as follows (similar to [241]):

$$\min_{\mathbf{d}} \min_{\mathbf{w}, \epsilon, \rho} \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{C}{2} \sum_{i=1}^n \epsilon_i^2 - \rho \quad (5.9)$$

$$s.t. \quad y_i \mathbf{w}^\top (\mathbf{x}_i \odot \mathbf{d}) \geq \rho - \epsilon_i, \quad (5.10)$$

where y_i denotes binary value, +1 if image i is liked by the user, -1 otherwise, $\mathbf{d} \in \mathcal{D} = \{\|\mathbf{d}\|_0 \leq B, \mathbf{d}_j \in \{0, 1\}\}$ and 1, 0 indicate that the feature is selected or not, respectively. $\mathbf{x}$ denotes the feature vector, $\mathbf{w}$ denotes the weight vector, ρ is the bias, ϵ_i is the i th instances loss incurred by classifiers, B is the number of features to be selected and C is the trade-off parameter. Solving the above optimization is an NP-hard problem with lots of possible combinations of features. To reduce the computation cost, we adopt the Feature Generation Machine (FGM) [241] method to learn $\mathbf{d}$ and $\mathbf{w}$ simultaneously, which represent the index of selected features and their corresponding weight. This helps us to do feature selection while achieving the task of classification. For more details of solving this problem, please refer to [241, 242].

While conducting the above experiment, we also note the features selected by FGM to answer the second question. We find that most users have preferences which can be captured by high-level semantic features. The weights learnt, along with random images liked by three users, are depicted in Figure 5.8. We apply SVD to convert sparse tag feature matrix into a compressed representation. We concatenate this representation with the visual features and conduct experiments.

We observe that User C likes images which have pictures with people (predominantly women), User B likes colorful images, and User A likes cluttered images which have a tinge of graphics in them. Also we note that there are users who have certain common preferences, for example, both Users A and B prefer high clutter in images. Similarly there were other instances in the data set where more than one user had similar preferences for the presence of people in the image (for example User's 3 and 4 in Figure 5.2). This observation can be used to build many interesting applications, of which one is providing recommendations to a group of people (by clustering them according to personality, culture, etc.). While applying feature selection, some of the dimensions of the tag feature representation are also selected, however we did not mention it in Figure 5.8 as they no longer have an intuitive meaning. However, their contribution in terms of performance can be seen in the later experiment where we report the performance of visual features, tag features and their combination (Figure 5.9).

5.5.3 Comparing bimodal deep representation with individual modalities

We then train multiple layers of features for each modality using the model and use the heterogenous feature mapping described in Section 5.4 to learn mapping from the tag features

Figure 5.8: Features weighted for three users and a sample of images they tagged as favorite. Users' preferences can be represented using high-level visual features

to the visual features. The intuition behind using a stacked auto-encoder is that pre-training multiple layers of features can help in capturing the characteristics of an image better (due to the non-linearities). Then, we concatenate features in all the layers to form the 'deep-feature' vector. To verify the intuition behind multi-layer pre-training and also to examine the efficacy of the trained bimodal (visual+tag) representation, we compare the performance of every layer of individual modalities in discriminating the users' likes with that of the combined bimodal representation. This is shown in Figure 5.9.

We randomly divide the data set into 25%/75% training and test sets and train a linear SVM model (using a 5-fold cross validation setting on the training set) on the combined (visual+tag) features. To see if the features are able to distinguish between the users' preferences, we test the recall of predicting the right user, given an image. Modeling user likes is a personalization problem which is user centric, where the number of images liked by each user might play a more important role than the number of users themselves. Keeping this issue in mind, we considered 200 images for each user, which is a similar number to that used by [2, 4]. And we test the method used in [2] on our dataset and use that as the baseline for comparing the performance of our method.

The results (in Figure 5.9) show a comparison of the performance of the above model under different settings (i.e. using visual features alone, using tag features alone, and using bimodal features). The following observations can be made based on Figure 5.9:

Figure 5.9: **Comparison of bimodal deep representation with individual modalities.** Each graph shows the recall of predicting the user who tagged test images as favorite in the top-k attempts. The baseline performance (with features used in [2]) is mentioned for each top-k setting. Learning multi-layer bimodal representation gives better performance than using shallow individual modalities. Layer 0 indicates the original features and Layers 1 & 2 the deep features for each modalities. 'Combined' indicates the performance of the bimodal representation at each layer.

- A 15%-20% increase in performance is achieved at each top-k setting by using the deep bimodal representation when compared to the baseline performance.
- A 5-8% increase in performance can be attributed to the use of new proposed features when compared to the baseline performance and 3-8% increase in performance is observed due to the inclusion of tags in the network.
- Hierarchical representations give a 5-10% increase in performance when compared to shallow models trained on corresponding individual modalities.
- From the top-1 to top-5 there is an improvement of over 35-40% using bimodal deep representation.
- Visual features model user likes better than tag features (by 20-30%), which is different from the applications like image search, etc. where tag features are more important.
- When compared with individual modalities, the combined representation improves the performance by 10-30% which shows that a combination of multiple modalities can model user likes better than individual modalities.

Figure 5.10: **WordClouds representations of Activating and Inhibiting Features for 2 neurons in the network.** Each of the words represents a feature we used as input to train the network. For 2 randomly chosen neurons, activating features are represented by WordClouds with a white background and inhibiting features are represented by those with a black background.

We also found in our experiments that simple concatenation of visual and tag features gives inferior performance (5%-10% at each the layer) when compared with learning a bimodal representation using feature mapping from the tag domain to the visual domain. This confirms that simple concatenation of features from different domains does not lead to a good representation [208] and feature mapping is necessary to learn effective bimodal representations to model user likes.

5.5.4 Interpreting the Mid-Level Representations

Based on the procedure defined in Section 5.4, we visualize the learnt mid-level representations.

Two sets of activating and inhibiting features for two random neurons are represented in Figure 5.10. The first set shows that the neuron gets activated by a component of Gray-Level Cooccurrence Matrix (GLCM) features, presence of upper bodies (or people), the colors white, purple and orange in the image, and so on. The corresponding inhibiting features show that the neuron is deactivated/inhibited by colorful images, Tamura features, the presence of a lot of edges, the color blue, and so on. This is represented for all neurons in Figure 5.11.

We thought of multiple ways to verify the semantic feasibility of co-occurrence of these features, like using the wordnet/sysnet hierarchies and other linguistic techniques. However, giving a semantic representation of texture and other low-level features which could be used by linguistic analysis modules was a challenge.

Therefore, we tried to examine which combinations of features activate different neurons. Essentially, this can be seen as a representation to see which features are activating different neurons and which are inhibiting them. While some neurons are activated by almost all the

Figure 5.11: **Matrix showing the features learnt by neurons at layer 1.** A white dash represents features activating the neurons and a black dash represents features inhibiting them. Weights have been thresholded to clearly visualize the strongest activating/inhibiting features. While some neurons seem to be activated by almost all the features, others are activated only for a particular set of features. Such neurons specifically might help in distinguishing the users' likes

features, there are specific neurons which are activated predominantly by features (and their combinations) of specific neurons. These neurons might help in distinguishing users likes. A specific example is shown in Figure 5.12. An example image ‘liked’ by a user is taken and the neurons getting activated by this image, the weights on features that these neurons have learnt and other images that these neurons get activated by are shown in the Figure.

We also identify two neurons, namely, the strongest and weakest neurons, and identify the images activating them to visually see the difference. The neuron with the largest difference between the mean and median of the feature weight distribution is denoted as strongest (w.r.t. the response it evokes) and the neuron with the smallest difference is called the weakest. The KL divergence is taken as the distance metric between these and the feature counts of the images to retrieve the top 10 images for both cases. It can be seen that the images activating the strongest neuron have more intra image diversity w.r.t. color and other image properties whereas the images activating the weakest neuron are more uniform.

A similar analysis can be done for neurons in layer 2. However, it is tough to present visualizations as we did for layer 1 as each neuron in layer 2 deals with a set of ‘set of semantic features’. Although there are works which try to explain deep representations learnt by networks (which take pixels as the input) by visualizing weights learnt by the hidden layers, to our

Figure 5.12: **Specific example to describe the visualizations in Figures 5.11.** The image in the example has the feature counts as shown by the barplot below it. This image activates 6 Neurons in layer 1. The boxplot of the features learnt by those 6 Neurons is shown to the right of the image.

Figure 5.13: **Top 10 images activating the strongest and weakest neurons with KL-Divergence as the distance metric.** Images activating the strongest neuron are have more intra image diversity wrt color and other image properties whereas the images activating the weakest neuron are more uniform.

knowledge this is the first work which analyzes mid-level representations learnt on semantic inputs. No doubt, a lot of study is yet to be done to analyse the deep representations learnt for semantic features to gain a deeper understanding into the kind of features that can be extracted, designed and learnt for very abstract user response tasks such as predicting likes.

5.6 Conclusion

In this work, we attempted to present a learnt feature representation which is effective in discriminating users based on images they like, and consequently which can capture the differences in semantics of images that the different users like. On a Flickr dataset, several syntactic, semantic, aesthetic and contextual features were used to build a deep knowledge representation for images (using visual and textual information). A feature selection strategy was applied to learn the most influential features for understanding the features which stand out to explain the differences in users' liking behavior. Deep bimodal representation was learnt using a novel approach for knowledge transfer between the tag domain and the visual domain of images to model user likes. Returning to the research questions posed at the outset of the paper, we can now state that:

- RQ1 We saw that while low-level features would need a lot of examples to identify the patterns in the users' preferences, the semantic nature of the high-level features ensures a stable modeling of user likes 5.5. Within the images ranked as top 100, high-level features can retrieve about 30% of the images 'liked' by the user while low-level features can retrieve only 15%. Also high-level features are able to retrieve all the images 'liked' by the user around 15% faster than low-level features 5.7.
- RQ2 We find that most users have preferences which can be captured by high-level semantic features – representing concepts such as people, colors, clutter, computer generated graphics and so on 5.8.
- RQ3 The performance of the proposed model was studied under different settings (i.e. using visual features alone, using tag features alone, and using bimodal features). A 15%-20% increase in performance was achieved when compared to shallow models trained on low-level features (and a 5%-10% increase when compared to shallow models trained on high-level features) 5.9.

RQ4 We also attempted to understand what the mid-level representations mean through the medium of wordclouds 5.10 and the feature co-occurrence matrix in the dataset 5.12.

In future work, it would be interesting to investigate how semantic feature representations of tags (say word vectors) can be exploited to explain the interaction between the combined modalities [243] at the task of predicting user likes on a larger dataset.

Chapter 6

Modeling Personality based on Image Likes

This work addresses the problem of computationally modeling users' personality based on images they 'like'. With the advent of social media, multimedia has become an interaction channel among users. Users convey innate preferences and tastes through implicit feedback ('likes' on Facebook, 'favorite' on Flickr etc.), which leave adequate cues to predict users' personality [188].

Images contain several concepts, which can be represented by features such as objects, scenes, faces, colorfulness, etc. Users' 'likes' on images can be attributed to the presence/absence of these features (and their combinations). For example, one user may prominently like colorful images with a lot of people while another user may like black and white images of natural landscapes. It would be interesting to investigate if there is a relationship between such preferences ('likes') and users' personality. Given a dataset of images liked by certain users, can we identify the concepts which influence the users with different personality traits to like certain images? If yes, can we computationally model this relationship thereby automating users' personality assessment? These are some of the questions that motivate this work.

To assess users' personality, researchers in psychology have developed several psychometric tests. However, the downside of such psychometric tests is that users have to explicitly give their feedback to questionnaires. By automating this process of personality assessment, several domains can be enhanced – personalizing recommender systems, increasing acceptance of HCI systems, adapting assistive technologies involving robots, and so on. These applications can be enabled by taking into consideration users' innate qualities that explain *why* they behave in certain ways.

There are two dimensions to automatic personality prediction [97]: a) *personality recognition* and b) *personality perception* (which was discussed in Chapter 4 Section 4.1). Often it is found that there is a lot of data related to users (such as videos/products liked/rated on social media and other websites), but no information about their personality profiles. It is also infeasible, in many cases, to ask users to answer the personality questionnaires. However, it is less troublesome to get assessors to build the personality profile of the users by asking them to examine users' data. This can be a huge step forward in non-intrusive personality acquisition techniques, which has been shown to be preferred the most by users in [244]. In this work, we focus on modeling the self-assessed personality traits.

