Informatics Approaches for Understanding Human Facial Attractiveness Perception and Visual Attention

by

Song Tong

A thesis submitted for the degree of

Doctor of Philosophy

in

Psycho-informatics

in the

Graduate School of Informatics

of the

Kyoto University

2021

Supervisors:
Prof. Taketsune KUMADA, Main Supervisor
Prof. Xuefeng LIANG, Co-Supervisor

Acknowledgments

Firstly, I would like to thank my advisors Xuefeng Liang and Takatsune Kumada. Prof. Liang brought me to Kyoto University, taught me how to be a scientist, and gave me a lot of help in research and life abroad. Prof. Kumada supported me financially over the past two years and lent me a great research vision for exploring frontier research questions in psychology. I'd also like to thank the members of my thesis and qualifying exam committees: Shohei Nobuhara, Ryoichi Nakashima, Shin'ya Nishida, and Jun Saiki.

My sincere thank also goes to my collaborators who made working on my research such a pleasure: Sunao Iwaki, Cheeseng Chan, Yuenpeng Loh, Yinghua Tan, and Yang Liu. I'd like to thank Ryoichi Nakashima, Ayumi Takemoto, Atsushi Nakazawa, Yong Yang, Yuming Fang, and Satoshi Tsujimoto, who gave me invaluable comments during my research. I am also thankful to the faculty members in the psycho-informatics lab, Takako Kowada, Natsuhiro Ichinose, who helped me a lot in daily life, Te Chi Huang, Qiang Zhang, Jumpei Yamashita, Byungju Kim, Xiangyue Zhao, Ran Wang, Wenqing Zhou, Ritsuko Iwaki, Ryuta Iseki, Saori Kajima, Ryuta Iseki, and Yoko Higuchi, who willingness to discuss research ideas, and for making it such a great hub to work. I especially thank Guanyu Chen, who collected data for my research, Jian Guo, and Yuyu Zhang, who encouraged me greatly during my study.

I got some life-long friends during this study experience. Zhicheng Huang, Yan Gu, Meng-Yu Kuo, Guochang Xu, Yun Zhou, and Siyang Yu, who colorful my university life. I especially appreciate Longfei Chen, who is my best friend, and his family. They accompanied my family to spend a lot of free time. I also enjoyed the Kyoto University and Japan, which provide related agencies to made my abroad life more accessible, such as KI-ZU-NA, KOKOKA, Townhall, Himawari Nursery, and St Mary's Kindergarten. I sincerely thank the staff in these agencies. Their hard work also inspired me a lot. In KI-ZU-NA, I recognized my Japanese teacher Machiko Ishimaru, who told me a lot about Japanese stories and cultures. Besides, I thank China Scholarship Council, which provided financial support for my study from

2016 to 2019, the Chinese embassy, and the Chinese association for international students, which cared about my life and mental health during the study.

Lastly, I'd like to thank my wonderful big family for indulging and supporting me, especially my wife, Zhili Wu, who is the love of my life and a staunch supporter of my research. Meanwhile, I appreciate the research topics, which improved my art of living, such as the broaden-and-build theory taught me to keep openness to face challenges. With these topics, I could travel worldwide to share my progress. I got many friends and mentors during the travel experience: Yan Zhang, Thomhert Siadari, Yifan Zuo, Xiaoqing Lyu, Liansheng Zhuang, Sung-En Chien, Zhengan Zhu, Hsiu-Ping Yueh, Weijane Lin, and Suling Yeh. They gave me valuable comments to improve my research.

Abstract

Informatics Approaches for Understanding Human Facial Attractiveness Perception and Visual Attention

by

Song Tong

Doctor of Philosophy in Psycho-informatics

Kyoto University

The majority of psychological theories have been investigated in the laboratory under a well-controlled experimental setting, such as using a small number of participants from a specific population and a limited number of well-controlled stimuli. However, the generalization of psychological theories to the outside of the laboratory is sometimes difficult because the real world is diverse, having a wide variety of individual differences and including complex stimuli. Recently, the advances in informatics approaches, e.g., machine learning and big data analysis, provide potential opportunities to complement the current psychological experiments and to investigate human cognition outside the laboratory. Firstly, deep neural networks (DNNs), a specific class of machine learning methods, learn effective representations from real-world data. Since some of the learned representations are biologically plausible, interpreting the representations gives a way to study human cognition outside the laboratory through the examination of DNNs. Secondly, social media platforms potentially contain data related to human action and cognition that are difficult to collect in laboratory settings. These large-scale behavior data, refers to as naturally occurring data sets (NODS), have the potential for understanding human cognition or complementing psychological theories.

However, a relatively small number of studies used the above informatics approaches to examine human cognition. Thus, this dissertation aims to forge these approaches into tools that psychologists can apply to complement and extend the current psychological experiments or theories. In Chapter 1, we discuss the issues of

external validity in the laboratory experiments that may be partially circumvented by examination of DNN and construction of NODS. In the following chapters, we propose two complementary frameworks to apply DNN and NODS to understand human cognition, respectively.

In the first study, we investigate facial attractiveness cognition through the internal representations learned by DNN. DNNs lock interpret-ability since they have millions of parameters, which is unexplained. Thus, how to train and evaluate the human-relevant representations of DNN is challenged. In Chapter 2, we train a DNN for facial attractiveness (namely FADNN) and interpret the representation learned by the FADNN. A dataset of facial images is constructed and used for training the FADNN to classify the attractiveness using four category-specific neurons (CSNs). Then, we calculate the activation values of CSNs for an external dataset and find a significant relationship of the values with human-rated scores of the dataset. Moreover, four face-like images are generated by deconvolution of the FADNN. The generated images depict attributes corresponding to the face categories, and some ideal arrangements of facial features related to facial attentiveness (putative ratios) are observed from these images. In Chapter 3, we measure the putative ratios on the generated images, revealing certain consistencies with results in previous psychological studies. In addition, simulated psychophysics-like experiments show that face images with varying putative ratios change the activity of the CSNs. These results are similar to those of human judgments reported in previous psychological studies. The results indicate that the FADNN can learn putative ratios as relevant features for the representation of facial attractiveness, based only on categorical annotation in which no annotated facial features for attractiveness were explicitly given. This finding advances our understanding of cognition of facial attractiveness via DNN-based approaches.

In the second study, we investigate the broaden-and-build theory by constructing and examining NODS. In general, there are many challenges in constructing the NODS, e.g., determine the target behaviors, minimize bias in the data, and structure the data into formats useful for testing an existing psychological theory, because the

online data may be incomplete and possibly erroneous. For the collected data, we develop machine learning methods to extract relevant features for human behaviors based on the theory. In concrete, we construct NODS from travel photos with tourist satisfaction and their attentional scopes (wide-view vs. narrow-view) detected by the machine learning model. Based on the theory, we assume that there is a causal relationship between tourist mood (i.e., satisfaction) and natural behaviors (i.e., attentional scope). In Chapter 4, we propose a machine learning model for classifying the attentional scopes (wide-view vs. narrow-view) from travel photos. In Chapter 5, we use the proposed model to detect the attentional scopes in a large number of travel photos and construct NODS. For 94 scenic spots, there is a significant correlation between the average tourist satisfaction and the percentage of wide-view photos. Tourists seem to prefer shooting wide-views at scenic spots with a higher rating of satisfaction, consistent with the broaden-and-build theory. We showed that the broaden-and-build theory is applied to real-world data with a variety of participants who perform photo-shooting in a natural setting.

In Chapter 6, we summarize the results and discuss the benefits of our framework to both psychology and informatics study. In addition, the limitations and the future directions of applying DNNs and NODS for understanding human cognition are discussed. To conclude, in this dissertation, we provide the two frameworks for implementing informatics approaches to study human cognition based on psychological theories. By two frameworks, we demonstrated possibilities to verify existing psychological theories in real-world settings, which can improve the external validity of the theory. In addition, informatics research can be benefited by psychological theories.

Contents

List of Figures

List of Tables

Nomenclature

Main Abbreviations

t-SNE	*t*-distributed stochastic neighbor embedding
ASM	Active shape model
BING	Binarized normed gradients
CNN	Convolutional neural network
CSN	Category-specific neuron
DNN	Deep neural network
DWT	Discrete wavelet transform
FADNN	Deep neural network model for facial attractiveness
FV	Fisher vector
HAF	Highly attractive female
HAM	Highly attractive male
HHFE	Histogram of the high-frequency energy
HVS	Human visual system
NODS	Naturally occurring data sets
NSCT	Non-sampled contourlet transform
PCA	Principal component analysis
SD	Standard deviation
SURF	Speeded-up robust features
UAF	Unattractive female
UAM	Unattractive male

Chapter 1

Introduction

The majority of psychological theories have been investigated in the laboratory under a well-controlled experimental setting, such as using a small number of participants from a specific population and a limited number of well-controlled stimuli. However, the generalization of psychological theories to the outside of the laboratory is sometimes difficult because the real-world is diverse, having a wide variety of individual differences and including complex stimuli. Recently, the advances in informatics approaches, e.g., machine learning and big data analysis, provide potential opportunities to complement the current psychological experiments and to investigate human cognition outside the laboratory. Two potentials are claimed. (1) Machine learning algorithms learn effective representations from real-world data and perform some perception tasks at the human performance level (Battleday et al., 2020; Paxton, 2020; Ritter et al., 2017). Moreover, several specific algorithms (especially deep neural networks, DNNs) learned representations with biological plausibility (Güçlü et al., 2016; Kagian et al., 2008; Rafegas & Vanrell, 2018), e.g., similar representation to visual object recognition in the human visual cortex (Cichy et al., 2017; Cichy et al., 2016). Thus, some psychologists claimed that the ability of feature representations gives a way to study human cognition outside the laboratory through the examination of DNNs (Kriegeskorte, 2015; Peterson et al., 2017; VanRullen, 2017; Y. Wang & Kosinski, 2018). (2) The social media platforms aggregate large amounts of data more richly and diversely than previously imagined. The new data streams potentially contain information about human action and cognition that are difficult to collect in laboratory settings (Adjerid & Kelley, 2018;

Griffiths, 2015). Paxton and Griffiths (2017) claimed to create naturally occurring data sets (NODS) with large-scale behaviors from social media platforms, which have the potential for providing insight into human cognition or refining existing psychological theories.

However, the process of applying the above informatics approaches to the studies of human cognition encounters some difficulties. One of the characteristics of data used for DNN training and big data analysis is *variety*. The data in classical psychological experiments are required to be controlled as well as possible. The critical reason is the *paradigm gap* between the experimental purposes of informatics and psychological studies. On the psychological side, the purpose is to understand human cognition and demonstrate what they measure for affording the highest internal validity with the control in laboratory settings. Thus, psychologists often encourage to control, carefully examine, resulting in the trade-off of the measurement results with high internal validity, while having to lose some parts of external validity. On the side of informatics research, especially for task-driven research, the purpose is to implement the data to specific tasks and functions, demonstrating the practical usability (ecological validity) with diverse data. For example, some DNNs are trained for specific tasks (e.g., face and object recognition) to obtain high accuracy. The recommendation systems target practical and accurate recommendations. Thus, informatics researchers often collect data with diverse features, while little of them look back to humans and provide insights into human cognition. The paradigm gap is the main barrier for applying these informatics approaches to psychology, which makes psychologists challenging to be deeply involved (e.g., of performing statistical data analysis and testing theories) in the analysis of a large amount of diverse data and makes the informatics researchers challenging to apply their computational and data resources for understanding human cognition.

To bridge the paradigm gap, we proposed two complementary frameworks, targeting to understand human cognition through the perspectives of *DNN* and *NODS*, respectively. (1) The first framework focuses on investigating how similar are the representations learned by DNN to human cognition. (2) The second framework focuses on testing the existing psychological theory through NODS created from social media platforms. These frameworks combined a large number of data, machine learning models (or DNNs), and psychological theories in two different ways and extended the scopes of psychological experiments based on these informatics approaches.

1.1 Current Issue to Laboratory Experiments

To increase the understanding of perception and cognition, psychological experiments in the laboratory often encourage controlling, carefully examining, and isolating the specific effects on the measures of task performance. Thus, a psychological experiment often uses a theory-driven approach. As shown in Figure 1.1, a theory-driven approach is usually described as (Maass et al., 2018): (1) formulating hypotheses from an existing theory; (2) designing experiments to minimize the confounding effects (stimuli, participants and tasks); (3) collecting the behavior data through experiments; and (4) analyzing data using statistical methods (e.g., analysis of variance and linear regression) to draw inferences. The majority of psychological theories were built, developed, and even rejected by insights from the results of such experiments. However, the generalization of psychological theories to the outside of the laboratory is sometimes difficult because most laboratory settings can not represent the natural environments, leading to a low external validity (Proctor & Xiong, 2020).

The typical laboratory experiments are restricted by time, space, effort, and cost. A laboratory experiment is often performed using a small number of participants and well-controlled stimuli, which are barriers to establish ecological validity. For example, using the Navon-letter task (Navon, 1977) shown in Figure 1.1, which is a widely-used task in the studies of visual attention, Navon (1977) found that, compared with local letters, global letters should be processed first. Such task leads to many interesting results, such as visual attention affects emotions (Fredrickson & Branigan, 2005; Srinivasan & Hanif, 2010). However, the stimuli have low

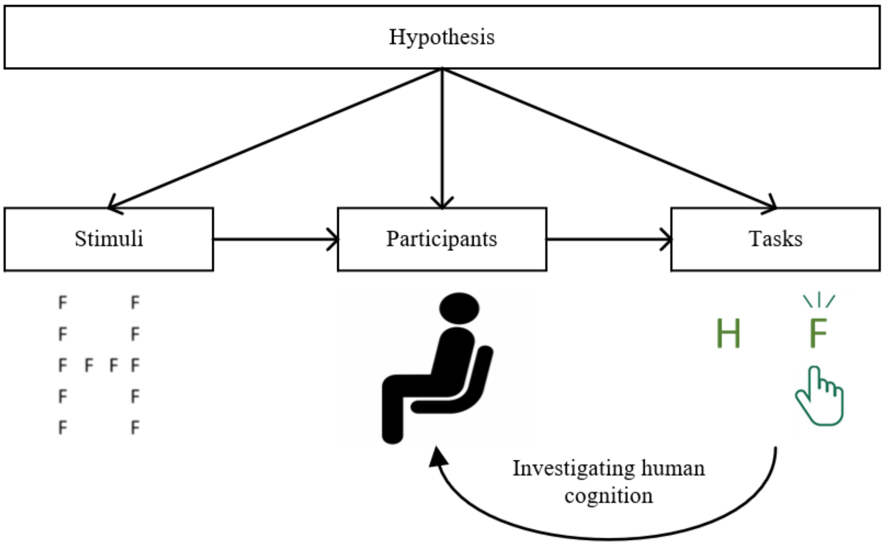


Figure 1.1: Investigate human cognition based on theory-driven approaches, for the Navon-letter task (Navon, 1977) as an example. The stimuli is a large letter made up of certain small letters. In the task, participants would press the key according which letter they recognized, e.g., 'H' or 'F'. Employ the actions to investigate the precedence between global processing and local processing.

generalizability, such as, Kinchla and Wolfe (1979) found the local letters would be processed prior to global letters, when the sizes of the stimuli change. To improve the generalizability, Ji et al. (2019) use the global-local of the landscape images (as shown in Figure 1.2) to replace the Navon-letter task in their study, which explores the association between emotion and global-local processing. In addition, psychological experiments usually are conducted by a small number of a specific population (such as college students), which provides a limited amount of human behavior data, which constrained the validity of the theories. Some researchers (Arnett, 2008; Henrich et al., 2010) criticized that most psychological theories of deriving from participants that are Western, educated, industrialized, rich, and democratic.

The advances in informatics approaches bring with the changes and opportunities. (1) The current DNNs learned object recognition have a similar level of perception as humans and make predictions from real-world stimuli with complex representations. (2) The social media platforms record human experiences and actions anytime,

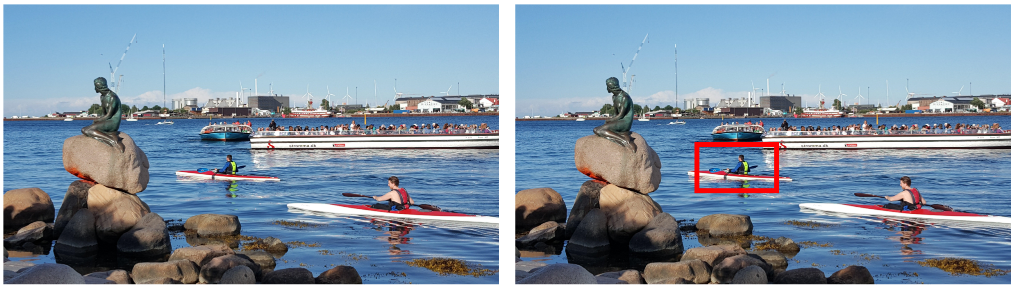


Figure 1.2: The redrawn landscape stimuli from a recent study (Ji et al., 2019) used for exploring the association between emotion and global-local processing with higher generalizability. In the task, the participants would give some descriptions about what they see on the entire landscape image (left), or the area with a bounding box. The image is downloaded from https://www.tripadvisor.com/LocationPhotoDirectLink-g189541-i265673288-Copenhagen_Zealand.html.

anywhere, making the data acquisitions breakthrough time and space limitations. A large-scale behavior data can be easily acquired from diverse participants, and the behavior is always naturally occurred. We will make these opportunities more concrete in the next section.

1.2 Opportunities and Challenges

This section provides opportunities and challenges to understanding human cognition via applying the approaches of DNN and NODS, respectively.

1.2.1 Deep neural network (DNN)

The DNN approaches or exceeds human performance in many essential perception tasks, such as object recognition (Krizhevsky et al., 2017) and face recognition (Parkhi et al., 2015; Y. Wang & Kosinski, 2018), because of their ability to learn relevant representations from natural images (real-world stimuli). These internal representations express the association between the real-world stimuli and outputs (target concepts). Some specific approaches can manipulate the representations (e.g., fully convolutional neural networks, FCN, and generative adversarial network, GAN), such as localizing the object concepts in the image space (e.g., semantic segmentation: Long et al., 2015) and generating the concepts within the images (e.g., manipulating the 'age' or 'expression' of a face image: Goodfellow et al., 2014; VanRullen, 2017). All the above indicate that the DNNs accurately distinguished the target concept and noise (confounding effects) in the real-world stimuli.

Moreover, some similarities have recently been observed between DNNs and biological vision systems in terms of their operations (Cadieu et al., 2014; Cichy et al., 2016; Kietzmann et al., 2019; O'Toole et al., 2018; Seeliger et al., 2018; P. Wang & Cottrell, 2017; Yamins & DiCarlo, 2016). For example, Cichy et al. (2016) showed that the DNN could capture the stages of human visual processing in both time and space, from the early visual areas to the dorsal and ventral streams. Rafegas and Vanrell (2018) reported that DNNs could capture hierarchical representations of color for object recognition correlating to human vision, such as the middle layers of the DNN show a denser sampling of hue than other layers, which suggests some correlation with hue maps in V2 (Conway, 2003; Shapley & Hawken, 2011). In addition, some research (Kriegeskorte, 2015; VanRullen, 2017) has proposed that DNNs can be used as a new tool to gain insight into biological visual perception, which provide practical and biological-plausible representations of real-world stimuli.

Therefore, it may contain some internal similarities between the representations of DNN and human perception. The relationship between the representation of DNN to human perception would extend the link from the real-world stimuli to human perception because representations learned by DNN are effectively correlated to real-world stimuli and the target concept.

However, comparing the representations of DNN with human perception has two challenges: (1) how to interpret the representations of DNN, and (2) how to evaluate the similarity between the representations of DNN and human. DNNs lock interpret-ability since they have millions of parameters, which is unexplained (Bau et al., 2017). One solution is inspired by neuroscience, which discovered brain neurons from response to specific concepts (A. Nguyen et al., 2016; Quiroga et al., 2005). Some informatics techniques (e.g., activation maximization: Erhan et al., 2009, and deconvolution method: Zeiler and Fergus, 2014) can generate the preferred images (stimuli) that highly active a specific neuron (DNN), which is accomplished by changing the input of DNN. The generated stimuli would make the implicit representation explicit and bring convenience to observation. In addition, to evaluate the similarity between the representation of DNN and human, we propose to implement psychological experiments in the specific neurons of DNN and then analyzing the behaviors of DNN. If DNN shows the human-like behaviors by manipulating stimuli as humans, this might evidence the similarity between the representations of DNN and human perception.

1.2.2 Naturally occurring data sets (NODS)

The social media platforms aggregate large amounts of data more richly and diversely than previously imagined. For example, TripAdvisor collects large numbers of experience ratings, reviews, and travel photos from worldwide, which features 859 million reviews by 2019 [1]. From a psychologist's perspective, the social media platforms provide diverse participants and massive actions that are difficult to acquire in laboratory experiments. In addition, these data usually occur spontaneously or describe people's daily life. In contrast, laboratory experiments essentially involve researchers intervening in the experimental process, which may cause some biases, e.g., social desirability and experimenter effects (Shah et al., 2015). For instance, the experimenters may unintentionally treat experimental subjects in different ways to shape their responses. Therefore, if the data collected from social media platforms can be used for study human cognition, it would advance the ecological validation of the measurements, because of the diverse and naturally occurring data.

In the past, most NODS might be called 'wild' data, typically gathered as observations of people, behaviors, or events for nonscientific purposes (Paxton & Griffiths, 2017). However, some recent works focus on the utility of naturally occurring data from the internet to complement the psychological experiments or refine the psychological theories (Goldstone & Lupyan, 2016; M. N. Jones, 2016; Paxton & Griffiths, 2017). Comparing to the past 'wild' data, the NODS collected from the internet is bigger, e.g., tera- or petabytes. In addition, the NODS for

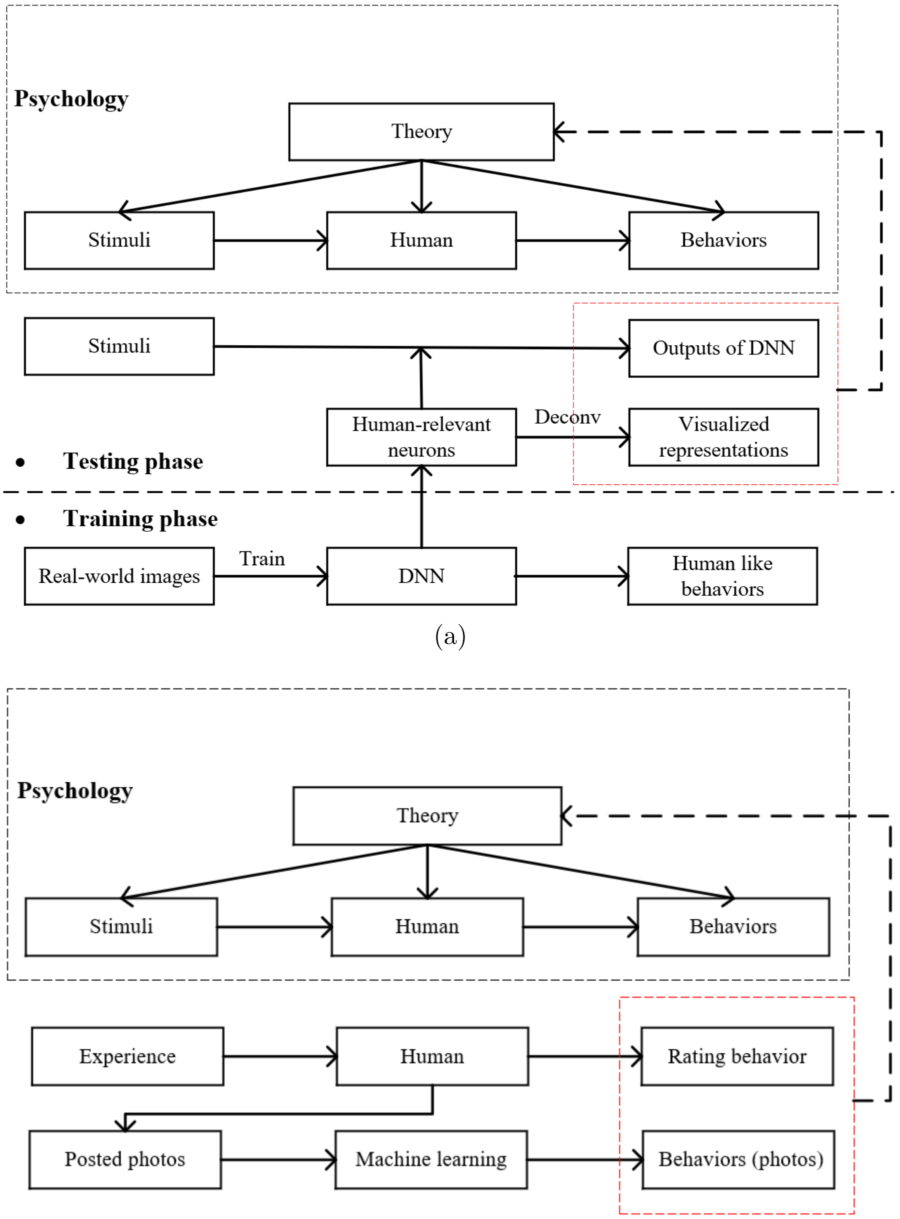
[1] https://en.wikipedia.org/wiki/Tripadvisor

psychological purposes is different from the dataset of *big data*. The big data required to create the dataset with 'Four Vs' (Data & Hub, 2013), i.e. *volume*, *variety*, *velocity*, and *veracity*. In contrast, Paxton and Griffiths (2017) augured that the NODS for psychological purposes only requires *veracity* and *variety*, and a medium size of *volume*. The reason is the NODS requires considerable controls of the main effects constrained by a psychological hypothesis, which would decrease the *volume* of data.

There are two main challenges to use the NODS from the social media platforms to study human cognition: (1) how to determine the main effects (target behaviors) and construct the NODS with minimizing bias in the data, and (2) how to analyze these NODS, such as to clean the data contents and to structure into formats usable for testing an existing theory (Endel & Piringer, 2015). To study human cognition using NODS, we have to look back to the theory-driven approach and start from hypotheses from an existing theory. The hypotheses would help us determine the target behaviors and acquire the data source. However, the collected data may be incomplete and possibly erroneous, which requires methods to extract meaningful information and prediction from the massive data (Proctor & Xiong, 2020). To solve this question, we pay more attention to computational modeling, i.e., training machine learning or DNN models to extract meaningful concepts or behaviors (main effects or confounding effects) from the collected data.

1.3 The Proposed Frameworks

This dissertation proposes two frameworks to study human cognition through the applications of DNN and NODS, respectively. The first framework, as shown in Figure 1.3a, investigates the similarity between the representations of DNN and human perception. The research process is divided into the training phase and testing phase of a DNN. The training phase aims to acquire the hypothesized human-relevant neurons by the DNN. In Chapter 2, we trained a DNN to detect face (un)attractiveness while evaluating the DNN in a dataset with continuous human-rated scores of attractiveness, which estimates whether the learned neurons include



(a)

(b)

Figure 1.3: The proposed complementary frameworks of studying human cognition through informatics approaches. (a) Deep neural network (DNN): The training phase would get the hypothesized *human-relevant neurons*. In the testing phase, we first interpret and visualize the human-relevant neuron, and then, we implement psychological experiments in the DNN by controlled stimuli as the input of DNN. (b) Naturally occurring data sets (NODS): Firstly, determine the main effects (behaviors included in photos and rating behavior) according to an existing psychological theory. Secondly, we develop machine learning-based models to refine the data to the main effects. Thirdly, analyzing the relationship between rating behavior and detected behaviors, then drawing an inference to extend the existing psychological theory.

informative representation to face attractiveness. The testing phase interprets and visualizes the human-relevant neurons' representations. The observation of the visualized representations may associate with some existing theories. For a further investigation, we implement experiments similar to psychological ones on the DNN and examine the neurons' output in Chapter 3, which estimates whether the association exists between theory and the learned representations.

Figure 1.3b shows the second framework of understanding human cognition through constructing NODS from social network platforms. This research process consists of three steps. Firstly, the existing psychological theory determines experience and main effects (target behaviors) and guides the data collection. In Chapters 4 and 5, we examine the hypothesis that tourist satisfaction correlates with visual attention according to an exiting psychological theory. Thus, we collect the tourism experience ratings and shared photos from social media platforms. Secondly, we focus on constructing machine learning-based models to identify the photo-shooting behaviors which reflect tourists' visual attention from the posted photos. Thirdly, we analyze the relationship between experience ratings and target photo-shooting behaviors, then drawing an inference to extend the exiting psychological theory.

1.4 An Overview of This Dissertation

The proposed complementary frameworks are proofs-of-concept designed to understand human cognition through informatics approaches. For the proofs-of-concept, this dissertation provides two case studies described as following.

1.4.1 Understanding facial attractiveness in deep feature representations

Chapters 2 and 3 provide the first study of applying representations learned by DNN for investigating human facial attractiveness perception. In this study, we investigate facial attractiveness for the following three reasons. (1) Facial attractiveness has significant social consequences, thus it stimulates great interests from diverse fields, e.g., arts, philosophy, biology, more recently, psychology (Perrett et al., 1994) and artificial intelligence (Eisenthal et al., 2006). (2) Although some individual differences and cultural differences have been noted, a large number of studies provide empirical evidence that people use similar criteria to determine facial attractiveness perception (Perrett et al., 1998; Perrett et al., 1994; Rubenstein et al., 2002). For example, it is showed that there is an ideal arrangement of facial features (ideal ratios) that can optimize the attractiveness of a person's face (Jefferson, 2004; Pallett et al., 2010; Valenzano et al., 2006). (3) Emerging research has successfully used DNNs for facial attractiveness prediction and enhancement (J. Li et al., 2015; Rothe et al., 2016; S. Wang et al., 2014), which indicates that DNNs can learn internal representations to formulate facial attractiveness. However, few studies have investigated whether the internal representation of DNN is interpretable based on the theory of human perception, leave much think of these DNNs as black boxes.

Chapter 2 trained a DNN for facial attractiveness and interpreting the representations learned by the DNN. Firstly, the DNN model is trained to recognize the attractiveness and gender (female/male × high/low attractiveness) of the face in the images using four category-specific neurons (CSNs) and achieves an accuracy of 0.9805. Secondly, the DNN model is tested on an external dataset with human-rated attractive scores, which showed a high correlation between the activate values of CSNs and human-rated scores. Thirdly, four face-like images are generated by using deconvolution of FADNN. The generated images depict intuitive attributes corresponding to the four face categories. In addition, the face-like images suggest that the putative ratios of facial attractiveness could be learned in the DNN. For further investigation, in Chapter 3, we measure the putative ratios in the face-like images, which reveal certain consistencies with reported evidence. In addition, simulated psychophysics-like experiments show that face images with varying putative ratios change the activity of the CSNs. These results are similar to those of human judgments reported in previous psychological studies. Thus, we conclude that the FADNN can learn putative ratios as relevant features for the representation of facial attractiveness, based only on categorical annotation in which no annotated facial features for attractiveness were explicitly given. This finding advances our understanding of human facial attractiveness perception via DNN-based perspective approaches.

1.4.2 Understanding tourist satisfaction through photo-shooting behaviors

Chapters 4 and 5 provide the second study to investigate the broaden-and-build theory (Fredrickson, 2004) by constructing and examining NODS from the travel experience. For the following three reasons, we focus on the travel experience in this study. (1) To remind their travel experiences, tourists often take photography to capture gaze and record attention (Osborne, 2000; Urry, 1992; Zeiler & Fergus, 2014). In addition, the broaden-and-build theory showed that visual attention affects emotion (Fredrickson & Branigan, 2005; Ji et al., 2019; Srinivasan & Hanif, 2010). Thus, it is possible to estimate the tourists' affective quality by the attentional scopes (wide-view vs. narrow-view) reflected from photos. (2) Current social networks (such as Flickr and TripAdvisor) record large numbers of user-generated content, such as travel photos and experience ratings (tourist satisfaction). (3) There are many computing resources for object recognition (Deng et al., 2009) and scene understanding (Zhou et al., 2017), which provide the potential for detecting behavior patterns from travel photos.

In this study, we propose the hypothesis that emotion affects photo-shooting behaviors. That is, positive emotion broadens visual attention, triggering the preference to take more wide-view photos. Then, we aim to the non-trivial task of the classification of narrow-view and wide-view images by machine learning algorithm in Chapter 4. In Experiment 1, we present a machine learning model to classify images inspired by the human visual system. We found two cues that can represent visual attention, i.e., focus cue and scale cue. The focus cue is modeled in the frequency domain. The scale cue is modeled by defining the spatial size and conceptual sizes of an object in the image. The experimental results on a newly established dataset with 5,050 images show the proposed model has better performance than the related methods. However, the proposed model is not suitable for extensive testing since it is difficult to implement in the GPU. Thus, we implement the focus cue and scale cue in an end-to-end DNN, which cumulative feature of the convolutional neural network (CFCNN) to learn both high-level and low-level features to represent the shooting patterns in Experiment 2. The newly proposed CFCNN obtains a considerable improvement in computational efficiency and accuracy.