Most of the works in automatic personality modeling take the approach of extracting features from data (related to users) and using them directly to model personality of users (termed as Features-to-Personality - F2P). However, research in psychology [84] shows that personality is better ascertained by asking users a set of well-designed questions. The answers to these questions are then converted into scores for five personality traits. Taking this approach (F2A+A2P: Features to Answers + Answers to Personality) with features extracted from data was shown to be more effective than predicting the personality scores directly [245]. In this work, we evaluate both these approaches towards personality modeling. *Based on the images that different users like, we aim to predict their personality along these five traits.*

Section 6.2 describes related works in personality computing literature and the prior work in personality modeling based on implicit feedback on images. Section 6.4 proposes a new set of features that complement feature sets in existing literature. A novel personality modeling approach based on these features is presented in Section 6.4. Section 6.5 presents an evaluation and analysis of both feature representation and personality modeling in the context of recommender applications.

6.1 Research Questions

Specifically, we attempt to answer the following questions:

RQ 1. Identify image feature representations that best address the problem of modeling users' personality based on images they 'like'. (Section 6.4)

Figure 6.1: **Proposed pipeline of Features to Answers and Answers to Personality:**

a-b) User likes images and answers self-assessed BFI-10 questionnaire; **c)** Assessors look at images liked by the user and answer ‘others-assessed’ BFI-10; **d)** Features are extracted from images liked by the user; **e-f)** F2A model (proposed FS-OR) is used to predict answers to BFI-10 questions and BFI-10 scoring scheme (A2P) is used to calculate scores on five personality traits. **g)** F2P model is used to predict personality directly from image features.

F2P approach is used by current computational models in Personality Computing. We compare its performance with our proposed F2A+A2P (FS-OR) approach.

RQ 2. How to model users’ personality based on these image representations (features)? (Section 6.4)

RQ 3. What concepts (features) are most important to model different personality traits? (Section 6.5)

The above questions are answered in the context of a recommender application that models user image association based on implicit feedback (‘likes’). The recommender models user image interactions to serve in two scenarios:

- Given a user (represented by personality profile), what kind of images will he/she like?
- Given an image, find the user/group (represented by a personality profile), who will like it.

6.2 Related Work

Most works in automatic personality modeling use explicit information about the users (such as their facial expressions, speech prosody features, textual features, etc. from users) to model their personality. To the best of our knowledge, works which use implicit feedback (for example, characteristics of images ‘liked’) to model users’ personality is [4, 188]. The factor that differentiates our work from [188] is that they just use the implicit feedback matrix by the users on several categories like movies, books, etc. However we attempt to model users’ personality based on the characteristics (features) associated with images they like. In [4], images liked by a group of Flickr users were collected in [4]. The users were asked to answer the BFI-10 questions [84] to get their self-assessed personality profile.

The process of automatically assessing the personality profile of users involved learning a regression model (LASSO [246]) mapping ‘low-level’ image features to their personality profile. The difference in self-assessment and others-assessment of personality was studied and the following observations were made [4]:

- Others-assessment depends on the consensus among different external assessors and therefore, with a reasonable level of accuracy, can be computationally predicted.
- Self-assessment depends on an individual assessment and the learned model does not generalize well.

Similar observations, regarding predicting self-assessed and others-assessed profiles, were also reported in other works [99, 247]. The reason for high performance at predicting others-assessed personality scores is that machines and assessors have access to the same information and can only work within the bounds of limited information provided to them, in this example the behavior of users’ ‘likes’ on images. However, this does not apply to predicting self-assessed scores as users’ assessment of their personality is based on factors which certainly go beyond the images they like (for instance their personal experiences). Therefore, if factors which are likely to better represent a users’ self-assessed personality profile are considered, it might improve the prediction performance, which was also our intuition in using interpretable mid-level cues for representing selfies, and large-scale image content analysis for representing image posts.

F2A+A2P Approach We take an F2A+A2P approach that has been shown [245] to model personality traits better than the F2P approach. However there are some differences with [245] distinguishing the application of the F2A+A2P approach to our situation. F2A+A2P was used in [245] where the F2A stage was implemented using SLoT for predicting the personality scores of people inside movies. This was done using explicit cues (such as facial expressions, prosodic features, etc.) about the subjects whose personality was being modelled from the movies. However, we propose a novel machine learning technique Feature Selection based Ordinal Regression: FS-OR (described in Section 6.4) to model the F2A stage using implicit cues (from images liked by users). This requires investigating a diverse set of features to represent the users’ personality expressed through their likes (which is more challenging than modelings users’ personality through explicit cues such as facial expressions, etc.).

Personality prediction is thus divided into two stages (shown in Figure 6.1), namely transforming features to answers (F2A) and mapping the answers to trait scores (A2P). For the first stage, i.e transforming the feature space into answers, FS-OR is used (which is an improvement over our earlier work [248]). The motivation for using this approach is two-fold. (a) Each personality trait (and answers corresponding to its calculation) might only be influenced by a few features (motivating the feature selection in the machine learning pipeline). (b) The values to answers are ordinal in nature (motivating the use of ordinal regression in formulating the problem). And for the second stage (A2P), BFI 10 scoring scheme [84] is used to predict the personality profiles.

6.3 Data

The PsychoFlickr dataset was collected by [4]. It consists of 60,000 images where 200 images each are liked by 300 ‘pro’ Flickr users. For every user, personality assessment was done in two ways: (a) Self-assessed scoring, where users were asked to fill BFI-10 [84] questionnaire, recording their personality in terms FFM model traits. The users responded to items like “I am a reserved person”, etc. (b) Others-assessed (attributed) scoring, where eight assessors were assigned to look at users’ liked images and answer the BFI-10 questionnaire. The assessors responded to items like “This person is reserved”, where the “person” is the user who ‘liked’ the images under examination. Thus every user has a collection of 200 ‘liked’ images and

Table 6.1: Krippendorff’s α measuring reliability of others-assessment, taken from [4]

Trait	O	C	E	A	N
α	0.07	0.16	0.23	0.19	0.20

Table 6.2: Spearman correlations (along with p-values) between self-assessed and others-assessed personality scores.

	O	C	E	A	N
ρ	0.171	0.112	0.049	0.088	0.02
p	0.003	0.054	0.393	0.131	0.730

also 2 sets of personality scores: one by the assessors and another by the users themselves. Apart from the dataset given by the authors of [4], we additionally downloaded the user profile information for all the users using the Flickr API.

The reliability of ‘others-assessed’ scores was measured using Krippendorff’s α [249] (shown in Table 6.1). α is a robust to small sample sizes and the values reported in Table 6.1 are statistically significant and in line with those in literature [250, 251, 252]. Spearmann correlations between corresponding traits of self-assessed and others-assessed personality scores are shown in Table 6.2. Self-assessments seen to reflect the actual personality of a user [84]. Assessment by others are a consequence of attributions and thus when provided by persons who are acquainted with users (like family members, etc.) tend to have high correlations [253]. However, in this dataset, the assessment is done by total strangers (to verify the effect of Brunswick Lens model [93]), and thus the low correlations on some of the traits is expected. Taking this into consideration, we focus on the prediction of self-assessed personality traits. However, we report the results on others-assessed personality traits using our method in comparison to the current state-of-the-art. The latter results on others-assessed traits warrant further validation due to the nature of the dataset – low reliability scores (α).

Context plays a very important role when analysing user behavior. Flickr, in particular, has differences w.r.t the influence of social circle (such as friends) when compared to other networks such as Instagram (and Facebook). The dataset considered in this study is a set of ‘Pro’ users on Flickr, who paid a fee to use the service which is an indicator of their passion for photography. Though some Instagram users also use that platform in this manner, it is largely found that people tend to like/react to images posted by others in their social network

Figure 6.2: Need for including Semantic Features. Left: High score on Conscientiousness; Center: Low score on Extraversion; Right: High score on Neuroticism.

(friends, family etc.) owing to the link with Facebook. Therefore, artefacts due to the social circle are relatively much lower on Flickr when compared to Instagram.

6.4 Methodology

6.4.1 Feature Extraction

Following the lessons learned in the previous work, we employ interpretable semantic features to represent images. The performance improvement seen in predicting likes due to the deep bi-modal representation of tags and visual features combined is expected to be present even when applied for modeling personality. However, we did not have the tags associated with images. However, most of the users' Flickr accounts from this database are no longer active and thus we could not crawl the tags associated with the images to experiment empirically and find out the exact improvement that the deep architecture could bring to personality prediction.

In this section, we first describe the intuition behind using 'high-level' semantic features especially from the view of capturing the differences in personality traits. The aim of these features is to provide a better representation of users' likes and how they relate to different personality profiles. The list of features used is similar to those described in the previous Chapter in Section 5.4. The list is in Table 5.1.

Need for Semantic features to capture individual preferences Depending on a person's psychophysical nature, one is drawn to different 'kinds' of images. To make this distinction, some amount of domain knowledge has to be included to capture the differences in what different people like. For example, Figure 6.2 shows images liked by users with high scores on

Figure 6.3: Sample of the images ‘liked’ by users with different personality profiles. Left: High score on Openness; Center: High score on Extraversion; Right side: High score on Agreeableness.

different personality traits in the dataset used in this study. The left of the collage shows images liked by a user with high score on conscientiousness. This trait reflects in socially prescribed impulse control and a sense of thoughtful behavior. Images liked by such users consist of carefully planned and timed shots and many black & white images (which is shown to be a sign of focus, subtlety of tones and versatility [254]). While users with a high score on extraversion have images with a lot of people, users with a low score have the opposite preference - consisting of scenic backdrops, without faces or objects to focus upon (images in the middle of the same collage). Also, people with a high score on neuroticism are characterised by anxious and tense behavior. This is conveyed by the images they like (images to the right of the collage), most of which have a high level of clutter [231].

Based on these observations, we hypothesize that personality traits should ideally be influenced by ‘high-level’ semantic concepts in images, each dealing with multiple facets of human behavior [255]. Therefore, we propose to use features such as identifying presence of people and understanding their attributes (gender, count, profile vs. front view of face, etc.) and also by representing the content of images by categorizing them as black & white vs. color, quantifying the amount of clutter and scene characteristics.

Relationship between images ‘liked’ and personality profile: Here, we describe how ‘high-level’ semantic concepts relate to likes of users with different personality profiles. Figure 6.3 shows images liked by three users. The left part of the collage show images ‘liked’ by a user with a high score on Openness. These images show a bias towards dance, music and artistic photography. Images liked by a user having a high score on Extraversion are shown in the

middle. This trait reflects an energetic approach towards the social and material world. This means that the user is expected to have a high affinity towards friends and partners - the images liked by the user can be seen to convey, in this case, a liking towards people of the opposite gender. Images on the right are liked by a user with a high score on Agreeableness, which reflects traits such as altruism, tender-mindedness, trust and, in behaviors like consoling friends who are upset, etc. The images liked by the user convey a similar inclination.

Each of the five personality traits convey specific innate tastes of users which lead to certain behavioral attributes. Features extracted from the data can be used to detect the above-described behavioral attributes (which are less subtler than personality traits themselves), thereby accurately predicting users' personality.

Therefore, we propose to use features which can represent the high-level semantics in the image (such as the presence of people in images and their attributes, along with other content-related characteristics) apart from aesthetics and other low-level features extracted from the images. These high-level semantics are expected to give a better representation of a user's likes and thereby his/her personality.

6.4.2 Feature Selection based Ordinal Regression (FS-OR) Approach

We define the notations we use for our formulation in Table 6.3. We are given the answers-images matrix, where the rows of the matrix represent the 10 answers to the BFI-10 questions and columns represent the different images. We can treat the classification of each answer as an ordinal regression (OR) problem. The advantage of OR over the normal regression is that OR considers the order information in the regression. Take the answer to the first question for example, 1 is least and 5 is strongest. In this case the decision function $f(C_1)$ should satisfy $f(C_1) < f(C_2) < f(C_3) < f(C_4) < f(C_5)$, where C_i represents the i^{th} scale of given answer.

Converting Ordinal Regression Problem into Binary Classification Problem To handle the proposed OR problem, we adopt the extended binary classification model [256]. In this model, an instance $\mathbf{x}$ is categorised as the i^{th} ordinal class when the predicted output $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ is in the range of $\theta_{i-1} < h(\mathbf{x}) \leq \theta_i$, where the ordered thresholds satisfying $\theta_1 < \theta_2 < \dots < \theta_{N-1}$. To learn these thresholds, we first transform the training data instances $\mathbf{x}_j \in \mathbb{R}^d$

and their corresponding class label $y_j \in \{C_1, C_2, \dots, C_{N_C}\}, j = 1, \dots, n$ as:

$$\begin{aligned} \mathbf{x}_j^k &= [\mathbf{x}_j^\top, \mathbf{e}_k^\top]^\top \\ y_j^k &= 1 - 2I[y_j \leq k] \\ \text{for } k &= 1, 2, \dots, N_C - 1 \end{aligned} \quad (6.1)$$

where $\mathbf{e}_k \in \mathbb{R}^{N_C-1}$ represents a vector with the k^{th} entry as one and remaining entries as zero. Each training instance $\mathbf{x}_j$ is duplicated $N_C - 1$ times and relabelled into $N_C - 1$ binary cases. Notice that an extended binary instance $\mathbf{x}_j$ has a dimension of $d + N_C - 1$. Therefore, with the duplicated training data, one can train a classifier with an augmented weight vector $\hat{\mathbf{w}} = [\mathbf{w}^\top, -\boldsymbol{\theta}^\top]^\top \in \mathbb{R}^{d+N_C-1}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_{N_C-1})$. This model renders the prediction in a straightforward way as follows:

$$f(\mathbf{x}^k) = \text{sign}((\hat{\mathbf{w}}^\top \mathbf{x}^k)) = \text{sign}(h(\mathbf{x}) - \theta_k)$$

where $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. Using Eq. 6.2, the predictive class label of instance $\mathbf{x}$ is given as follows:

$$\sum_{k=1}^{N_C-1} I[f(\mathbf{x}^k) = 1] + 1, \quad (6.2)$$

where $I(\cdot)$ is indicator function. To obtain the label prediction for an unlabeled data instance from the OR classifier, one projects the data instance onto the weight vector, and assigns a label by comparing the confidence value with N_{C1} threshold values. From the definition of $\hat{w}$, we know that θ is absorbed into $\hat{w}$ such that the optimization of SVM can automatically optimize θ and w simultaneously. For more detailed optimization proof, we request the readers to refer to reference [257].

Selecting Relevant Features After extracting the high-level features, only a subset of features are related to each question. Therefore, we need to select related features which benefits most for OR in each question. We modify the Feature Generation Machine (FGM) [258, 259] to select the relevant features. Note that in Section 6.4, we have converted the OR problem into a binary classification problem. So we can directly apply FGM here with an additional

Table 6.3: Notations used in FS-OR formulation

N_C	number of classes
d	dimension of vector
$\mathbf{x}_j$	data instance j
$\mathbf{y}_j$	label of instance j
I	indicator function
x_j^k	k^{th} duplication of instance x_j
y_j^k	label of x_j^k
$\mathbf{w}$	weight vectors for SVM
$\hat{\mathbf{w}}$	augmented weighted vectors for OR
$\boldsymbol{\theta}$	threshold vectors for SVM
ρ	SVM margin
ϵ_i^k	loss of x_i^k
$\mathbf{d}$	Feature selection indicator vectors
$\boldsymbol{\alpha}$	Lagrange variables
$\mathcal{C}$	selected features index set
$\mathcal{D}$	feasible set of feature selection indicator vectors

constraint of indicator vector $\mathbf{d}$ specific for this application.

$$\begin{aligned}
 \min_{\mathbf{d} \in \mathcal{D}} \min_{\hat{\mathbf{w}}, \epsilon, \rho} & \frac{1}{2} \|\hat{\mathbf{w}}\|_2^2 + \frac{\lambda}{2} \sum_k \sum_{i=1}^n \epsilon_i^{k^2} - \rho \\
 s.t. & y_i \hat{\mathbf{w}}^\top (\hat{\mathbf{x}}_i^k \odot \mathbf{d}) \geq \rho - \epsilon_i^k, \\
 & i = 1, \dots, n; k = 1, \dots, N_C - 1,
 \end{aligned} \tag{6.3}$$

where ρ denotes SVM margin, ϵ_i^k denotes loss of data $\mathbf{x}_i^k$, $\mathcal{D} = \{\mathbf{d} | \sum_{j=1}^d \mathbf{d}_j \leq B + N_C - 1, \mathbf{d}(d - N_{C+1} : \text{end}) = 1, \mathbf{d}_j \in \{0, 1\}\}$, $\mathbf{d}$ is the feature selection vector with entries 0, or 1 and B is parameter to control the number of selected features and $\mathbf{d}(d - N_{C+1} : \text{end}) = 1$ denotes last N_C element in $\mathbf{d}$ must be selected for thresholds for OR. In the following, we denote the proposed model feature selection based ordinal regression in short for FS - OR.

6.4.3 Algorithm for Feature Selection based Ordinal Regression

Observe that the above problem is an NP-hard problem. However, following the derivation of [258], we can convert the original problem into the following dual problem:

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\mathbf{d} \in \mathcal{D}} -\frac{1}{2} \left\| \sum_{i=1}^{(N_C-1)n} \boldsymbol{\alpha}_i y_i (\mathbf{x}_i \odot \mathbf{d}) \right\|^2 - \frac{1}{2\lambda} \boldsymbol{\alpha}' \boldsymbol{\alpha} \tag{6.4}$$

Algorithm 2 Feature Selection based Ordinal Regression (FS-OR).

Input: Training data $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, where $y_i \in \{C_1, C_2, \dots, C_{N_C}\}$.

Data Conversion: Converting Ordinal Regression Problem into binary class according to formulation .

1: Initialize α_i as $\frac{1}{n}$; Find the most violated features $\mathbf{d}$ and let $\mathcal{C} = \{\mathbf{d}\}$.

repeat

2: Initialize $\mu = 1$;

3: Update α by MKL.

4: Find the most violated features $\mathbf{d}$ and let $\mathcal{C} = \mathcal{C} \cup \{\mathbf{d}\}$

until Converged

Output: Selected features set $\mathcal{C}$.

Extended OR Training: We shrink the features as $\mathbf{x}^k(\mathcal{C})$ and feed data into SVM to train the model.

Prediction: For any test data $\mathbf{x}$, we first select features according to $\mathcal{C}$, namely $\mathbf{x}(\mathcal{C})$. Then substitute it into the following form:

$$f(\mathbf{x}) = \{k : \theta_{k-1} < h(\mathbf{x}) \leq \theta_k, \quad \text{where } h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}(\mathcal{C})\} \quad (6.5)$$

where $\mathcal{A} = \{\alpha \mid \sum \alpha_i = 1, \alpha_i \geq 0, i = 1, \dots, n\}$.

Further details on the optimization can be seen in Appendix 1. Following the derivation of [258], this problem can be reduced into a Multiple Kernel Learning (MKL) problem, which can be solved by simple MKL [260] or its corresponding primal problem directly [261]. However, we observe that problem itself is very expensive due to the huge number of base kernels (see appendix). Fortunately, only small subset of $\mathcal{D}$ is active at optimality. With this the problem can be solved by the cutting plain algorithm [262], which can be divided into iterative two steps: 1) fixing $\mathcal{D}$ and updating α with MKL 2) fixing α and finding the active subset of $\mathcal{D}$.

Now the problem is how to find the most violated $\mathbf{d}$ from $\mathcal{D}$ each time with fixed α . To find active subset of $\mathcal{D}$, we have to solve an equivalent problem as mentioned below:

$$\max_{\mathbf{d} \in \mathcal{D}} \frac{1}{2} \left\| \sum_{i=1}^{(N_C-1)n} \alpha_i y_i (\mathbf{x}_i \circ \mathbf{d}) \right\|^2 - \frac{1}{2\lambda} \alpha' \alpha = \frac{1}{2} \sum_{j=1}^n c_j^2 \mathbf{d}_j \quad (6.6)$$

where $c_j = \sum_{i=1}^{(N_C-1)n} \alpha_i y_i \mathbf{x}_{ij}$, α is obtained by MKL. Different from [258], our feasible set $\mathcal{D}$ of $\mathbf{d}$ has an additional constraint $\mathbf{d}(d - N_C + 1 : \text{end}) = 1$. Fortunately, we can still update the $\mathbf{d}$ by sorting first $d - N_C$ entries with fixing the last $N_C - 1$ entries as 1.

After learning the selected features index set $\mathcal{C}$, we further shrink the features according to $\mathcal{C}$ and feed data into SVM to train the model. For any test data $\mathbf{x}$, we first select features

according to $\mathcal{C}$, namely $\mathbf{x}(\mathcal{C})$. Then substitute it into the following form

$$f(\mathbf{x}) = \{k : \theta_{k-1} < h(\mathbf{x}) \leq \theta_k, \text{ where } h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}(\mathcal{C}).\} \quad (6.7)$$

The algorithm is depicted in Algorithm 2.

6.5 Results and Discussion

We describe results from experiments conducted to answer the research questions [RQ1, RQ2, RQ3] raised after the Introduction section. Section 6.4.1 deals with questions, RQ1 and RQ2 and Section 6.4.2 deals with question RQ3. Section 6.4.3 investigates the model predictive performance for a recommender application. To recapitulate, the research questions pertain to

- a.) investigating the efficacy of proposed features set,
- b.) evaluating the F2A+A2P personality modeling approach, and
- c.) evaluating our contributions on a personality modeling based recommender application.

Experimental Set Up For all the experiments, unless specified otherwise, the data split is 75% training with 5-fold cross validation and 25% testing. Results are reported on test data. In the experiments, ‘low-level features’ refers to the features used in [4] and ‘combined features’ refers to the combined set of low-level features and high-level features (Table 5.1). This notation is used just for the sake of better discrimination of results between current state-of-the-art and our approach and not to define them as such. The traits, for which performance of models is being reported, are: conscientiousness (C), openness to experience (O), agreeableness (A), extroversion (E) and neuroticism (N).

6.5.1 Evaluating Modeling Approaches and Features

F2P approach: LASSO regression was taken as the baseline (to compare with [4]). Secondly, to account for sparsity constraint (i.e. only a few features contributing to each answer as explained in Section 6.2), a sparse SVM modeling approach was chosen.

F2A+A2P approach: Recent results [245] have shown that F2A+A2P gives better performance for modeling personality profiles of individuals inside video segments. We use our proposed FS-OR method in the context of F2A+A2P. FS-OR can also be used in the context of F2P and we evaluate its performance in the experiments below. In a previous version of our

work [248], we used Sparse and Low-rank Transformation (SLoT) proposed by [245] in the context of F2A+A2P.

We trained F2P (Lasso and sparse SVR) and F2A+A2P (FS-OR) models using the low-level features as the input and self-assessed personality scores as the output. Table 6.4 shows the performance (in terms of Spearman Correlation Coefficient) of our approach (FS-OR in the framework of F2A+A2P) when compared to the current state-of-the-art [4].

Table 6.5 shows that the F2A+A2P outperforms F2P approach with significant performance gains. Even the RMSE of traits using sparse SVR is lesser than that using LASSO. It also shows that F2A+A2P can bring huge gains in modeling personality scores, reducing the error by 40-50%. The same experiment was repeated using combined features. In Table 6.5, it is observed that in case of F2P approach, the RMSE for E using lasso with low-level features was 2.077 and the same using new features is 1.859. The improvement is about 11%. Similarly for other traits, adding new features led to a 3-12 % increase in performance. There are cases where the performance has degraded as well (trait N, for instance), but the degradation was very small. We performed PCA on the set of features but the performance was less than the proposed approach - the correlation and RMSE between predicted and actual traits was lower by about 0.10-0.15 and 0.04-0.1 respectively for personality perception (predicting others-assessed scores) and by 0.30-0.50 and 0.1-0.2 respectively for personality recognition (predicting self-assessed scores). Here, it is seen that SLoT outperforms FS-OR in terms of RMSE while using combined features. However, in terms of correlation and accuracy (as seen in Table 6.7, FS-OR is seen to outperform SLoT.

Since the personality profiles are histograms over quantized bins it needs to be understood that the actual value of prediction error (deviation) can have a significant bearing on the above claim. More specifically, if the error is less than the quantization interval we can expect no error. The scores were divided into high, neutral and medium levels. Table 6.7 shows that F2A+A2P improves the performance by about 7-15% of predicting all the traits except for Openness. FS-OR is seen to outperform SLoT [245] in predicting all traits except Openness, where it is comparable. It should be noted that the images were ‘liked’ by Flickr ‘pro’ users, who pay a fee for using privileged features on Flickr. This means that the users’ interests would be highly inclined towards art, photography and the like, as mentioned in Section 6.4, thereby skewing the traits’ distribution causing low performance on Openness.

F2P approaches used in previous personality modeling literature show good performance

Table 6.4: Automatic personality recognition performance comparison between our approach and [4] in terms of Spearman Correlation Coefficient between actual and predicted trait. All correlations for our approach are found to be significant ($p < 0.05$).