In Chapter 5, we use the CFCNN to detect the attentional scopes detected from travel photos and construct the NODS with attentional scopes and tourist satisfaction. Firstly, we collect 39,099 photos from 30 popular scenic spots. We find a significant difference of wide percentage exists in 30 spots between high-rated scenic spots and low-rated scenic spots. At high-rated scenic spots, tourists prefer to take wide-view photos to capture wide landscapes (e.g., mountains, lakes, tall buildings). Secondly, we collect 549,772 photos from 94 popular scenic spots. In addition, to rule out the confounding factors, we use subgroup analysis to divide the scenic spots into different continents and scenic types. Moreover, we divide the photos into different visual contents, i.e., green, water, and building perceptions. Under the scenic spot and visual content conditions mentioned above, the experimental results still indicate a significant correlation between tourist satisfaction and photo-shooting behaviors, which is consistent with the broaden-and-build theory. At last, we discuss the cognitive process in the scenario of tourist experience evaluation. We hope that the results revealed in this study would provide a new perspective for psychologists to study human behaviors and cognition by employing the advantages of NODS in tandem with machine learning methods.

Chapter 6 summarizes the results reported in Chapters 2-5, and discusses implication of them in terms of the paradigm gap of research fields. Furthermore, we discuss the limitations and future directions for applying the informatics approaches for understanding human cognition.

Part I

Understanding Facial Attractiveness in Deep Feature Representations

Chapter 2

Learning Facial Attractiveness Representations Using DNN

This part aims to study facial attractiveness through the representations learned by DNN, demonstrating a case study of the complex facial attractiveness cognition applying the proposed framework shown in 1.3a. In Chapter 2, our purpose is to investigate the following issues: (1) learning the facial attractiveness representation using DNN and (2) interpreting the representations learned by DNN.

2.1 Introduction

Facial attractiveness has significant social consequences, thus stimulating great interests in diverse fields, including art, philosophy, biology, psychology (Burleson et al., 2016), and artificial intelligence (Gan et al., 2014). In psychology, for example, there is a debate that attractiveness is explained by some physical facial features or not (Cunningham, 1986). Some researchers believed that there were no cross-culturally universal standards for what constituted an attractive face. Some studies suggested that the local cultures affect the facial attractiveness evaluation across various social organizations, schools, families, and media(Coetzee et al., 2014). In addition, the individual differences for facial preference were also found across nations and ages (Cooper et al., 2006; Geldart et al., 1999; Penton-Voak et al., 2004). For instance, Penton-Voak et al. (2004) found the Jamaican preferred masculine faces more than Britain and Japanese due to their particular living conditions. Marcinkowska et al. (2014) showed that the facial attractiveness judgment varied across 28 countries and found a correlation between the femininity preference index and the national health index. Moreover, Perrett et al. (2002) showed that attractiveness judgment reflected the learning of parental characteristics. Reber et al. (2004) assumed that people would form a set of facial attractiveness ‘prototypes’ and consciously match the perceived face to their prototypes for attractiveness judgment.

However, variability in some aspects of facial attractiveness judgment does not preclude the possibility of other universally attractive characteristics. For example, Perrett et al. (1994) found that the average face is more attractive than individuals, and suggested the average face includes the attractive shape across different cultures. Cunningham et al. (1995) found even if the faces were judged by the participants from same or different cultural groups (i.e., Asians, Spaniards, white Americans and black Americans), the ratings by the groups showed strong consistency (Pearson correlation: $r > 0.9$). Notably, some studies (Langlois et al., 1987; Samuels & Ewy, 1985) showed that infants could distinguish attractive faces from unattractive faces rated by adults. Thus, some universal characteristics of facial attractiveness are proposed based on feature cues (e.g., eye size and chin shape) (Rhodes, 2013), the configurational cue (ratios between facial landmarks) (Danel & Pawlowski, 2007; Pallett et al., 2010), the skin cue (i.e., skin health and color) (Fink et al., 2001), and sexual characteristics (masculinity and femininity) (Little et al., 2011). For instance, Danel and Pawlowski (2007) proposed the eye-mouth-eye angle measuring the vertex in the middle of the mouth and the sides of the vertex crossing the centers of pupils as a configurational feature of attractiveness. Marquardt and Stephen (2002) also developed a front line contour map based on the golden decagon matrices for the standard of facial attractiveness, named golden facial mask, which was widely applied in plastic surgery (Bashour, 2006; Holland, 2008).

Most informatics researchers were inspired by universal characteristics of facial attractiveness and designed geometric features (Eisenthal et al., 2006; Kagian et al., 2007) and appearance features (Altwaijry & Belongie, 2013; Gray et al., 2010; Whitehill & Movellan, 2008) to compute the facial attractiveness. The geometric structure can represent the harmonious parts of the face and their ratios based on the

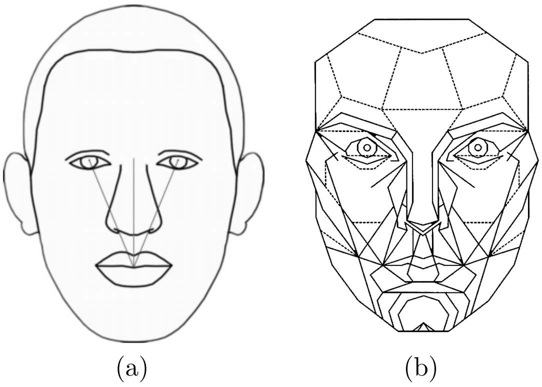


Figure 2.1: Geometric features for facial attractiveness: (a) Eye-Month-Eye angle (Danel & Pawlowski, 2007); (b) golden facial mask (Marquardt & Stephen, 2002).

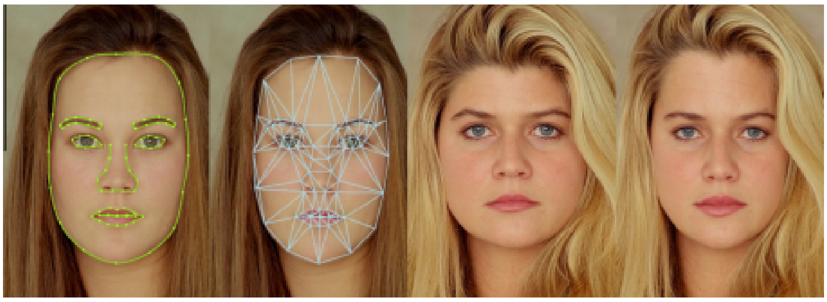
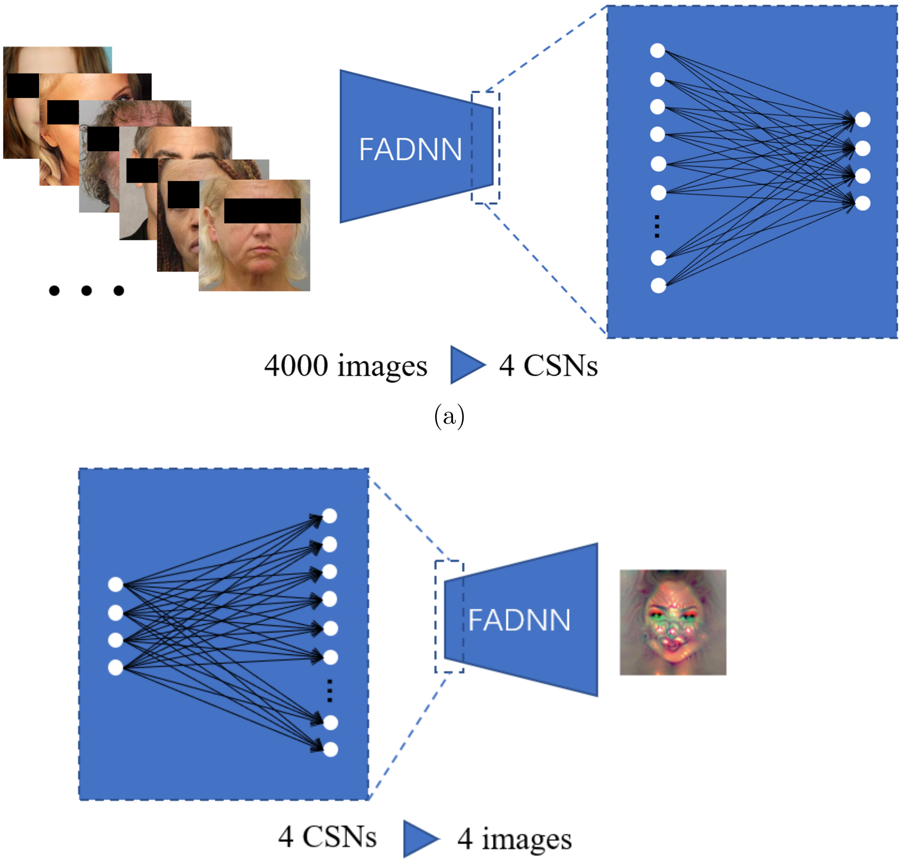


Figure 2.2: The geometric features based method for the enhancement of facial attractiveness. From left to right: face feature points; face triangulation; original and beautified face (Leyvand et al., 2008).

distances among some key fiducial points, and ratios between distances. Eisenthal et al. (2006) used 36-dimensional feature vectors and ratio vectors extracted from 92 face images to measure facial attractiveness, achieving a correlation of 0.6 with manual measurement. Kagian et al. (2007) utilized 90 principal components of 6972 distance vectors among 84 fiducial points of a face image and reported a correlation of 0.82 between the attractive scores rated by human and predicted scores. In addition, Leyvand et al. (2008) proposed a face triangulation to beautify faces based on 84 landmark points and 234 vectors with normalized lengths, as shown in Figure 2.1. However, geometric features require a large amount of manual workload for marking the points and lack the fine details, e.g., rippling muscles and structure transition of organs and expressions.



(a)

(b)

Figure 2.3: Summary of the methods. (a) Four category-specific neurons (CSNs) were trained using thousands of face images and their attractiveness-related annotations. (b) The feature representations learned by the CSNs were projected to four face-like images by deconvolution method.

Additionally, appearance features usually refer to the texture information, e.g., local binary patterns, the histogram of oriented gradients, color histograms, and scale-invariant feature transform, representing the local appearances of faces. Whitehill and Movellan (2008) employed Gabor filters, eigenface projections, and edge orientation histograms as features for a support vector regression to predict facial attractiveness. However, the method only reached a correlation between the attractive scores rated by human and predicted scores with 0.45 . Altwaijry and Belongie (2013) implemented a personalized beauty ranking system that utilized the histogram of oriented gradients, gist features, color histograms, and dense scale-invariant feature transform, and resulted in an average accuracy of 0.63.

Recently, emerging studies applied deep neural networks (DNNs) for facial attractiveness analysis due to the success of deep learning for visual recognition (LeCun et al., 2015; Rothe et al., 2016; S. Wang et al., 2014). In contrast to methods using handcrafted features, DNNs learn higher-level features from a large number of face images to yield more precise predictions of attractiveness (F. Chen et al., 2016; S. Wang et al., 2014). For example, Rothe et al. (2016) constructed a convolutional neural network (CNN) to learn facial attractiveness from thousands of images by relying on millions of ratings from the internet. The computational performance of DNNs has led to their use in smartphones, such as for face beautification (J. Li et al., 2015) and facial makeup recommendation (T. V. Nguyen & Liu, 2017). All the above suggest that DNNs could learn highly abstract and implicit feature representations to formulate facial attractiveness better than conventional features. However, little research investigated the intuitive and interpretable representations learned by DNNs, which are consistent with human attractiveness cognition, which leaves much thought of these models as black boxes.

In this chapter, we investigate interpretable representations learned by DNNs for facial attractiveness judgment. In concrete, we address the following issues: (1) training DNN for learning representation for the facial attractiveness judgment; (2) visualizing and interpreting the feature representations learned by DNNs.

To this end, we conducted two experiments. In Experiment 1, we trained a DNN model for gender and facial attractiveness (FADNN) classification (female/male × (un)attractiveness). In this model, four category-specific neurons (CSNs) were learned to represent four attractiveness categories from 4000 annotated face images, as shown in Figure 2.3a. The experimental results demonstrated that the FADNN learned more discriminating features and better performance for attractiveness classification than the traditional methods. In Experiment 2, to evaluate the robustness of the feature representations, we tested the FADNN using the other image dataset with attractiveness scores rated by human subjects. The results showed that the FADNN predicted human-rated scores with high accuracy, showing learned intended features contribute to the facial attractiveness classification. Furthermore, we used the deconvolution method (Mahendran & Vedaldi, 2015; Zeiler & Fergus, 2014) to visualize the learned representations, as shown in Figure 2.3b. The implicit features of the CSNs were projected to four face-like images by reversing FADNN (i.e., deconvolution: Yosinski et al., 2015). The face-like images made the implicit features of attractiveness explicit. This suggested that some configurational cues of facial attractiveness exist on the face-like images, which are consistent with the findings in previous psychological studies (e.g., Feser et al., 2007; Marquardt and Stephen, 2002; Valenzano et al., 2006).

2.2 Experiment 1: Train a DNN for Facial Attractiveness

In Experiment 1, we trained a DNN model for gender and facial attractiveness classification (FADNN). Figure 2.3a shows the concept of this experiment. The goal of FADNN is to extract feature representation across categories using its four CSNs. To train FADNN for facial attractiveness, the training dataset of FADNN must have sufficient representative instances and be diverse, i.e., large amounts of high and low attractive face images that contain a variety of attributes specific to each category. High and low attractive samples ensure the DNN model focusing on learning the intended features of facial attractiveness. The diverse data would prevent some shortcut features (Geirhos et al., 2020) from learning for classification. Here, the shortcut features denote the image properties correlated to the task while they are irrelevant to facial attractiveness. However, most open datasets are either too small (Altwaijry & Belongie, 2013; Eisenthal et al., 2006; Kagian et al., 2008) or contain mostly neutrally attractive faces (Rothe et al., 2016; Zhang et al., 2017). To address this issue, we collected the attractive set from celebrities while collected the unattractive set from mugshots, which provide sufficient representative instances and diverse data sources. The constructed dataset comprises 4,512 face images divided into four categories (female/male × high/low attractiveness).

However, images collected from two distinct data sources may introduce some confounding factors, which may also be learned by the FADNN model. To alleviate this issue, we took the following two steps to mitigate the influence of potentially confounding factors. Firstly, a strict selection process has been carried out to exclude low-quality images, such as cropping faces using a unified identification method (face localization) and three steps of subjective evaluations (see section 2.2.1.1 for details). Secondly, we applied transfer learning based on the VGG-Face model (Parkhi et al., 2015), which trained on millions of face images for face identification. The VGG-Face model can be used as a feature extractor and is similar to the traditional scoring keys that accompany psychometric tests. A traditional scoring key involves converting responses to test questions into a set of psychometric scores, while here the VGG-Face translates a face image into a score with 4,096 dimensions which well-trained for face identification. Each dimension might subsume differences in multiple features in the convolution layers. The VGG-Face offers two main advantages in the context of this study. (1) Successful facial recognition depends on robust facial features. The VGG-Face aims to represent a given face as a feature vector of identity, which is invariant under various changes in the facial expression, background, lighting, and head orientation, and such properties of the image as brightness and contrast. (2) Transferring the parameters of a well-trained DNN to a different but related task reduces the risk of overfitting (Y. Wang & Kosinski, 2018).

We designed the output layer of FADNN to consist of four neurons corresponding to the four categories (CSNs). Because the VGG-Face translates a face image into a feature vector, the training process on the collected dataset involved optimizing the weights of the connections between the feature vector and the four CSNs. Furthermore, each of CSNs needs to be strongly activated by face images from the corresponding category. We compared the feature representation of FADNN with two traditional features used for facial attractiveness computing, i.e., geometric (Kagian et al., 2008) and Gabor features (Y. Chen et al., 2010).

2.2.1 Materials and methods

2.2.1.1 Dataset

To collect candidates for the attractive sets, we focused on celebrities and their publicly available images (Zhang et al., 2017). The image data was collected from two websites, TC Candler [1] and TB World [2]. These sites listed millions of suggestions in terms of images of attractive people every year. In this study, a name list with 406 females and 247 males was created according to the rankings of these websites in recent years, and 65,300 photos of these person were crawled via Google Image Search (100 photos/person).

Considering the accessibility of the dataset, we select the unattractive set from criminal mugshot images, because the mugshot images is public and contain relative larger variety of face images from attractive to unattractive (MacLin & MacLin, 2004), comparing with celebrities. In this study, we collected 39,764 mugshot images from the "St. Louis Arrests and Mugshots website" [3] as the resource for the unattractive set.

Because the locations of the facial regions varied in images, at first, they were detected and cropped using the Deformable Part Model proposed in Mathias et al. (2014) and normalized to 224×224 (Parkhi et al., 2015). To ensure image quality, we cleaned up the raw dataset using the following two steps: (1) Two informatics researchers (male: aged 25 and 41 years) scrutinized all face images and filtered out erroneously tagged, side-view, and duplicate photos. (2) Three males and three females (average age: 27.3 years, SD = 2.7) rated their agreement (agree or disagree) with the attractiveness of the faces in the sets of highly attractive and unattractive face images. Finally, we separated the two sets into four categories: highly attractive females (HAFs), unattractive females (UAFs), highly attractive males (HAMs), and unattractive males (UAMs). For each category, we chose 1,128 face images that had secured the agreement of at least three participants to construct a dataset of facial

[1] https://www.facebook.com/independentcritics/

[2] https://www.facebook.com/topbeautyworld.d/

[3] http://www.stlmugshots.com/

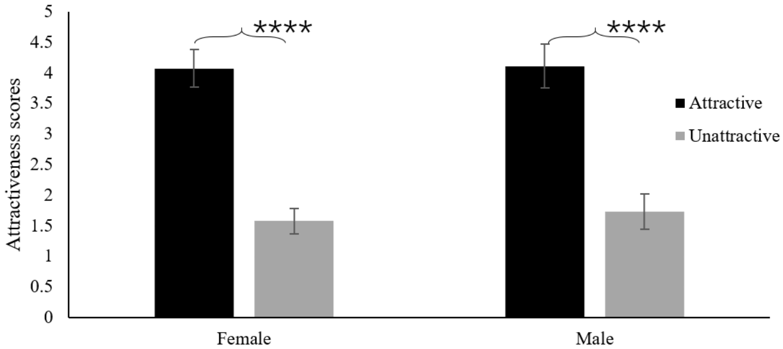


Figure 2.4: Attractiveness ratings of images in the constructed dataset. ****: $p < 0.00001$.

attractiveness, consisting of 4,512 face images with four categories.

To ensure that the images are extremely (un)attractive, we conducted an experiment where participants rated the attractiveness of these data. First, 240 images (60 per category) were randomly selected from the dataset. Each image was displayed on the screen for 1.5 s, and then participants were given 3.5 s to rate it by one of five categories (as “Very Attractive”, “Attractive”, “Neutral”, “Unattractive”, or “Very Unattractive”). Fifteen participants were recruited for this experiment (all male, average age: 22.9 years, SD = 2.4). The rating results of 13 participants were valid, and these results are shown in Figure 2.4. A significant difference (unpaired t-test: $t(118) = 52.002$, $p = 3.387 \times 10^{-83}$) exists between attractive females (Mean ± SD: 4.073 ± 0.308) and unattractive females (1.576 ± 0.208). Analogously, a significant difference (unpaired t-test: $t(118) = 39.769$, $p = 3.399 \times 10^{-70}$) also exists between attractive males (4.110 ± 0.359) and unattractive males (1.731 ± 0.293). This study was approved by the Ethics Committee in the Unit for Advanced Studies of the Human Mind, Kyoto University. Participants provided their written informed consent to participate in this study.

2.2.1.2 Training the category-specific neurons (CSNs)

The VGG-Face has six convolutional layers and three fully connected layers. Hierarchical computation was used to repeatedly stack the following operations: (1) a set of learned two-dimensional convolution operators that extract features hierarchically from the input stimuli and the output of previous convolutional layers; (2) spatial maximum pooling applied to a small local region of the feature maps obtained from the convolutional operation, which reduces their dimensionality and gains some degree of translational invariance; and (3) nonlinear activation functions applied to the pooled responses. The learned low-level features (i.e., edges, corners, and contours) are represented in the initial layers whereas the subsequent layers represent high-level object-related representations (i.e. parts of objects or entire objects) (Zeiler & Fergus, 2014). The fully connected layers (i.e., fc6, fc7, and fc8) are typically on top of these stacked convolutional operations and generate more abstract and task-dependent features. The fc6 layer transforms the two-dimensional feature map into a 4,096-dimensional feature vector. The fc7 layer is a hidden layer with the same number of dimensions, and could be considered as a middle-level feature vector. The fc8 layer has 2,622 neurons, each of which represents an identity. Finally, the output layer is a softmax layer with a cross-entropy objective function that calculates the probability of the target category of the input. In this study, we replaced the fc8 layer of VGG-Face with a layer consisting of four neurons for attractiveness classification, that is, FADNN.

The collected dataset (4,512 face images) was split into a training set and a testing set, where the former consisted of 4,000 images, (1,000 per category), and the latter contained 512 images (128 per category). FADNN was fine-tuned on the last fully connected layer while other parameters were fixed. The model was trained by minibatch gradient descent with a batch size of 20 (owing to memory-related constraints) for 30,000 iterations to ensure convergence. The learning rate of the last layer was set to 0.001, the weight decay was set to 5×10^{-4}, and the momentum was 0.9. The process was conducted using the Caffe deep learning framework (Y. Jia et al., 2014) on an NVIDIA Tesla K40 GPU 12 GB.

2.2.1.3 Traditional feature-based models for facial attractiveness

Geometric features are defined by a set of landmarks on the face (Zhang et al., 2017) that include shape information. The face landmarks were automatically

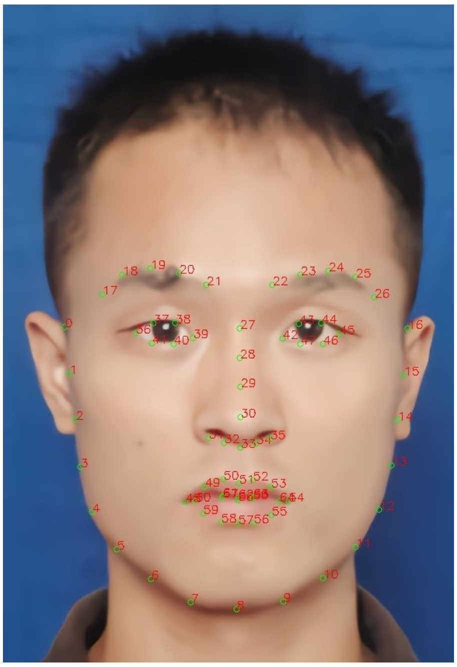


Figure 2.5: Graphical illustration of the facial landmarks (green dots, n=68) and their indices produced by the Active Shape Model (ASM). (The sample image, used for illustration only, is of S. Tong and is presented with his full consent.)

extracted using the Active Shape Model (ASM) (Cootes et al., 1995; B. C. Jones et al., 2007; Sagonas et al., 2013). Figure 2.5 shows the output of the ASM in a graphical format. Here, the geometrical features were generated from the extracted landmarks, composed of 2,346 pairwise distances. A Support Vector Machine (SVM) was then used as the classifier to distinguish images using these extracted features (Mao et al., 2009). The SVM is a robust classifier that constructs an n-dimensional hyperplane to optimally separate the data into different categories. The radial-basis function kernel (Chang & Lin, 2011) was used as well, $K(x, y) = \exp(-\gamma \|x - y\|), \gamma > 0$, where γ is a kernel parameter. Gabor features are appearance features encoding facial shape and texture information over a range of spatial scales (M. Yang et al., 2013) and are formed by convolution of the image with a family of Gabor kernels. These kernels had five scales and eight orientations with a size of 39 × 39 pixels (L. Shen & Bai, 2006; Zou et al., 2007). The results of convolution were vectorized as a 2,560-D vector (M. Yang et al., 2013) and classified by the SVM.

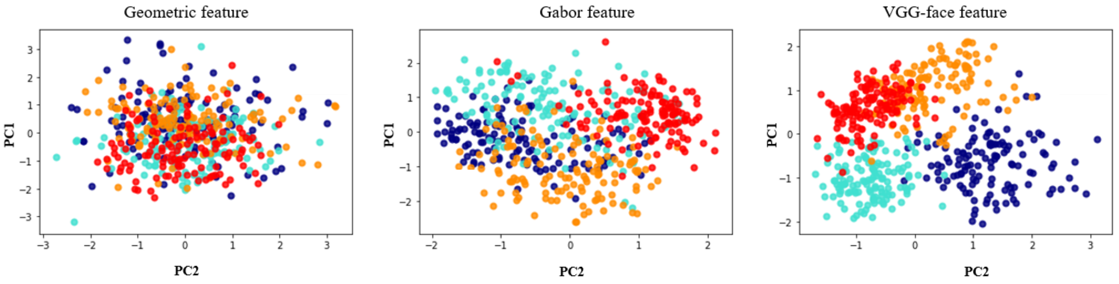


Figure 2.6: Comparison of the separability of models using three kinds of features through PCA-based dimensional reduction. The visualization of the extracted features illustrates the distribution of the facial features in the four categories (marked by different colors). Blue dots represent highly attractive female faces; green dots represent unattractive female faces; yellow dots represent highly attractive male faces; and red dots represent unattractive male faces.

Table 2.1: Comparison of separability of the three feature models using within-class and between-class distances.

Methods	Within-class distance (WD)	Between-class distance (BD)
Geometric feature	1.660	0.633
Gabor feature	1.284	1.527
VGG-Face feature	0.923	1.910

2.2.2 Results and discussion

To evaluate the feature representations learned by FADNN, we compared the FADNN with traditional feature-based models for facial attractiveness, including those based on geometric and appearance-based features. We then used Principal Component Analysis (PCA) to visualize the high-dimensional features extracted by the models for intuitive analysis of the separability of the 512 test images (Barshan et al., 2011). These test images were not presented to FADNN in the training phase. Figure 2.6 shows the distributions of the extracted features in the two-dimensional space of PCA. In the model that used geometric features, the four categories were mixed with one another, thus having low separability. The model using Gabor features has better separability than the geometric feature model, but it was difficult to show enough separability. In contrast, the VGG-Face feature (fc6) exhibits the best separability, which indicates it has the ability to classify different attractiveness categories.

Table 2.2: Comparison of the classification accuracy of different models.

Methods	Accuracy
Geometric features model	0.906
Gabor features model	0.928
FADNN	0.981

In addition, the within-class and between-class distances (WD and BD) were used to quantitatively evaluate the separability of these features. WD and BD denote the distances between features within a category and between categories respectively. Thus, smaller the values of WD and larger the values of BD indicate that the features are more separable. The DW and DB are and defined as follows:

$$\begin{aligned} DW &= \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{d}(x_{j,k},c))}{J \times K}, \\ DB &= \frac{\sum_{j=1}^{J} d(\bar{x}_j,c)}{J}, \end{aligned} \tag{2.1}$$

where $x_{j,k}$ represents the k-th instance in the j-th class, $\bar{d}(x_{j,k},c)$ represents the average Euclidean distance between instance $x_{j,k}$ and all other $k-1$ instances in the j-th class, and $d(\bar{x}_j,c)$ represents the mean Euclidean distance between the center of the j-th class and the center of the other $j-1$ classes.

Table 2.1 shows the quantitative results of the models. The VGG-Face feature model obtains the smallest WD and largest BD, which suggests that it is more discriminating than the other feature models and more representative of the attractiveness categories. Moreover, we compared the models in terms of the predictive accuracy of the 512 test images, as shown in Table 5.1. Both the geometric feature-based and Gabor feature-based models achieve accuracies over 0.90, but FADNN outperforms them with an accuracy of 0.981. These results indicate that FADNN can learn more discriminative information from the different categories than the traditional feature-based models.

2.3 Experiment 2: Interpret the Deep Feature Representations

Experiment 1 indicated that the FADNN learned compelling features to discriminate the facial attractiveness categories, outperforming traditional feature-based models. However, the FADNN is in a data-driven way, which may capture whatever information that was varying across given categories in the training images (M. Q. Hill et al., 2019). To examine the possibility that the captured representation consists of physiognomic features contributing to (un)attractiveness, we designed an additional experiment that interprets feature representations learned by FADNN, discovering the association between the feature representations with attractiveness.

Firstly, we tested the FADNN on another dataset (L. Liang et al., 2018) without using faces from the dataset for training. This dataset constructed using face images with continuous attractiveness scores given by humans, the same race and neutral expression without backgrounds. The result demonstrates that the scores of CSNs correlated with attractiveness scores, which is a shred of converging evidence that FADNN learns feature representations correlating with facial attractiveness.

Secondly, to intuitively interpret the feature representations for attractiveness categories, we used the deconvolution method (Mahendran & Vedaldi, 2015; Zeiler & Fergus, 2014) to visualize the learned representations of CSNs as face-like images. Figure 2.3b shows the flowchart of the visualization. The deconvolution method was proposed to interpret the intrinsic functions of DNNs. The face-like image showed in Figure 2.3b is an example generated by the neuron representing the category of attractive females. This method can generate an image that maximizes the activation of a specific neuron by inputting a content-free (noise) image (X. Liang et al., 2019). Because the CSNs are located in the fully connected layer, the position information of the preferred patterns is lost (Horikawa & Kamitani, 2017). Instead of a content-free image, we used the mean of all faces as input. Thus, the features learned by a CSN can be projected back into a facial space (Grant et al., 2016).

2.3.1 Methods

2.3.1.1 Dataset with attractiveness scores

The SCUT-FBP5500 dataset (J. Zhao et al., 2019) contains 5,500 facial images. Each image was rated by 60 volunteers. The average score was used as the ground truth to remove personal preference bias. The dataset has been used for studies of facial attractiveness analysis in both computer vision studies (Shi et al., 2019; Xu et al., 2019) and the psychological studies (J. Zhao et al., 2019), which was divided into four subsets with diverse races and both genders, including 2,000 Asian females, 2,000 Asian males, 750 Caucasian females, and 750 Caucasian males. Because the FADNN trained using dataset with mostly Caucasian faces, only face images of 750 Caucasian females and 750 Caucasian males were used for this experiment. Note that the training data included none of the faces in the SCUT-FBP5500 dataset.

2.3.1.2 Method to generate mean face images

The mean face images were generated by three sequential processes: face landmark detection, coordinate transforms, and averaging. First, the facial landmarks of all faces were extracted by the ASM, as shown in Figure 2.5. Second, landmarks on the coordinate system of the input face were transferred to the average face coordinate system by the following similarity transform matrix:

$$S = \begin{bmatrix} s_x cos(\theta) & sin(\theta) & t_x \\ -sin(\theta) & s_y cos(\theta) & t_y \end{bmatrix}, \tag{2.2}$$

where the first two columns of this matrix encode rotation (θ) and scale (s_x,s_y), respectively, and the last column encodes translation (t_x,t_y) in the x and y directions. Third, the mean face image was generated by taking the mean of the pixel intensities of all warped face images.

2.3.1.3 Generating face-like images for CSNs

The deconvolution method (Mahendran & Vedaldi, 2016; A. Nguyen et al., 2016) was employed to visualize the learned representation for the CSNs of FADNN. The technique was used to project the learned representation of a particular neuron onto an input image by maximizing the activation of this neuron (Erhan et al., 2009) (as an optimization problem). A mean face image was first propagated through FADNN, and the derivative of a target CSN activation was then calculated with respect to each pixel of the input image [4]. Then, changed the pixel values of the input image to increase the activation of the target CSN. The total variation and jitter (described in the study: Mahendran and Vedaldi, 2016) were used in the optimization process to improve the quality of the generated images. The optimizer was implemented using the source code provided by A. Nguyen et al. (2016). Finally, four face-like images corresponding to the four CSNs (categories) were generated by the deconvolution method.

2.3.2 Results and discussion

2.3.2.1 Test deep feature representations on another dataset

To evaluate the robustness of our model, we used PCA to visualize the high-dimensional features extracted by the model on the data in the SCUT-FBP5500 dataset. However, the SCUT-FBP5500 dataset lacks categorical labels, which is different from our constructed dataset. Thus, we chose the 10% highest/lowest scoring faces for testing, i.e., 75 images for each category. Figure 2.7 shows the distribution of the VGG-Face features of these 300 images, which is highly similar to the distribution shown in Figure 2.6.

In addition, we compared the correlation between the activation of the CSNs and the attractiveness scores. Before the analysis, it should be noted that the distribution of the attractiveness scores of the SCUT-FBP5500 dataset is unbalanced (see Figure 2.8), which may affect the correlation analysis. Thus, we re-balanced the dataset as follows. First, the female and male facial sets were each divided

[4] We only use one mean face as input without cluster one category into various aspects, which is different from the original setting in A. Nguyen et al. (2016). A. Nguyen et al. (2016) aims to observe the variance within the category from various aspects. However, we target to compare the variance between categories. Thus, we average the faces of images from different categories and use the same image as input for a fair comparison.

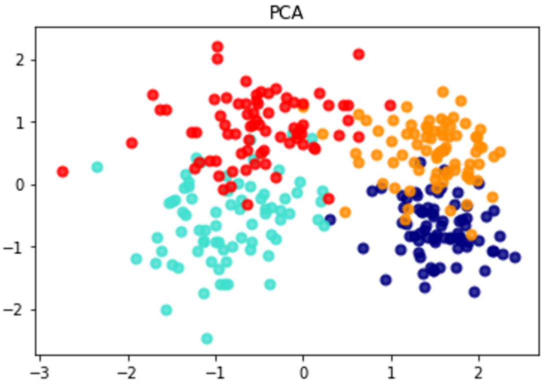


Figure 2.7: Distribution of the facial features of the SCUT-FBP5500 dataset. Blue dots represent the 75 female images with the highest attractiveness scores (top 10%); green dots represent the 75 female images with the lowest attractiveness scores (bottom 10%); yellow dots represent the 75 male images with the highest attractiveness scores (top 10%); and red dots represent the 75 male images with the lowest attractiveness scores (bottom 10%).

into 20 subsets according to the attractiveness scores (i.e., from lowest score to highest score in 20 steps). Second, 10 face images were randomly selected from every subset. All images would be selected if the subset had no more than 10 faces. We performed 10 repeated sampling phases. In each sampling phase, 173 female face images and 180 male images were selected. Figure 2.8 (right) shows the distribution of the attractiveness scores of a sampling phase (353 face images). Finally, the face stimuli were fed to FADNN, which outputted the scores of the four CSNs in the last fully connected layer (fc8): the highly attractive female neuron (Neuron-HAF), unattractive female neuron (Neuron-UAF), highly attractive male neuron (Neuron-HAM), and unattractive male neuron (Neuron-UAM). For each sampling phase, the scores of Neuron-HAF and Neuron-UAF were recorded for the 173 Caucasian females. Analogously, the scores of Neuron-HAM and Neuron-UAM were recorded for the 180 Caucasian males.