Trait	Type	[4]	Our Approach
O	Self	0.25	0.39
	Attributed	0.35	0.49
C	Self	0.24	0.37
	Attributed	0.57	0.70
E	Self	0.28	0.43
	Attributed	0.62	0.77
A	Self	0.20	0.36
	Attributed	0.52	0.73
N	Self	0.14	0.29
	Attributed	0.60	0.75

Table 6.5: Results with low-level features of predicting self-assessed personality-profiles for both F2A+A2P (FS-OR) and F2P (LASSO and sparse SVR) approaches, represented in terms of (RMSE, correlation). Significant correlations ($p < 0.05$) are marked with *

With low-level features					
Trait	F2P			F2A+A2P	
	LASSO	sparse SVR	FS-OR	SLoT	FS-OR
O	1.69 , 0.20*	1.77 , 0.22*	1.76 , 0.20*	0.98 , 0.26*	0.97 , 0.29*
C	1.78 , 0.18	1.68 , 0.16	1.67 , 0.19*	0.90 , 0.24*	0.89 , 0.28*
E	2.07 , 0.19*	1.83 , 0.21*	1.82 , 0.24*	0.97 , 0.27*	0.97 , 0.33*
A	1.66 , 0.17	1.39 , 0.18	1.27 , 0.18*	0.83 , 0.30*	0.83 , 0.30*
N	2.20 , 0.21	2.35 , 0.13	2.38 , 0.21*	1.20 , 0.21*	1.19 , 0.24*

for Extraversion and Conscientiousness traits as they can be easily perceived [263]. But when the approach of F2A+A2P, it results in significant gains in performance for all traits, as seen in the experiments. This can be attributed to the fact that the relationship between features and answers are less abstract. The performance gains can also be attributed to the usage of BFI scheme, a standard method used in psychology for personality assessment.

6.5.2 Features important for Self-assessment

Table 6.8 reports the contribution of combined features that were learnt corresponding to

Table 6.6: Results with combined features of predicting self-assessed personality-profiles for both F2A+A2P (FS-OR) and F2P (LASSO and sparse SVR) approaches, represented in terms of (RMSE, correlation). Significant correlations ($p < 0.05$) are marked with *

With combined features					
	F2P			F2A+A2P	
Trait	LASSO	sparse SVR	FS-OR	SLoT	FS-OR
O	1.79 , 0.24*	1.71 , 0.23*	1.70 , 0.26*	0.91 , 0.32*	0.93 , 0.39*
C	1.70 , 0.23	1.65 , 0.20*	1.63 , 0.23*	0.83 , 0.33*	0.89 , 0.37*
E	1.85 , 0.26*	1.82 , 0.26*	1.81 , 0.27*	0.90 , 0.32*	0.92 , 0.43*
A	1.44 , 0.22*	1.37 , 0.24*	1.28 , 0.26*	0.72 , 0.34*	0.78 , 0.36*
N	2.29 , 0.25	2.31 , 0.20	2.32 , 0.22*	1.11 , 0.25*	1.17 , 0.29*

Table 6.7: Measuring actual deviations in prediction of personality scores (which are converted into High, Medium, Low categories based on thresholding)

Classification Accuracy (in %)				
	F2P		F2A+A2P	
Trait	sparse SVM	FS-OR	SLoT	FS-OR
O	62.4	62.7	66.10	63.1
C	58.1	70.9	70.50	79.5
E	53.7	72.4	69.70	84.7
A	64.3	74.5	72.30	87.5
N	50.7	61.2	61.50	65.5

each personality trait (based on the questions used to calculate the score on the corresponding trait) with the combined feature set and F2A+A2P modeling approach. These represent the weights allocated to the different features by our FS-OR model. For features which are multi-dimensional (vectors), this value represents the average of the weight on each dimension of the feature vector. Features omitted for different traits obtained a value of 0 in the trained models.

Users with a high score on the openness trait prefer colorful images [264] but do not like to have clutter or profile faces. Those on conscientiousness prefer black and white images with people - who are not facing towards them. Those with a high score on extraversion prefer images with a lot of people in them [265, 266], who are of the opposite gender (the negative weight on males is probably due to the higher number of males in the dataset). Users with high score on Agreeableness prefer spread out images with faces which are not profile. Neuroticism

Table 6.8: Weights allocated to features by the trained model, indicating their influence on personality traits along with Pearson correlation ($p < 0.05$) in *italics* of the feature values with the self-assessed traits.

Trait	Positive	Negative
O	Scene(1.6 , 0.08) , Color(1.34 , 0.07), #Upperbody(1.25 , 0.05)	Clutter(-1.49 , 0.03), #Profilefaces(-0.24 , 0.01)
C	#Heads(1.41 0.04)	Color(-1.02 0.05), #Faces(-0.25 , 0.01)
E	#Upperbody(2.79 , 0.09), Female(1.45 , 0.07), Saliency(1.25 , 0.06)	Male(-0.38 , 0.06)
A	Saliency(2.21 , 0.10), #Faces(0.8 , 0.04), Scene(0.63 , 0.04)	#Profilefaces(-0.43 , 0.03), Female(-0.13 , 0.03)
N	Clutter(2.06 , 0.11), Graphics(0.35 , 0.04)	Color (-0.28 , 0.03), Saliency(-0.329 , 0.04)

high score users prefer images with high clutter [267], some amount of graphics but not in color. The cited works also report similar findings by psychology researchers in other domains, verifying our results in the multimedia domain.

These insights can help recommender systems to make quick and relevant decisions, in a cold-start scenario, as to what kind of images should be shown to a new user, whose personality is known. For example neurotic people were seen to ‘like’ images with clutter and graphics. To a new user, the system can show images whose scores on these two features are high. Then as the user keeps using the system, it can fine-tune the other aspects such as the user liking cluttered images with people or without people, etc. which are more subjective in nature.

6.5.3 Recommender Application

The recommendation problem has two aspects:

1. Given a user, predict which image will the user like.
2. Given an image, predict which user (or group of users) will like it.

For the first aspect, our goal is, given a user profile, evaluate recall at top-K for all test images. Note that in the given dataset, we only have information about images that every user liked. We do not know if a user has seen the images liked by remaining users. Therefore, although we cannot reject the possibility that the images liked by other users could be his to

Figure 6.4: Top 10 Images retrieved by F2A+A2P for a sample of users based on their Personality Profile. The images with green border are ‘liked’ by the user. Retrieved images not belonging to the user’s liked set can be seen to have lot of similarity w.r.t cues associated with the corresponding personality traits. Explained in Section 6.5.

his liking as well, we prefer to be more strict in our evaluation and the performance evaluation forms a lower bound on accuracy.

For these experiments, we used the models for both F2P (sparse SVM) and F2A+A2P as described in previous sections. We took a total of 2000 images (for which user assignments are known and are different from the training set) as the test set and ranked the images based on RMSE between the predicted profile and users profile. The models were trained using 5-fold cross-validation setting on the training set of images and then used on the test set, aforementioned. Performance evaluation is based on recall for top-K (where $K \in \{1, 2, \dots\}$).

Figure 6.5 shows that F2A+A2P is able to retrieve a higher number of images (in top-K) when compared with the F2P model. It is definitely a possibility that some images that were highly ranked by the model but were not counted towards recall might have been liked by the user. To verify this we visually inspected the top 10 images retrieved by the model for a few users, out of which 3 are shown along with the corresponding user profile in Figure 6.4. The images bounded by green are ‘liked’ by the users. The first user has a high score on Openness and Neuroticism traits and the images are very colorful (which was found to be a common characteristic of images depicting custom work/home environments (image 2 from the left)

Figure 6.5: Mean Recall (for all users) using top k images retrieved by F2P and F2A+A2P models. Some images that were highly ranked by the model but were not counted towards recall might have been liked by the user, had the user viewed them.

Table 6.9: Given a set of images, performance of using combined features at predicting the right user

		low-level features		
Train/Test split:		75%/25%	50%/50%	25%/75%
Correct user hit	top 5	75.60%	70.9	69.3
	top 3	61.10%	57.2	54
		combined features		
Train/Test split:		75%/25%	50%/50%	25%/75%
Correct user hit	top 5	78.30%	71.9	73
	top 3	65.20%	58.2	58.8

and artistic photography). Some of the images retrieved have high clutter (images 1 and 10 from the left), which is a characteristic of images ‘liked’ by neurotic users. The second user has a very low score on extraversion and corresponding images retrieved have scenic backdrops without any explicit objects.

We also examined the difference in performance, given a set of images, of predicting the right user before and after adding our proposed features. The results are shown in Table 6.9 - with combined features, performance of 25/75% data split is better than 50/50% split of low-level features.

This confirms that a) representing images through semantic features improves performance in mapping images to users and b) the F2A+A2P approach is highly effective in modeling

Table 6.10: Confusion matrix for mapping images to users' ethnic profiles.

	Europeans	Americans	Asians
Europeans	93.13	6.87	0.00
Americans	16.67	83.33	0.5
Asians	0.00	18.00	82.00

Table 6.11: Confusion matrix for mapping images to gender

	Male	Female
Male	95.2	4.80
Female	10.00	90.00

personality profiles from images with high generalization.

For the second aspect of image recommendation, our goal is, given a set of users and an image, to study what kind of user (or group of users) would like an image. This involves, first predicting the profile from the image and then comparing with all user profiles to compute the RMSE. The image gets assigned to the user with the lowest RMSE. Then we can check if the image is correctly assigned. This is done for every user and we find out the accuracy. Note that images can portray multiple traits and hence it is more useful to target user groups rather than individual users.

To test for user group assignments we use the Flickr API to extract the profile information (gender, ethnicity) of the users using their User IDs/User names and grouped them accordingly. When the data was collected by [4], 75% of the users were males and 25% were female. 75% of the users were European, 15% were Americans and 10% were Asians. The performance (confusion matrix) of the model in predicting the gender of the user who would like the images is shown in Table 6.11 and the same for ethnicity is shown in 6.10. We see that there is a considerable misclassification between American and European ethnicities; however, none of the them were misclassified as Asian, possibly indicating the similarities between American and European cultures and the distance from Asian cultural preferences [268, 269, 270].

As an alternate way of grouping, we clustered the users into 5 groups based on the users' personality profiles (by taking the correlation between the scores of 5-personality scores as the distance metric). Clusters 1 through 5 consisted of 5%, 30%, 15%, 30% and 20% users respectively. We performed the same experiments (as described above with gender and ethnicity) to

Table 6.12: Confusion matrix for mapping images to Personality Profile Clusters

	C1	C2	C3	C4	C5
C1	7.00	0.00	0.00	93.00	0.00
C2	0.00	50.33	0.5	26.67	22.5
C3	0.00	0.00	76.67	22.00	1.33
C4	0.00	0.16	21.33	78.16	0.33
C5	0.00	0.00	0.00	49.5	50.5

check if the images go to the right cluster of users and checked the assignment accuracy. The results are reported in 6.12. Certainly the personality clusters could be formed in a better manner than we did, but this is a first attempt to understand if personality modeling based image recommendation is a promising approach.

6.6 Conclusion

In this paper, the problem of computationally predicting users' self-assessed personality profiles along the Big-5 traits, based on images they 'like' was studied. Three research questions which were raised in the beginning of the chapter are answered as follows:

RQ1 Results of personality recognition improved considerably from the current state-of-the-art by using combined features. This can be attributed to the observation that high Availability criteria [237], which indicates that traits perceivable by others help improve the personality recognition performance. This was possibly a consequence of using features focused on the content and people in the image. This gives an indication that predicting users' personality, which is a challenging task, we need more human-interpretable and semantic representations.

RQ2 It was seen that the approach of predicting answers to the questions to personality inventory (F2A+A2P), BFI-10 in this case, and then converting them to personality scores outperforms the approach of predicting personality scores directly from the image features (F2P). This can be used to facilitate non-intrusive personality acquisition techniques, which have been shown to be highly preferable [244].

RQ3 Experiments showed that features related to scene, color, #upperbody have a positive influence and clutter, #profilefaces have a negative influence in predicting the openness trait. #Heads contributes positively and color, #faces contribute negatively to conscientiousness. #Upperbody, female, saliency contribute positively and male contributes negatively to extraversion. Saliency, #faces, scene have a positive influence and #profilefaces, female have negative influence on agreeableness. Clutter, graphics contribute positively and color, salience negatively to neuroticism. These insights could be used in the application of image recommendation to use personality as a good way of providing quick ‘likable’ recommendations to users. This could address the cold-start situation where information about users is missing. However, in this scenario, further work is warranted with datasets that have both ‘like’ and ‘dislike’ annotations so as to have a more concrete evaluation.

Further, in this work all the features focused on the visual content and they were seen to model users’ preferences quite well. However, the tags associated with images might complement visual features by conveying information which visual features might fail to convey [207]. The place in the image or the relationship with the person in the image, or the scene in the image might trigger user’s memory, and this context is often present in tags [208]. It might be worthwhile to explore the contribution of tags to personality prediction.

Attributed traits are often predictive of important aspects of social life, and can be useful in multiple domains [172, 188]. Therefore, another aspect of future work would involve collecting more ‘reliable’ survey data for personality profiles using larger and possibly multiple inventories, especially with regards to others-assessed personality profiles due to low reliability scores (Table 6.1). This makes it difficult to compare the prediction performance between the self-assessed and others-assessed profiles. Additionally, research has shown that emotion plays a crucial role in users’ ‘likes’ [271, 272] - exploring this as a latent feature for understanding the relationship between personality and ‘likes’ might give deeper insights into ‘who likes what and, why’.

Chapter 7

Conclusion and Future Work

In this thesis, we examined Anais Nin’s quote – *We don’t see things as they are, we see them as we are* – holds in the domain of multimedia. While several studies examined the role of content on multimedia perception, the user who is the perceiver was largely left out with studies taking the route of modeling ‘average user’s’ perception. To fill this gap, we conducted a series of studies to answer two fundamental questions: 1) Do individual factors play a significant role in influencing users’ perception of multimedia affect and enjoyment? and 2) How to automatically build predictive models for users’ personality based on social media data?

7.1 Main Contributions

This thesis makes the following main contributions in modeling users’ perception:

1. We analysed the influence that two primordial dimensions of user factors - personality and culture - in conjunction with system factors - frame rate, resolution, and bit rate - have on user affect and enjoyment of multimedia presentations. Specifically, if and how well a model based on these factors can predict perceived affect and enjoyment; which system and individual characteristics influence the outcomes of perceived affect and enjoyment the most; the relationship between both outcomes across stimuli [273]; and, how predictive models perform at automatic assessment. Towards this goal, we built a suitable dataset of 144 video clips rated by 114 participants, since most of the available datasets collect ground truth on users’ perception (be it of affect or enjoyment) by averaging or taking the majority vote from users. This dataset is released [128] and

is already being taken up by researchers in the community [274] for extending cross-cultural studies across multiple cohorts.

2. While it was seen that personality factors have significant correlations with users behavior in a variety of fields, acquisition of users traits was possible only through questionnaires. To unobtrusively assess users' personality, from data sources which are available widely and freely – such as social media, we leveraged both the dimensions of posted and liked content on social media. This is based on the finding that personality is strongly expressed on a platform which offers users sufficient self-expression and freedom of control. In modeling users personality based on posted content, we used selfies from Sina Weibo and image posts from Twitter.
3. In the first study on selfies, we attempt to computationally model users' personality based on their selfies. In particular, we examine if computer models can predict users' self-assessed personality (personality recognition); we also examine if computer models can predict how others assess users' personality (personality perception); and what cues aid in both personality recognition and perception. Correlation analysis b/w mid-level cue detectors and personality traits showed some behavioral insights that were corroborated by previous findings in Psychology literature, while others that were novel [190].
4. In a second study on online image posting in general, we extracted interpretable semantic image features (color aesthetics and objects) using large-scale content analysis on an order of magnitude larger dataset than previous studies from Twitter to give insights into the personality differences in people posting them, and build predictive models personality traits from image posts that showed reliable accuracy on held-out data using image features.
5. In modeling users personality based on liked content, we used images tagged as 'Favorite' on Flickr. In addressing this problem, two questions were addressed - a.) how best to represent users' likes on images? and b.) how to predict users' personality based on image they like?

- (a) To address the first part, backed by intuition and previous studies, we represented images using semantic features which are interpretable - for instance, is the image colored or black & white, is the image rendered by graphics or a natural image, are there people in the image and what their characteristics are and so on. We further proposed a deep bi-modal representation exploiting the visual content along with the tags associated with images to better the state-of-the-art on user 'likes' [275], and present an approach to understand what the hidden layers of a neural network are learning when presented with semantic inputs [276].
- (b) To address the second part, we applied the semantic interpretable features for predicting users' personality [277] and a novel Feature Selection based Ordinal Regression for personality prediction [278]. Extensive experiments showed that our proposed method significantly improves personality prediction performance.

7.2 Future Work

Several prospective extensions to the technologies explored in this thesis along with interdisciplinary applications can be explored, a few of which are detailed below.

1. Extensions

- Though personality and culture significantly improved the performance of modeling users' perception, several other human factors, apart from these can be taken into consideration in this study such as user behavior in a particular domain of study (e.g., movies rated or images liked) or other psychological constructs like mood [164, 165, 166] can be explored.
- In modeling users' personality, the features focused on the visual content and they were seen to model users' preferences quite well. However, the tags associated with images might complement visual features by conveying information which visual features might fail to convey [207]. The place in the image or the relationship with the person in the image, or the scene in the image might trigger user's memory, and this context is often present in tags [208]. It would be interesting to explore the contribution of tags to personality prediction and would require creation of a suitable dataset for the same.

- Owing to the time period and resources under which the different studies were undertaken, the cohorts of users were different - selfies, image posts, images liked all belonged to different users. It would be interesting to explore the variance in personality prediction performance of different sources used and explore on which platform do users reveal themselves the most which has several potential big data social science implications [279, 280]. Moreover, every study focused on users from a similar culture - it would be interesting to study how cross-cultural variations interact with personality.

2. Interdisciplinary Applications

- *Targeted Marketing*: With multimedia data becoming a form of communication b/w users on social media, modeling users' perceptual judgements can help improve ad diffusion, virality and impact specificity on social media [281].
- *Social Science*: Most of the studies in social science utilise the language of users to understand their behavior online [280]. However, with the methods presented in this thesis, multimodal approaches involving both images and text can be exploited. Language especially might constrain the analysis to certain regions of the world where a specific language is spoken, however with images, which are universal in nature, users' expressions can be understood on a much diverse scale.
- *Art and Aesthetics*: While a lot of studies on computational aesthetics and art are being taken up, relatively few focus on modeling the perception of art and aesthetics in a computational manner. The personality modeling techniques presented in this thesis could be used to derive relevant outcomes and analyse their influence on perception.
- *Neurological Science*: With an increasing effort to understand the neurological basis for various activations in the human mind, and active strive to model them using deep neural networks [282], this thesis can help understand the role of neurological processes in users' perception of multimedia and also of cognitive processes such as personality perception based on users' multimedia traces.

- *Embodied Agents*: Several technologies like social robots, avatars, conversational agents which interact with users can be improved to react with more ease by taking the personality of users' into account and the ability to interact in a manner which is amiable to users' behavior in a certain context [283].

Publications

List of published works during Ph.D. candidature.

Journal Articles

1. **S.C. Guntuku**, J. T. Zhou, S. Roy, W. Lin, I. W. Tsang, “Understanding Deep Representations Learned in Modeling User ‘Likes’ ”, IEEE Transactions on Image Processing, 2016
2. **S. C. Guntuku**, J. T. Zhou, S. Roy, W. Lin, I. W. Tsang, “Who Likes What and Why? Insights into Image based Personality Modeling”, IEEE Transactions on Affective Computing, 2016
3. **S. C. Guntuku**, M.J.Scott, G. Ghinea, W. Lin “Personality, Culture, and System Factors - Impact on Affective Response to Multimedia”, Under Review, 2016

Conferences

1. **S. C. Guntuku**, J. Carpentar, W. Lin, W.K. Ng, L. H. Ungar, D. Preotiuc-Pietro, Studying Personality through Posted Image Content, Under Review, 2016
2. **S. C. Guntuku**, S. Roy, W. Lin, K. Ng, W.K. Ng, V. Jakhetiya, “Personalizing User Interfaces for improving Quality of Experience in VoD Recommender Systems”, IEEE QoMEX, 2016
3. **S. C. Guntuku**, L.Qiu, S. Roy, W. Lin, V. Jakhetiya, “Do others perceive you as you want them to? Modeling Personality based on Selfies”, ACM Multimedia 2015 Workshop on Affect and Sentiment in Multimedia

4. M.J.Scott[†], **S. C. Guntuku**[†], Y. Huan, W. Lin, G. Ghinea, “Modelling Human Factors in Perceptual Multimedia Quality: On The Role of Personality and Culture”, ACM Multimedia 2015 ([†] – denotes equal contribution)
5. **S. C. Guntuku**, M.J.Scott, G. Ghinea, W. Lin “Modelling the Influence of Personality and Culture on Affect and Enjoyment in Multimedia”, ACII 2015
6. **S. C. Guntuku**, S. Roy, W. Lin, “Evaluating Visual and Textual Features for Modeling User ‘Likes’”, IEEE ICME, 2015.
7. **S. C. Guntuku**, M.J.Scott, Y. Huan, G. Ghinea, W. Lin, “The CP-QAE-I: A Video Dataset for Exploring the Effect of Personality and Culture on Perceived Quality and Affect in Multimedia”, IEEE QoMEX, 2015.
8. **S. C. Guntuku**, S. Roy, W. Lin, “Personality Modeling Based Image Recommendation”, MultiMedia Modeling. Springer International Publishing, 2015.
9. **S. C. Guntuku**, J. T. Zhou, S. Roy, W. Lin, I. W. Tsang, “Deep Representations to Model User ‘Likes’”, 12th Asian Conference on Computer Vision, 2014

Bibliography

- [1] “Pinterest,” <https://www.pinterest.com/LokidByLoki/angry/>.
- [2] P. Lovato, A. Perina, N. Sebe, O. Zandonà, A. Montagnini, M. Bicego, and M. Cristani, “Tell me what you like and I’ll tell you what you are: discriminating visual preferences on flickr data,” in *Computer Vision—ACCV 2012*, pp. 45–56, Springer, 2013.
- [3] L. R. Goldberg, “An alternative” description of personality”: the big-five factor structure,” *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.
- [4] M. Cristani, A. Vinciarelli, C. Segalin, and A. Perina, “Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis,” in *21st ACM international conference on Multimedia*, ACM, 2013.
- [5] G. Lekakos and G. M. Giaglis, “Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors,” *Interacting with computers*, vol. 18, no. 3, pp. 410–431, 2006.
- [6] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, “Factors influencing quality of experience,” in *Quality of Experience*, pp. 55–72, Springer, 2014.
- [7] S. D. Gosling, S. J. Ko, T. Mannarelli, and M. E. Morris, “A room with a cue: personality judgments based on offices and bedrooms,” *Journal of personality and social psychology*, vol. 82, no. 3, p. 379, 2002.
- [8] S. Duncan and L. F. Barrett, “Affect is a form of cognition: A neurobiological analysis,” *Cognition and emotion*, vol. 21, no. 6, pp. 1184–1211, 2007.

BIBLIOGRAPHY

- [9] G. R. VandenBos, *APA dictionary of psychology*. American Psychological Association, 2007.
- [10] L. R. Goldberg, “The structure of phenotypic personality traits.,” *American psychologist*, vol. 48, no. 1, p. 26, 1993.
- [11] G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality traits*. Cambridge University Press, 2003.
- [12] C. Mora, “Cultures and organizations: Software of the mind intercultural cooperation and its importance for survival,” *Journal of Media Research*, vol. 6, no. 1, p. 65, 2013.
- [13] G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival*. McGraw-Hill, 2010.
- [14] W. Wundt, *Outlines of psychology*. Springer, 1980.
- [15] P. Ekman and W. V. Friesen, “Constants across Cultures in the Face and Emotion,” *Journal of Personality and Social Psychology*, vol. 17, no. 2, 1971.
- [16] L. F. Barrett, M. Gendron, and Y.-M. Huang, “Do discrete emotions exist?,” *Philosophical Psychology*, vol. 22, no. 4, pp. 427–437, 2009.
- [17] C. Harmon-Jones, B. Bastian, and E. Harmon-Jones, “The discrete emotions questionnaire: A new tool for measuring state self-reported emotions,” *PloS one*, vol. 11, no. 8, 2016.
- [18] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
- [19] T. F. Cootes, G. J. Edwards, C. J. Taylor, *et al.*, “Active appearance models,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
- [20] M. Kenji, “Recognition of facial expression from optical flow,” *IEICE TRANSACTIONS on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.

- [21] P. Chiranjeevi, V. Gopalakrishnan, and P. Moogi, “Neutral face classification using personalized appearance models for fast and robust emotion detection,” *Image Processing, IEEE Transactions on*, vol. 24, no. 9, pp. 2701–2711, 2015.
- [22] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews, “Predicting movie ratings from audience behaviors,” in *WACV*, IEEE, 2014.
- [23] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, “Robust representation and recognition of facial emotions using extreme sparse learning,” *Image Processing, IEEE Transactions on*, vol. 24, no. 7, pp. 2140–2152, 2015.
- [24] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, “Micro-expression recognition using color spaces,” *Image Processing, IEEE Transactions on*, vol. 24, no. 12, pp. 6034–6047, 2015.
- [25] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, “Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields,” *Image Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 1386–1398, 2015.
- [26] A. Kleinsmith and N. Bianchi-Berthouze, “Affective body expression perception and recognition: A survey,” *Affective Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 15–33, 2013.
- [27] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE TPAMI*, 2009.
- [28] J. Han, X. Ji, X. Hu, D. Zhu, K. Li, X. Jiang, G. Cui, L. Guo, and T. Liu, “Representing and retrieving video shots in human-centric brain imaging space,” *Image Processing, IEEE Transactions on*, vol. 22, no. 7, pp. 2723–2736, 2013.
- [29] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis; using physiological signals,” *Affective Computing, IEEE Transactions on*, 2012.
- [30] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

- [31] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, *et al.*, “Query by image and video content: The qbic system,” *Computer*, vol. 28, no. 9, pp. 23–32, 1995.
- [32] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 837–842, 1996.
- [33] J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei, “Content-based image indexing and searching using daubechies’ wavelets,” *International Journal on Digital Libraries*, vol. 1, no. 4, pp. 311–328, 1998.
- [34] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li, “The role of attractiveness in web image search,” in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 63–72, ACM, 2011.
- [35] S. Dhar, V. Ordonez, and T. L. Berg, “High level describable attributes for predicting aesthetics and interestingness,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1657–1664, IEEE, 2011.
- [36] M. Solli and R. Lenz, “Color based bags-of-emotions,” in *Computer Analysis of Images and Patterns*, pp. 573–580, Springer, 2009.
- [37] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223–232, ACM, 2013.
- [38] P. Isola, J. Xiao, A. Torralba, and A. Oliva, “What makes an image memorable?,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 145–152, IEEE, 2011.
- [39] J. Kim, S. Yoon, and V. Pavlovic, “Relative spatial features for image memorability,” in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 761–764, ACM, 2013.