Pearson correlation was used as the metric to evaluate the relationship between the CSN scores and human-rated attractiveness scores. For the female faces, the average correlation coefficient (of the ten repeated sampling phases) between the scores of Neuron-HAF and the attractiveness scores was $r = 0.7583 \pm 0.0198$, all

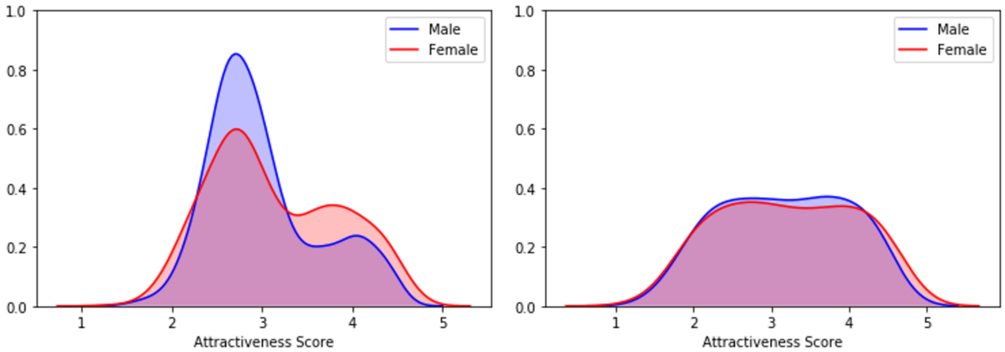


Figure 2.8: Distribution of the attractiveness scores of the SCUT-FBP5500 dataset. The left figure illustrates the distribution of the attractiveness scores (3,000 face images). The right figure illustrates the distribution of the attractiveness scores after re-balancing (353 face images).

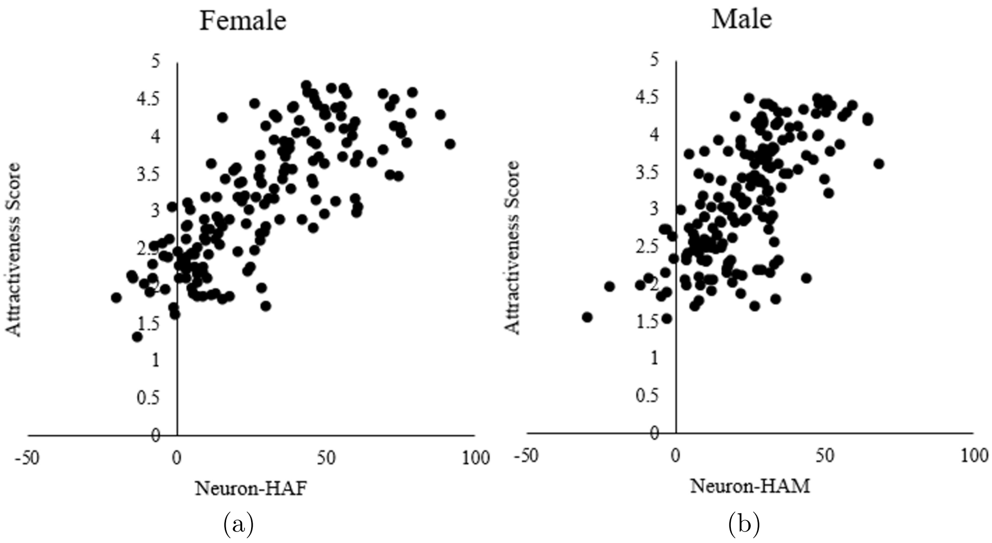


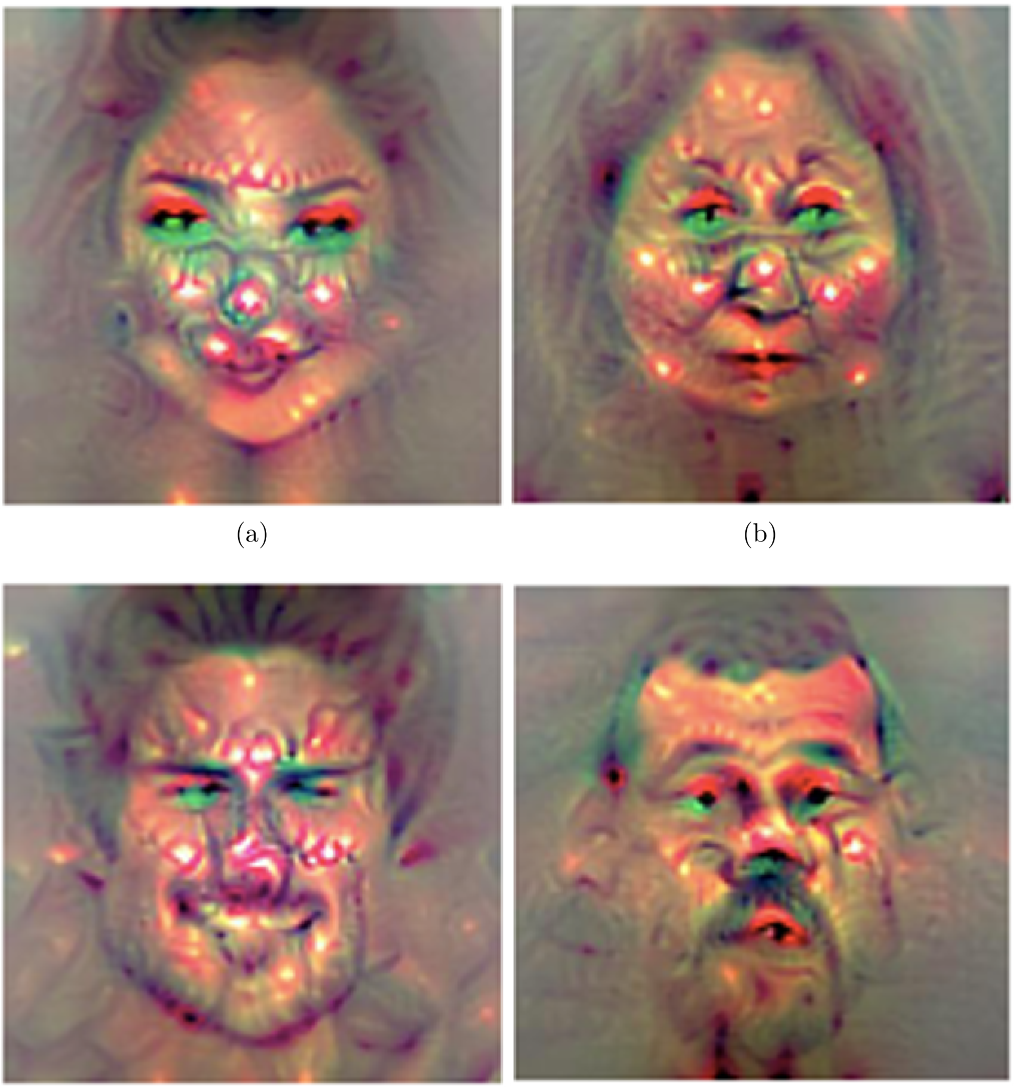
Figure 2.9: The correlation between the activation of the CSNs and the attractiveness scores. (a) Correlation between the Neuron-HAF scores and attractiveness scores of 173 female images. (b) Correlation between the Neuron-HAM scores and attractiveness scores of 180 male images.

p <0.00001 (see Figure 2.9a), whereas the average correlation coefficient between the scores of Neuron-UAF and the attractiveness scores was $r = -0.5655 \pm 0.0217$, all $p < 0.00001$. For the male faces, the average correlation coefficient between the scores of Neuron-HAM and the attractiveness scores was $r = 0.6653 \pm 0.0161$, all $p < 0.00001$ (see Figure 2.9b), whereas the average correlation coefficient between the scores of Neuron-UAM and the attractiveness scores was $r = -0.6221 \pm 0.0202$, all $p < 0.00001$. Xu et al. (2019) found that the handcrafted feature-based models can predict attractiveness scores with Pearson correlations of [0.6, 0.7], whereas the DNN-based models can predict attractiveness scores with Pearson correlations of [0.8, 0.9]. Unlike these models, even though that our FADNN was not trained on the SCUT-FBP5500 dataset, Neuron-HAF and Neuron-HAM achieve performances that are comparable with those of handcrafted feature-based models. This result demonstrates that the four trained CSNs subsume the main features associated with facial categories and can be used as metrics to determine the membership of an image in a category.

2.3.2.2 Visualize the deep feature representations

The four generated face-like images are shown in Figure 2.10 that present mainly facial features. In addition, these images illustrate the different visual features of the four categories. For instance, the generated image for Neuron-HAF has a pointed chin and standard arched eyebrows. On the contrary, the generated image for Neuron-UAF has a short chin and short eyebrows. These features are consistent with the shape analysis in terms of the attractiveness of female faces (Feser et al., 2007; Valenzano et al., 2006). Moreover, the image generated for Neuron-HAM has a square chin, which correlates with the analysis of attractive male faces (Rhodes, 2006).

Moreover, Figure 2.11 shows the face-like image generated from Neuron-HAF covered with the golden facial mask, which was proposed by Marquardt and Stephen (2002) that derived from Golden Decagon Matrices composed of the complicated relationship of the Golden ratio *phi*. It seems that the attractive face generally conforms to the golden facial mask. However, the claim that attractive faces should conform to the golden facial mask could not be tested in real-world faces due to the difficulty in defining the facial landmarks to make measurements corresponding to the geometrical lines of the golden facial mask. The result suggests that the FADNN capture golden facial mask from the real world face images as features for classification.



(a) (b) (c) (d)

Figure 2.10: Generated face-like images illustrating the discriminating features learned from raw face data. The images were generated using methods using different CSNs: (a) highly attractive female neuron (Neuron–HAF), (b) unattractive female neuron (Neuron-UAF), (c) highly attractive male neuron (Neuron–HAM), and (d) unattractive male neuron (Neuron-UAM).

Therefore, the generated images showed the learned feature representation of the CSNs for different categories of attractiveness. Notably, Some configurational features observed in the generated images were consistent with the insights of psychophysical research (Marquardt & Stephen, 2002; Rhodes, 2006; Valenzano et al., 2006).

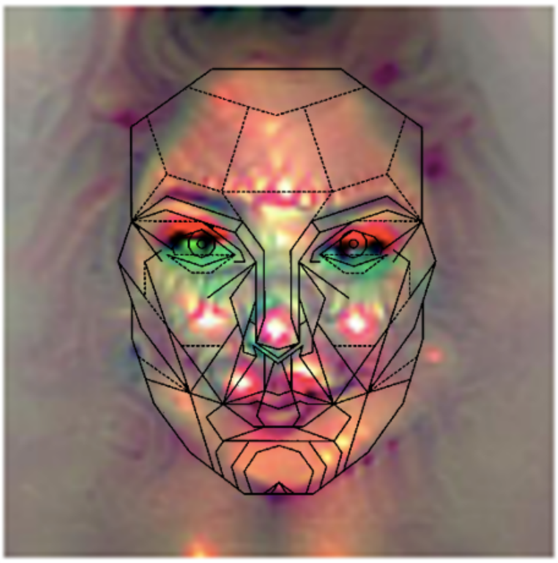


Figure 2.11: The face-like image generated from Neuron-HAF is found covered with golden facial mask (Marquardt & Stephen, 2002).

2.4 Discussion

In this chapter, we proposed FADNN to learn facial attractiveness and investigated the implicit representations that could be learned by FADNN associating with facial attractiveness. In Experiment 1, we built a dataset containing 4512 extremely attractive/unattractive face images and trained the FADNN only using gender and attractiveness (2 × 2) categories. The FADNN used four CSNs as feature representations of facial attractiveness and gender in each category. Comparing with traditional features (Geometric features and Gabor features), the feature representations learned by FADNN showed the best discrimination performance, which achieved accuracy up to 0.981 on classifying attractive/unattractive faces. In Experiment 2, we tested and interpreted the implicit representations learned by FADNN. To test the robustness of the FADNN, we used the image dataset out of the training set to test the FADNN. The experimental results showed that the activation of four CSNs was correlated to the subjective rating scores by human participants ranged between 0.5655 and 0.7583. This suggested that the four CSNs subsume the main features associated with facial categories and learn the intended features to represent different facial attractiveness categories. Furthermore, the CSNs projected the attractive features of different categories onto four face-like images, which make the hidden representations to intuitive representations. We observed that some configurational cues of facial attractiveness exist on the generated images consistent with the suggested cues by previous studies, such as pointed chin and standard arched eyebrows for attractive female (Feser et al., 2007; Valenzano et al., 2006), square chin (Rhodes, 2006) for attractive male and golden facial mask (Marquardt & Stephen, 2002).

This chapter makes the following three contributions. (1) We created a novel dataset with 4,512 face images that belong to two extreme categories of facial attractiveness. It is different from others that either has a small set or include a relatively large proportion of neutral faces. Using the new dataset, FADNN learns features that represent extreme categories, which makes it possible for us to compare the differences between categories from the feature domain. (2) We proposed the FADNN model to recognize facial attractiveness, which directly learns the facial attractiveness representation directly from the data without manual intervention. The manual intervention such as annotated the landmarks aims to use landmarks to drive the model but also constrains the model to only learning the landmarks. However, data-driven DNN provides opportunities to observe the strategy of DNN extracting from the real world images for the task and explore new representations for psychology. (3) We interpreted the feature representation learned by FADNN. The results suggested some visible features associated with the configurational cues of facial attractiveness found in the previous psychological studies.

Chapter 3

Putative Ratios in DNN for Facial Attractiveness

Empirical evidence (Pallett et al., 2010) has shown that there is an ideal arrangement of facial features (ideal ratios) that can optimize the attractiveness of a person's face. These putative ratios describe facial attractiveness in terms of spatial relations and provide important rules for measuring the attractiveness of a face. In Chapter 2, the experimental results suggested that the proposed FADNN learned effective representations for attractiveness categories, including configurational cues of facial attractiveness. In Chapter 3, we examine whether the putative ratios are learned by FADNN, where no annotated facial features for attractiveness are explicitly given. Two experiments are conducted to compare the output of the FADNN with the performance of human cognition to answer this question.

3.1 Introduction

Attractiveness is one of the significant properties of human faces (Frieze et al., 1991; Pashos & Niemitz, 2003). A large body of studies suggested that people implicitly use the same criteria for determining facial attractiveness, although some individual and cross-cultural differences have been noted (detailed information can be found in section 2.1). One criterion of facial attractiveness is measured using ideal ratios (Valenzano et al., 2006), such as neoclassical canons (Jayaratne et al., 2012; Schmid et al., 2008), golden proportions (Borissavliévitch & Hautecœr, 1958; Jefferson, 2004), facial thirds (Farkas & Kolar, 1987; Farkas & Schendel, 1995), and new golden ratios (Bóo et al., 2013; Pallett et al., 2010). These putative rules describe a physical measure for facial attractiveness using spatial relations between parts of the face. For example, the ancient Greeks believed that the golden ratio represents the essence of beauty (Rhodes, 2006). Leonardo da Vinci's painting *Mona Lisa* also illustrates the golden ratio of facial beauty (Atiyeh & Hayek, 2008; Jefferson, 2004).

Empirical evidence provides support for the association between putative ratios and the human perception of facial attractiveness. For example, the golden ratio is an irrational value (0.618) and a commonly accepted metric for measuring the harmoniousness of proportions in aesthetics. The golden ratio has also been proposed as a universal standard of facial attractiveness (Jefferson, 2004), especially as a standard in plastic surgery (Holland, 2008), and has been used to compute facial attractiveness (Schmid et al., 2008). In addition, Pallett et al. (2010) has reported that ideal facial ratios, called the new golden ratios, can optimize the attractiveness of face images. These new golden ratios consist of two ratios: a vertical ratio of the eye–mouth distance to the height of the face is approximately 0.36, and a horizontal ratio of the distance between the eyes to the width of the face is approximately 0.46. Furthermore, H. Shen et al. (2016) found that the facial attractiveness induced by facial proportions evokes a neural response in some regions related to the reward system in the brain, such as the orbitofrontal cortex and amygdala. This suggests that the putative ratios contain high-level information about facial attractiveness.

The rules of putative ratios have also been used to construct computer models for the assessment of the attractiveness of face images (F. Chen & Zhang, 2014; Gunes & Piccardi, 2006; Schmid et al., 2008). Some of these models relied on a set of putative ratios such as the golden proportions and facial thirds as features of the face images and used them to calculate attractiveness scores. For instance, Schmid et al. (2008) used a feature vector comprising 77 putative ratios to build a computer model for facial attractiveness. Results of experiments revealed a significant correlation between the scores predicted facial attractiveness by the computer model and those rated by human subjects. Other approaches (Kagian et al., 2008; Zhang et al., 2017) used only geometric features based on the automatic detection of facial landmarks and all the ratios among them as discussed in Section 2.1. For example, Kagian et al. (2008) used 6,972 distance vectors of 84 facial landmarks as features to predict the attractiveness of faces. They reported that these geometric features implicitly capture the basic human characteristics for the perception of facial attractiveness, such as averageness (Langlois & Roggman, 1990) and symmetry (Rhodes et al., 1999). However, these ratio-based models require the annotation of the facial landmarks, which incurs high labor costs. These landmarks also lack hierarchical, high-level semantic information (S. Wang et al., 2014).

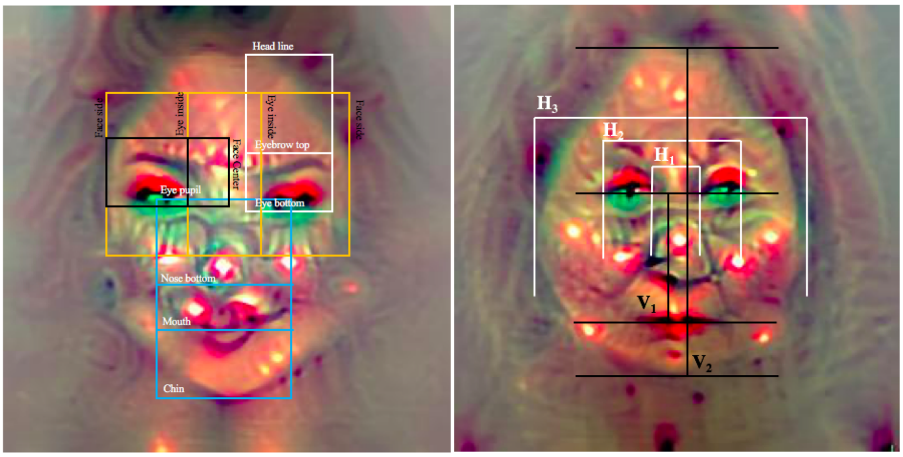
In Chapter 2, we proposed a DNN model for facial attractiveness (FADNN) and found the model captured more effective representation, comparing with handcrafted features. The DNNs learn higher-level features from a large number of face images. Recently, some similarities have been observed between DNNs and biological vision systems in terms of their operation (Cadieu et al., 2014; Cichy et al., 2016; Kietzmann et al., 2019; O'Toole et al., 2018; Seeliger et al., 2018; P. Wang & Cottrell, 2017; Yamins & DiCarlo, 2016). Cichy et al. (2016) showed that the DNN could capture the similar stages of human visual processing in both time and space, from the early visual areas to the dorsal and ventral streams. Seeliger et al. (2018) reported that DNNs could capture hierarchical representations of color for object recognition that rivals the performance of primates. For instance, the middle layers of the DNN show a denser sampling of hue that suggests some correlation with hue maps in V2 (Conway, 2003). Thus, some researchers (Kriegeskorte, 2015; VanRullen, 2017) have proposed using DNNs as a new tool to gain insights into visual perception in human. However, the above studies mainly concentrated on object recognition tasks to explain the association between DNNs and human visual perception. In addition, little research has explored the implicit feature representations learned by DNNs that are associated with high-level perception. For example, the studies of McCurrie et al. (2018) and Parde et al. (2019) are examples in which implicit feature representations learned by DNNs are used for understanding subjective social traits (e.g., trustworthiness, dominance, and anxiety) of human face perception.

In the previous chapter, we visualized the learn representations of FADNN. The generated images suggest that the configurational cues of facial attractiveness might be learned by FADNN. This chapter examines whether the putative ratios have been learned by FADNN, which is based only on categorical annotation, i.e., where no annotated facial features for attractiveness are explicitly given. We conducted two experiments to address the issue of how to evaluate the implicit feature representations in FADNN contain the putative ratios of facial attractiveness[1]. In Experiment 1, we measured the putative ratios on these face-like images, which were generated by the specific neurons representing the features subsume from raw images. In Experiment 2, a "psychophysical-like" experiment was conducted on the trained FADNN to examine whether the responses of the CSNs match the results of human judgments of the putative ratios (e.g., Pallett et al., 2010). The experimental results demonstrated that the FADNN learned putative ratios as key features for the representation of facial attractiveness. This finding advances our understanding of facial attractiveness via DNN-based perspective approaches.

3.2 Experiment 1: Putative Ratios in Neuron Interpretation

The generated face-like images in Chapter 2 contained essential features (such as the mouth, eyes, and chin), thus we can assess the information using FADNN on the basis of physical measures of facial attractiveness such as the golden ratio. In Experiment 1, we employed the putative ratios of facial attractiveness, including the golden ratio and new golden ratios, to quantify the generated images, estimating the association between the generated images and human perception of facial attractiveness. The rules of the golden ratio can be used to measure ratios between facial elements to obtain the local putative ratios of facial attractiveness (Holland, 2008; Jefferson, 2004), whereas the new golden ratios give global features of these putative ratios (Bóo et al., 2013; Pallett et al., 2010).

[1] It should be noted here that we examined feature representations of faces only based on photos. These do not necessarily cover all inherent features of individual faces.



(a) (b)

Figure 3.1: Measurement of the putative ratios on the generated face-like images. The boxes and lines represent the rules of the golden ratio and new golden ratios, respectively. A detailed description of the measurement of the rules of the golden ratio is provided in Table 3.1.

3.2.1 Methods

The four generated face-like images show mainly facial features. The observation results illustrated different visual features of the four categories. For instance, the generated image for Neuron-HAF has a pointed chin and standard arched eyebrows. Here, we calculated ratios between landmarks, which suppose to represent the putative ratios (i.e., the golden ratio and new golden ratios), for the generated images to provide quantitative measures of facial attractiveness. To measure the rules of the golden ratio for facial attractiveness, the six ratios (r_m) described in Table 3.1 were measured, as shown by the boxes in Figures 3.1a. In addition, the new golden ratios of facial attractiveness included a horizontal and a vertical ratio, mathematically defined as $R_h = \frac{0.5(H_1+H_2)}{H_3}$ and $R_v = \frac{V_1}{V_2}$, respectively. The measurements of H_1, H_2, and H_3, and V_1 and V_2 are shown in Figures 3.1b.

3.2.2 Results and discussion

Tables 3.2 and 3.3 show the results of the measurements according to the rules of the golden and new golden ratios, respectively, for the generated images. As

Table 3.1: Rules of the golden ratio for attractive faces (Schmid et al., 2008). The color names in brackets represent the colors of the rectangles in Figures 3.1a. Here, r_m represents the m-th ratio for an attractive face with the golden ratio.

No.	Description of Ratio
r_1	Distance between the top and bottom of eyebrows to that between eyebrows and hairline (White)
r_2	Distance between the mouth and bottom of nose to that between bottom of nose and pupil of eye (Blue)
r_3	Distance between the pupil of eye and bottom of nose to that between bottom of nose and chin (Blue)
r_4	Distance between the chin and mouth to that between mouth and pupil of eye (Blue)
r_5	Distance between the side of face side and inside of eye to that between inside of eye and side of face (Golden)
r_6	Distance between the center of nose and inside of eye to that between inside of and outside of eye (Black)

shown in Table 3.2, the numbers in the brackets describe the absolute error between the golden ratio and the measured values. These five ratios (except for r_5) of the generated image for Neuron-HAF are closer to the golden ratio than those of the image generated for Neuron-UAF. An analogous difference was found between the generated images for Neuron-HAM and Neuron-UAM. The mean absolute error (MAE), which reflects the difference between the measured ratios and the golden ratio. Table 3.2 shows that the generated image for Neuron-HAF has a lower MAE (0.066) than that generated for Neuron-UAF (0.183). Thus, the image for Neuron-HAF is closer to the putative ratios of facial attractiveness according to the rules of the golden ratio. This tendency also holds for the image of Neuron-HAM (0.175), which has rations closer to the golden ratio than does Neuron-UAM (0.258).

Table 3.3 shows that the values of R_v of Neuron-HAF and Neuron-HAM are close to the standard value (0.36) of the new golden ratio. Both R_h and R_v for the image for Neuron-HAM (0.470, 0.354) are close to the new golden ratios (0.46, 0.36). An analogous consistency was observed in the image of Neuron-HAF (0.494, 0.357). However, neither of these consistencies appears in the images for Neuron-UAF and Neuron-UAM. For instance, the image for Neuron-UAF has ratios of (0.339, 0.393),

Table 3.2: Measured ratios for generated images for different CSNs. The values in the brackets indicate the absolute error between the measured values and the golden ratio. MAE represents the mean absolute error.

Golden Ratio	r_1	r_2	r_3	r_4	r_5	r_6	MAE
Neuron-HAF	0.681 (0.063)	0.536 (0.082)	0.761 (0.143)	0.619 (0.001)	0.585 (0.033)	0.694 (0.076)	0.066
Neuron-UAF	0.755 (0.137)	0.416 (0.202)	1.032 (0.415)	0.551 (0.067)	0.639 (0.021)	0.360 (0.258)	0.183
Neuron-HAM	0.346 (0.272)	0.381 (0.237)	0.660 (0.042)	0.426 (0.192)	0.668 (0.050)	0.363 (0.255)	0.175
Neuron-UAM	1.020 (0.402)	0.907 (0.289)	0.379 (0.239)	1.094 (0.476)	0.530 (0.088)	0.669 (0.051)	0.258

Table 3.3: Results of the new golden ratios for generated images for the CSNs. Here, R_h and R_v respectively represent the horizontal and vertical ratios defined by the new golden ratios for attractive faces. The values in the brackets indicate the absolute error between the measured values and the new golden ratios.

New Golden Ratio	$R_h(0.46)$	$R_v(0.36)$
Neuron-HAF	0.494 (0.034)	0.357 (0.003)
Neuron-UAF	0.339(0.121)	0.393(0.033)
Neuron-HAM	0.470(0.010)	0.354(0.006)
Neuron-UAM	0.415(0.045)	0.367(0.007)

which are far from the new golden ratios. In addition, the R_v of Neuron-HAM (0.354) is close to the standard value. However, the standard for new golden ratio is supposed to have a good fit for the combination of R_h and R_v, no matter which one measurement (R_h or R_v) doesn't fit standard value would be considered as lower attractiveness to standard face (Pallett et al., 2010).

The quantitative result showed that the images for Neuron-HAF and Neuron-HAM contain arrangements of facial parts close to the golden ratio and new golden ratios. However, we did not explicitly present these shape features and rules of the putative ratios in the learning process. Therefore, this result suggests that the rules of the putative ratios were learned by FADNN without any cues.

3.3 Experiment 2: Putative Ratios in Neuron Activation

Experiment 1 showed that the rules of putative ratios might be encoded in feature representations of four CSNs. To further explore the association between the putative ratios and the CSNs, we conducted a psychophysical-like experiment for FADNN. In a psychophysical experiment with human observers, Pallett et al. (2010) distorted the original face using different horizontal and vertical ratios, and found that faces with the optimal ratios (i.e., new golden ratios) rated as the highest facial attractiveness. Similarly, we distorted the horizontal and vertical ratios of faces while keeping other factors constant, and used these faces as stimuli for FADNN. Figure 3.2 shows examples of facial stimuli with different degrees of distortion. The test stimuli were made by distorting the face images from a new dataset. When the modified face images were fed into FADNN, if it had learned the golden ratio as feature for attractiveness, its CSNs would be expected to generate lower attractiveness scores for distorted face images.

3.3.1 Methods

3.3.1.1 Method of distorting face shapes

To obtain the stimuli of the distorted faces, we used a computer vision method to manipulate the horizontal and vertical proportions of original images. First, facial landmarks were relocated using the active shape model. Second, new landmarks were computed by simultaneously increasing the eye–mouth distance and the distance between the eyes to a certain extent. Third, Delaunay triangulation (W. Hong et al., 2015) was employed to distort the faces in face images according to the new landmarks.

For each original face image, the eye–mouth distance and the distance between the eyes were simultaneously expanded by 4%, 8%, 12%, and 16% compared with those in the original image. The examples of the expanded faces are shown in

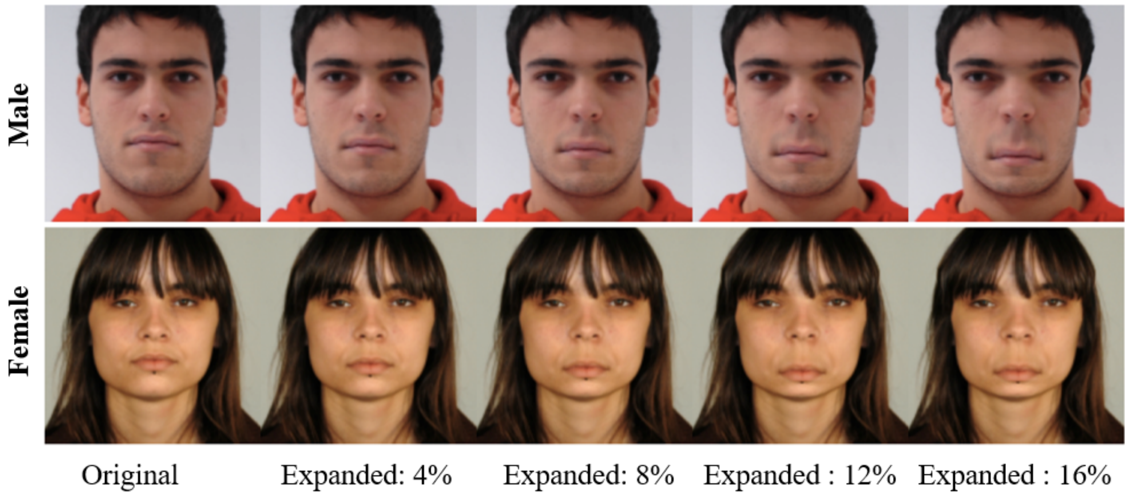


Figure 3.2: Examples of face stimuli. Original face image and four degrees of expanded face images.

Figure 3.2. In addition, we contracted the distance between the eyes by 16%, as shown in Figure 3.3. We selected images from two public face image datasets: colorFERET [2] and GTAV (Tarrés, 2012). After removing low-quality images, 887 images (527 males and 360 females) were selected. In addition, we measured the R_h and R_v (described in Experiment 2) for every original image. The mean values and standard deviations of the R_h and R_v of the female faces are 0.465 ± 0.024 and 0.358 ± 0.020, while those of the R_h and R_v of the male faces are 0.465 ± 0.026 and 0.365 ± 0.020, respectively. Both were approximated as 0.46 and 0.36, respectively, which is analogously consistent with the new golden ratios. Each image was cropped and resized to 224×224 pixels.

Although the ColorFERET dataset consists of 994 face images, some of them were images of the same person. We chose one image per person, resulting in 504 images of males and 347 images of females. The GTAV face dataset comprises 44 face images. Again, by discarding multiple images of the same person, 23 images of males and 13 of females were obtained. In total, 887 face images were used in this experiment.

[2] http://face.nist.gov/colorferet/request.html

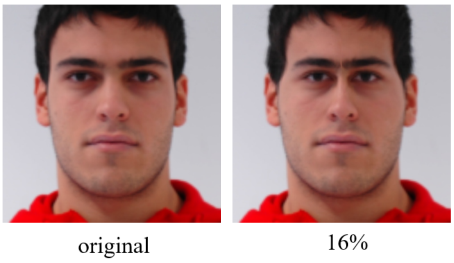


Figure 3.3: Examples of original and contracted face stimuli.

3.3.1.2 Experimental procedure

The face stimuli were input into FADNN. For each image, the model assigned scores (activations) for each of the four CSNs. The scores were used as metrics to evaluate the extent to which an input image belongs to one of the experimental categories. The expanded stimuli consisted of five levels of distortion: the original photograph and four levels of expansion (887 original images: 527 males and 360 females). The contracted stimuli include two levels, the original photograph and a level of contraction. The dependent variables were scores in the CSNs. The results were analyzed using the SPSS 25 statistical software.

3.3.2 Results and discussion

Figure 3.4 shows the scores (mean values and standard deviations, SDs) of the four CSNs for different degrees of expansion. As the faces were expanded, both the scores of Neuron-HAF and Neuron-HAM decreased whereas those of Neuron-UAF and Neuron-UAM increased.

For the original faces, the paired t-test was used to compare the scores of Neuron-HAF with those of Neuron-UAF, as well as those of Neuron-HAM with those of Neuron-UAM. For male faces, there was no significant difference between either pair — Neuron-HAF vs. Neuron-UAF: $t(526) = 1.385$, p=0.167; Neuron-HAM vs. Neuron-UAM: $t(526) = 1.522$, p=0.129. This indicates that the original images of the male faces are close to neutral faces, and belong to neither HAM nor UAM. For female faces, there were significant differences in both pairs — Neuron-HAF vs. Neuron-UAF: $t(359)$ = -7.587, $p < 10^{-12}$; Neuron-HAM vs. Neuron-UAM: $t(359)$

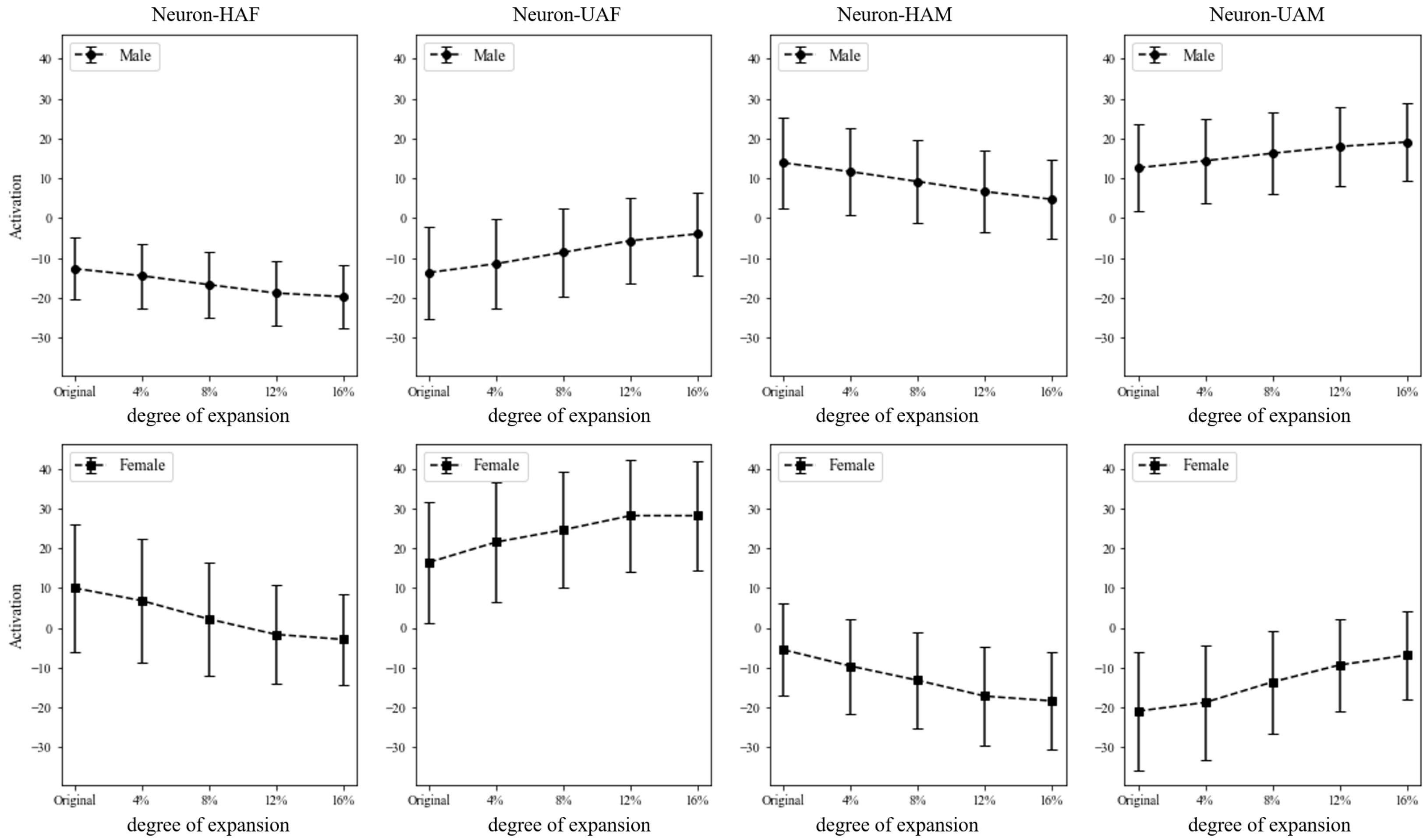


Figure 3.4: Scores (mean values and standard deviations) of the four CSNs for different degrees of expansion. The top row illustrates the neuron scores (showed by four dependent sub-figures) of 2,635(5 levels distortion of 527 faces) male facial stimuli. The bottom row illustrates the neuron scores of 1,800 female facial stimuli.

$= 13.483$, $p < 10^{-33}$. The scores for Neuron-UAF were 10.44 higher than those for Neuron-HAF on average, which indicates that the original images of females were mostly predicted to be unattractive.

To examine the effects of the degree of expansion, an analysis of variance (ANOVA, degree of expansion × gender) was used independently for each CSN, . The main effects of the degree of expansion were significant in all neurons — $F(4, 4425) = 126.815$, 102.354, 132.332, and 116.690, respectively; p<0.00001. Multiple comparisons (Bonferroni) of the degree of expansion were significantly different in all neurons for the original photographs, and those expanded by 4%, 8%, and 12% (all p<0.00001). In particular, in a comparison of images expanded by 12% and 16%, the activations of Neuron-HAM and Neuron-UAM changed significantly — p=0.011 and p=0.029, respectively, whereas the changes in Neuron-HAF and Neuron-UAF were more moderate — p=0.624, and p=0.720, respectively. The scores of Neuron-HAF and Neuron-HAM decreased when the faces expanded, whereas those of Neuron-UAF and Neuron-UAM increased, indicating that FADNN is more likely to recognize a face as unattractive when its measures are more than the putative ratios of facial attractiveness.

In addition, the interaction (degree of expansion × gender) is significant in Neuron-HAF, Neuron-HAM, and Neuron-UAM — $F(4, 4425)$=11.405, $p < 10^{-8}$, $F(4, 4425) = 3.935$, p=0.003, and $F(4, 4425) = 17.331$, $p < 10^{-13}$, respectively. However, in Neuron-UAF, the interaction (degree of expansion × gender) is only marginally significant — $F(4, 4425)$=2.113, p=0.077.

Figure 3.5 compares the scores of the four CSNs for contracted stimuli with those of the original photographs. For male stimuli, the scores of Neuron-HAM, Neuron-HAF, and Neuron-UAF change significantly — $t(526) = 3.322$, 13.231 and -11.698, respectively; all p<0.00001. Specifically, the contracted stimuli obtain lower Neuron-HAM scores than do the original photographs, which means the confidence that the image was of an attractive male decreased as the distance between the eyes contracted. For the female stimuli, the scores of Neuron-HAF, Neuron-UAF, Neuron-HAM, and Neuron-UAM change significantly — $t(359) = 15.780$, -6.953, -3.662,

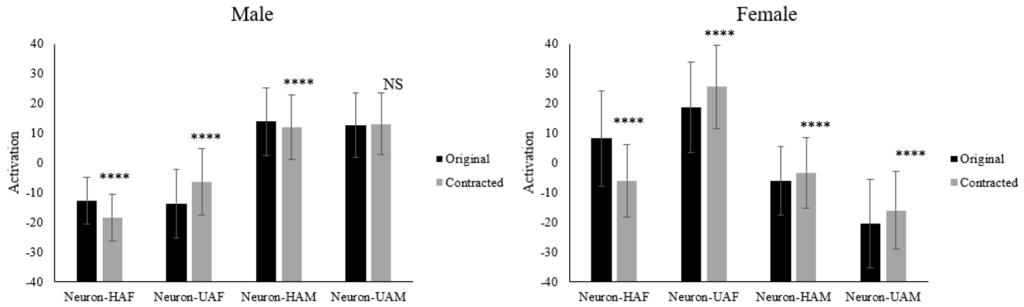


Figure 3.5: Scores of the four CSNs obtained by testing the contracted stimuli compared with the original photographs. The left figure illustrates the neuron scores of 527 male photographs and 527 contracted male stimuli. The right figure illustrates the neuron scores of 360 female photographs and 360 contracted female stimuli. NS: $p > 0.05$; ****: $p < 0.00001$.

and -4.892 respectively; all p<0.00001. Analogously, the scores of Neuron-HAF decrease and the scores of Neuron-UAF increase when the distances between eyes are contracted. These results show that FADNN recognizes a face as unattractive when its measures are less than the putative ratios of facial attractiveness.

Pallett et al. (2010) reported that human judgment of attractiveness decreases when the facial arrangement strays from the new golden ratios. The above experiment demonstrated a similar trend in the attractiveness determination of FADNN, although it was not explicitly trained using putative ratios. This result confirms that the putative ratios were encoded into FADNN. Moreover, the new golden ratios had a significantly stronger effect for female faces than male faces, which suggests that rule of the putative ratios impacts judgment of faces of females more than those of males in FADNN. This suggest that the female faces may more sensitive than male faces to putative ratios for human facial attractiveness perception. Although some psychophysical studies of putative ratios with human observers have shown that the new golden ratios can be employed to judge both female and male faces (Bóo et al., 2013; Yoo et al., 2013), to the best of our knowledge, the difference in size effect between female and male face images has not been studied. This suggestion can be verified in a further rigorous psychological experiment by the researchers who are interested in the functional significance or origin of attractiveness perception.

3.4 Discussion

In Chapter 2, we trained FADNN only using gender and attractiveness (2 × 2) categories, and it accurately classified images into these categories of facial attractiveness. The FADNN used four CSNs represented the features of facial attractiveness in each category. The CSNs projected the facially attractive features onto four face-like images, demonstrating the configurational cues of facial attractiveness. In addition to the suggestions, this chapter offered a demonstration that features learned by a FADNN include the putative ratios while only the attractiveness category is given as a supervisory signal. In Experiment 1, we found that the generated face-like images representing high attractiveness contain the putative ratios whereas images representing low attractiveness do not through quantitatively analyses. To validate this finding, we designed a psychophysical-like experiment for FADNN in Experiment 2, where face images were distorted away from the new golden ratios. The scores given by the four CSNs showed tendencies that are similar to those observed in human judgments of facial attractiveness in Pallett et al. (2010). These results suggest that the rules of putative ratios are learned as the latent distinctive features that represent facial attractive categories in FADNN.

The FADNN learned the putative ratios for attractiveness judgment similar to the human perception of attractiveness for two reasons. First, attractiveness was annotated by human participants in the training images. The rules of the putative ratios are shared criteria for the perception of facial attractiveness across humans and across participants in this study. Thus, FADNN might learn such shared criteria of attractiveness. Second, the DNN has the capacity to learn abstract features from raw inputs (A. Nguyen et al., 2016). For example, Zeiler and Fergus (2014) found that, in object recognition, the initial layers of the DNN capture low-level features (i.e., color, edges, and corners) whereas subsequent layers capture high-level object-related representations. In particular, the last layer captures the representation of the entire object. The rules of the putative ratios relate to a high-level representation of facial attractiveness (H. Shen et al., 2016). Thus, it is natural to argue that FADNN can learn high-level representations similar to these rules in the human perception of attractiveness.

There is a limitation in the dataset. It is composed of non-standardized images and the (un)attractive categories were collected from two sources (celebrity and mugshots) even though we excluded low-quality images before the training phase. These may introduce confounding factors to FADNN, because recent studies (Colón et al., 2020; M. Q. Hill et al., 2019) showed that DNN retains certain features irrelevant to the tasks (annotations). Thus, the FADNN may learn some shortcut features, which could influence the results of this study. For example, the quality of makeup in the face images of attractive females might have been different from those of unattractive females. To alleviate learning shortcut features from each source, we applied transfer learning on VGG-face model, specifically, fine-tune FADNN based on the high-level representation of face identity. Then, we tested FADNN on the face images out of the distribution of training data in Chapters 2 and 3, and importantly, a highly controlled dataset with varied putative ratios was used in Chapter 3. Under such complicated conditions, the results still suggest that FADNN has learned putative ratios from the dataset. However, it does not mean that the attractive scores are solely determined by the putative ratios. The other features (e.g., symmetry and sexual dimorphism) that have been learned by FADNN require further investigation.

Another limitation is the FADNN, which is trained for attractiveness classification. Suppose we can train a DNN model for attractiveness rating and show some similarities to human facial attractiveness perception. In that case, the DNN model might be used to derive humans in a more general way. However, to train this model is still an open question for informatics study. The difficulty is collect enough data (facial images and attractiveness rating) to train the model. The informatics research always focuses on rating prediction based on data-driven results while providing little insight into human perception. In addition, it is still difficult to know how much FADNN can be extended for studying human facial attractiveness perception. To know this requires knowing more about the DNN. However, the

DNNs are always considered as the black boxes, since the interpretation of DNNs is an underdetermined problem. In the future, we will focus on the relationship or difference between human perception and DNN perception. However, it may require more technological innovation for the interpretation of DNNs to be possible.

Many social factors that can be abstracted in human faces, such as first impressions (Vernon et al., 2014), health (Russell et al., 2016), and sexual orientation (Y. Wang & Kosinski, 2018), which are otherwise difficult to measure. The neural network interpretation (the generation of face-like images) and neuronal selections with known links to such factors and psychological traits can provide an avenue for generating hypotheses that can be verified in experimental studies. For example, in Experiment 2, we found that the new golden ratios impact female face in images more than those of males in FADNN. However, there is no empirical evidence for this in human cognition studies, even with the new golden ratios employed for both female and male face images (Bóo et al., 2013; Yoo et al., 2013). Further research is required to clarify this issue.

3.5 Summary of This Part

It has been suggested that some machine learning models can implicitly capture basic human psychological characteristics, such as averageness, symmetry (Kagian et al., 2008), and sexual dimorphism (Said & Todorov, 2011). However, they are based on either handcrafted features or artificial stimuli. For example, the input facial stimuli in (Said & Todorov, 2011) were generated by specific features (50 dimensions for shape and 50 dimensions for reflectance), which were well controlled. In this part, we showed that FADNN could learn the putative ratios without knowing any prior features annotated in the face images. This result opens up opportunities for elucidating the origins of facial attractiveness, from basic perceptual mechanisms to high-level human perception, via a DNN-based perspective approach. Moreover, the discussion in this part advances the interpretation of DNN, which uncovers the links between complex stimuli and specific tasks and demonstrates the potential of our framework shown in 1.3a, which leverages the representations learned by DNN and combines them with psychological theories to shed light on understanding human cognition.

Part II

Understanding Tourist Satisfaction Through Photo-shooting Behaviors

Because this study (including Chapter 4 and Chapter 5) refers to multidisciplinary, such as psychology, sociology (tourism), and informatics, we hope to clarify the related works about visual attention, photo shooting behavior, and affection and their relationships in related fields before going to the main works about this study.

It is crucially essential for humans to adjust the scopes of visual attention. A fundamental distinction is whether they attend to the gestalt of a stimulus or focus on the specific details of a stimulus, i.e., broad vs. narrow attention (Navon, 1977). Generally, broadening attention refers to processing information in a more big-picture way and seeing the forest, whereas narrowing attention refers to processing information in a more detail-oriented way and seeing the trees (Förster, 2012; Navon, 1977). Narrower attention can improve the perception of spatial detail (Goodhew et al., 2017), such as detecting the presence of a small spatial gap in a ring (Goodhew, 2020). In contrast, broader attention can improve the visual process of those invoked in the opening of selection, such as scanning a crowd to locate a friend (Gao et al., 2011; Macrae & Lewis, 2002) and the processing of summary scene statistics (Chong & Treisman, 2005).

A plethora of psychological studies (Fredrickson, 2004; Ji et al., 2019; Kuhbandner et al., 2011; Tamir & Robinson, 2007) has documented that visual attention moderates emotional states, whereas emotional states can influence the level of attentional scopes. Specifically, broadening attention has an affective advantage over narrowing attention. The representative theory for describing this human's characteristic is the broaden-and-build theory in positive psychology (Fredrickson, 2004; Fredrickson & Branigan, 2005). According to Fredrickson, there is a causal relationship between emotion and attentional scopes. On the one hand, positive emotions can broaden the individual attentional scope, while negative emotions shrink it (Gasper & Clore, 2002). For example, when participants are asked to attend to a central targets, induction of positive emotion results in more significant to precept irrelevant surroundings than negative emotion (Fenske & Raymond, 2006; Rowe et al., 2007; Schmitz et al., 2009). That is, positive emotion leads to the processing of a wider area even if the distractors appear in the area. On the other hand, broadening attention can induce participants to improve positive emotions or reduce negative emotions, while narrowing attention can reduce positive emotions or induce more negative emotions (Gu et al., 2017; Ji et al., 2019). For example, Gu et al. (2017) demonstrated that, the depressed participants (n=44) reduced negative emotion, and increased the positive emotion significantly after watched distant scenes (i.e., broadening training) for eight weeks, comparing with watching proximal scenes (i.e., narrowing training).

The intrinsic factor of the relationship between visual attention and emotion may also influence human social behaviors. In this study, we focus on the shooting behavior of travel photos. Because tourists often take photos to capture their gaze, and to record the selections in travel experience (Osborne, 2000; Urry, 1992). The travel photos can be the notations that reflect the photographers' feelings (Pan et al., 2014). According to the tourism research (Ittelson, 1976; Pan et al., 2014), when travelers experience the sightseeing environments, affect would response firstly, then the quality of the affect would governs the following behaviors toward these environments. In other words, the natural beauty stimulates tourists to the emotional moment, which then directs the photo-shooting behaviors to record the moment. Specifically, Pan et al. (2014) found that travel photos featured natural resources or taken in a distant view may related to arousing and pleasant feelings toward the sightseeing places. Thus, there may be an internal relationship that the tourists' emotions are influenced by natural resources, leading the specific photo-shooting behaviors (e.g., shooting a distant view).

Photo-shooting behaviors may relate to the social cognitive theory (Bandura, 2001; Lewin, 1951; Pan et al., 2014), which connect the human behaviors, environment, and person's affect with a triadic reciprocal causation. A travel photo is the product of travelers' behavior, which may be influenced by this theory. In tourism research (Lawson & Baud-Bovy, 1977; Pan et al., 2014), travel photos are the condensation of a tourism destination image — "the expression of all objective knowledge, impressions, prejudices, imaginations, and emotional thoughts with which a person or a group judges a particular object or place". The action of photo-shooting is a process to record their gaze and visual attention (Urry, 1992). Behind the camera, tourists gaze in the beautiful sightseeing, which may arouse momentary feelings. This is like the movie directors who use the landscape to reflect the feelings of characters (Elsaesser & Buckland, 2002), tourists may catch the landscape to represent their inner feelings. Thus, the landscape can transform the affective feelings and vice versa. Similarly, there are two basic approaches (Jacobsen, 2007) in research related to landscape aesthetic estimation: one is to measure the beauty based on the actual setting of landscape; the other is to assume the beauty of the landscape is in the eyes of the tourists. These factors (e.g., inherent factors and tourist gaze) may be recorded by travel photos.

Although some recent informatics research (Bartie & Mackaness, 2016; Zhuang et al., 2014) focused on the triadic reciprocal relationships (photo-shooting behavior, affect, and environmental characteristics), they mainly examined the relationship between the photo-shooting behaviors (related to photographers' attention) and the environmental characteristics. For example, Bartie and Mackaness (2016) suggested that visual exposure could be used to measure the scene aesthetic. The visual exposure is defined to measure the field-of-view of the photographer occupied by an object (Winter, 2003). Zhuang et al. (2014) applied the distinction of close-up view and distant view as an important cue of their sightseeing discovery system. These works applied factors in the viewing distance to measure the scenes' quality but lacked consideration of the internal factors of photographers. To our knowledge, work applying the photo-shooting behaviors in explore internal factors of photographers has been relatively limited; we only are aware of Cao and O'Halloran (2015), who used the viewing distances detected from travel photos to explore insight into personal characteristics. They suggested that tourists prefer to take more distant views (lakes, wide landscape, and mountains) to remind the travel experience in unfamiliar places, while residents use more close-up views (food, pets, and humans) to catch the elements related to their daily life.

Considering the bidirectional emotion-attention relationship, interesting question can be raised. Does the viewing distance or angle in photos reveals the emotional states of photographers, or is the tendency of the field of view of photos influenced by broaden-and-build theory? Currently, social media platforms collected a large number of experiences and photos from many tourists and various tourism destinations. The star of the experience rating can be used to represent the tourist satisfaction or sentiment (Alaei et al., 2019; Broß, 2013). In addition, the contents of travel photos can reflect attentional scopes (wide-view vs. narrow view). Thus, we assume this relationship between the attentional scopes detected from travel photos and tourist rating of satisfaction is one of the real-world examples of broaden-and-build theory. Thus, we intend to construct the NODS with photos expressed attentional scopes and tourist satisfaction for complementing the broaden-and-build theory. To construct the NODS, computer vision is a reasonable area to be engaged, thus triggering the image view-type classification problem, which determines the photo-shooting behaviors of a single photo as the research aim of Chapter 4. By using the proposed computer vision algorithm, we examine photo-shooting behaviors by analyzing a large number of travel photos and the association between tourists' satisfactions and their photo-shooting behaviors in Chapter 5.

Chapter 4

Wide or Narrow: View-type Classification for Travel Photos

In Chapter 4, our goal is to examine the issue of detecting the photographers' attentional scopes (wide-view vs. narrow-view) from travel photos, namely view-type classification.

4.1 Introduction

The view-type of photos (wide-view vs. narrow-view) is defined as a measure of the field of view of the photographer, i.e., what the photographer is seeing, or how much of the given scene is seen by the photographer. As shown in Figure 4.1, a wide-view photo expresses the photographer sees a distant area or more of scenes, while a narrow-view photo expresses the photographer catches a close-up area or less of scenes. The view-type of photos has been used to manipulate the participants' attentional scopes in a psychological study (Gu et al., 2017). Concretely, Gu et al. (2017) used the zoomed-in scenes (the photo transited from a wide-view to a narrow-view) and zoom-out scenes (the photo transited from a narrow-view to a wide-view) as broadened and narrowed attention task, respectively. The study demonstrated that the participants in the broadened condition would significantly relieve the negative emotion, which induced by before task, while the ones in narrowed condition increase the negative emotion.

We aim to clarify the concept of view-type photos and their distinction with the depth of view (e.g., close-up and distant photos) in computer vision and the angle of

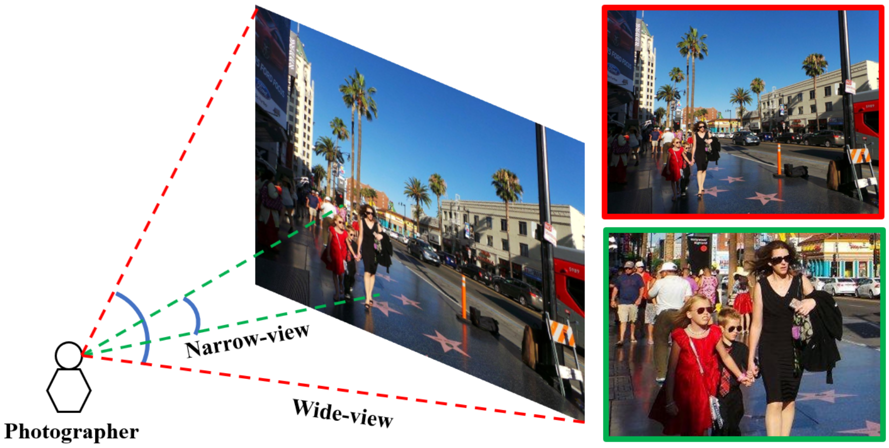
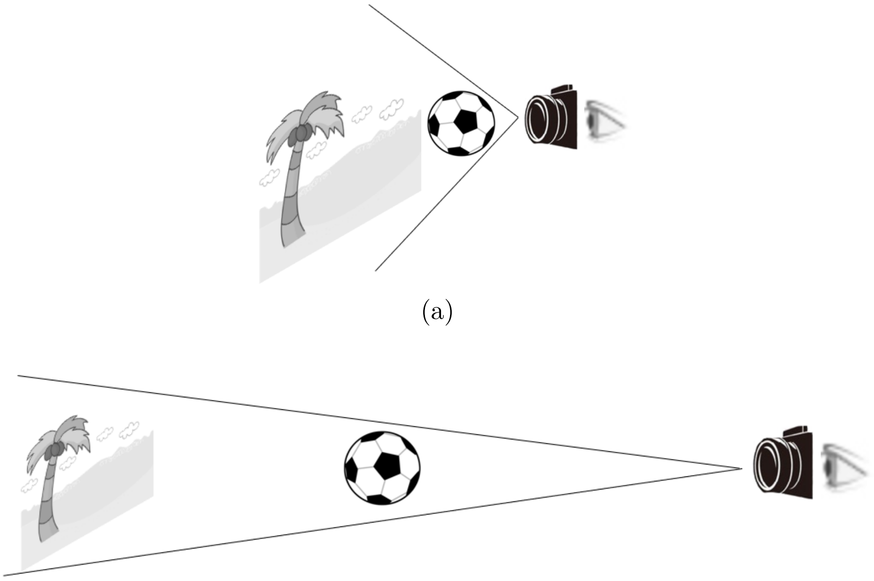


Figure 4.1: View-type perspectives of a photo (wide-view denoted by the red lines, and narrow-view denoted by green lines). Narrow-view only records two pedestrians, while the wide-view records the pedestrians and the entire street-view surrounding them. The image is downloaded from https://www.flickr.com/.

view (e.g., wide- and narrow-angle) in digital photography. Depth of view is defined as "the absolute distance between the photographer and a scene" (Torralba & Oliva, 2002). In the computer vision community, the depth estimation or reconstruction of an image is a popular research topic (Cao & O'Halloran, 2015; Torralba & Oliva, 2002). Most depth estimation algorithms required multiple images, such as binocular vision (Garcia & Solanas, 2004). Because the cues of absolute depth measurements (e.g., binocular disparity, motion, and defocus) are absent, monocular depth estimation is challenging. One possible solution is to measure the structure of images (Torralba & Oliva, 2002). For large-scale depth estimation, Cao and O'Halloran (2015) simplified the depth estimation as an image classification problem, i.e., classifying far-distant and close-up photos according to the absolute depth. They proposed an image classification algorithm based on texture features (Gabor features). This algorithm only took a little time to classify these two specific photo-shooting behaviors and was used for large-scale photo analysis. In addition, the angle of view is defined as the ways to use the camera lens, i.e., zoom-in corresponds to the narrow-angle of view, and zoom-out corresponds to the wide-angle of view [1]. The skilled use of the angle of view mainly focuses on improving the quality of

[1] https://inst.eecs.berkeley.edu//~ee198-4/fa07/week11_assignment.shtml



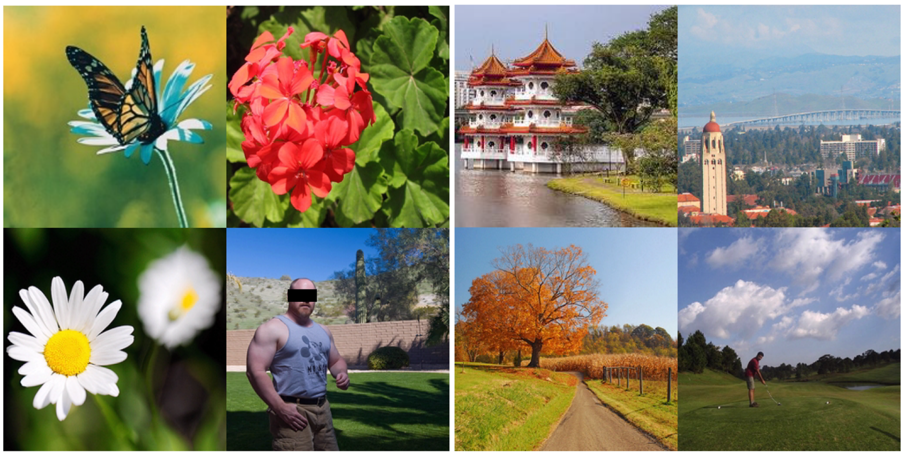
(a)

(b)

Figure 4.2: The distinction between view-type, depth of view, and angle of view. Shooting senses apply (a) wide-angle, or (b) narrow-angle. The depth of view represents the the absolute distance between the photographer (eye) and a scene. View-type represents the field of view of the photographer.

the taken photo. For example, when the angle of view and distance is adjusted, it will result in different photo quality such as perspective and restoration effects to the same object. Specifically, when the object is very close to the camera, a wide-angle would distort the object but restorative its three-dimensional information, while a narrow-angle with suitable distance would reflect the actual two-dimensional information of the target object. Because the camera is in the middle of the eye and the shooting scene, the view-type does not solely depend on angle of view or depth of view, while it may be related to the combined effects of them. As shown in Figure 4.2, the photographer may attend to football in a wide-angle with a close-up view (As shown in Figure 4.2a) or a narrow-angle with a distant view (as shown in Figure 4.2b), while both of them result in a narrow-view.

View-type classification is a relatively new and challenging problem in the field of computer vision since estimating photographers' attentional scope is *a subjective property of photography*. Different from the objective tasks (e.g., face recognition and object detection), it's difficult for algorithms to find specific patterns for this



(a) (b)

Figure 4.3: Image samples, (a) narrow-view photos; (b) wide-view photos. The narrow-view or wide-view photos contain a variety of objects, locking specific patterns for each category.

subjective task, because the natural images contain a wide variety of contexts. As shown in Figure 4.3, similar objects may exist in both narrow- and wide-view photos, e.g., plants and humans. For view-type classification, we conducted two experiments by proposing two different classification algorithms. In Experiment 1, we present a classification framework inspired by the human visual system (HVS). We propose two cues that can represent the HVS, i.e. *focus cue* and *scale cue*. The focus cue is modeled in the frequency domain. The scale cue is modeled by defining the spatial size and conceptual sizes of the main objects in the photo. A newly established dataset with 5050 natural images annotated by humans is used to evaluate the proposed framework. The experimental results show that the proposed framework is better than related algorithms used for depth estimation and close-up/distant view classification. However, we note that the proposed framework contains two drawbacks. Firstly, there are some ambiguous photos between narrow-view and wide-view, such as the photo contains narrow-view objects during a wide-view background. These photos may not be well detected by the proposed focus cue and scale cue. Secondly, the framework requires expensive computing resources since it is implemented on the CPU.

Thus, we propose an end-to-end CNN model for view-type classification, namely CFCNN, in Experiment 2. A larger dataset with 43,906 travel photos is established for CFCNN training. In addition, a dataset with 3,000 images (labeled as wide-view and narrow-view by five participants) has different view-type levels for testing CFCNN and the performance for photos with various ambiguous levels. The results show that the proposed CFCNN outperforms the framework proposed in Experiment 1 for classification accuracy and time-efficiency. In addition, the classification results are similar to the human cognition for ambiguous photos. Therefore, the CFCNN can get a considerable performance for estimating the subjective attentional scopes of large-scale travel photos.

4.2 Experiment 1

In Experiment 1, we proposed a framework based on the focus cue and the scale cue for view-type classification. The focus cue and the scale cue are inspired by the human visual system because the view-type classification task is trivial for the human vision that involves sufficient and robust features learned by effective mechanisms. According to the visual attention mechanisms, the broaden and narrow attention would stimulate a trade-off between the area of focus and the resolution of this focus (Eriksen & James, 1986). Specifically, with narrowing attention, only a small region of a scene is visible in sharp detail, while with expanding attention, a wider field of view becomes visible at the expense of perceptual resolution (Goodhew, 2020). This process has been likened to a zoom lens of the camera (Eriksen & James, 1986), i.e., the limited focus range has been replicated by modern imaging systems, particularly professional cameras. In addition, human vision can also estimate the sizes of familiar objects like the human body, trees, and buildings. Therefore, we defined two cues for modeling the human visual system mechanisms for view-type classification, as shown in Figure 4.4. (1) *Focus cue*: Some narrow-view images tend to focus on objects (in the focal point as shown in Figure 4.4a) at a short distance, while the areas surrounding the object are out of focus and appear to be blurred

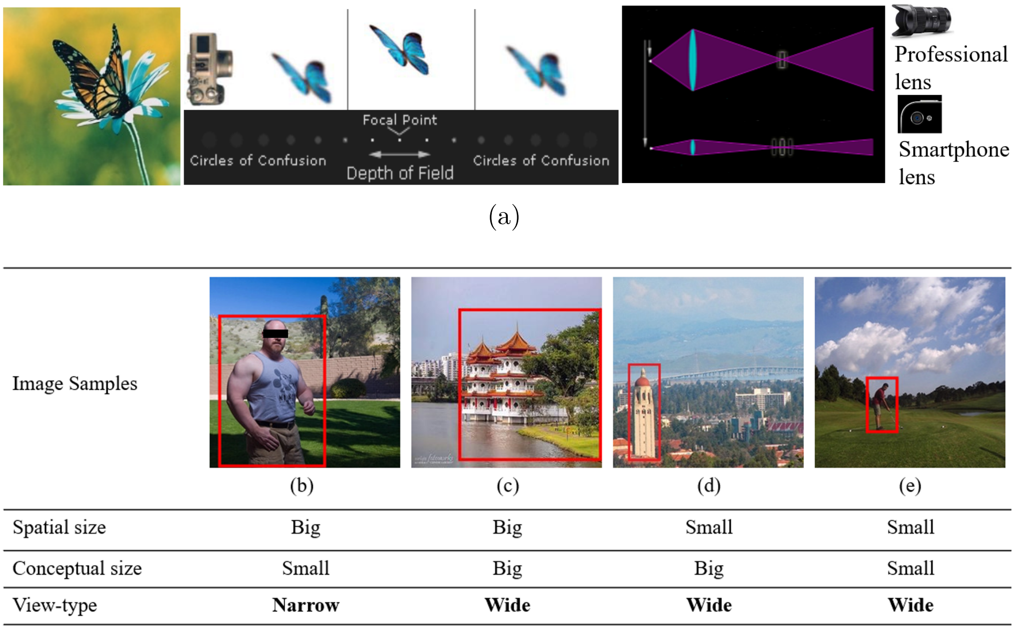


Image Samples	(b)	(c)	(d)	(e)
Spatial size	Big	Big	Small	Small
Conceptual size	Small	Big	Big	Small
View-type	**Narrow**	**Wide**	**Wide**	**Wide**

Figure 4.4: Focus cue and scale cue for view-type classification, (a) focus cue: focal point caused by the optical imaging for narrow-view images; and (b-e) scale cue: the difference of the spatial size and conceptual size between narrow-view and wide-view images.