- [40] A. Khosla, A. Das Sarma, and R. Hamid, “What makes an image popular?,” in *Proceedings of the 23rd international conference on World wide web*, pp. 867–876, International World Wide Web Conferences Steering Committee, 2014.
- [41] J. Machajdik and A. Hanbury, “Affective image classification using features inspired by psychology and art theory,” in *Proceedings of the International Conference on Multimedia*, MM ’10, (New York, NY, USA), pp. 83–92, ACM, 2010.
- [42] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Studying aesthetics in photographic images using a computational approach,” in *Computer Vision—ECCV 2006*, pp. 288–301, Springer, 2006.
- [43] P. Valdez and A. Mehrabian, “Effects of color on emotions.,” *Journal of Experimental Psychology: General*, vol. 123, no. 4, p. 394, 1994.
- [44] C. E. Osgood, *The measurement of meaning*. No. 47, University of Illinois press, 1957.
- [45] Ö. Axelsson, “Towards a psychology of photography: Dimensions underlying aesthetic appeal of photographs 1, 2, 3,” *Perceptual and motor skills*, vol. 105, no. 2, pp. 411–434, 2007.
- [46] J. O’Doherty, J. Winston, H. Critchley, D. Perrett, D. M. Burt, and R. J. Dolan, “Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness,” *Neuropsychologia*, vol. 41, no. 2, pp. 147–155, 2003.
- [47] G. Peters, “Aesthetic primitives of images for visualization,” in *Information Visualization, 2007. IV’07. 11th International Conference*, pp. 316–325, IEEE, 2007.
- [48] Y. Ke, X. Tang, and F. Jing, “The Design of High-level Features for Photo Quality Assessment,” *CVPR*, 2006.
- [49] J. Li and J. Z. Wang, “Studying digital imagery of ancient paintings by mixtures of stochastic models,” *Image Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 340–353, 2004.
- [50] M. Del Fabro and L. Böszörményi, “State-of-the-art and future challenges in video scene detection: a survey,” *Multimedia systems*, vol. 19, no. 5, pp. 427–454, 2013.

- [51] D. M. Thounaojam, A. Trivedi, K. M. Singh, and S. Roy, “A survey on video segmentation,” in *Intelligent Computing, Networking, and Informatics*, pp. 903–912, Springer, 2014.
- [52] E. Acar, “Learning representations for affective video understanding,” in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 1055–1058, ACM, 2013.
- [53] M. K. Abadi, S. M. Kia, R. Subramanian, P. Avesani, and N. Sebe, “User-centric affective video tagging from meg and peripheral physiological responses,” in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 582–587, IEEE, 2013.
- [54] L. Canini, S. Benini, and R. Leonardi, “Affective recommendation of movies based on selected connotative features,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 4, pp. 636–647, 2013.
- [55] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, “Towards a comprehensive computational model for aesthetic assessment of videos,” in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 361–364, ACM, 2013.
- [56] C.-Y. Wei, N. Dimitrova, and S.-F. Chang, “Color-mood analysis of films based on syntactic and psychological models,” in *ICME*, IEEE, 2004.
- [57] B. Li, W. Xiong, O. Wu, W. Hu, S. Maybank, and S. Yan, “Horror image recognition based on context-aware multi-instance learning,” *Image Processing, IEEE Transactions on*, vol. 24, no. 12, pp. 5193–5205, 2015.
- [58] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, “Context-sensitive learning for enhanced audiovisual emotion classification,” *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.
- [59] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, “Avec 2011—the first international audio/visual emotion challenge,” in *Affective Computing and Intelligent Interaction*, Springer, 2011.

- [60] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, “A bayesian framework for video affective representation,” in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–7, IEEE, 2009.
- [61] K. Sun and J. Yu, “Video affective content representation and recognition using video affective tree and hidden markov models,” in *Affective Computing and Intelligent Interaction*, pp. 594–605, Springer, 2007.
- [62] R. M. A. Teixeira, T. Yamasaki, and K. Aizawa, “Determination of emotional content of video clips by low-level audiovisual features,” *Multimedia Tools and Applications*, vol. 61, no. 1, pp. 21–49, 2012.
- [63] D. Borth and et. al, “Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content,” in *ACM Multimedia*, 2013.
- [64] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in *Advances in Neural Information Processing Systems*, pp. 487–495, 2014.
- [65] K. Winter and N. Kuiper, “Individual differences in the experience of emotions,” *Clinical psychology review*, 1997.
- [66] W. Choe, H.-S. Chun, J. Noh, S.-D. Lee, and B.-T. Zhang, “Estimating multiple evoked emotions from videos,” in *Annual Meeting of the Cognitive Science Society*, 2013.
- [67] J. Wache, “The secret language of our body: Affect and personality recognition using physiological signals,” in *ICMI*, ACM, 2014.
- [68] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.
- [69] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. S. Jin, “Mutual information-based emotion recognition,” in *The Era of Interactive Media*, pp. 471–479, Springer, 2013.
- [70] Y.-G. Jiang, B. Xu, and X. Xue, “Predicting emotions in user-generated videos.,” in *AAAI*, pp. 73–79, 2014.

- [71] S. Benini, L. Canini, and R. Leonardi, “A connotative space for supporting movie affective recommendation,” *Multimedia, IEEE Transactions on*, vol. 13, no. 6, pp. 1356–1370, 2011.
- [72] Y. Cui, J. S. Jin, S. Zhang, S. Luo, and Q. Tian, “Music video affective understanding using feature importance analysis,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 213–219, ACM, 2010.
- [73] S. Zhang, Q. Tian, Q. Huang, W. Gao, and S. Li, “Utilizing affective analysis for efficient movie browsing,” in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 1853–1856, IEEE, 2009.
- [74] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, “Affective visualization and retrieval for music video,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 510–522, 2010.
- [75] S. Arifin and P. Y. Cheung, “Affective level video segmentation by utilizing the pleasure-arousal-dominance information,” *Multimedia, IEEE Transactions on*, vol. 10, no. 7, pp. 1325–1341, 2008.
- [76] L. Canini, S. Benini, P. Migliorati, and R. Leonardi, “Emotional identity of movies,” in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 1821–1824, IEEE, 2009.
- [77] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, “Affective audio-visual words and latent topic driving model for realizing movie affective scene classification,” *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 523–535, 2010.
- [78] Y. Baveye, J.-N. Bettinelli, E. Dellandréa, L. Chen, and C. Chamaret, “A large video data base for computational models of induced emotion,” in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 13–18, IEEE, 2013.
- [79] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, “Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,” *Cognition and Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

- [80] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, and S. Geva, “Social event detection at mediaeval 2013: Challenges, datasets, and evaluation,” in *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013*, 2013.
- [81] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, “Content-based prediction of movie style, aesthetics and affect: Data set and baseline experiments,” 2014.
- [82] J. Rusten, “Theophrastus: characters,” *Cambridge, MA*, 1993.
- [83] G. Saucier and L. R. Goldberg, “The language of personality: Lexical perspectives on the five-factor model,” *The five-factor model of personality: Theoretical perspectives*, pp. 21–50, 1996.
- [84] B. Rammstedt and O. P. John, “Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german,” *Journal of research in Personality*, vol. 41, 2007.
- [85] R. R. McCrae and P. T. Costa Jr, “A contemplated revision of the neo five-factor inventory,” *Personality and Individual Differences*, vol. 36, no. 3, pp. 587–596, 2004.
- [86] P. T. Costa Jr and R. R. McCrae, “Domains and facets: Hierarchical personality assessment using the revised neo personality inventory,” *Journal of personality assessment*, vol. 64, no. 1, pp. 21–50, 1995.
- [87] C. J. Kemper and B. Rammstedt, “Psychometric properties of the bfi-k: A cross-validation study anastassiya kovaleva department of didactics of biology, university of bielefeld, germany constanze beierlein gesis–leibniz institute for the social sciences, mannheim, germany,” *International Journal*, vol. 13, p. 1, 2013.
- [88] F. R. Lang, D. John, O. Lüdtke, J. Schupp, and G. G. Wagner, “Short assessment of the big five: robust across survey methods except telephone interviewing,” *Behavior Research Methods*, vol. 43, 2011.

- [89] A. G. Thalmayer, G. Saucier, and A. Eigenhuis, “Comparative validity of brief to medium-length big five and big six personality questionnaires.,” *Psychological Assessment*, vol. 23, no. 4, p. 995, 2011.
- [90] Y. Zhuyu, “The relationship between personality and facebook usage,” 2015.
- [91] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, “The voice of personality: Mapping nonverbal vocal behavior into trait attributions,” in *Proceedings of the 2nd international workshop on Social signal processing*, pp. 17–20, ACM, 2010.
- [92] T. Polzehl, K. Schoenenberg, S. Moller, F. Metze, G. Mohammadi, and A. Vinciarelli, “On speaker-independent personality perception and prediction from speech,” 2012.
- [93] E. Brunswik, *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.
- [94] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, “Lexical predictors of personality type,” 2005.
- [95] G. Mohammadi and A. Vinciarelli, “Humans as feature extractors: Combining prosody and personality perception for improved speaking style recognition,” in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pp. 363–366, IEEE, 2011.
- [96] J. Golbeck, C. Robles, and K. Turner, “Predicting personality with social media,” in *CHI’11 Extended Abstracts on Human Factors in Computing Systems*, pp. 253–262, ACM, 2011.
- [97] A. Vinciarelli and G. Mohammadi, “A survey of personality computing,” 2014.
- [98] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.
- [99] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text.,” *J. Artif. Intell. Res.(JAIR)*, vol. 30, pp. 457–500, 2007.

- [100] A. Minamikawa and H. Yokoyama, “Personality estimation based on weblog text classification,” in *Modern approaches in applied intelligence*, pp. 89–97, Springer, 2011.
- [101] A. J. Gill, S. Nowson, and J. Oberlander, “What are they blogging about? personality, topic and motivation in blogs.,” in *ICWSM*, 2009.
- [102] S. Nowson, “Identifying more bloggers: Towards large scale personality classification of personal weblogs,” in *In Proceedings of the International Conference on Weblogs and Social*, Citeseer, 2007.
- [103] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the international conference on Multimedia*, pp. 1459–1462, ACM, 2010.
- [104] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, “Recognition of personality traits from human spoken conversations.,” in *INTERSPEECH*, pp. 1549–1552, 2011.
- [105] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, “Multimodal recognition of personality traits in social interactions,” in *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 53–60, ACM, 2008.
- [106] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, “Multimodal recognition of personality traits in human-computer collaborative tasks,” in *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 39–46, ACM, 2012.
- [107] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, “Please, tell me about yourself: automatic personality assessment using short self-presentations,” in *Proceedings of the 13th international conference on multimodal interfaces*, pp. 255–262, ACM, 2011.
- [108] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, “A multimodal annotated corpus of consensus decision making meetings,” *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2007.
- [109] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, “Predicting personality from twitter,” in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 149–156, IEEE, 2011.

- [110] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, “Our twitter profiles, our selves: Predicting personality with twitter,” in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 180–185, IEEE, 2011.
- [111] S. Bai, T. Zhu, and L. Cheng, “Big-five personality prediction based on user behaviors at social network sites,” *arXiv preprint arXiv:1204.4809*, 2012.
- [112] F. Celli, “Unsupervised personality recognition for social network sites,” in *ICDS 2012, The Sixth International Conference on Digital Society*, pp. 59–62, 2012.
- [113] T. Nguyen, D. Q. Phung, B. Adams, S. Venkatesh, *et al.*, “Towards discovery of influence and personality traits through social link prediction.,” in *ICWSM*, 2011.
- [114] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli, “Face-Based Automatic Personality Perception,” *MM*, 2014.
- [115] F. Celli, E. Bruni, and B. Lepri, “Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures,” *MM*, 2014.
- [116] L. Liu, D. Preoȃiuc-Pietro, Z. Riahi Samani, M. E. Moghaddam, and L. Ungar, “Analyzing Personality through Social Media Profile Picture Choice,” *ICWSM*, 2016.
- [117] B. Ferwerda, M. Schedl, and M. Tkalčič, “Using Instagram Picture Features to Predict Users’ Personality,” *MMM*, 2016.
- [118] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl, “Fusing Social Media Cues: Personality Prediction from Twitter and Instagram,” *WWW Companion*, 2016.
- [119] S. Butt and J. G. Phillips, “Personality and self reported mobile phone use,” *Computers in Human Behavior*, vol. 24, no. 2, pp. 346–360, 2008.
- [120] G. Chittaranjan, J. Blom, and D. Gatica-Perez, “Who’s who with big-five: analyzing and classifying personality traits with smartphones,” in *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*, pp. 29–36, IEEE, 2011.