(in the circles of confusion as shown in Figure 4.4a), i.e., exhibiting the focus and fringe attributes. (2) *Scale cue*: Scale cue contains two parts, namely spatial size evaluation and conceptual size evaluation. Spatial size is defined as the object's size in the observed image, whereas the conceptual size is its actual size in reality. Figures 4.4b-4.4e illustrate the scale cue concepts, where images can be divided into four groups by two spatial sizes (large/small) × two conceptual sizes (large/small). Only the images where the spatial size of the object is large but its conceptual size is small belong to narrow-view images, while the remaining cases belong to wide-view images.

We proposed the scene focus model and size scale model to formulate the focus cue and scale cue, respectively, as shown in Figure 4.5. The scene focus model transforms the images into the frequency domain to extract the focus features from the high-frequency coefficients. The frequency transformation methods, such as discrete wavelet transform (DWT), non-sampled contourlet transform (NSCT), are commonly used to represent the focus information (Y. Yang et al., 2015). The

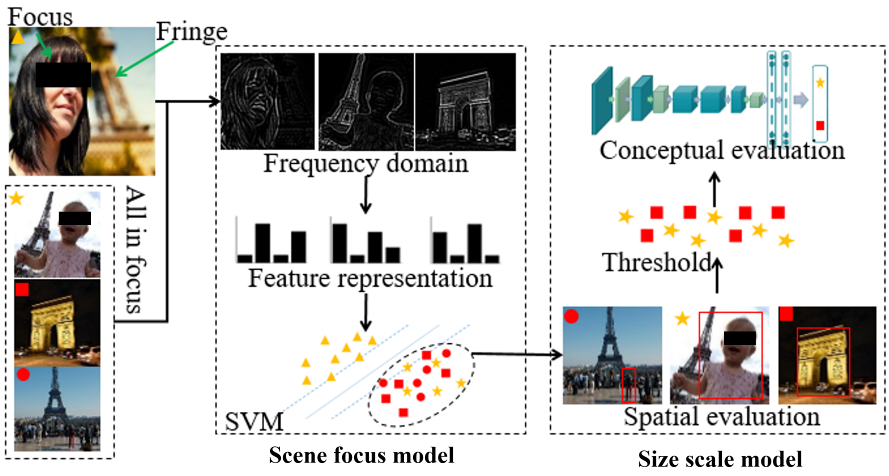


Figure 4.5: The framework of the proposed computational algorithm for view-type classification inspired by human visual system, which contains a scene focus model and a size scale model. (1) The *scene focus model* aims to distinguish whether photos have focus and fringe attributes and filter out these narrow-view photos (yellow triangles). (2) The *size scale model* aims to detect the spatial and conceptual size of the objects of photos. The images with small spatial sizes (red circles) will be classified as wide-view and filtered out in object evaluation. However, the narrow-view and wide-view bounded by large boxes (yellow stars and red rectangle) are classified by conceptual evaluation based on a convolutional neural network (CNN).

high-frequency coefficients are beneficial to detect high contrast edges. Following the obtained high-frequency coefficients, a feature representation method (e.g., speeded-up robust features, SURF: Bay et al., 2006) and a binary support vector machine (SVM) for filtering out photos has focus and fringe attributes. The size scale model aims to detect the spatial and conceptual size of the objects of photos. An object proposal method (e.g., EdgeBoxes: Zitnick and Dollár, 2014) is used to estimate the spatial size of the objects of interest in the remaining images after the scene focus model. For example, Edgeboxes model proposes object bounding boxes based on the grouping of edges and uses the edge content of the bounding box to compute the objectness (likelihood it is an object) score. Thus, the objects in the images would be located with a bounding box. After filtering out the wide-view with small spatial sizes, we classify objects within bounding boxes according to their conceptual size. We perform this phase using a CNN model, which can learn features that represent the concept of objects, i.e., object recognition (Krizhevsky et al., 2017). The large objects (e.g., buildings and mountains) in concept indicate the wide-view, while small objects (e.g., flowers, animals, and humans) indicate the narrow-view.

To facilitate the view-type classification, we collect a dataset, *AttentionShoot*, with human annotations (wide-view vs. narrow-view) that contains 5050 natural images. We conducted experiments on this dataset to evaluate the performance of the proposed framework and to estimate the effectiveness of the scene focus model and scale size model, respectively. The experimental results show that the proposed framework achieves better accuracy for view-type classification than the current methods for depth estimation, including distant/close-up view classification algorithms.

4.2.1 Materials and methods

4.2.1.1 Datasets

To construct the dataset, we first selected natural photos from different sources, i.e., Google Image Search, Flickr, and the ImageNet (Russakovsky et al., 2015) dataset, ensuring the object variety of the photos. Specifically, we used the keywords, e.g., 'insect', 'maple', 'tree', and 'prairie' by Google Image Search and collected travel photos from Flickr. ImageNet dataset has more than one million photos with 1000 categories. We selected photos from the natural classes of this dataset, e.g., animals, trees, and birds. We collected 5050 natural images to construct the dataset named *AttentionShoot*.

Each photo was annotated narrow-view and wide-view by two leading informatics researchers (male, aged 26 and 42 years) and then again by two independent persons (a male aged 22 years and a female aged 28 years) who had been fully briefed about the definition of view-type. The *AttentionShoot* contains 2598 narrow-view and 2452 wide-view photos, termed as the full-set. In addition, the narrow-view set is separated as 1,522 narrow-view images containing the focus and fringe attributes by an informatics researcher (a male, aged 26 years) and 1,315 randomly selected wide-view images, termed as the focus-set, and used for the scene focus model, i.e., training, and testing.

4.2.1.2 Proposed methods

Following the proposed framework inspired by the human visual system as shown in Figure 4.5, we proposed two computational algorithms, namely HVS method and HVS+ method.

HVS method. The scene focus model's critical component is DWT (Pajares & De La Cruz, 2004), while the size scale model has two components, the spatial size evaluation based on EdgeBoxes (Zitnick & Dollár, 2014), and the conceptual size evaluation using a CNN model. Concretely, the photos are first transferred as high-frequency and low-frequency components using DWT. Secondly, the histogram of the high-frequency energy (HHFE) smoothed and refined by two different windows is used to construct feature vectors to represent the photos. Thirdly, the SVM is trained for filtering out photos with focus and fringe attributes. Fourthly, the EdgeBoxes model is used to locating the objects by a bounding box in the photos remaining photos after the SVM phase. Fifthly, a threshold is used to filter out photos with small boxes on objects. Finally, Alexnet (Krizhevsky et al., 2012) re-trained in our dataset is used to classify the remaining photos after spatial size evaluation. HVS method is implemented as follows.

Scene focus model of HVS method. We perform the single-level discrete two-dimensional wavelet transform on the images to obtain the high-frequency details to build the histogram of features with a 63-dimensional vector. We then train an SVM classifier with the radial basis function kernel.

Size scale model of HVS method. To obtain relevant object bounding box proposals, the box search is performed through sliding windows where the step size, scale, and aspect ratio are determined by the overlap ratio of the current window and the previous window. We set the ratio as 0.30 and also eliminate lower-scoring windows that have 0.25 overlap on a higher scoring window for the low amount of box proposals and exclude low-scoring big boxes. The box size ratio threshold is then set as 0.20, which means objects that take up less than 20% of the image would be considered small spatial size (i.e., wide-view as shown in Figure 4.4). For the Alexnet fine-tuning, we implement the model using MatConvNet toolbox (Vedaldi & Lenc, 2014). The learning rate (η) is set to logarithmically reduce at each training epoch from 0.001 to 0.00001 across 30 epochs. We set the learning rate for the last layer, $\eta_l = 10\eta$.

HVS+ method. The key components of the scene focus model are the NSCT and SURF, while the size scale model has two components, the spatial size evaluation based on AdobeBING, and the conceptual size evaluation using a fine-tuned CNN classifier. Concretely, the photos are first transferred as high-frequency and low-frequency components using NSCT. Secondly, the SURF is used to construct feature vectors to represent the photos. Thirdly, the SVM is trained for filtering out photos with focus and fringe attributes. Fourthly, because the EdgeBoxes model calculates the edge feature based on pixel information only. However, humans can "catch" an object quickly before identifying it, and capture several object parts of significant appearance difference from the background before the whole object is effectively located according to these perceived parts, without necessarily knowing all the object components. Following this principle, we apply AdobeBING, which is Binarized Normed Gradients (BING) (Cheng et al., 2019) refined by the AdobeBoxes model (Fang et al., 2016), to locate the objects by a bounding box in the photos remaining photos after the SVM phase. BING is a significantly fast object proposal algorithm based on the correlation between object boundaries and the norm of image gradients. In contrast, AdobeBoxes model uses groups of superpixels with high contrast from the background as the representation of object parts, named adobes, to propose object bounding boxes. This combination may give a fast and efficient evaluation mechanism for spatial size. The rest two steps (i.e., steps of filtering out small box and retraining the Alexnet) are the same as HVS method.

Scene focus model of HVS+ method. For the NSCT, the decomposition scale directions used is $\{1, 2, 8, 16\}$ (NSCT-3), with the '9-7' pyramidal filter and 'pkva' ladder directional filter (Y. Yang et al., 2015). The SURF produces feature representation vectors of 64-dimensions. The SURF features are quantized Fisher Vector (FV), where the vocabulary size of Gaussian Mixture Model cluster amount for FV was set to 50. We then train an SVM classifier with the radial basis function kernel.

Size Scale model of HVS+ method. For the AdobeBING, the normed gradients in horizontal and vertical directions of BING were obtained using a one-dimensional mask $[-1, 0, 1]$, whereas, for AdobeBoxes, the minimum size for superpixel generation is 128 pixels. The box size ratio threshold and the Alexnet in HVS+ method are the same as HVS method.

We randomly select 2000 images from the focus-set for SVM training. The remaining 1450 photos (focus-set) are used to build the test set applied for the evacuation of HVS methods. For the Alexnet retraining, we select 3600 images from full-set, which includes the training images for SVM. The remaining photos are used to test size scale model. HVS and HVS+ methods are implemented by MATLAB2016a on an NVIDIA K40 GPU and a 12 × 3.30GHz Intel i7-5820K CPU with 32 GB RAM.

4.2.1.3 Related algorithms

To evaluate the performance of the proposed framework, we compared it with the following five baselines. (1) *FCRN-Depth (Laina et al., 2016):* FCRN-Depth is a Depth estimation model using FCRN, and reports remarkable results for depth estimation. We involve depth estimation for the view-type classification task due to the close relationship between depth of view and view-type, such as a larger depth being associated with wide-view images and vice versa for narrow-view images. The FCRN-Depth is used to obtain a depth map for each test image in our experimental dataset. We represent the features by the histogram of oriented gradients, and SVM was used as the classifier. (2) *Mono-Depth (Godard et al., 2017):* Mono-Depth is another depth estimation model using fully convolutional network (Mayer et al., 2016). Mono-Depth uses a novel training loss to model the ambiguous mapping between monocular images, leading to improved depth estimation performance. Similar to FCRN-Depth, the Mono-Depth was used to obtain the depth map of the testing images, and then the features were extracted by the histogram of oriented gradients with SVM as the classifier. (3-4) *Sobel-SVM and Canny-SVM:* Zhuang et al. (2014) proposed to use edge histogram and SVM to classify the two views automatically. We reimplemented this algorithm using "Sobel" and "Canny" edge detectors as the edge detection methods, respectively. We refer to them as Sobel-SVM and Canny-SVM, which serve as two baselines. (5) *Gabor-SVM:* Cao and O'Halloran (2015) proposed to use a textured pattern based on the Gabor feature to represent the image structure for the classification of distant and close-up images. We reimplemented this algorithm, and the structure of an input image was represented by a $1,536$ dimensional feature vector (3 color channels × 4 scales × 8 orientations × 16 sub-regions).

4.2.1.4 Evaluation Metrics

The prediction accuracy $\in [0,1]$ is the common assessment of a classification algorithm. Larger value of accuracy indicates better classification performance. In addition to the accuracy, we also applied Precision and Recall, which are based on True Positives, False Positives, True Negatives and False Negatives, for more insights of the evaluation of our proposed framework. These metrics have been applied in many studies to evaluate the performance of classification algorithm (Bardou et al., 2018; Le et al., 2019; Ning et al., 2018). All of them differentiate the correct classification of labels within different classes (Ma et al., 2019; Yu et al., 2019), which are defined as follows,

$$\begin{aligned} \text{Precision} &= \frac{\text{True Positive}}{\text{Predicted Positive}}, \\ \text{Recall} &= \frac{\text{True Positive}}{\text{Actual Positive}}, \end{aligned} \tag{4.1}$$

where Precision, Recall $\in [0,1]$; 'Predicted Positive' is the sum of True Positive and False Positive; and 'Actual Positive' is the sum of True Positive and False Negative. Here, we set the narrow-view as 'Positive'. Thus, the 'True Positive' is the number of photos correctly predicted as narrow-view, 'Predicted Positive' is the number of photos predicted as narrow-view, 'Actual Positive' is the number of ground truth

Table 4.1: The performance of the proposed framework compared with the related algorithms on the full-set of *AttentionShoot*. The precision and recall are defined as indicators of the narrow-view class.

Classification algorithms	Accuracy	Precision	Recall
FCRN-Depth (Laina et al., 2016)	0.7034	0.7507	0.6905
Mono-Depth (Godard et al., 2017)	0.8110	0.8639	0.7794
Sobel-SVM (Zhuang et al., 2014)	0.8055	0.8351	0.8058
Canny-SVM (Zhuang et al., 2014)	0.8193	0.8583	0.8045
Gabor-SVM (Cao & O'Halloran, 2015)	0.8740	0.8652	0.8860
HVS method	0.8400	0.8538	0.8559
HVS+ method	**0.9317**	**0.9040**	**0.9799**

(narrow-view). Generally, the larger values of Precision, Recall indicate the better classification performance.

4.2.2 Results and discussion

4.2.2.1 Compared with related algorithms

We compared the proposed HVS method and HVS+ method with five baselines on the full-set of *AttentionShoot*. As shown in Table 4.1, most algorithms achieve accuracies over 0.8 (expect FCRN-Depth), but only the HVS+ method achieves accuracy over 0.9. The HVS+ method performs the best in all metrics among competing algorithms. In addition, the depth models (FCRN-Depth and Mono-Depth) have high precision, while low recall for narrow-view, which means the number of predicted narrow-view images is smaller than actual narrow-view images. This result indicates that the depth models have a considerable bias for predicting more photos as wide-view. Oppositely, the proposed HVS+ method also has a certain bias, and obtains a low precision for narrow-view, indicating it predicts more photos as narrow-view. Although existing this bias, the HVS+ method shows considerable improvements in accuracy, precision, and recall by 0.0577, 0.0388, and 0.0939, compared with the second-best algorithm, Gabor+SVM (Cao & O'Halloran, 2015). Thus, we conclude that the HVS+ method is the most suitable algorithm for view-type classification, comparing with the five baselines and the HVS method.

Table 4.2: The performance of the proposed scene focus model compared with depth models (FCRN-Depth and Mono-Depth) on the focus-set.

Classification algorithms	Accuracy	Precision	Recall
FCRN-Depth (Laina et al., 2016)	0.7145	0.8166	0.6992
Mono-Depth (Godard et al., 2017)	0.8662	0.9399	0.8391
Scene focus model of HVS method	0.7634	0.9219	0.6782
Scene focus model of HVS+ method	**0.9510**	**0.9781**	**0.9425**

Since the scene focus model extracts the focus and fringe attributes from original images, we compare the proposed scene focus models with the depth models on the focus-set of *AttentionShoot*, as shown in Table 4.2. We note that the Mono-Depth achieves second place in the classification task, indicating the potential of depth maps representing the focus characteristics of images under certain cases. However, the HVS+ method (scene focus model) performs best, outperforming Mono-Depth with considerable improvements of 0.0848, 0.0382, and 0.1043 for accuracy, precision, and recall, respectively. By comparing with Table 4.1, it can be found that drops of accuracy exist in the depth models. For example, the performance of Mono-Depth dropped from 0.8662 to 0.8110 (accuracy). This drop may be caused by the weaker capability of the depth models in representing the view-type of the photos without the focus and fringe attributes. In addition, the scene focus model of the HVS method performs worse than Mono-Depth on focus-set. However, the HVS method performs better than Mono-Depth on full-set. This result suggests that the size scale model is an important trait to be considered in view-type classification. The reason for the importance of the proposed framework (HVS and HVS+ methods) is that the focus cue and scale cue are crucial attributes of the human visual systems, whereas depth information alone is insufficient for the view-type classification task.

4.2.2.2 Ablation studies

We conduct ablation studies on the following designs to verify the effectiveness of varied (1) scene focus models and (2) size scale models.

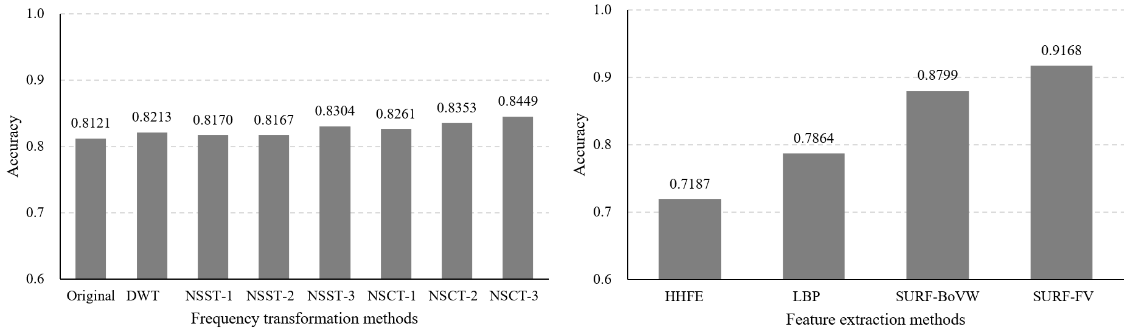
The effects of different scene focus models. To evaluate the performance of the proposed scene focus models, we compared it against the various methods of

Table 4.3: Comparison of different scene focus models, i.e., (8 frequency transformation methods × 4 feature extraction methods). The number is the classification accuracy of different models on the focus-set.

	HHFE	LBP	SURF-BoVW	SURF-FV
Original	0.5152	0.9092	0.8765	0.9474
DWT	0.7634	0.7649	0.8783	0.8784
NSST-1	0.7654	0.7008	0.8901	0.9116
NSST-2	0.7692	0.7367	0.8757	0.8853
NSST-3	0.7305	0.8024	0.8617	0.9271
NSCT-1	0.7544	0.7424	0.9092	0.8984
NSCT-2	0.7474	0.7964	0.8619	0.9355
NSCT-3	0.7042	0.8387	0.8857	**0.9510**

frequency transformation and feature extraction. (1) *Frequency Transformation:* The DWT and NSCT are compared with the original domain of image and Nonsubsampled Shearlet Transform (NSST) (Easley et al., 2008). The NSST uses $\{1, 8\}$ (NSST-1), $\{1, 8, 16\}$ (NSST-2), and $\{1, 8, 16, 16\}$ (NSST-3) decomposition scale directions with the 'maxflat' pyramidal filter (Moonon & Hu, 2015). In addition, the NCST performs different decomposition scale directions, i.e., $\{1, 2\}$ (NSCT-1) and $\{1, 2, 8\}$ (NSCT-2), are also added to the comparison. (2) *Feature Extraction:* The HHFE and SURF-FV is compared with the Local Binary Pattern (LBP) (Ojala et al., 2002) and SURF quantized by bag of visual words (BoVW). LBP produces a feature representation vector of 10-dimensions. The vocabulary size of the BoVW codebook is set to 50. Thus, 32 different scene focus models (8 frequency transformation methods × 4 feature extraction methods) are compared in this phase.

The classification accuracies of different scene focus models are shown in Table 4.3. In addition, the averaged accuracies of different frequency transformation and feature extraction methods are shown in Figure 4.6a and 4.6b, respectively. Figure 4.6a shows that the proposed NSCT can get the best performance among the varied frequency transformation methods except for NSCT-1. Specifically, NSCT-3 shows a considerable improvement of the accuracy of 0.0245 from DWT, indicating that the NSCT-3 can get a better representation to detect the blur and fringe information than other frequency transformation models. Figure 4.6b shows that the SURF



(a) (b)

Figure 4.6: The averaged accuracy of (a) different frequency transformation and (b) feature extraction methods. Each averaged accuracy is achieved by averaging the row or column of Table 4.3 for the corresponding scene focus model.

performs the best out of the other investigated feature extraction methods. For example, the accuracy achieved by SURF-FV shows an improvement of 0.1985 compared to the accuracy achieved by HHFE. Thus, the SURF-FV can better represent the frequency domain than the others. In addition, Table 4.3 shows that the LBP, SURF-BoVW, and SURF-FV perform well with an accuracy above 0.8, where the SURF-FV applied on the NSCT-3 (i.e., NSCT-3+SURF-FV) is the best with the averaged accuracy of 0.9510. The HHFE performs the worst likely due to the insufficiency of representation by solely relying on the spatial summation of high-frequency signals compared to the higher-level representation provided by SURF. From the analysis above, the NSCT-3+SURF-FV used in HVS+ method performs better for the scene focus model than DWT+HHFE used in HVS method with an accuracy increase of 0.1911.

In addition, we tested the selected methods, whose classification accuracies are over 0.9, on the full-set. Table 4.4 shows a large drop, 0.1312 on average, which may contribute by the narrow-view without focus and fringe attributes in the full-set into the test. This result suggests the necessity of a size scale model for better classification performance.

The influence of different scale size models. We investigated the ability of the Edge boxes and AdobeBING to locate the spatial size of the main objects of

Table 4.4: The difference of the selected models' performance between testing on the focus-set and full-set. The selected models (accuracy > 0.9) are shown in grayed cells of Table 4.3.

	Focus-set	Full-set	Difference
Original + LBP	0.9092	0.8152	0.0940
Original + SURF-FV	0.9474	0.8200	0.1274
NSST-1 + SURF-FV	0.9116	0.7724	0.1392
NSST-3 + SURF-FV	0.9271	0.7814	0.1457
NSCT-1 + SURF-BoVW	0.9092	0.7890	0.1202
NSCT-2 + SURF-FV	0.9355	0.7931	0.1424
NSCT-3 + SURF-FV	0.9510	0.8014	0.1496
Mean	0.9273	0.7961	0.1312

a photo. Two object proposal methods are compared, i.e., *AdobeBoxes (Fang et al., 2016)* and *RPN (Ren et al., 2015).* RPN is an object proposal approach using Fully Convolutional Networks, which is designed by sharing the convolution parameters of a specified object detection network. Two different CNN detectors were tested here: one is proposed by Zeiler and Fergus (2014), the other is a VGG16 model proposed by Simonyan and Zisserman (2014), namely RPN1 and RPN2, respectively. The box size ratio threshold of the competing methods is set as $\alpha = 0.2$, which is the same as the setting in HVS+ method. Size scale model serves as the secondary classifier to address images that do not fit into the scene focus model. In other words, the size scale model is used to pick out the narrow-view images that were misclassified as wide-view images by the scene focus model due to the lack of depth-of-focus attribute. Thus, all images classified as wide-view by the scene focus models are used for the testing of the scale model, regardless of right or wrong. Additionally, we compare the performances of different scale size models based on the selected scene focus models with accuracy above 0.9. Thus, 35 scale size models in total (7 scene focus models × 5 object proposal models) are compared in this phase. The subsequent conceptual size classification uses the same CNN of HVS+ method.

The accuracies of different size scale models are shown in Table 4.5, and the averaged accuracies of size scale models are shown in Figure 4.7. Figure 4.7 shows clearly that AdobeBING model can get the best performance out of all the size scale

Table 4.5: Comparison of different scale size models, i.e., 7 scene focus models × 5 object proposal models. The number is the classification accuracy of the corresponding combination of method testing on the full-set.

	EdgeBoxes	AdobeBoxes	RPN1	RPN2	AdobeBING
Original+LBP	0.8641	0.8221	0.8048	0.8717	**0.8938**
Original+SURF-FV	0.8807	0.8297	0.8290	**0.8869**	0.8807
NSST-1+SURF-FV	0.8559	0.7993	0.8117	0.8717	**0.9138**
NSST-2+SURF-FV	0.8690	0.8041	0.8221	0.8848	**0.9228**
NSCT-1+SURF-BoVW	0.8759	0.8200	0.8200	0.8883	**0.9248**
NSCT-2+SURF-FV	0.8703	0.8097	0.8193	0.8814	**0.9179**
NSCT-3+LBP	0.8097	0.7345	0.7917	0.8497	**0.8838**
NSCT-3+SURF-FV	0.8807	0.8214	0.8283	0.8945	**0.9317**

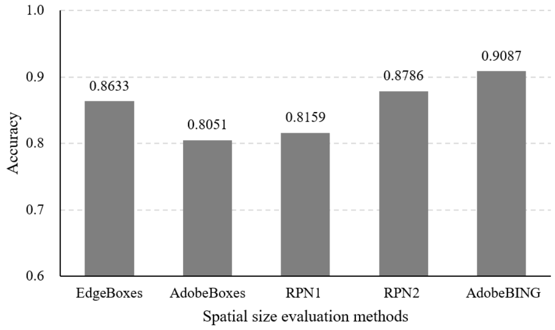


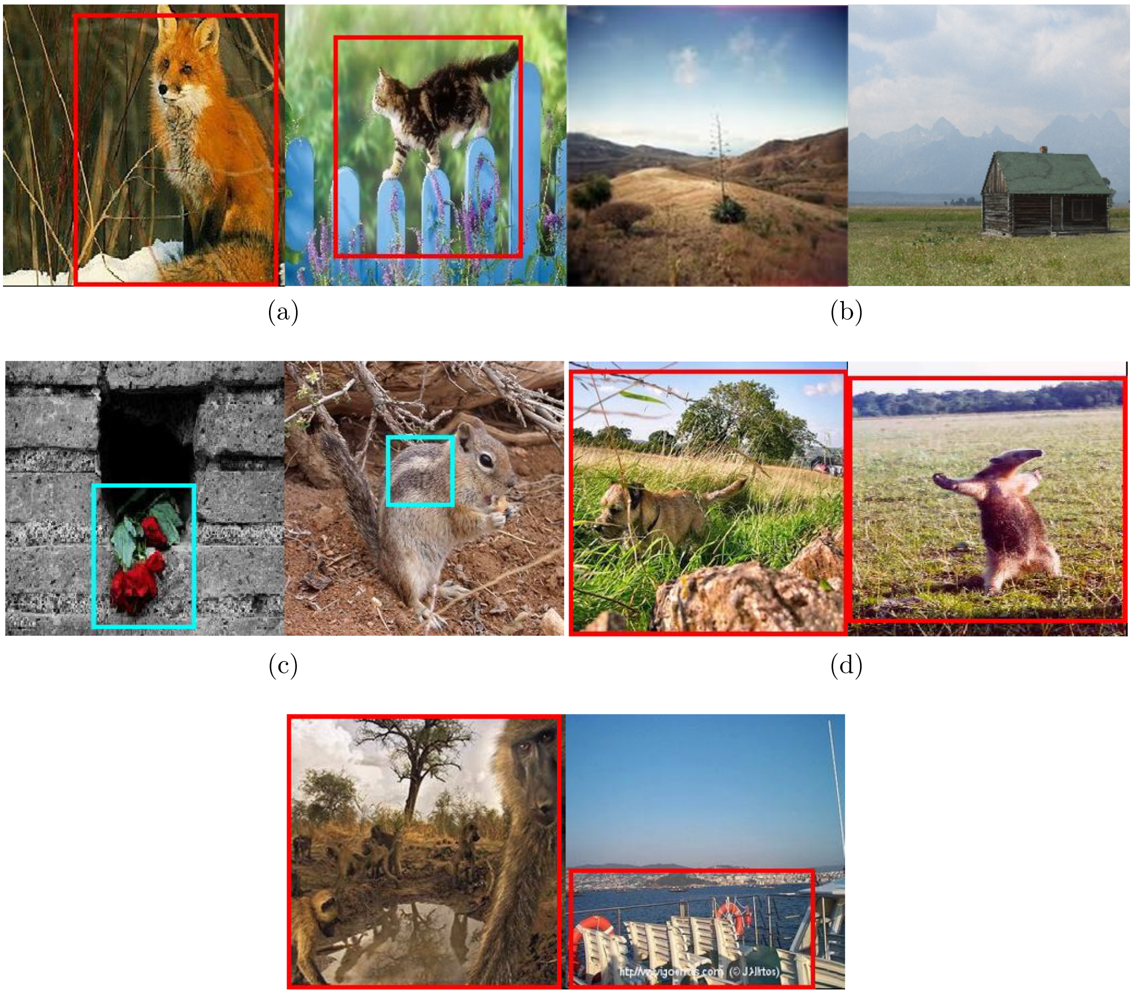
Figure 4.7: The averaged accuracies of different size scale models. The averaged accuracy of each size scale model is achieved by averaging the corresponding column of Table 4.5.

models. For example, the accuracy achieved by AdobeBING shows an increase of 0.0414 compared to the accuracy achieved by Edge boxes. This result suggests that AdobeBING is more effective for spatial size evaluation. Table 4.5 shows that the best performing joint model is the NSCT-3+SURF-FV and AdobeBING+CNN, at 0.9317 accuracy, and thus these settings are used in HVS+ method.

4.2.2.3 Failure cases and ambiguities

Due to the imperfections of HVS+ method and ambiguities in the constructed dataset, the following two aspects are found to cause failures.

Algorithm errors: Figure 4.8a-4.8b show example failures of the scene focus model. The failure cases may be caused by the confusion in the depth-of-focus attribute of the photos. The background of the narrow-view photos in Figure 4.8a



(a)

(b)

(c)

(d)

(e)

Figure 4.8: Examples of failure cases. (a) Narrow-view while misclassified by scene focus model, (b) Wide-view while misclassified by scene focus model, (c) Narrow-view while misclassified by spatial size evaluation, (d) Narrow-view while misclassified by conceptual size evaluation, (e) Wide-view while misclassified by conceptual size evaluation.

are relatively sharp, hence could generate inaccurate features, even though the fringe is still apparent to human observation. Nevertheless, these two images were amended by the scale model, as seen from the precise object bounding box. On the other hand, the wide-view images in Figure 4.8b contain smooth sky regions which could have been confused as the fringe by the model. Figure 4.8c shows examples of narrow-view photos filtered as wide-view by AdobeBING, where there are two

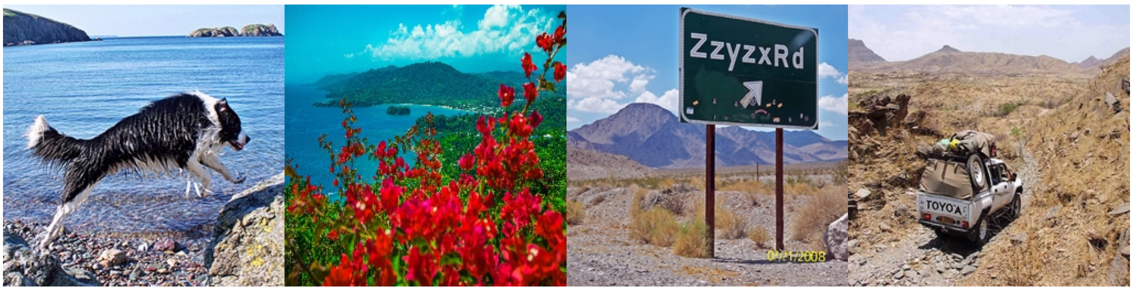


Figure 4.9: The samples of ambiguous photos on the *AttentionShoot*. These photos are considered as ambiguous because they include a close-up object with a wide-view background.

possible reasons for the error. The first reason is that the localized object has a size that is below the threshold, while another reason is incorrect localization, shown by the blue bounding boxes proposed.

Ambiguities: We find the majority of the misclassification is due to the ambiguous nature of the images. In the narrow-view photos, as shown in Figure 4.8d, the background of the image can be considered as wide-view. Moreover, the spatial size evaluation was unable to localize the object of interest to assist the classification. In contrast, the misclassification of narrow-view images in Figure 4.8e can be due to the interference by objects that were not the point of interest in the images, such as the narrow-view of the monkey found on the right of the left photo, and the chairs localized by the AdobeBING in the right photo. Other examples of ambiguous photos on the constructed dataset are shown in Figure 4.9. These misclassified and ambiguous photos contain close-up objects while the background shows the wide-view of its surroundings, which is difficult to determine the attentional scope of the photographer and may challenging even to humans.

4.3 Experiment 2

Experiment 1 proposed a framework inspired by the human visual system (HVS and HVS+ methods) and showed that the HVS+ method gets a considerable result for view-type classification. However, there are two drawbacks to the HVS+ method.

Firstly, there is a certain number of failure cases and ambiguities caused by the imperfect setting of the HVS+ method and *AttentionShoot*. Secondly, the scene focus model (e.g., NSCT) is difficult to implement in the GPU. These drawbacks become obstacles to implement HVS+ method for estimating the attentional scopes of extensive travel photos.

In Experiment 2, we aim to solve these drawbacks by looking into a CNN, and using a single CNN to perform view-type classification as opposed to the hand-designed HVS+ method. This is on account of the success shown by CNN at covering both low- and high-level features for a variety of tasks (Donahue et al., 2014; Lee et al., 2017; Yosinski et al., 2015; Zeiler & Fergus, 2014). However, conventional CNNs only utilize high-level features after multiple layers of convolution. According to the investigation of Experiment 1, both focus and scale information is crucial for view-type classification, suggesting feature represent view-type (e.g., focus feature) may vanish after multiple convolutional and pooling operations in conventional CNNs. Therefore, we designed a cumulative feature CNN (CFCNN) to extract features from each stage and accumulate them into one representation, thus incorporating both low and high-level features for view-type classification. Figure 4.10 illustrates the architecture of the CFCNN model, where the inputs are photos and the outputs are their respective narrow- and wide-view photos. Specifically, we introduce additional convolutional paths on each existing convolutional layer to produce 1024-dimension features. The new convolutional layers are placed after pooling layers, and if a convolutional layer is not followed by a pooling layer, an added pooling layer would be set before the corresponding convolutional layer. These new convolutional layers are to be trained end-to-end with all the other convolutional layers. The kernels are expected to focus on significant features from different levels, and then directly summed up to obtain the cumulative feature and proceed to the subsequent fully connected layers for classification. This cumulative feature extraction scheme is similar in spirit to the well-recognized ResNet (He et al., 2016) bottleneck architecture. The proposed CFCNN, nevertheless, is different since CFCNN features are summed up from multiple layers.

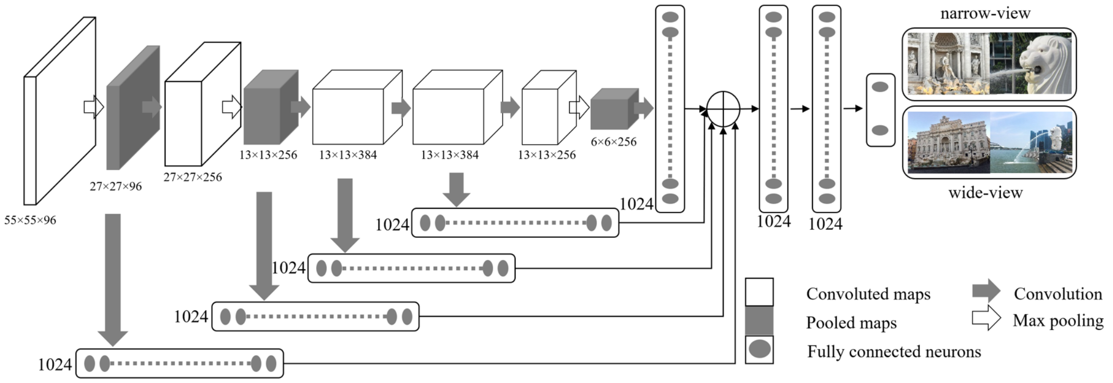


Figure 4.10: The architecture of the proposed cumulative feature convolutional neural network (CFCNN) to learn the view-type representation. The features from each convolution layer are cumulated into one representation.

In this experiment, we constructed two datasets. Considering the potential application of the view-type classification in tourism research, we use travel photos to reconstruct a dataset with view-type annotations, termed as the *TravelShoot* dataset. The *TravelShoot* contains 43,906 travel photos collected from the famous sightseeing spots on Flickr. We use this dataset to train the CFCNN for learning the view-type representations of travel photos. In addition, the travel photos will inevitably be mixed with ambiguous photos. To investigate the percentile of ambiguous photos, we randomly collected 3000 travel photos, and five participants annotated the view-types. When two or three participants label the photo as wide-view, while others annotated it as narrow-view, this photo will be considered ambiguous. Since this dataset contains view-type cognition from different persons, we name it as *CognitionShoot* and use it as a test set for classification performance evaluation of CFCNN. Experimental results indicate that the CFCNN improves the classification performance and has less computational complexity compared with the HVS and HVS+ methods proposed in Experiment 1. Furthermore, the results suggest that the CFCNN might learn the shared patterns with human cognition for different ambiguous levels. Lastly, we visualize the activation maps of CFCNN and find that the CFCNN has learned some other features beyond the focus cue and scale cue of our proposed framework in Experiment 1.

4.3.1 Materials and methods

4.3.1.1 Datasets

We collected raw photos from Flickr according to the geographic locations of famous sightseeing spots provided by TripAdvisor. Firstly, we collected over 100,000 travel photos from Flickr. Secondly, two informatics researcher (male, aged 26 and 42 years) scrutinized all photos to remove the duplicated photos and meaningless photos. Thirdly, the remaining 46,906 is automatically detected as wide-view and narrow-view by using the HVS+ method then again by two independent persons (a male aged 24 years and a female aged 28 years) who had been fully briefed about the definition of view-type. Finally, we randomly selected 3,000 travel photos to construct the *CognitionShoot* dataset. The remaining 43,906 travel photos, 25,776 wide-view and 18,130 narrow-view, are composed of the *TravelShoot* dataset.

To label the *CognitionShoot* dataset, we recruited five participants (3 male and 2 female, mean age $= 23.2 \pm 1.2$) and designed a binary classification task according to their cognition to view-type. Before the task, ten wide-view and ten narrow-view photos were demonstrated to let participants understand the definition of view-type. In the task, 20 photos (four rows and five columns) were simultaneously shown on the screen to give a better visual comparison, and each participant selects photos into wide-view or narrow-view. This procedure was iteratively carried out until all photos were checked. The *CognitionShoot* dataset consists of 1480 wide-view and 1520 narrow-view photos. According to the different ambiguous levels of the photos, the *CognitionShoot* dataset has six groups. For example, if three participants annotated a photo as wide-view, others labeled it as narrow-view, this photo would be regarded as an ambiguous wide-view. If only four participants annotated a photo as wide-view, this photo would be a normal wide-view. If one photo is labeled as wide-view by all five participants, this photo would be an obvious wide-view. The size of each group is shown in Table 4.6.

Table 4.6: The number of different groups of *CognitionShoot*. Obvious wide-view, normal Wide-view, and ambiguous wide-view represent such photo annotated as wide-view by five, four, and three participants, respectively, where other participants annotated as narrow-view. The same definition for obvious narrow-view, normal narrow-view, ambiguous narrow-view.

	Obvious	Normal	Ambiguous
Wide-view	929	300	251
Narrow-view	1031	250	239

4.3.1.2 Proposed algorithm

We design CFCNN on the basis of AlexNet (Krizhevsky et al., 2012). We use the pooling size 3×3 for every additional pooling layer. The kernel sizes of the convolutional layers are transferred from the AlexNet, while the additional convolutional layers' kernels follow the size of the feature map that is to be convolved. For example, the feature map after the first pooling layer is $27 \times 27 \times 96$. Hence the kernel size of the first additional convolutional layer is $27 \times 27 \times 96$ and mapped to 1024 neurons. These additional convolutional layers are to be trained end-to-end with all the other convolutional layers. Hence, the feature maps from different layers are directly summed up to obtain the cumulative feature map and then proceed to the subsequent fully connected layers for classification. Thus the cumulative feature representation is expected to focus on significant features from different levels (layers).

We used 3/4 of photos in *TravelShoot* for training while the remaining for validation. During the training process, each training image is augmented by resizing the shorter side to 256 dimensions while maintaining aspect ratio, and then random cropping and flipping are performed, followed by normalization by subtracting with the average image of the dataset. Finally, a $227 \times 227 \times 3$ dimension image is fed to the network. CFCNN uses the stochastic gradient descend approach with a training batch size of 230, weight decay of 0.0005, and the learning rate that logarithmically reduces from $\eta = 10^{-5}$ after every training epoch. In addition, we transferred ImageNet pre-trained weights from the AlexNet for the five convolutional layers

Table 4.7: The accuracy and time performance of different algorithms testing on *CognitionShoot*. The Obvious, normal and ambiguous represent the accuracy in sets of obvious, normal, and ambiguous wide-/narrow-view. The accuracy is compute from *CognitionShoot* except for ambiguous set. Execution time is the time for classifying 3000 photos on *CognitionShoot* with the methods' own implements. The HVS methods are implemented on MATLAB2016a with parallel processing of 6 compilers, while CFCNN is implemented PyTorch with a batch size of 10 photos.

Classification algorithm	Accuracy	Obvious	Normal	Ambiguous	Execution time (h)
HVS method	0.7454	0.7985	0.5564	0.5367	0.3896
HVS+ method	0.8434	0.8959	0.6564	0.5408	1.0129
CFCNN	**0.8813**	**0.9367**	**0.7182**	**0.5612**	**0.0158**

to improve the generalization of the main feature extraction layers of our model. To prevent overfitting, the training was stopped at 200 epochs, where there was no significant reduction in the trend of the validation error. The difference between the validation and training errors was 0.047, an acceptable range of over-fitting as the validation performance achieved over 0.8. The CFCNN is implemented by PyTorch on an NVIDIA K40 GPU and a 12 × 3.30GHz Intel i7-5820K CPU with 32 GB RAM.

4.3.2 Results and discussion

To evaluate the CFCNN, we use *AttentionShoot* and *CognitionShoot* for testing. The classification accuracies of the HVS, HVS+, and CFCNN on *AttentionShoot* are 0.8400, 0.9317, and 0.9752, respectively. In addition, the classification accuracies and time performance of these algorithms on *CognitionShoot* are shown in Table 4.7. Firstly, we can find that the CFCNN gets the highest accuracy and a considerable improvement of 0.0379 compared with HVS+ method. Secondly, these algorithms show different accuracies in the different ambiguous levels. For example, from the obvious set (obvious wide-view vs. narrow-view) to the ambiguous set, all algorithms get manifest drops. For example, the CFCNN decreases 0.2185 from the obvious set to the normal set. Thirdly, the computational time of CFCNN is lower than HVS and HVS+ methods. The reason is HVS and HVS+ methods are implemented on MATLAB with CPU, while CFCNN is implemented on PyTorch

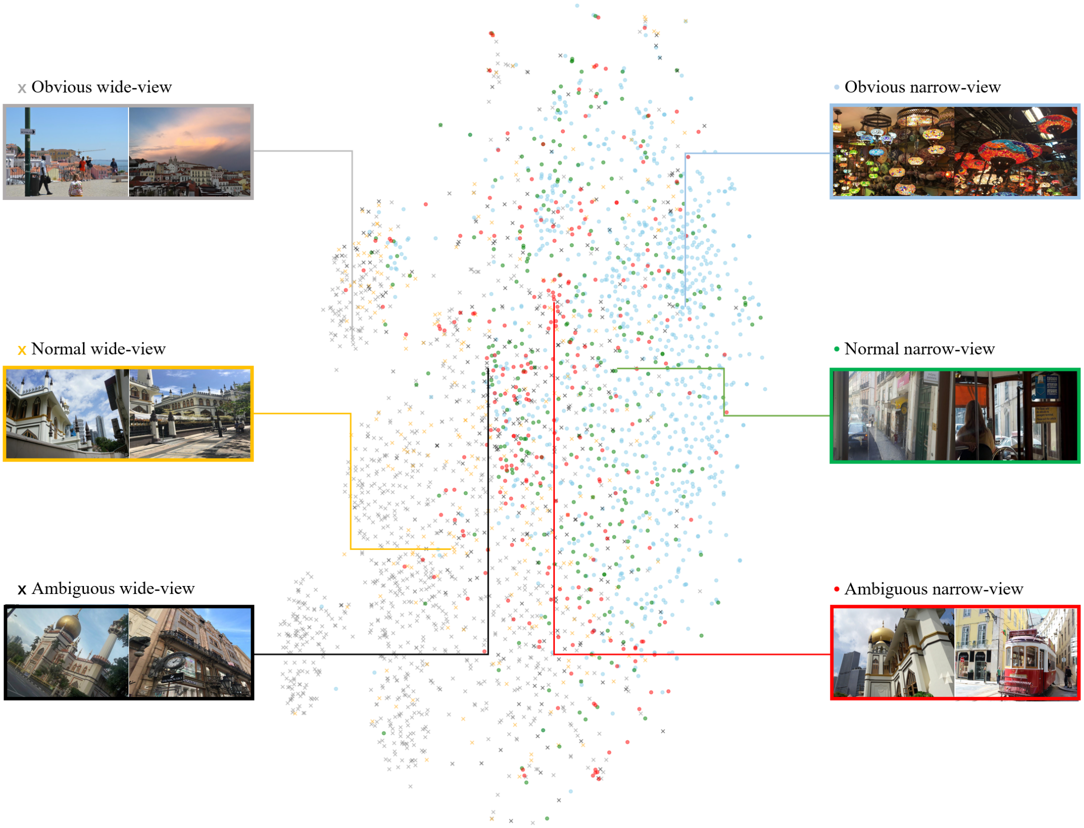


Figure 4.11: The *t*-SNE visualization of the last hidden layer representations in the CFCNN for view-type classification. Here we show the CFCNN's internal representation of view-type by applying *t*-SNE , a method for visualizing high-dimensional data, to the last hidden layer representation in the CFCNN of *CognitionShoot* dataset (3000 photos). The 'X' and point clouds represent the wide-view and narrow-view, respectively, showing how the algorithm clusters the photos. Insets show images corresponding to various cognition groups with different colors.

with GPU. In addition, HVS+ method is most time-consuming because the NSCT has more computational complexity than DWT. Specifically, the CFCNN only takes 0.0156 times as long as the execution time of HVS+ method to complete the test of *CognitionShoot* (3000 photos). Therefore, we can conclude that the CFCNN model is superior to the HVS+ method we previously proposed in Experiment 1 for view-type classification and more suitable for estimating the attentional scopes of extensive travel photos.

To better understand the view-type classification of CFCNN, we examined the internal features learned by the CFCNN using *t*-SNE (*t*-distributed stochastic

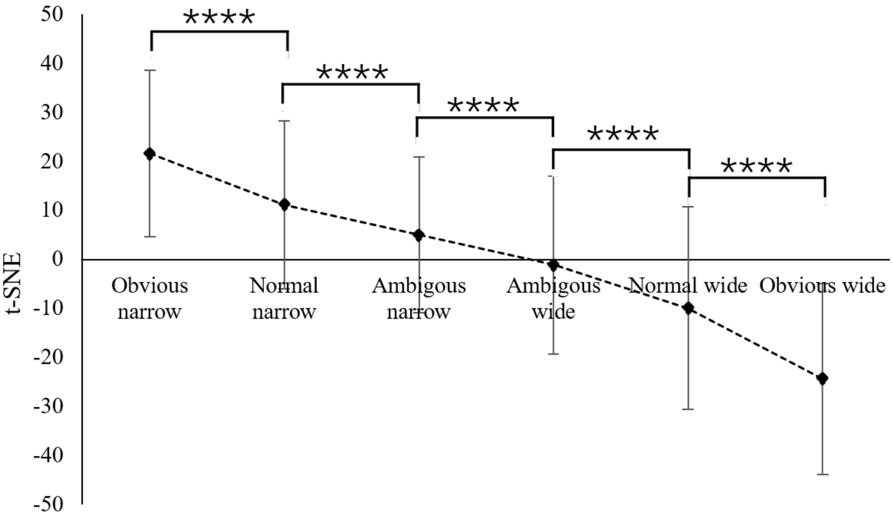
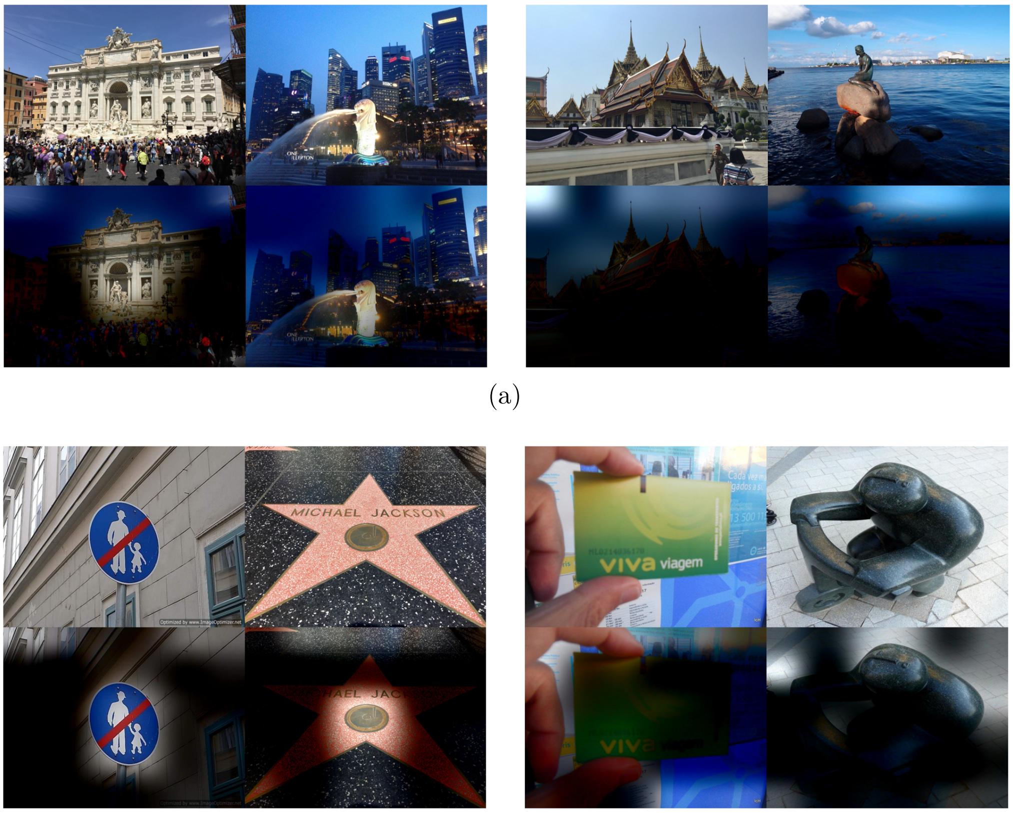


Figure 4.12: The statistical results of the t-SNE's dimensions for different ambiguous levels. The vertical value represents the first of t-SNE, which is the horizontal axis as shown in Figure 4.11, respectively. * p, **** p <0.00001.

neighbor embedding) as shown in Figure 4.11. Each point or 'X' represents a travel photo projected from the 1024-dimensional output of the CFCNN's last hidden layer into two dimensions. We see clusters of points and 'X's of the same view-type classes. The obvious wide-view photos cluster in the left and opposite to obvious narrow-view photos (right). The normal wide-view photos spread in the obvious wide-view photos, while cluster in the right part. Similarly, the normal narrow-view photos spread in the obvious narrow-view photos, while cluster in the left part. The ambiguous set spread in the center. In addition, we showed the averaged values of the t-SNE's first component between different ambiguous groups in Figure 4.12. There are significant differences (unpaired t-test) between every two groups. The result also shows a correlation may exist between humans and CFCNN for view-type cognition, and the t-SNE's first component represents the attentional scopes from narrow to wide to some extent.

For additional insight into the view-type classification of CFCNN, following Loh and Chan (2019). We apply receptive field analysis to highlight the regions of the image that are most salient in generating the output distribution. Specifically, we extract the receptive field produced by the last pooling operation for each test image,



(a)

(b)

Figure 4.13: Network receptive field analysis of photo examples, (a) wide-view photos and (b) narrow-view photos. Given an input image (top), the output mask (bottom) highlights image regions with the most contribution to the view-type classification by CFCNN.

where the dimension of the map is 6 × 6 × 256. Max pooling is again performed on the extracted maps in the third dimension to obtain an aggregated map with 6 × 6 dimension, where it is then resized to the size of the original image. This final map is used to mask the luminance channel of the original image to obtain a visualization of the area in which the features are used for classification. This process leads to a salient map over the input image. The bright areas show the most contribution to the classifier. Figure 4.13 shows several examples of receptive field analysis. Each pair of images shows the input and the corresponding masked photos. It can be found that the CFCNN looking at the object of interest (e.g., as shown in left examples of Figure 4.13) for some photos, while others only focus on the image fringe ((e.g., as shown in right examples of Figure 4.13). However, the proposed framework in Experiment 1 is only looking at the main object of interest — focus cue estimates whether focus or defocus of the main object; scale cue estimates spatial size and conceptual size of the main objects. Therefore, CFCNN goes beyond the focus cue and scale cue designed in the HVS+ method, which may be one of the components the CFCNN obtain a better classification performance than HVS methods.

4.4 Discussion

This chapter discovered a novel and challenging computer vision problem — the view-type classification, which expresses the photographers' attentional scopes and shooting behavior. This task is challenging because estimating the attentional scope of the photographer is subjective processing and it is difficult to estimate objectively. Furthermore, we found the natural photos contain ambiguous photos that are difficult to be defined as wide-view and narrow-view even by human. To solve the classification problem, we designed two methods corresponding to two experiments. In experiment 1, we proposed the focus and scale cues inspired by the human visual system for view-type classification and a framework that contains scene focus and size scale models. In addition, we established a view-type classification benchmark (*AttentionShoot*) with 5,050 natural photos to evaluate the performance of the proposed framework. Through a large number of comparative experiments, the results show that NSCT+SURF and AdobeBING+CNN are essential and optimal for scene focus and size scale models, whereby the proposed algorithms remarkably outperform the existing algorithms for related works. However, the separated models in the proposed framework could produce classification bias, and the ambiguous photos on the *AttentionShoot* may influence the classification performance. Thus, we proposed a CFCNN which combines the low-level feature with high-level image interpretations in Experiment 2. We established two datasets, i.e., *TravelShoot* and *CognitionShoot*. The *TravelShoot* is used to train the CFCNN, which makes the training data expanded to 10 times compared with *AttentionShoot*.

The *CognitionShoot* had finer classifications with six levels of view-type cognition and was used for CFCNN evaluation. The experimental results showed that the CFCNN is outperformed the algorithms proposed in Experiment 1 with higher classification accuracy and time effectiveness. In addition, we showed the CFCNN might learn similar patterns from humans for different levels of view-type cognition based on the feature visualization.

The proposed computational algorithm, CFCNN, for view-type classification is the basis for our future works. This classification task is motivated by such visual attention, existing relationships with sightseeing values (Bartie & Mackaness, 2016), and emotion (Fredrickson, 2004). We intend to extend this research in two directions, i.e., data science and psychological science. (1) Data science: The travel photos may represent the behavior and visual attention of tourists. This chapter proposed an algorithm that can automatically estimate the tourists' attentional scopes from their shared photos. Thus, the proposed algorithm can potentially be used to measure sightseeing aesthetics and tourists' sentiments of travel locations. We intend to implement the proposed algorithm in a travel recommendation system in our future work. (2) Psychological science: According to the broaden-and-build theory, scopes of visual attention indicate the observers' emotions. With the proposed algorithm, we intend to analyze more data and construct NODS with attentional scopes for expanding the broaden-and-build theory in the tourism scenario. We will make this point more concrete in the next Chapter.

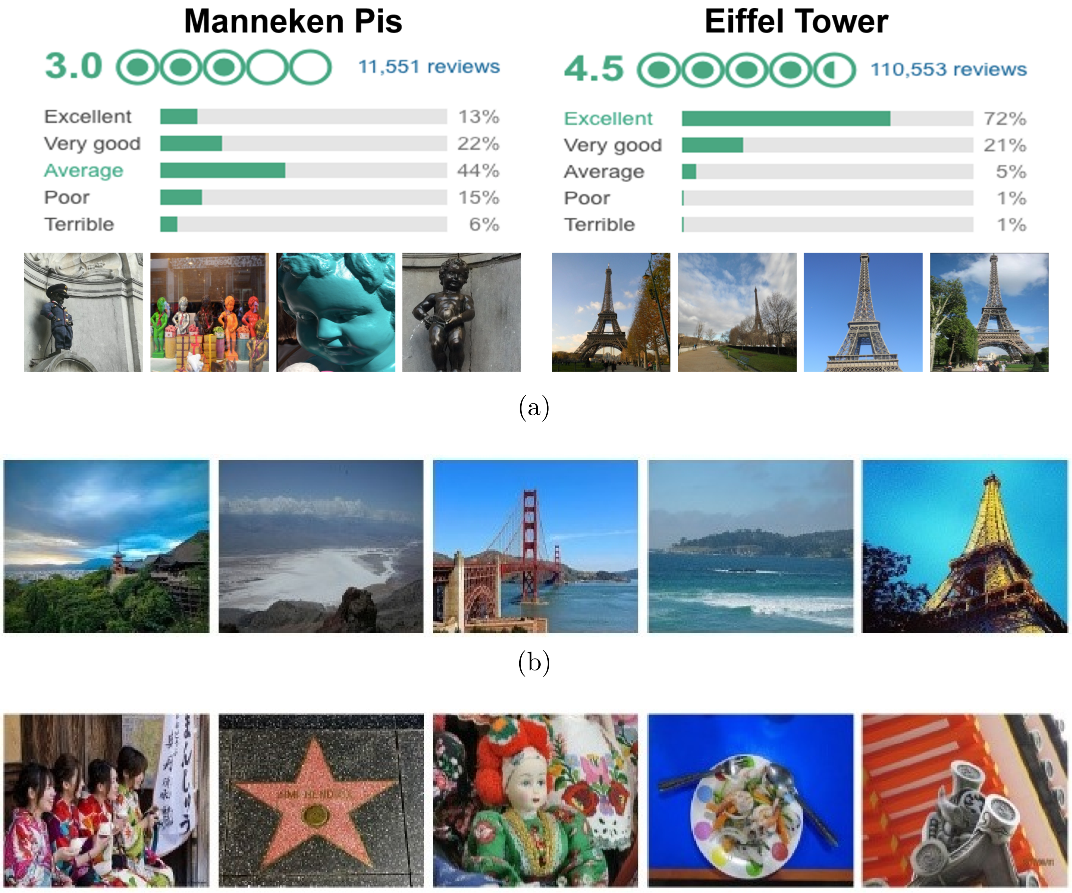
Chapter 5

The Relationship Between Tourist Satisfaction and Photo-shooting Behaviors

This part aims to estimate the broaden-and-build theory (Fredrickson, 2004; Fredrickson & Branigan, 2005) by constructing NODS. The tourism social media platforms (e.g., *Flickr* and *TripAdvisor*) collect big data of rating of travel experience (satisfaction) and the travel photos related to the experience. The tourist satisfaction with their attentional scopes revealed by travel photos is assumed as a living example of the broaden-and-build theory, as shown in Figure 5.1a. Thus, the target behavior of such NODS is detecting the wide or narrow shooting behaviors which reveal the attentional scopes, named wide-view vs. narrow-view. In Chapter 4, we proposed CFCNN model to classify the view-types from travel photos. The experimental results showed CFCNN obtained a sufficient accuracy and time performance for this classification task from travel photos. In Chapter 5, the CFCNN is used to examine the attentional scopes from large numbers of travel photos and construct the NODS including tourist satisfaction and photo-shooting behaviors for refining the broaden-and-build theory in a natural setting.

5.1 Introduction

Recent advances in machine learning technologies in conjunction with big data from social media platforms offer psychologists an unprecedented opportunity to test psychological theories outside the laboratory. For example, the sequential



(a)

(b)

(c)

Figure 5.1: Examples of travel experience ratings (satisfaction) and the travel photos with different shooting behaviors. (a) Two examples of scenic spots with experience ratings and travel photos. The travel photo examples of (b) wide-view and (c) narrow-view, i.e., wide or narrow shooting behaviors which reveal the different attentional scopes of photographers.

dependence functions in higher-order cognition were investigated on millions of online reviews posted on Yelp (Vinson et al., 2016), a machine learning model trained on a standard corpus of online text can result in human-like semantic biases (Caliskan et al., 2017). Emerging studies suggested that NODS could be used as a complement to traditional laboratory paradigms and could refine theories (Goldstone & Lupyan, 2016; Griffiths, 2015; M. N. Jones, 2016; Paxton & Griffiths, 2017). In this chapter, we aim to demonstrate a case study to construct NODS for testing the broaden-and-build theory (Fredrickson, 2001, 2004) in the wild.

According to the broaden-and-build theory, positive emotions globalize the attentional scope of the observer and result in the processing of a global scene picture, i.e., a big picture (Fredrickson & Branigan, 2005; Gu et al., 2017; Ji et al., 2019), while negative emotions localize attentional scope and induce the processing of local elements. Conversely, broadening attention leads to positive emotions, and narrowing attention leads to negative emotions (Niedenthal & Kitayama, 2013; Srinivasan & Hanif, 2010). The individual mood (internal factor) might influence different perceptions and prompt different photo-shooting behaviors, i.e., wide-view vs. narrow-view. Nonetheless, direct studies to verify and practically adopt such psychological theory in photo-shooting behaviors are scarce, with only a few recent attempts to suggest image view-type for scene aesthetic evaluation (Bartie & Mackaness, 2016) and sightseeing quality assessment (Zhuang et al., 2014), i.e., image view-type related to the environment (external factors). Moreover, the broaden-and-build theory is supported by many laboratory experiments with restricted environments, such as a limited number of subjects and simple stimuli, forming a low external validity. Thus, we intend to test this theory in the tourism scenario with NODS. A good fit of the theory using natural behaviors would lead to demonstrable progress (Roberts & Pashler, 2000), i.e., improving the persuasion of the theory and building the conjunction with practical applications.

For testing the relationship between emotion and attentional scopes, the measurement or manipulation of the attentional scopes is crucial for experimental design. Manipulating the attentional scope often aims to study the effect of visual attention affects emotion (Gu et al., 2017; Ji et al., 2019), where the attentional scope is the independent variable. When the attentional scope is the dependent variable, it often requires measuring the attentional scope. The methods for measuring or manipulating the attentional scopes sometimes are similar. The most commonly used methods for both measurement and manipulation are Navon stimuli (Ji et al., 2019; Navon, 1977) and Kimchi and Palmer stimuli (Fredrickson & Branigan, 2005; Kimchi & Palmer, 1982), as shown in see Figures 1.1 and 5.2a, respectively. For Kimchi and Palmer stimuli, the participants were asked to choose one from the two comparison figures (bottom), which is more similar to the target figure (top).

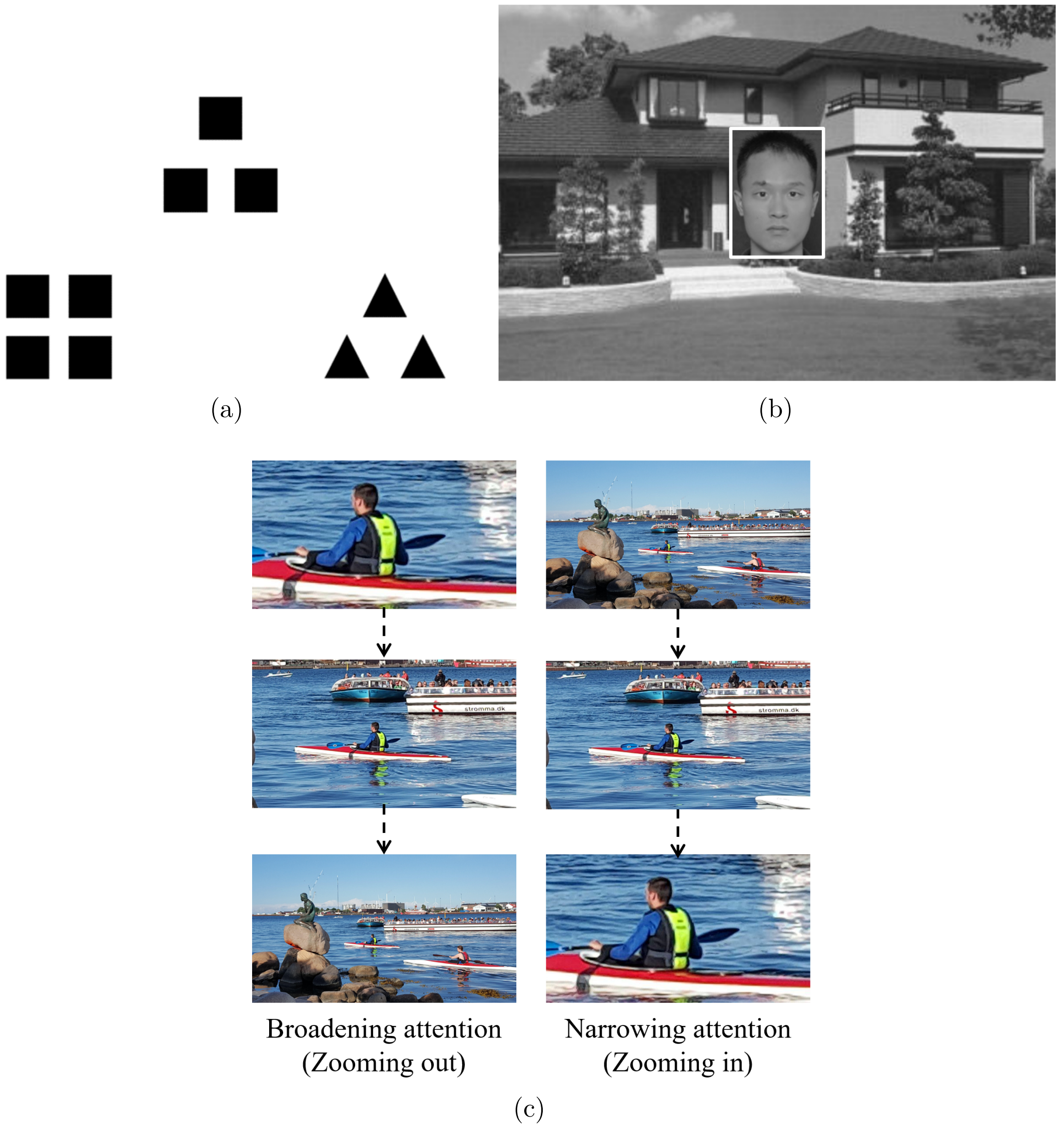


Figure 5.2: The redrawn psychological stimuli used for manipulating or measuring the attentional scopes in the psychological studies, which intended to explore the relationship between emotional states and attentional scopes. (a) The Kimchi and Palmer stimuli used for measuring the attentional scopes in a behavior study (Fredrickson & Branigan, 2005); (b) a visual stimulus applied for measuring the attentional scopes in a brain imaging study (Schmitz et al., 2009); and (c) scene-viewing stimuli employed for manipulating the attentional scopes in Gu et al. (2017). Another two stimuli (i.e., Novan stimuli and global-local landscape stimuli) can be found in Figures 1.1 and 1.2.