- [121] D. O. Olguin, P. A. Gloor, and A. S. Pentland, “Capturing individual and group behavior with wearable sensors,” in *Proceedings of the 2009 aaai spring symposium on human behavior modeling*, SSS, vol. 9, 2009.
- [122] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, “Friends don’t lie: inferring personality traits from social network structure,” in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 321–330, ACM, 2012.
- [123] S. T. Fiske, A. J. Cuddy, and P. Glick, “Universal dimensions of social cognition: Warmth and competence,” *Trends in cognitive sciences*, 2007.
- [124] J. Smith, “The most unforgettable ad campaigns of 2013,” December 2013.
- [125] R. Jain, “Quality of experience,” *IEEE MultiMedia*, vol. 11, no. 1, pp. 96–95, 2004.
- [126] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, “Quality of experience in distributed interactive multimedia environments: toward a theoretical framework,” in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 481–490, ACM, 2009.
- [127] M. J. Scott, S. C. Guntuku, Y. Huan, W. Lin, and G. Ghinea, “Modelling human factors in perceptual multimedia quality: On the role of personality and culture,” in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pp. 481–490, ACM, 2015.
- [128] S. C. Guntuku, M. J. Scott, H. Yang, G. Ghinea, and W. Lin, “The CP-QAE-I: A video dataset for exploring the effect of personality and culture on perceived quality and affect in multimedia,” in *QoMEX*, IEEE, 2015.
- [129] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, “G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences,” *Behavior research methods*, vol. 39, no. 2, pp. 175–191, 2007.
- [130] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, “A very brief measure of the big-five personality domains,” *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.

- [131] G. J. McHugo, C. A. Smith, and J. T. Lanzetta, "The structure of self-reports of emotional responses to film segments," *Motivation and Emotion*, vol. 6, no. 4, pp. 365–385, 1982.
- [132] G. Hofstede, G. J. Hofstede, M. Minkov, and H. Vinken, "Values survey module 2013," URL: <http://www.geerthofstede.nl/vsm2013>, 2013.
- [133] G. Hofstede and M. Minkov, "Vsm 2013," *Values survey module*, 2013.
- [134] "PASW Statistics, <http://www.spss.com.hk/statistics/>," 2017.
- [135] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.
- [136] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [137] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pp. 197–206, ACM, 2007.
- [138] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM Multimedia*, 2010.
- [139] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.
- [140] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [141] R. Plutchik, *Emotion: A psychoevolutionary synthesis*. Harper & Row New York, 1980.
- [142] F. Eyben, M. Woellmer, and B. Schuller, "the munich open speech and music interpretation by large space extraction toolkit," 2010.

- [143] M. Müller and S. Ewert, “Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features,” in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011. *hal-00727791*, version 2-22 Oct 2012, Citeseer, 2011.
- [144] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, “Robust prosodic features for speaker identification,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1800–1803, IEEE, 1996.
- [145] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a completely blind image quality analyzer,” *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 209–212, 2013.
- [146] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *Image Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [147] G. Ghinea and J. P. Thomas, “Qos impact on user perception and understanding of multimedia video clips,” in *Proceedings of the sixth ACM international conference on Multimedia*, pp. 49–54, ACM, 1998.
- [148] N. Yeung and A. G. Sanfey, “Independent coding of reward magnitude and valence in the human brain,” *The Journal of Neuroscience*, 2004.
- [149] C. E. Izard, D. Z. Libero, P. Putnam, and O. M. Haynes, “Stability of emotion experiences and their relations to traits of personality,” *Journal of personality and social psychology*, 1993.
- [150] L. Pessoa, “On the relationship between emotion and cognition,” *Nature Reviews Neuroscience*, 2008.
- [151] R. Bosker and T. Snijders, “Multilevel analysis: An introduction to basic and advanced multilevel modeling, 2nd ed,” *New York*, 2012.
- [152] M. T. Soto Sanfiel, L. Aymerich Franch, F. X. Ribes Guardia, and J. R. Martinez Fernandez, “Influence of interactivity on emotions and enjoyment during consumption of audiovisual fictions,” *International Journal of Arts and Technology*, 2011.

- [153] E. Diener, S. Oishi, and R. E. Lucas, “Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life,” *Annual review of psychology*, 2003.
- [154] R. J. Larsen and T. Ketelaar, “Personality and susceptibility to positive and negative emotional states.,” *Journal of personality and social psychology*, 1991.
- [155] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, “Emotion representation, analysis and synthesis in continuous space: A survey,” in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 827–834, IEEE, 2011.
- [156] B. Jou, S. Bhattacharya, and S.-F. Chang, “Predicting viewer perceived emotions in animated gifs,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 213–216, ACM, 2014.
- [157] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez, “Inferring mood in ubiquitous conversational video,” in *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, p. 22, ACM, 2013.
- [158] B. Siddiquie, D. Chisholm, and A. Divakaran, “Exploiting multimodal affect and semantics to identify politically persuasive web videos,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 203–210, ACM, 2015.
- [159] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3–10, ACM, 2013.
- [160] S. Bardzell, J. Bardzell, and T. Pace, “Understanding affective interaction: Emotion, engagement, and internet videos,” in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–8, IEEE, 2009.
- [161] H. Bilandzic and R. W. Busselle, “Enjoyment of films as a function of narrative experience, perceived realism and transportability,” *Communications*, vol. 36, no. 1, pp. 29–50, 2011.

- [162] V. T. Visch, E. S. Tan, and D. Molenaar, “The emotional and cognitive effect of immersion in film viewing,” *Cognition and Emotion*, vol. 24, no. 8, pp. 1439–1445, 2010.
- [163] G. Ghinea and S. Y. Chen, “Perceived quality of multimedia educational content: A cognitive style approach,” *Multimedia systems*, vol. 11, no. 3, pp. 271–279, 2006.
- [164] S. Boll, “Multitube—where web 2.0 and multimedia could meet,” *IEEE MultiMedia*, no. 1, pp. 9–13, 2007.
- [165] P. A. Madden, A. C. Heath, N. E. Rosenthal, and N. G. Martin, “Seasonal changes in mood and behavior: the role of genetic factors,” *Archives of General Psychiatry*, vol. 53, no. 1, pp. 47–55, 1996.
- [166] F. T. O’Donovan, C. Fournelle, S. Gaffigan, O. Brdiczka, J. Shen, J. Liu, and K. E. Moore, “Characterizing user behavior and information propagation on a social multimedia network,” in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, pp. 1–6, IEEE, 2013.
- [167] S. Vranica, “Behind the preplanned oscar selfie: Samsung’s ad strategy,” *Wall Street Journal*, vol. 6, p. 41, 2014.
- [168] P. Borkenau, S. Brecke, C. Möttig, and M. Paelecke, “Extraversion is accurately perceived after a 50-ms exposure to a face,” *Journal of Research in Personality*, vol. 43, no. 4, pp. 703–706, 2009.
- [169] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, “Personality judgments based on physical appearance,” *Personality and Social Psychology Bulletin*, 2009.
- [170] S. Nestler, B. Egloff, A. C. Küfner, and M. D. Back, “An integrative lens model approach to bias and accuracy in human inferences: hindsight effects and knowledge updating in personality judgments.,” *Journal of personality and social psychology*, vol. 103, no. 4, p. 689, 2012.
- [171] J. A. Hall and N. Pennington, “Self-monitoring, honesty, and cue use on facebook: The relationship with user extraversion and conscientiousness,” *Computers in Human Behavior*, 2013.

- [172] N. C. Krämer and S. Winter, “Impression management 2.0,” *Journal of Media Psychology: Theories, Methods, and Applications*, vol. 20, no. 3, pp. 106–116, 2008.
- [173] S. Mehdizadeh, “Self-presentation 2.0: Narcissism and self-esteem on facebook,” *Cyberpsychology, Behavior, and Social Networking*, vol. 13, no. 4, pp. 357–364, 2010.
- [174] Z. Papacharissi, “A networked self,” *A Networked Self*, p. 304, 2011.
- [175] Y.-H. Wang, *Taiwanese girls? self-portraiture on a social networking site*. PhD thesis, London School of Economics, 2012.
- [176] L. Qiu, J. Lu, S. Yang, W. Qu, and T. Zhu, “What does your selfie say about you?,” *Computers in Human Behavior*, vol. 52, pp. 443–449, 2015.
- [177] O. P. John and S. Srivastava, “The big five trait taxonomy: History, measurement, and theoretical perspectives,” *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.
- [178] J. Nie, P. Cui, Y. Yan, L. Huang, Z. Li, and Z. Wei, “How your portrait impresses people?: Inferring personality impressions from portrait contents,” in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 905–908, ACM, 2014.
- [179] N. Al Moubayed, Y. Vazquez-Alvarez, A. McKay, and A. Vinciarelli, “Face-based automatic personality perception,” in *ACM Multimedia*, 2014.
- [180] Y. Jia, “Caffe: An open source convolutional architecture for fast feature embedding.” <http://caffe.berkeleyvision.org/>, 2013.
- [181] F. X. Yu and et.al, “Designing category-level attributes for discriminative visual recognition,” in *CVPR*, IEEE, 2013.
- [182] I. A. Junglas, N. A. Johnson, and C. Spitzmüller, “Personality traits and concern for privacy: an empirical study in the context of location-based services,” *European Journal of Information Systems*, vol. 17, no. 4, pp. 387–402, 2008.
- [183] J. E. Bono and T. A. Judge, “Personality and transformational and transactional leadership: a meta-analysis,” *Journal of applied psychology*, vol. 89, no. 5, p. 901, 2004.

BIBLIOGRAPHY

- [184] A. Burdick, J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp, *Digital Humanities*. MIT Press, 2012.
- [185] P. T. Costa and R. R. MacCrae, *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-factor Inventory (NEO FFI): Professional Manual*. Psychological Assessment Resources, 1992.
- [186] R. R. McCrae and O. P. John, “An Introduction to the Five-factor Model and its Applications,” *Journal of Personality*, vol. 60, 1992.
- [187] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, “Workshop on computational personality recognition (shared task),” in *Proceedings of the Workshop on Computational Personality Recognition*, 2013.
- [188] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.
- [189] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, and M. E. Seligman, “Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach,” *PloS ONE*, vol. 8, no. 9, 2013.
- [190] S. C. Guntuku, L. Qiu, S. Roy, W. Lin, and V. Jakhetiya, “Do others perceive you as you want them to?: Modeling personality based on selfies,” in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pp. 21–26, ACM, 2015.
- [191] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, “Automatic Personality Assessment through Social Media Language,” *Journal of Personality and Social Psychology*, vol. 108, no. 6, pp. 934–952, 2014.
- [192] G. J. Meyer, S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, E. J. Eisman, T. W. Kubiszyn, and G. M. Reed, “Psychological Testing and Psychological Assessment: A Review of Evidence and Issues,” *American Psychologist*, vol. 56, no. 2, pp. 128–165, 2001.

BIBLIOGRAPHY

- [193] L. B. Wexner, "The Degree to which Colors (hues) are Associated with Mood-tones," *Journal of Applied Psychology*, vol. 38, no. 6, 1954.
- [194] K.-Q. Huang, Q. Wang, and Z.-Y. Wu, "Natural Color Image Enhancement and Evaluation Algorithm Based on Human Visual System," *Computer Vision and Image Understanding*, vol. 103, no. 1, 2006.
- [195] J. Bigun, *Vision with Direction: A Systematic Introduction to Image Processing and Vision*. Springer, 2006.
- [196] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. MIT Press, 1974.
- [197] J. A. Russell and A. Mehrabian, "Evidence for a Three-Factor Theory of Emotions," *Journal of Research in Personality*, vol. 11, no. 3, 1977.
- [198] V. R. K. Garimella, A. Alfayad, and I. Weber, "Social Media Image Analysis for Public Health," CHI, 2016.
- [199] V. Lampos, N. Aletras, D. Preotiuc-Pietro, and T. Cohn, "Predicting and Characterising User Impact on Twitter," EACL, 2014.
- [200] G. Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," in *Biennial GSCL Conference*, 2009.
- [201] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," NIPS, 2002.
- [202] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, 2000.
- [203] U. von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, 2007.
- [204] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*.

- [205] H. Zou and T. Hastie, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society, Series B*, 2005.
- [206] C. DeYoung, L. Quilty, and J. Peterson, “Between facets and domains: 10 aspects of the big five,” *Journal of Personality and Social Psychology*, vol. 93, no. 5, pp. 880–896, 2007.
- [207] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, “How flickr helps us make sense of the world: Context and content in community-contributed media collections,” in *Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA ’07*, (New York, NY, USA), pp. 631–640, ACM, 2007.
- [208] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A multi-view embedding space for modeling internet images, tags, and their semantics,” *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.
- [209] N. Srivastava and R. R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” in *Advances in neural information processing systems*, pp. 2222–2230, 2012.
- [210] J. M. Hughes, D. N. Rockmore, and Y. Wang, “Bayesian learning of sparse multi-scale image representations,” *Image Processing, IEEE Transactions on*, vol. 22, no. 12, pp. 4972–4983, 2013.
- [211] A. Wang, J. Lu, J. Cai, G. Wang, and T.-J. Cham, “Unsupervised joint feature learning and encoding for rgb-d scene labeling,” *Image Processing, IEEE Transactions on*, vol. 24, no. 11, pp. 4459–4473, 2015.
- [212] M. R. Azimi-Sadjadi, J. Salazar, and S. Srinivasan, “An adaptable image retrieval system with relevance feedback using kernel machines and selective sampling,” *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1645–1659, 2009.
- [213] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, “Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification,” *Image Processing, IEEE Transactions on*, vol. 24, no. 12, pp. 4766–4779, 2015.

- [214] L. Wang, Y. Qiao, and X. Tang, “Latent hierarchical model of temporal structure for complex activity classification,” *Image Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 810–822, 2014.
- [215] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang, “Video tracking using learned hierarchical features,” *Image Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 1424–1435, 2015.
- [216] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, “Rapid: Rating pictorial aesthetics using deep learning,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 457–466, ACM, 2014.
- [217] X. Yang, T. Zhang, and C. Xu, “Cross-domain feature learning in multimedia,” *Multimedia, IEEE Transactions on*, vol. 17, no. 1, pp. 64–78, 2015.
- [218] W. Zheng, X. Zhou, C. Zou, and L. Zhao, “Facial expression recognition using kernel canonical correlation analysis (kcca),” *Neural Networks, IEEE Transactions on*, vol. 17, no. 1, pp. 233–238, 2006.
- [219] J. T. Zhou, S. J. Pan, I. W. Tsang, and S.-S. Ho, “Transfer learning for cross-language text categorization through active correspondences construction,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [220] J. T. Zhou, X. Xu, S. J. Pan, I. W. Tsang, Z. Qin, and R. S. M. Goh, “Transfer hashing with privileged information,” in *The 25th International Joint Conference on Artificial Intelligence*, 2016.
- [221] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, “Hybrid heterogeneous transfer learning through deep learning,” in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.
- [222] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, “Heterogeneous domain adaptation for multiple classes,” in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 1095–1103, 2014.
- [223] Flickr, “Flickr, <https://www.flickr.com/>,” 2014.

- [224] M. H. Johnson, “Subcortical face processing,” *Nature Reviews Neuroscience*, vol. 6, no. 10, pp. 766–774, 2005.
- [225] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, “2d articulated human pose estimation and retrieval in (almost) unconstrained still images,” *International Journal of Computer Vision*, vol. 99, pp. 190–214, September 2012.
- [226] M. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari, “Detecting people looking at each other in videos,” *International Journal of Computer Vision*, vol. 106, no. 3, pp. 282–296, 2014.
- [227] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2879–2886, June 2012.
- [228] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [229] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [230] X. Qi, R. Xiao, J. Guo, and L. Zhang, “Pairwise rotation invariant co-occurrence local binary pattern,” in *Computer Vision – ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), vol. 7577 of *Lecture Notes in Computer Science*, pp. 158–171, Springer Berlin Heidelberg, 2012.
- [231] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin, “Feature congestion: a measure of display clutter,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 761–770, ACM, 2005.
- [232] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, “Learning color names for real-world applications,” *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1512–1523, 2009.

- [233] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
- [234] T.-T. Ng and S.-F. Chang, “Classifying Photographic and Photorealistic Computer Graphic Images using Natural Image Statistics,” tech. rep., ADVENT Technical Report, No. 220-2006-6, Columbia University, Oct 2004.
- [235] A. Langmeyer, A. Guglhör-Rudan, and C. Tarnai, “What do music preferences reveal about personality?,” *Journal of Individual Differences*, 2015.
- [236] M. Park, K. Hennig-Fast, Y. Bao, P. Carl, E. Pöppel, L. Welker, M. Reiser, T. Meindl, and E. Gutyrchik, “Personality traits modulate neural responses to emotions expressed in music,” *Brain research*, vol. 1523, pp. 68–76, 2013.
- [237] A. G. Wright, “Current directions in personality science and the potential for advances through computing,” *Affective Computing, IEEE Transactions on*, vol. 5, no. 3, pp. 292–296, 2014.
- [238] Flickr, “Freecg, <https://www.flickr.com/groups/freecg/>,” 2014.
- [239] M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha, “Marginalized denoising autoencoders for domain adaptation,” in *ICML*, 2012.
- [240] D. Kelly and J. Teevan, “Implicit feedback for inferring user preference: a bibliography,” in *ACM SIGIR Forum*, vol. 37, pp. 18–28, ACM, 2003.
- [241] M. Tan, L. Wang, and I. W. Tsang, “Learning sparse svm for feature selection on very high dimensional datasets,” in *ICML*, pp. 1047–1054, 2010.
- [242] M. Tan, L. Wang, and I. W. Tsang, “Learning sparse svm for feature selection on very high dimensional datasets,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1047–1054, 2010.
- [243] S. C. Guntuku, S. Roy, and L. Weisi, “Evaluating visual and textual features for predicting user ?likes?,” in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pp. 1–6, IEEE, 2015.

- [244] R. Hu and P. Pu, "Acceptance issues of personality-based recommender systems," in *Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09*, (New York, NY, USA), pp. 221–224, ACM, 2009.
- [245] R. Srivastava, J. Feng, S. Roy, S. Yan, and T. Sim, "Don't ask me what i'm like, just watch and listen," in *Proceedings of the 20th ACM International Conference on Multimedia, MM '12*, (New York, NY, USA), pp. 329–338, ACM, 2012.
- [246] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [247] F. Mairesse and M. Walker, "Words mark the nerds: Computational models of personality recognition through language," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 543–548, 2006.
- [248] S. C. Guntuku, S. Roy, and W. Lin, "Personality modeling based image recommendation," in *MultiMedia Modeling*, pp. 171–182, Springer, 2015.
- [249] K. Krippendorff, "Reliability in content analysis," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.
- [250] J. C. Biesanz and S. G. West, "Personality coherence: Moderating self–other profile agreement and profile consensus.," *Journal of personality and social psychology*, vol. 79, no. 3, p. 425, 2000.
- [251] D. A. Kenny, "Person: A general model of interpersonal perception," *Personality and Social Psychology Review*, vol. 8, no. 3, pp. 265–280, 2004.
- [252] D. A. Kenny, L. Albright, T. E. Malloy, and D. A. Kashy, "Consensus in interpersonal perception: acquaintance and the big five.," *Psychological bulletin*, vol. 116, no. 2, p. 245, 1994.
- [253] G. J. Boyle and E. Helmes, "Methods of personality assessment," 2009.
- [254] D. Rowse, "Why black and white photography, <http://digital-photography-school.com/why-black-and-white-photography/>," 2014. Retrived.

- [255] O. P. John and S. Srivastava, “The big five trait taxonomy: History, measurement, and theoretical perspectives,” *Handbook of personality: Theory and research*, vol. 2, pp. 102–138, 1999.
- [256] L. Li and H.-T. Lin, “Ordinal regression by extended binary classification,” in *NIPS*, pp. 865–872, 2006.
- [257] C.-W. Seah, I. W. Tsang, and Y.-S. Ong, “Transfer ordinal label learning,” *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, no. 11, pp. 1863–1876, 2013.
- [258] M. Tan, L. Wang, and I. W. Tsang, “Learning sparse svm for feature selection on very high dimensional datasets,” in *ICML*, pp. 1047–1054, 2010.
- [259] P. Liu, J. T. Zhou, I. W.-H. Tsang, Z. Meng, S. Han, and Y. Tong, “Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis,” in *Computer Vision–ECCV 2014*, pp. 151–166, Springer, 2014.
- [260] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, *et al.*, “Simplemkl,” *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.
- [261] M. Tan, I. W. Tsang, and L. Wang, “Towards ultrahigh dimensional feature selection for big data,” *Journal of Machine Learning Research*, vol. 15, pp. 1371–1429, 2014.
- [262] J. E. Kelley, Jr, “The cutting-plane method for solving convex programs,” *Journal of the Society for Industrial & Applied Mathematics*, vol. 8, no. 4, pp. 703–712, 1960.
- [263] C. M. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima, “Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth,” *Journal of personality and social psychology*, vol. 89, no. 6, p. 899, 2005.
- [264] A. Bjerstedt, “Warm-cool color preferences as potential personality indicators: preliminary note,” *Perceptual and Motor Skills*, vol. 10, no. 1, pp. 31–34, 1960.
- [265] R. E. Lucas, E. Diener, A. Grob, E. M. Suh, and L. Shao, “Cross-cultural evidence for the fundamental features of extraversion,” *Journal of personality and social psychology*, vol. 79, no. 3, p. 452, 2000.

- [266] A. Nadkarni and S. G. Hofmann, “Why do people use facebook?,” *Personality and individual differences*, vol. 52, no. 3, pp. 243–249, 2012.
- [267] C. Massey, S. TenBrook, C. Tatum, and S. Whittaker, “Pim and personality: what do our personal file systems say about us?,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3695–3704, ACM, 2014.
- [268] C. Chen, S.-y. Lee, and H. W. Stevenson, “Response style and cross-cultural comparisons of rating scales among east asian and north american students,” *Psychological Science*, 1995.
- [269] T. Masuda, R. Gonzalez, L. Kwan, and R. E. Nisbett, “Culture and aesthetic preference: Comparing the attention to context of east asians and americans,” *Personality and Social Psychology Bulletin*, vol. 34, no. 9, pp. 1260–1275, 2008.
- [270] J. D. Fearon, “Ethnic and cultural diversity by country*,” *Journal of Economic Growth*, vol. 8, no. 2, pp. 195–222, 2003.
- [271] O. Vartanian and V. Goel, “Emotion pathways in the brain mediate aesthetic preference,” *Bulletin of Psychology and Arts*, vol. 5, no. 1, pp. 37–42, 2004.
- [272] T. Adelaar, S. Chang, K. M. Lancendorfer, B. Lee, and M. Morimoto, “Effects of media formats on emotions and impulse buying intent,” *Journal of Information Technology*, vol. 18, no. 4, pp. 247–266, 2003.
- [273] S. C. Guntuku, W. Lin, M. J. Scott, and G. Ghinea, “Modelling the influence of personality and culture on affect and enjoyment in multimedia,” in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 236–242, IEEE, 2015.
- [274] Y. Zhu, A. Hanjalic, and J. A. Redi, “Qoe prediction for enriched assessment of individual video viewing experience,” in *Proceedings of the 2016 ACM on Multimedia Conference, MM ’16*, (New York, NY, USA), pp. 801–810, ACM, 2016.
- [275] S. C. Guntuku, J. T. Zhou, S. Roy, L. Weisi, and I. W. Tsang, “Deep representations to model user “likes,”” in *Computer Vision—ACCV 2014*, pp. 3–18, Springer, 2015.

- [276] S. C. Guntuku, J. T. Zhou, S. Roy, W. Lin, and I. Tsang, "Understanding deep representations learned in modeling users 'likes'," *IEEE Transactions on Image Processing*, 2016.
- [277] S. C. Guntuku, S. Roy, and L. Weisi, "Personality modeling based image recommendation," in *MultiMedia Modeling*, pp. 171–182, Springer, 2015.
- [278] S. C. Guntuku, J. T. Zhou, S. Roy, L. Weisi, and I. W. Tsang, "Who likes what, and why? insights into personality modeling based on imagelikes'," *IEEE Transactions on Affective Computing*, 2016.
- [279] E. T. Higgins, "Self-discrepancy: a theory relating self and affect.," *Psychological review*, vol. 94, no. 3, p. 319, 1987.
- [280] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language.," *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.
- [281] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.
- [282] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS Comput Biol*, vol. 10, no. 12, p. e1003963, 2014.
- [283] A. Vinciarelli, "Computing technologies for social signals," in *Emotions and Personality in Personalized Services*, pp. 101–118, Springer, 2016.