Suppose participants select the left one as most similar, that is to say, they are focusing on the local detail elements, i.e, narrowed attention. In one sense, Kimchi and Palmer stimuli are essentially a variant of Navon stimuli (Goodhew, 2020). Navon stimuli often use letters as the structural elements, whereas they are always shapes for Kimchi and Palmer stimuli. The metric using Navon stimuli or Kimchi and Palmer stimuli is often the selection of the behavioral response. Recently, psychologists designed physiological metric for measuring attentional scopes. For example, in a brain image (fMRI) study, Schmitz et al. (2009) designed a visual stimulus containing a scene of a house (place information) with a face image like Figure 5.2b is given to the participants, and the task is to answer if the face is male or female. The reason for such design is that the fusiform face area of the human brain specifically activates to face information while the parahippocampal place area specifically activates to place information. When participants attended to central face information, the surrounding place would be unattended. Thus, the physiological metric of different attentional scopes was a valence-dependent change in the magnitude of parahippocampal place area response to place information.

In addition, psychologists designed the stimuli (for manipulating attentional scopes) using realistic images or scenes (Gu et al., 2017; Ji et al., 2019) to improve the generalizability of the emotion-attention relationship. Ji et al. (2019) designed global-local landscape stimuli as shown in Figures 1.2. The participants were being asked to look at the entire landscape image (left) to inducing the broadening attention, while focus on the area delimited by a box (and ignore the rest of the image) to inducing the narrowing attention. After this phase, participants were being instructed to describe in a few sentences what they saw. This study suggested that the landscape stimuli showed similar affective valence to Navon stimuli for manipulating attentional scopes. Moreover, Gu et al. (2017) designed two scene-viewing stimuli. For the first scene-viewing stimuli, as shown in Figure 5.2c, researchers isolated original shots from several documentaries and then converted the zoom-in (or zoom-out) scenes into corresponding zoom-out (or zoom-in) scenes. The zoom-in scenes are expected to narrow the attention since more details would be present in the central area, while the zoom-out scenes would broaden the attention since new scenes would be present in the peripheral area. Furthermore, they designed another scene-viewing stimuli — the real-world distant and close-up scenes — for manipulating attentional scopes. Specifically, the participants (depressed) were asked to watch distant or proximal scenes 20 minutes a day and write a description of the scenes. After eight weeks, the participants significantly reduced depressed mood in broadening attention training.

In Chapter 4, we proposed CFCNN to classify the attentional scopes of tourist (narrow-view vs. wide-view) from travel photos. Similar to the study (Gu et al., 2017) which used scene-viewing stimuli to manipulate the attentional scope, we use image view-type to measure the attentional scope of tourist. In this chapter, we construct two NODS from two famous social media platforms — *Flickr* and *TripAdvisor*. The NODS includes the attentional scopes and experience rating. That is to say, CFCNN is used to measure the attentional scopes, and the experience rating represents the tourist satisfaction. We estimate the relationship between proportion of wide-view and the experience rating. For 94 scenic spots, there is a significant relationship between average experience rating and proportion of wide-view. This suggests that emotional states may influence the decision-making of tourists in composing specific camera viewpoints. Tourists seem to prefer shooting wider-views at scenic spots with higher experience rating. Our finding is consistent with the broaden-and-build theory. Meanwhile, in contrast to the traditional laboratory paradigms, our method substantially increases the numbers and diversity of participants by exploiting machine learning methods to construct NODS with the vast amount of behaviors from social media platforms. Furthermore, we show the found relationship has potential significance for real-world applications, i.e., point-of-interest (POI) discovery and sightseeing value assessment.

5.2 Experiment 1

In Experiment 1, we construct a NODS for testing the broaden-and-build theory. The orginal dataset includes 39,099 geotagged images collected from *Flickr* of 30 popular scenic spots across Asia, Europe, North America, and Oceania. To construct NODS, we use CFCNN to detect the natural photo-shooting behavior, wide- or narrow-view, inferred from a single photo. Then, we calculate the percentage of wide-view (named wide percentage $\in [0, 1]$) concerning each scenic spot, and analyse the relationship between wide percentage and tourist satisfaction. The tourist satisfaction of each scenic spot is represented by the average score of travel experience ratings for each spot provided by *TripAdvisor*.

In this experiment, we find a significant difference of wide percentage exists between 15 high-rated and 15 low-rated scenic spots (unpaired t-test: $t(28) = 8.460$, $p < 0.0001$). At high-rated scenic spots, tourists prefer to take wide-view photos to capture wide landscapes (e.g., mountains, lakes, tall buildings), while for low-rated scenic spots, the preference appears to be moderate or inclined towards narrow-view, typically containing small elements (e.g., food, people, activities).

In the tourism industry, large geotagged image collections are currently used to discover, recognize, and reconstruct the landmarks of scenicness and scenic spots (Arase et al., 2010; Workman et al., 2017; Zhuang et al., 2017). For example, Arase et al. (2010) presented an effective method to detect tourists' frequent trip patterns as well as produce descriptive tags that characterize the trip pattern using 5.7 million geotagged photos. Thus, we map the travel photos with different shooting behavior into the scenic spot location, and discuss the potential application of point of interest (POI) discovery, based on found relationship between tourist satisfaction and their photo-shooting behaviors.

5.2.1 Datasets

We made use of *Flickr* to gather the required image data, and *TripAdvisor* to obtain the required tourist satisfaction. The outcome was a collection of 39,099 photos and tourist satisfaction taken from 30 scenic spots. The data collection was consisted of the following three steps.

Scenic spots collection. We collected 12,482 scenic spots from *TripAdvisor* with tourist satisfaction. The tourist satisfaction ($R^c \in [1, 5], c \in [1, 12482]$) of each scenic

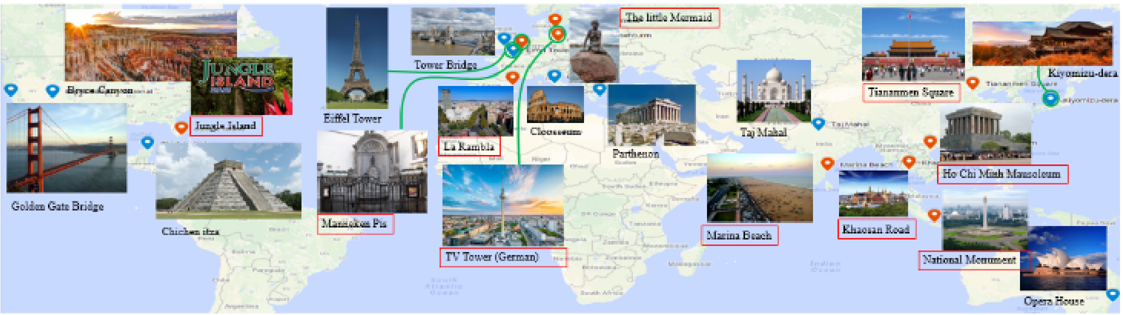


Figure 5.3: The geo-locations and representative photographs of 20 scenic spots over 4 continents, i.e., Asia, Europe, North America, and Oceania. The high-rated scenic spots are denoted by red, while the low-rated scenic spots are denoted by blue.

spot is the average score of experience rating by various tourists, who provide review comments. The R^c is defined as,

$$R^c = \frac{\sum^{N_r^c} R_r^c}{N_r^c}, \tag{5.1}$$

where N_r^c represents the review comments number of c-th candidate, and $R_{rm}^c \in [1, 5]$ (5: 'Excellent, 4: 'Very good', 3: 'Average', 2: 'Poor', and 1: 'Terrible') represents the experience rating of different tourists at c-th scenic spot. Figure B.1 shows the number distribution of 12,482 scenic spots w.r.t. R^c. We note that most R^c range from 3 to 5 with a median of 4.275.

Scenic spots selection. We selected the 30 scenic spots based on the following five criteria: (1) Popularity: Recommended by top search engines, i.e., *TripAdvisor*, National Geographic, and Travel + Leisure; (2) Objectivity: Having at least 1,000 experience ratings regardless of language, age, gender, nationality, *etc.*; (3) Diversity: Keeping spot types as diverse as possible, but avoid religious and political places. (4) Independence: Having an appropriate distance from other spots to avoid cross-rating. Among the selected spots, 10 are high-rated, and 10 are lower-rated taken from 4 continents (named worldwide spot), whereas the remaining 5 high-rated and 5 lower-rated spots are specifically selected from within the state of California, USA (named regionwide spot) to facilitate a study based on a local region. The threshold of rating for a high-/low-rated spot was 4.275 (the median of R^c). Figure 5.3 shows the geolocations of the selected worldwide spots.

Photo collection. Thousands of geotagged photos from each spot were downloaded using the public *Flickr* API. Its geolocation and specified zone that was defined based on the type of structures of the spot. For instance, we used a circling zone for the Eiffel Tower, whose center is located at the tower, with a radius of 350 meters, whereas for the Hollywood Walk of Fame, a long strip covering the entire street is more reasonable. Firstly, we downloaded all the photos in these specified zones (306,000 photos in total). However, since the number of photos in each spot is unbalanced. Secondly, a subset with 39,099 photos for 30 spots was randomly selected. The details about the photo number of 30 scenic spots (N_p^s, where $s \in [1, 30]$ is the serial number of these spots) can be found in Table B.1.

To test the validity of the proposed CFCNN for attentional scope analysis in scenic spots, the 9,923 photos of 10 regionwide spots were manually annotated as narrow-/wide-view by a leading informatics researcher (a male aged 27 years) and then again by one independent person (a male aged 23 years) who had been fully briefed about the definition of view-type. The annotations results of the 10 regionwide spots can be found in Table B.1, where N_{wp}^s (Human) represent the number of wide-view photos in s-th scenic spots. The experience ratings (R^s) of the 10 regionwide spots and 20 worldwide spots are shown in Table 5.1 and Table 5.2, respectively.

5.2.2 Results and discussion

5.2.2.1 Machine vs. human

To test the validity of the CFCNN for photo-shooting behavior analysis in scenic spots, we employed the CFCNN on the annotated photos of the regionwide scenic spots to compare machine (CFCNN) performance with human's. The photos were fed to the CFCNN for classification and gathered $P_w^s \in [0, 1]$, which represents percentage of the wide-view photos of s-th scenic spots and defined as,

$$P_w^s = \frac{N_{wp}^s}{N_p^s} \tag{5.2}$$

Table 5.1: The statistical results of 10 regionwide scenic spots from California, USA. The R^s represents averaged experience rating of the corresponding spot. The P^s_{wp} of Human or CFCNN represents the percentage of wide-view photos annotated by participants or classified by CFCNN. The accuracy represents CFCNN classification performance for each scenic spot using human annotation as ground-truth.

No. (s)	Spot Name	R^s	Human	CFCNN	
			P^s_{wp}	P^s_{wp}	Accuracy
1	Hollywood Walk of Fame	3.383	0.182	0.209	0.9029
2	Haight-Ashbury	3.950	0.291	0.312	0.8468
3	Cannery Row	3.926	0.379	0.419	0.8340
4	Fisherman's Wharf	4.074	0.389	0.431	0.8543
5	Venice Beach	3.798	0.486	0.482	0.8499
6	Torrey Pines State Natural Reserve	4.733	0.627	0.643	0.8383
7	Fort Rosecrans Cemetery	4.760	0.648	0.631	0.8519
8	Carmel City Beach	4.716	0.735	0.727	0.8764
9	Yosemite Valley	4.708	0.864	0.751	0.8297
10	Dante's View	4.796	0.872	0.878	0.9282

where N^s_p and N^s_{wp} represent the number of photos and wide-view photos in s-th scenic spots.

Firstly, using human annotation as the ground-truth, the classification accuracy of the model records a notable 0.8297-0.9282 (as shown in Figure 5.1). These results for different scenic spots are close to the testing accuracy (0.8813) on *CognitionShoot*. Considering there are certain ambiguous photos, these results suggest the CFCNN has considerable validity for calculating the wide percentage for different scenic spots. Secondly, we obtained a wide percentage for each spot as human annotations (as shown in Table 5.2). Figure 5.4 shows the comparison of the wide percentage for regionwide scenic spots between human and CFCNN. Remarkably, the performance of the CFCNN model closely resembles human annotation (Pearson's correlation coefficient: $r = 0.987, p = 1.05 \times 10^{-7}$; Mean absolute error: $MAE = 0.052$). This result suggests the CFCNN is effective for photo-shooting behavior analysis for different scenic spots. Therefore, we use the CFCNN in the experiments of this chapter.

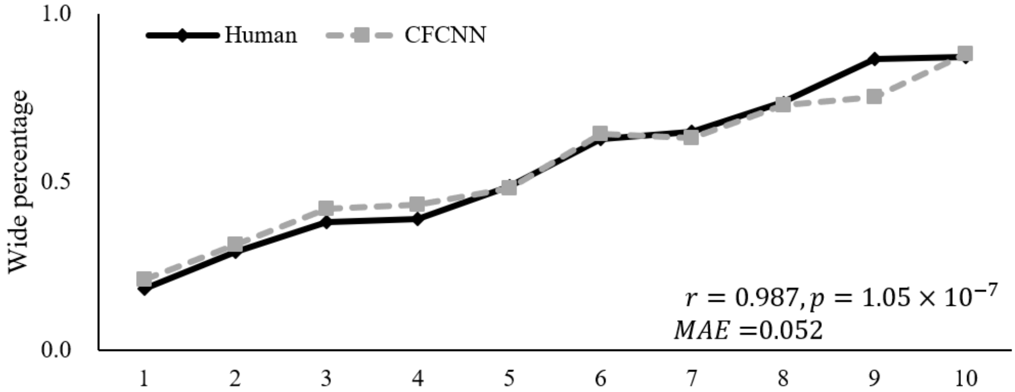


Figure 5.4: Comparison of calculating the wide percentages (P_{wp}^{s}) for 10 regionwide scenic spots between human and machine (CFCNN).

5.2.2.2 Photo-shooting behavior analysis

The CFCNN model is used to predict the photo-shooting behaviors (wide- or narrow-view) of a single image. We analyze the predicted results for the images of 20 worldwide spots and get P_{wp}^{s} for each spot as shown in Tables 5.2. It is noted that the spots with different ratings have different ratios. Especially, the high-rated scenic spots have a larger wide percentage than low-rated scenic spots. For example, scenic spots selected from worldwide, such as Khaosan Road ($R^{15} = 3.810$; $P_{wp}^{15} = 0.173$) has a lower wide percentage than Bryce Canyon National Park ($R^{30} = 4.908$; $P_{wp}^{30} = 0.836$) with a difference of 0.663. Similar results can be found in regionwide spots, such as Danta's View ($R^{10} = 4.796$), with the wide percentage that is 0.689 higher than the Hollywood Walk of Fame ($R^{1} = 3.383$). More importantly, such gaps between high and low-rated spots exist in comparing spots of similar nature. For example, seaside type locations such as Marina Beach ($R^{14} = 3.774$) and Carmel City Beach ($R^{8} = 4.716$) show a difference of 0.214 in favor of the higher-rated spot. Likewise, the gap between Television Tower ($R^{18} = 4.003$) and Eiffel Tower ($R^{24} = 4.594$) is 0.163 inclined towards the Eiffel Tower. Some other same scenery type comparisons can be found in Figure 5.5, including square areas, towers, and seasides.

Our hypothesis is that the photo-shooting behaviors are influenced by the emotional states of tourists. For comparison, we access the preferences of photo-shooting behaviors in an entirely random mode. An independent dataset of 1,000

Table 5.2: The statistical results of 20 scenic spots from worldwide. The P^s_{wp} represents the percentage of wide-view photos classified by CFCNN.

No. (s)	Spot name	Country	R^s	P^s_w
11	Manneken Pis	Belgium	3.260	0.196
12	The Little Mermaid Den Lille Havfrue	Denmark	3.418	0.460
13	Jungle Island	USA	3.769	0.264
14	Marina Beach	India	3.774	0.522
15	Khaosan Road	Thailand	3.810	0.173
16	Las Ramblas	Spain	3.901	0.333
17	Ho Chi Minh Mausoleum	Vietnam	3.914	0.551
18	Television Tower	Germany	4.003	0.608
19	Tiananmen Square	China	4.027	0.601
20	National Monument	Indonesia	4.113	0.367
21	Kiyomizu dera Temple	Japan	4.407	0.605
22	Tower Bridge	England	4.587	0.741
23	Chichen Itza	Mexico	4.589	0.730
24	Eiffel Tower	France	4.594	0.771
25	Colosseum	Italy	4.674	0.801
26	Sydney Opera House	Australia	4.674	0.797
27	Parthenon	Greece	4.687	0.719
28	Golden Gate Bridge	USA	4.697	0.779
29	Taj Mahal	India	4.790	0.819
30	Bryce Canyon National Park	USA	4.908	0.836

photos was collected from the YFCC100M dataset (Thomee et al., 2016) without using geotag or any other keywords. YFCC100M is a subset of *Flickr* containing 100 million data, which has travel photos and a super-wide diversity. A close investigation of those random photos revealed that the vast majority of narrow-view photos were cliché photos of everyday life. The random photos represent the “normal behavior” in composing wide vs. narrow view photos. Thus, the wide percentage of these photos serves as a baseline and will be compared against the photos collected from scenic spots. The 1,000 photos are annotated by two informatics researchers (male aged 27 years and 43 years) and obtain a wide percentage of 0.397.

The wide percentages of different sightseeing groups are compared, as shown in Figure 5.6. A significant difference (unpaired t-test: $t(28) = 8.460$, $p < 0.0001$) exists between 15 low-rated spots (Mean±SD: 0.387 ± 0.148) and 15 high-rated spots (Mean±SD: 0.756 ± 0.082). Analogously, as shown in Figure 5.6, a significant

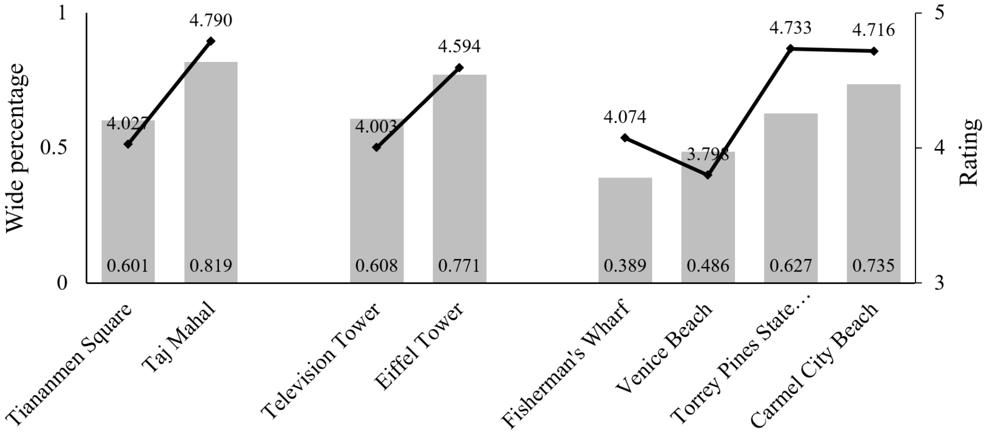


Figure 5.5: The comparison of wide percentage and experience ratings of same scenery type, i.e., square areas (Tiananmen Square vs. Taj Mahal), landmarks (Television Tower vs. Eiffel Tower), and seasides (Fisherman's Wharf and Venice Beach vs. Torrey Pines State Natural Reserve and Carmel City Beach).

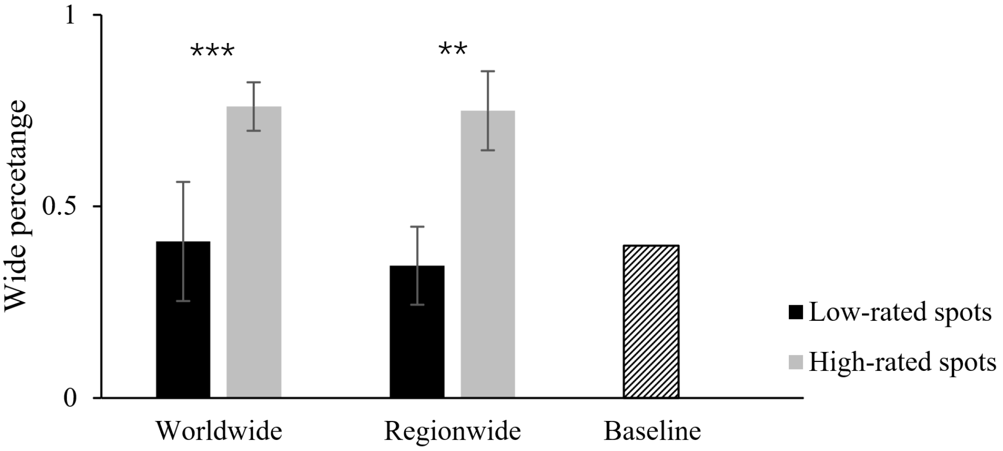


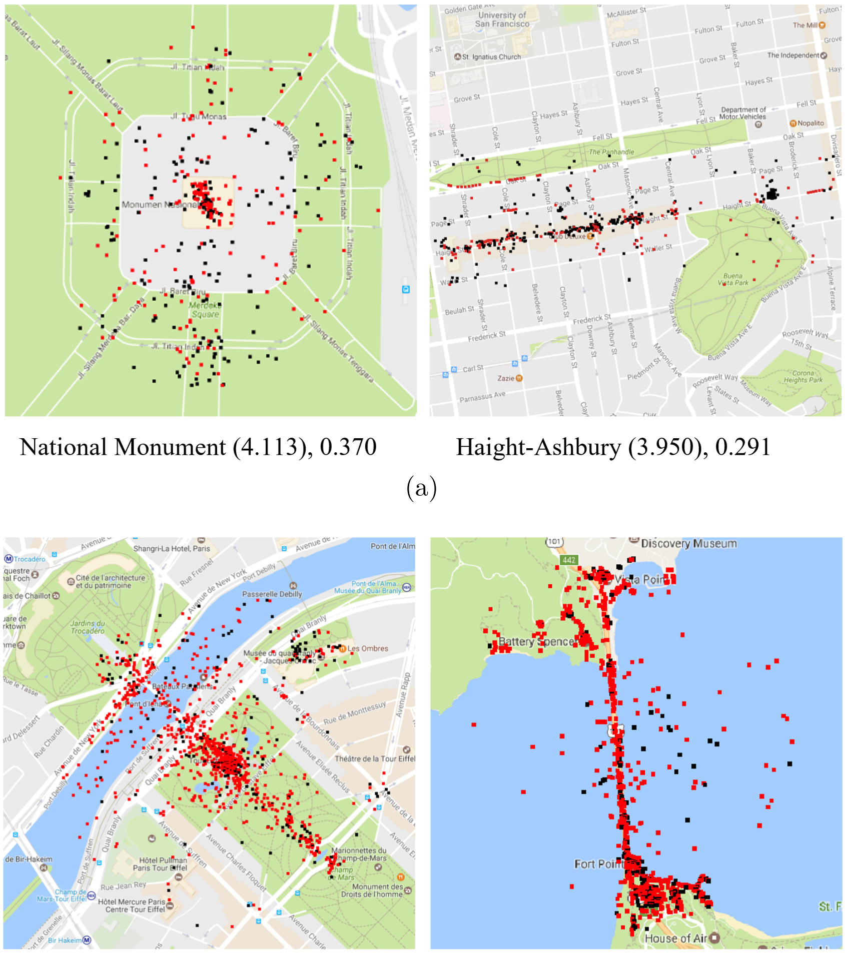
Figure 5.6: Compare the wide percentage at different sightseeing groups, 20 worldwide spots, and 10 regionwide spots. The wide percentage at regionwide scenic spots used the annotations of participants for analysis. The baseline is the wide percentage of 1,000 photos annotated by participants, which were randomly selected from the YFCC100M dataset (Thomee et al., 2016). ***: $p < 0.0001$; **: $p < 0.001$.

difference exists in the each group of worldwide (unpaired t-test: $t(18) = 6.302$, $p < 0.0001$) and regionwide (unpaired t-test: $t(8) = 5.548$, $p < 0.001$) spots, respectively. Moreover, the wide percentage of the baseline is 0.397 closely resembles the wide percentage of low-rated spots. We conjecture that this similarity might be ascribed to when a tourist is unsatisfied that travel experience, they would show the similar photo-shooting behavior or attentional scopes to their daily life. The alternative reason is that the data collection from *Flickr* does not distinguish between local people and tourists. The low-rated spots may contain more photos shared by local people because of the relatively lower popularity.

We note that wide-view photos usually capture large-scale landscape (e.g., sea-views, tall buildings, and mountains), while narrow-view photos typically contain small elements (e.g., food, people, activities). The intuitive belief for the chosen perspective in photography is based on location; for example, wide natural landscapes like beaches inspire more wide-view photos. However, our results suggest that the spot with a high rating would inspire people to take more wide-view photos to record the overall scenery, including the landscape and structures, while at low-rated spots, people tend to take more narrow-view photos to record more mundane elements such as food and activities, which is similarly done on a regular non-vacation day and location. This finding is coincidentally consistent with the broaden-and-build theory (Fredrickson & Branigan, 2005), suggesting that tourists' behavior in capturing photos while traveling to scenic spots somehow adheres to this theory.

5.2.2.3 Mapping the photo-shooting behaviors

To discover the piratical application of the relationship between photo-shooting behavior and experience rating, we analyzed the spatial distribution of photo-shooting behaviors across sub-regions of a scenic spot. We use the geotagged information of a photo to find out the location in which the tourist had taken and then "place" the shooting behaviors of this photo into a regional map of the spot. Figure 5.7 shows the photo-shooting behavior maps of different spots (red and black for wide- and narrow-view photos, respectively). Based on these behavior maps, we noted that the behavior map exhibits the tourists' attention distribution of the spots and may partly reflect the "spatially emotion" area in the spot. For example, low-rated spots have some locals with a high density of wide-view photos, whereas the high-rated spots have some locals with a high density of narrow-view photos. For example, in the map of the National Monument in Figure 5.7b, there is a concentration of red



National Monument (4.113), 0.370   Haight-Ashbury (3.950), 0.291

(a)

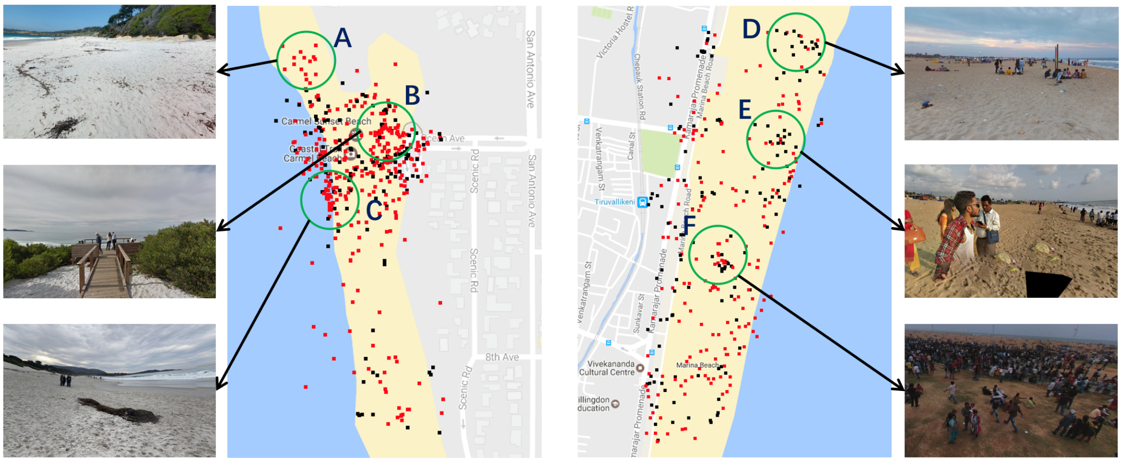
Eiffel Tower (4.003), 0.771   Golden Gate Bridge (4.697), 0.779

(b)

Figure 5.7: The photo-shooting behavior mapped on the scenic spot's location: (a) 2 low-rated scenic spots, and (b) 2 high-rated scenic spots. Each colored dot represents a geotagged photo with a color indicator of its shooting behaviors (red: wide-view, black: narrow-view).

dots next to the central area of the main structure, which means this area provides a right scenic view for tourists, and the emotional states of the tourists when they visit here may be more positive than visit other areas.

Furthermore, a density analysis of photo-shooting behaviors by comparing two spots of the same scenery type also yields more insights, ie., a comparison of the beach scenic spots, namely Carmel City beach (4.716) and Marina Beach (3.774), as shown in Figure 5.8. Both spots have famous sea-views and can provide people with



Carmel City Beach (4.716), 0.735

Marina Beach (3.774), 0.521

Figure 5.8: The photo-shooting behavior mapped on the scenic spot's location with Point-of-interests (POI). The circles represent the POI determined by the density of the wide percentage. POI-A, POI-B, and POI-C at Carmel City beach have high wide percentages (> 0.80); POI-D, POI-E, and POI-F at Marina Beach have low wide percentages (< 0.45). Each circle directs an example photo cropped from Google Street View (https://www.google.com/streetview/) are showed for each corresponding POI.

enough areas of opportunity to capture wide-view photos of the beaches. However, their wide percentages are entirely different from each other, with a difference of 0.214. The density analysis on the map allows us to identify three popular point-of-interests (POIs) for each map, as shown in Figure 5.8. In these POIs, we found that the three regions of Carmel City beach (POI-A, POI-B, and POI-C) have high percentages of wide-view photos (> 0.80), while the three regions of Marina Beach (POI-D, POI-E, and POI-F) have low percentages of wide-view photos (< 0.45). To get an unbiased understanding of the environment around these three areas, we surveyed the regions using Google Street View and cropped sample images on a random point at the corresponding POI. As shown on both sides of Figure 5.8, we can see that Carmel City beach is clean and tidy with white sand and an Observation deck (POI-C). These examples photos show that Carmel City beach provides people the right places to enjoy the beach scenery. On the other hand, Marina Beach is crowded with people (POI-E and POI-F) and a polluted environment (POI-D and POI-E). These elements give a bad impression to tourists and distract their attention even though the seascape and beach are beautiful.

In conclusion, this experiment found a significant difference of shooting behavior (wide- and narrow-view) choice between high-rated and low-rated scenic spots. Concretely, people visiting high-rated scenic spots tend to shoot wide-view photos, while at low-rated spots, they are more likely to capture elements from a narrow-view. This preference may be influenced by the psychological notion that positive emotion broadens visual attention and triggers a choice of wide-view photos. Moreover, we map the photo-shooting behaviors of each spot and found it has potential significance to tourism, such as for sightseeing planning and POI discovery.

5.3 Experiment 2

Experiment 1 suggested that the photo-shooting behaviors are correlated to the experience ratings of scenic spots provided by tourists. However, there are two limitations. Firstly, when collecting the dataset, we did not distinguish between local people and tourists who took photos. That is, the people who shared the photos are not the tourists who give experience ratings. In addition, photo-shooting behaviors and tourist satisfaction may correlate to other confounding factors, e.g., environmental factors. To alleviate these issues, we collect an alternative NODS only from *TripAdvisor*, where the photos and experience ratings are shared by the same tourists. The new dataset consists of 549,772 photos and experience ratings from a collection of 94 scenic spots. To rule out the confounding factors, we use subgroup analysis to divide the scenic spots into different continents and scenery types. Moreover, we divide the photos into different visual contents, i.e., green, water, and building perceptions. Under the scenic spot and visual content conditions mentioned above, the experimental results still indicate a significant correlation between tourist satisfaction and photo-shooting behaviors.

To evaluate the application of the photo-shooting behaviors, we compare its performance for the 'visiting worth' assessment of scenic spots with the current image content-based methods. The visiting worth assessment of sightseeing spots is first proposed by (Zhuang et al., 2014), which aimed to “discover the obscure scenic spots that are less well-known while still worth visiting”. The visiting worth is different from the popularity, consisting of many aspects such as “visual attractiveness, history, related government policies, and the surrounding facilities” (Ge et al., 2019). The current methods mainly focus on the visual attractiveness of the visual object (Ge et al., 2019; Y. Shen et al., 2018) based on the visual content analysis. The experimental result shows that the wide percentage is a better metric compared to recent image content-based methods for inferring tourist satisfaction. This suggests the association between tourist satisfaction and the wide percentage could be used for sightseeing value assessment.

5.3.1 Materials and methods

5.3.1.1 Photo dataset

We collected the travel data from *TripAdvisor* where the hosted scenic spots provide travel experience ratings, travel photos, and reviews. The scenic spots were selected based on the same criteria in Experiment 1 — popularity, objectivity, generality, and independence. In addition, we added a criterion the photo number of each spot required more than 300. Thus, based on the available 12,000 tourism destinations provided by *TripAdvisor*, 94 scenic spots were selected. Then, we created a dataset consisted of 947,174 travel photos associated with these spots from *TripAdvisor*. After removing the photos without experience rating, we gathered a subset with 549,772 images to construct the new NODS. The details of the location information of each scenic spot and photo number can be found in Table B.2 and Table B.3. These data were collected in March 2020.

To construct the NODS, we applied the CFCNN to predict the view-type of a single image. The tourist satisfaction of scenic spots is calculated by the ratings of the photo providers only. Table B.3 shows the experience rating (R^i) and the wide percentage (P^i_{wp}) of each scenic spot. The $R^i \in [1,5]$ and $P^i_{wp} \in [0,1]$ is defined as,

$$R^i = \frac{\sum^{N^i_{pp}} R^i_{pp}}{N^i_{pp}}, P^i_{wp} = \frac{N^i_{wp}}{N^i_p}, \tag{5.3}$$

where N^i_{pp} and R^i_{pp} represent the photo providers' number and their rating of i-th scenic spots ($i \in [1,94]$), respectively. The N^i_p and N^i_{wp} represent the number of photos and wide-view photos at i-th scenic spots, respectively.

5.3.1.2 Image content-based methods for visiting worth assessment

The general way of using the image-content to estimate the visiting worth of sightseeing spots consists of the following steps (Ge et al., 2019). (1) Selected (manually (G. Li et al., 2019) or automatically (Y. Shen et al., 2018)) the representative images of the sightseeing spots, removing the elements that have no relation with the attractiveness of a sightseeing spot, such as the sky, cars; and then (2) Estimated the image quality (Y. Shen et al., 2018) or attractiveness (Ge et al., 2019) based on image-content based features.

We compare the wide percentage with two image-content based methods for visiting worth assessment as follows. (1) *Visual depth estimation*: Visual depth reflects the visibility of the landscape, which is proposed by Y. Shen et al. (2018) for visiting worth assessment. The visual depth score is calculated based on the GIST-based method (Oliva & Torralba, 2001) to estimate the visibility of a single image. The GIST-based method is implemented by the source code provided by (Oliva & Torralba, 2001). The score range of visual depth is from 1 to 5. (2) *Natural attractiveness estimation*: (Workman et al., 2017) proposed a method to predict a single image's attractiveness. The training dataset includes 217,000 outdoor images with crowdsourced ratings of natural beauty. The method for natural attractiveness estimation is implemented by the code provided by Workman et al. (2017). The attractiveness score is ranged from 1 to 10.

To capture the representative images of each sightseeing spot, we reproduced the method introduced by Y. Shen et al. (2018) to get the selected images with frequency features in the high-level layer of a VGG model (Simonyan & Zisserman,

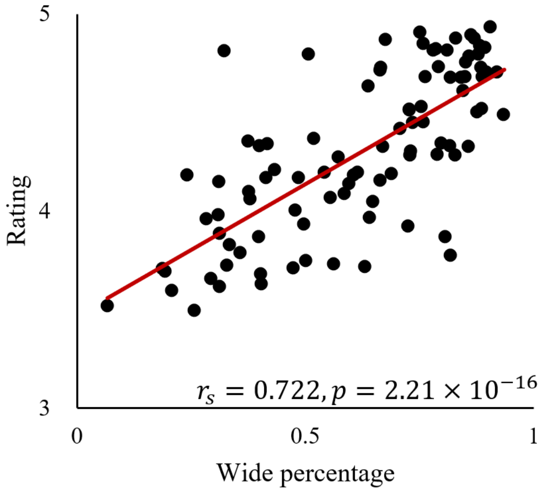


Figure 5.9: The correlation analysis between the wide percentages and the experience ratings of 94 scenic spots. The details of the experience rating and the wide percentage of each scenic spot can be found in Table B.3.

2014). For each spot, we got 20 representative images for visual-depth and visual attractiveness estimation and used the average visual-depth or visual-attractiveness score of these images to represent the estimations of visiting worth of corresponding spots.

5.3.2 Results and discussion

5.3.2.1 Photo-shooting behavior analysis

Spearman's rank-order correlation coefficient is applied to analyze the relationship between the wide percentage and experience ratings. Spearman's rank-order correlation (r_s) is a nonparametric measure, indicating the agreements between the two sets of rankings. Figure 5.9 shows a notable positive correlation between wide percentage and experience ratings (r_s = 0.722, p <0.00001). This preference is ascribed to the broaden-and-build theory that positive emotions broaden visual attention and trigger wide-view photo compositions.

On the one hand, the affective quality attributed to a place governs subsequent behavior and tourist gaze (Ittelson, 1976; Pan et al., 2014). On the other hand, tourists use photography to capture relationships with other people, places and

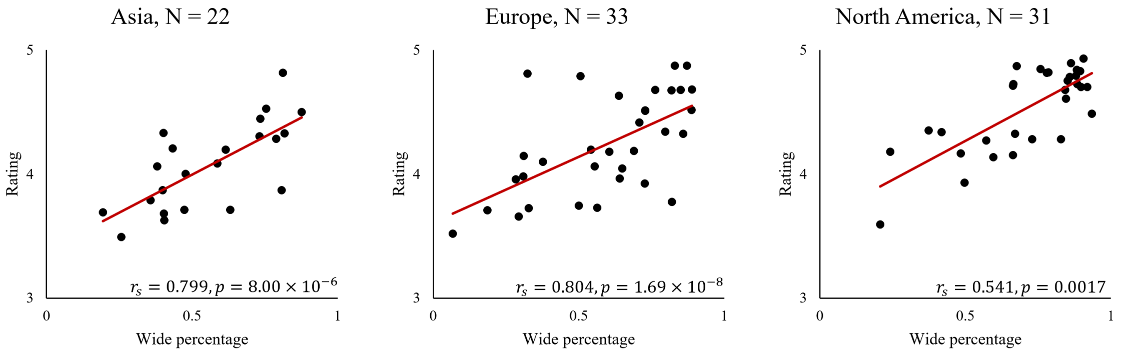


Figure 5.10: The correlation analysis between the experience rating and the wide percentage of scenic spots grouped by different continents, i.e., Asia (22), Europe(33), and North America (31). The details of the location information of each scenic spot can be found in Table B.2.

cultures (Edensor, 2000). Thus, some external factors, such as cultural and environmental factors, may influence photo-shooting behaviors. To rule out the confounding factors of cultural and environmental factors, we applied the subgroup analysis where the scenic spots were divided into different subgroups by various continents and scenery types. Furthermore, Pan et al. (2014) showed that the image dimension influences the tourist affect, such as travel photos feature natural resources are frequently associated with pleasant feelings. Thus, we classify the travel photos into conditions with different image contents (i.e., green perception, water perception, and building perception).

Firstly, the scenic spots are selected according to their locations, i.e., Asia (22), Europe (33), and North America (31), to control the cultural factor. The details of the location information of each scenic spot can be found in Table B.2. As shown in Figure 5.10, the spots in Asia, Europe showed significantly positive correlations between the wide percentages and experience ratings (Asia: $r_s = 0.799$, Europe: $r_s = 0.804$, both $p < 0.00001$). However, the spots in North America showed a relatively lower correlation between them ($r_s = 0.541$, $p < 0.005$).

Subsequently, the scenic spots are selected according to their types of attractions, i.e., parks (14), historical sites (15), landmarks (29), and neighborhoods (16). The classification of the scenery type can be found in Table B.2. As shown in Figure

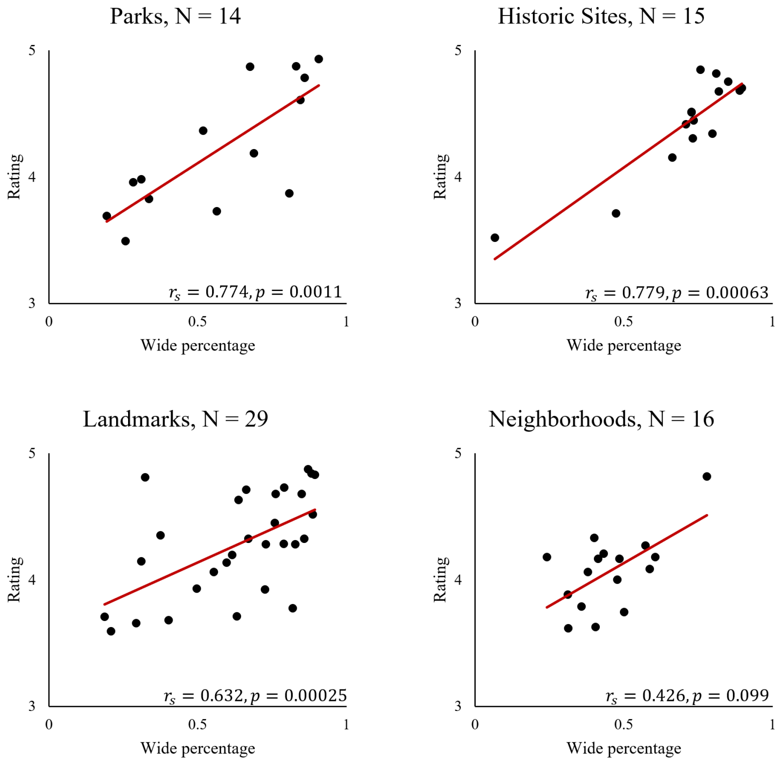


Figure 5.11: The correlation analysis between the experience rating and the wide percentage of scenic spots grouped by different scenery types, i.e., Parks (14), Historic Sites (15), Landmarks (29), and Neighborhoods (16). The classification of the scenery type can be found in Table B.2.

5.11, the historical sites and parks had significantly positive correlation between the wide percentage and experience rating (historical sites: $r_s = 0.779$, p<0.001; parks: $r_s = 0.774$, $p < 0.005$). However, landmarks and neighborhoods showed relatively lower correlation between them (landmarks: $r_s = 0.632$, $p < 0.0005$; neighborhoods: $r_s = 0.426$, $p = 0.099$). This result suggests that the environmental characteristics may affect the association between photo-shooting behaviors and experience rating.

Lastly, we divided the travel photos into various visual contents, i.e., green perception, water perception, and building perception. The green perception and water perception represent the natural resources, while the building perception represents the cultural resources. The method uses a well-trained CNN model provided by (L. Liu et al., 2016) to detect the 102 SUN attributes (Patterson et al., 2014) and then groups are divided based on the related attributes to green perception,

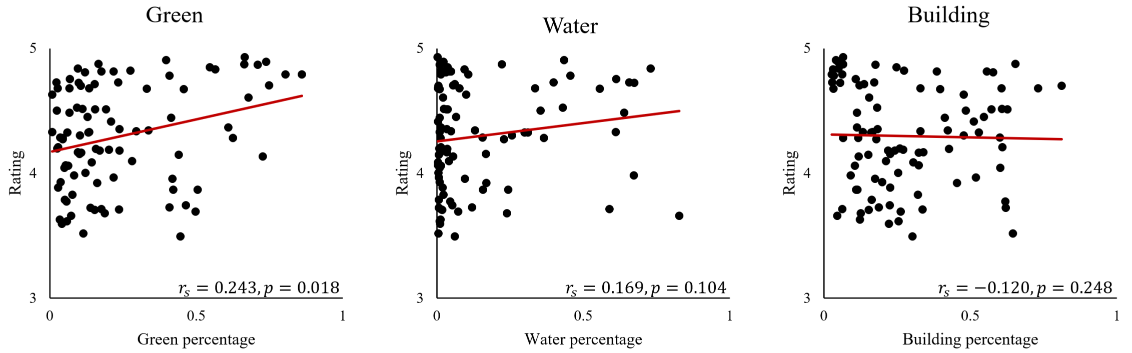


Figure 5.12: The correlation analysis between the experience rating and the percentages of different visual contents, i.e., green, water, and building perceptions. The method for the visual content analysis was introduced by L. Liu et al. (2016).

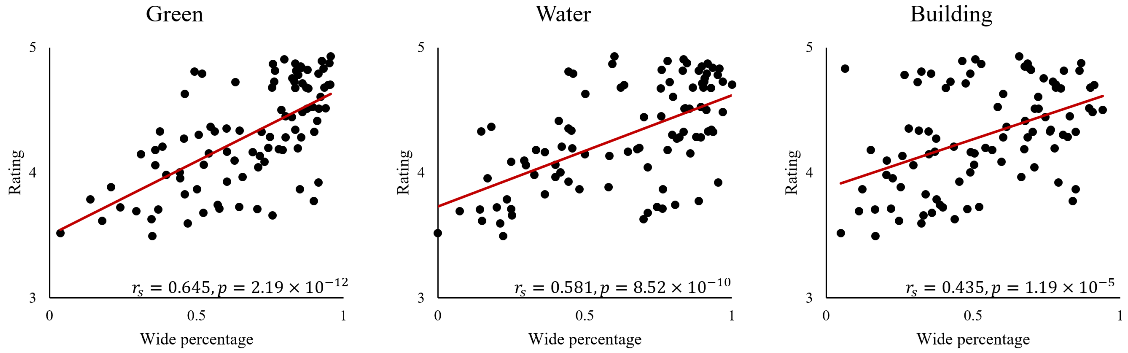


Figure 5.13: The correlation analysis between the experience rating and the wide percentages in different visual contents.

water perception, and building perception (the details of the method can be found in (L. Liu et al., 2016; Zhou et al., 2014)). Using this method, all 549,772 images had been detected as green perception, water perception, building perception, or others. Thus, we obtained the percentage of green perception (P^{ig}), water perception (P^{iw}), and building perception (P^{ib}) for each spot. The P^{ig}, P^{iw}, P^{ib}, are defined as,

$$P^{ig} = \frac{N_p^{ig}}{N_p^i}, P^{iw} = \frac{N_p^{iw}}{N_p^i}, P^{ib} = \frac{N_p^{ib}}{N_p^i}, \tag{5.4}$$

where N_p^{ig}, N_p^{iw}, and N_p^{ib} represents the numbers of photos are detected as green perception, water perception, and building perception at i-th scenic spots, respectively. The result of correlation analysis between the experience rating and the percentages of different visual contents is showed in Figure 5.12. We found the percentages of water perception and building perception do not correlate with experience ratings (water perception: $r_s = 0.169$, $p = 0.104$; $r_s = -0.120$, $p = 0.248$). While the percentage of green perception weakly correlate with experience rating ($r_s = 0.243$, $p = 0.018$). Then, we obtain the wide percentages of diffident visual contents, which are defined as,

$$P_{wp}^{ig} = \frac{N_{wp}^{ig}}{N_p^{ig}}, P_{wp}^{iw} = \frac{N_{wp}^{iw}}{N_p^{iw}}, P_{wp}^{ib} = \frac{N_{wp}^{ib}}{N_p^{ib}}, \tag{5.5}$$

where P_{wp}^{ig}, P_{wp}^{ig}, and P_{wp}^{ig} represents the wide percentage of green perception, water perception, and building perception at i-th scenic spots, respectively; N_{wp}^{ig}, N_{wp}^{iw}, and N_{wp}^{ib} represent the number of wide-view photos in N_p^{ig}, N_p^{iw}, and N_p^{ib} at i-th scenic spots, respectively. Figure 5.13 shows the wide percentages of diffident visual contents correlate with the experience rating (wide percentages of green perception: $r_s = 0.645$, $p < 0.00001$; water perception: $r_s = 0.581$, $p < 0.00001$; building perception: $r_s = 0.435$, $p < 0.00001$). This result reveals a notable correlation between the preference for wide-view photos and the high rating of scenic spots by considering the conditions of different visual contents.

5.3.2.2 Comparing with visiting worth assessment

In the previous research, informatics research often predicted the visiting worth of sightseeing based on the visual contents of the photos, i.e., the objective or inherent characteristics of scenic spots. In our proposal, we measure the attentional scopes of photos and refer them to tourist satisfaction. Considering that tourist satisfaction is an aspect of visiting worth, the attentional scopes may be a subjective metric to measure the visiting worth. Thus, we compare the wide percentage to visual attractiveness (Workman et al., 2017) and visual depth (Y. Shen et al., 2018) for inferring tourist satisfaction. As shown in Figure 5.14, the visual attractiveness and visual depth are significantly correlated to the tourist satisfaction (visual attractiveness: $r_s = 0.561$, visual depth: $r_s = 0.614$, both $p < 0.00001$). However, the wide percentage outperforms both methods which has a correlation of $r_s = 0.772$. This result suggests that the the attentional scope reflected by the shooting behavior is more related to tourist satisfaction than current methods for visiting worth assessment. Thus, the relationship between photo-shooting behavior and tourist

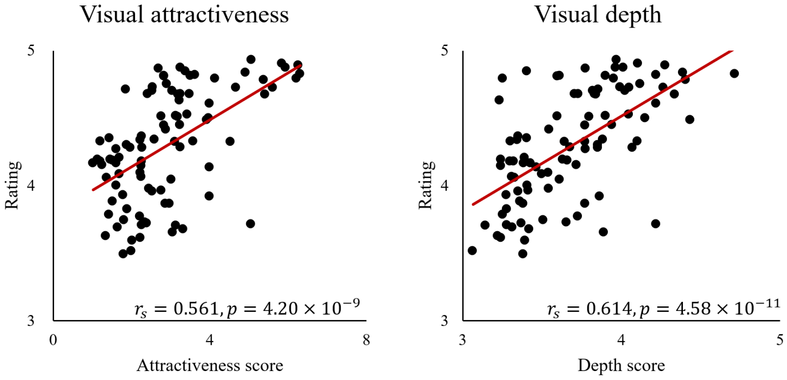


Figure 5.14: The correlation analysis between the experience rating and visiting worth scores based on current visual content-based methods, i.e., natural attractiveness (Workman et al., 2017) and visual depth (Y. Shen et al., 2018).

satisfaction is potentially applied for sightseeing value assessment.

5.4 Discussion

The broaden-and-build theory (Fredrickson, 2004; Fredrickson & Branigan, 2005) bridges emotion and attention. Specifically, the emotional states can influence the attentional scopes (Fredrickson, 2004; Kuhbandner et al., 2011; Tamir & Robinson, 2007), and broadened attention has an affective advantage over narrowed attention (Ji et al., 2019; Srinivasan & Hanif, 2010). This chapter aims to test the broaden-and-build theory in the tourism scenario by constructing NODS from social media platforms in tandem with the advantages of machine learning. We focused on the photo-shooting behaviors and hypothesized that positive affect would induce tourists to capture more wide-view photos than narrow-view ones and to give a high rating for their travel experience. In Experiment 1, we applied the CFCNN proposed in Chapter 4 for view-type classification and then analyzed 39,099 travel photos. The experimental results indicated that people visiting high-rated spots tend to capture wide-view photos, while at low-rated spots they are more likely to capture elements from a narrow-view. In Experiment 2, we found a significant correlation between the proportion of wide-view photos posted by tourists and their experience rating. This is the first evidence, to our knowledge, that has demonstrated the broaden-and-build theory using NODS. Moreover, we showed two potential significance to real-world applications by using the association between photo-shooting behaviors and tourist satisfaction, i.e., POI discovery and sightseeing value assessment.

Figure B.2 showed the questionnaire survey provided by *TripAdvisor*. Firstly, tourists make an overall rating of the travel experience — 'Excellent, 'Very good', 'Average', 'Poor', 'Terrible'. Secondly, they will be asked to leave a review to describe — "What are your favorite part of your experience? Any tips for future travelers". Thirdly, they will post the photos to share this travel experience. Many tourism studies (Alaei et al., 2019; Bjørkelund et al., 2012; Broß, 2013; Gindl et al., 2010; Y. Liu et al., 2019) demonstrated that the content of the reviews could be used to predict the experience rating. However, few studies used the visual contents extracted from images to inferring the rating. In this chapter, we found that the view-type of the images inducing tourist attention provides a significant cue to reveal the experience rating. During the travel experience, the affective quality attributed to the environment governs subsequent photo-shooting behaviors (Ittelson, 1976). During the experience rating, the possible cognitive process is as follows: first, tourists recall the scenes or emotional moments during traveling, then write reviews and post photos according to the recalled scenes and emotional moments. Thus, the posted photos may be related to the emotional moments of the travel experience. Our results suggest the positive emotional moments prefer to take more wide-view photos. In addition, the experience rating is always used to represent the subjective sentiment or satisfaction of tourists in tourism studies (Alaei et al., 2019; Broß, 2013; Y. Liu et al., 2019). However, how the experience rating of tourists reverts the definition of emotion in psychology requires further investigation.

Does positive affect trigger the shooting behavior of wide-view, or do the characteristics of scenic spots (e.g., the object of interest) trigger wide-view shooting behavior? The evidence is inconclusive. The assumption that fits the broaden-and-build theory most is that a high-quality spot inspires positive emotion, which then causes them to take more wide-view photos. However, an alternative assumption is characteristics of scenic spots (e.g., the object of interest is a big size) directly inspire people to take wide-view photos to record the location. For example, in Experiment 1, we showed that Carmel City Beach has an observation deck for tourists (as shown in Figure 5.8), which provides convenience to take a wide-view photo. To rule out the external factors (e.g., cultural and environmental factors), we used the visual content analysis to control the conditions of green, water, and building perception of tourists and divided the scenic spots into different subgroups to the various continents and scenery types, respectively. The results still suggested that there exists a relationship between experience rating and the wide percentage. There may have other factors that influence the relationship (satisfaction vs. photo-shooting behaviors). However, controlling factors related to travel experience using NODS would be an open question for informatics study. In the future, we intend to purely examine this relationship by conducting a psychological study to control the environmental factors. In addition, there may be individual or cultural differences in the relationship. We will design the experiments using NODS for within-subject analysis, examining the causal relationship between the emotional states and photo-shooting behaviors detected from individual text and photo data.

With rapid development, machine learning becomes a promising methodology to examine emotions aroused by visual contents. Most informatics scientists are apt to explore specific features in images to infer emotions, which is objective recognition, rather than subjective cognition, unlike our study. For example, X. Lu et al. (2012) investigated the shape features, which could affect human emotional responses and modeled a dimensional emotion aroused by roundness and angularity. J. Jia et al. (2012) proposed a factor graph model using color features from the images and their social relationships for emotion prediction. Machajdik and Hanbury (2010) designed a richer feature set (including color, texture, composition, and content features) to represent the emotional content of an image. An emerging work (S. Zhao et al., 2014) explored the principles-of-art, then defined more robust and invariant visual features, such as balance, variety, and gradation. These handcrafted features had reported a good performance on several public visual sentiment datasets. Recently, deep learning has demonstrated a robust and accurate ability to feature learning.

You et al. (2016) suggested a fine-tuned CNN could outperform the state-of-the-art handcrafted features for visual sentiment analysis. Moreover, there were studies about image memorability (Isola et al., 2011) and image interestingness by taking into account psychological theories (Gygli et al., 2013). Although citing psychological inspirations, these works were invariably treated as image classification tasks and rarely redounded upon the psychology research on observing large-scale human behaviors.

Recent research (Zhuang et al., 2014) has focused on discovering new tourism resources through mining text or evaluating picture quality in social media platforms. However, few studies tried to link tourist natural behaviors and mood states with these data, particularly the image data. In this chapter, the broaden-and-build theory with the support of NODS may serve as a new measure for such a task and boost tourism economics. In the mental and welfare healthcare field, researchers are also reviewing big data resources and their use to characterize applications to address mental illness, e.g., depression (Reece & Danforth, 2017). Our approach may help such special populations better live by mining their provided data on social media platforms.

Chapter 6

General Discussion

This dissertation attempts to forge the advances in informatics approaches into tools that psychologists can apply to complement and expand the current psychological experiments or theories. On the one hand, the DNNs provide valuable representations of complex stimuli with diverse effects, which can complement the simple stimuli on classical psychological experiments. On the other hand, social media platforms offer extensive behavior data (e.g., photos and comments posted on a message board, number of likes, and stars of experiences), which can be applied to extend the limited behavioral data on classical psychological experiments. In any case, it is natural to ask how psychologists can apply the above informatics approaches. This dissertation provided two frameworks that use the potentials in the complementary new paradigms to study human cognition. Two case studies demonstrate the applications of frameworks for respectively understanding human facial attractiveness perception and visual attention.

6.1 Summary of Current Work

In chapter 1, we discussed the issues of external validity in classic psychological experiments due to experimental purpose and demands to the time, space for experiments. The recent informatics approaches may partially circumvent the issues of external validity, e.g., DNN effective representations from real-world stimuli, and NODS collected from social media platforms contain human actions that broke through the constraints of time and space. However, a relatively small number of studies used these informatics approaches to examine human cognition. Thus, we domesticate the applications of these approaches in psychological studies through two case studies, which (1) understanding human facial attractiveness perception from representations learned by DNN, and (2) understanding visual attention by constructing NODS from travel photos, respectively.

Chapters 2 and 3 focused on how to train and evaluate the human-relevant representations of DNN. In this study, we interpreted the DNN and compared the DNN to humans in the behavior-level and theory domain. In chapter 2, we trained the DNN for facial attractiveness (FADNN) using category annotations (discrete) while testing specific neurons using continuous rating scores. The results demonstrated the specific neurons subsume the main features associated with the categories and can be used as metrics to determine an image's membership in a category. Furthermore, we visualized the specific neurons by deconvolution of FADNN and generated four face-like images to interpret the representations learned by neurons. From observation, we suggested that these neurons might learn putative ratios of facial attractiveness. For further investigation, we measured the generated images and implemented psychophysical-like experiments to test the specific neurons' activation with different putative ratios in Chapter 3. The results showed that the FADNN learned the putative ratios as critical features for the specific neurons and suggested that DNN models can capture essential psychological characteristics from complex stimuli. This study opens up opportunities for elucidating the origins of facial attractiveness perception via the DNN-based approach for face recognition.

Chapters 4 and 5 aimed to test the broaden-and-build theory by constructing NODS from social media platforms. In this study, the experiment constructed NODS with main effects related to the theory. Specifically, the broaden-and-build theory (Fredrickson, 2004; Ji et al., 2019; Schmitz et al., 2009) showed the relationship between emotion and visual attention, where the positive emotion broadens the attentional scope. We investigate the association between tourist satisfaction and photo-shooting behaviors (narrow/wide-view) triggered by attentive processing.

Chapter 4 focused on narrow-view and wide-view classification, which is a novel and subjective task in computer vision. We proposed two cues to model this problem — focus cue, and scale cue. The experimental results on a newly created dataset showed that the proposed model's performance (0.9317) outperformed other traditional models based on related works (e.g., Gabor-SVM, 0.8740: Cao and O'Halloran, 2015). In addition, we implemented focus cue and scale cue in the CFCNN model, which gets better performance (0.9752) than the model based on focus cue and scale cue. Using CFCNN, we investigate the association between tourist satisfaction and photo-shooting behaviors from newly scraped data from Flickr, TripAdvisor in chapter 5. The experimental results showed that a significant relationship between photo-shooting behaviors and tourist satisfaction, which is consistent with the broaden-and-build theory. We showed that the broaden-and-build theory is applied to real-world data with a variety of participants who perform photo-shooting in a natural setting. This finding provides a new perspective for psychologists to observe human behaviors and cognition by building NODS from social media platforms in tandem with the advantages of machine learning.

6.2 Bridge the Paradigm Gap

The dissertation started with the motivations to apply the emerging informatics approaches to expand psychological studies and bridge the paradigm gap between experimental purposes of informatics and psychological studies. In this section, we discussed the effects of the gap-filling and the contributions to both fields.

The paradigm gap is the main barrier to apply the emerging informatics approaches to psychology. Psychology aims to understand the complex human mind and demonstrate the highest internal validity with a well-controlled experimental setting. On the contrary, informatics studies aim to implement the data to specific tasks and functions and confirm the high ecological validity with data from diverse situations. Thus, we proposed two complementary frameworks to bridge the paradigm gap, i.e., applying DNNs and NODS to study human cognition, as shown

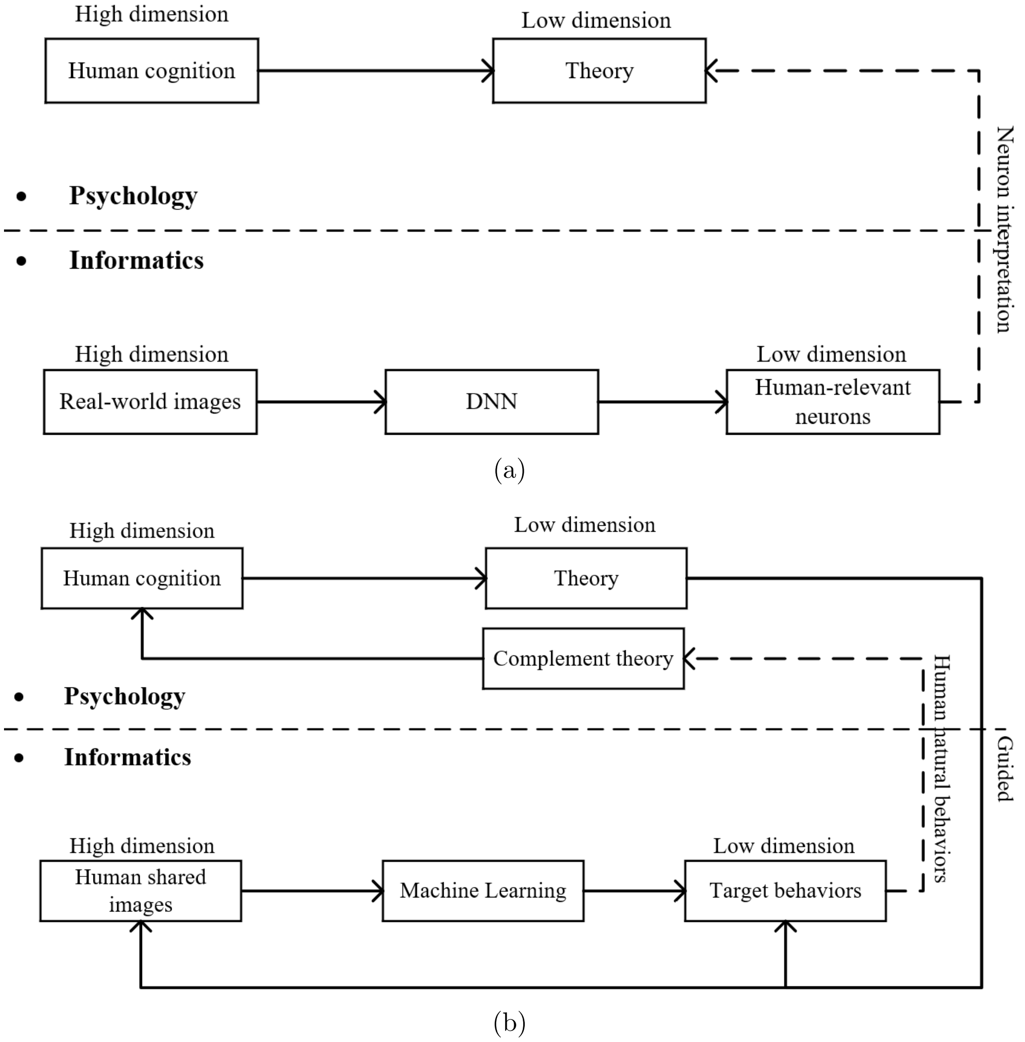


Figure 6.1: The proposed frameworks of study human cognition through informatics approaches. On the side of psychology, the theories of cognition provide low dimension descriptions of high dimensional human cognition. On the side of informatics, we provided two pathways to implement cognition theories for the evaluations of frameworks. (a) The first framework interpreted the representations learned by DNN to contain human-relevant neurons or not. The DNN trained by real-world images, including diverse factors, while evaluating the neurons contains human-relevant representations via neuron visualization and psychophysical-like experiments. (b) The second framework has applied the current theory to determine the NODS construction and target behaviors. We trained a specific machine learning model to recognize the target behaviors from human shared images. The goal of this framework is to obtain an extended theory from the constructed NODS.

in Figure 6.1. The first framework interpreted the representations learned by DNN to contain human-relevant neurons or not. The DNN trained by real-world stimuli, including diverse factors, and then we evaluated the neurons contains human-relevant representations through neuron visualization and psychophysical-like experiments. Thus, the real-world stimuli are partially connected to psychological characteristics, mediated by human-relevant representations. The second framework has applied an existing theory to construct NODS and target behaviors derived from the theory. We developed machine learning models to recognize the target behaviors (main effects) from human shared images, and then we evaluated the relationship between main effects. Thus, the found relationship would complement the existing theory to a natural setting.

The case study using the first framework showed that features learned by a DNN for facial attractiveness include the putative ratios while the attractiveness category is given as a supervisory signal. The results suggested the DNN can learn putative ratios from complex stimuli. On the other side, psychologists design experiments with well-controlled stimuli based on extensive observations of human behaviors and their rich knowledge or theoretical background. These experiments look simple while accumulating with extensive observations. The putative ratios for facial attractiveness are found by this experimental setting, e.g., (Pallett et al., 2010). However, we showed that the well-trained DNN mirrors the psychologists, who extract high-level representations from complex stimuli and extensive observations. We used the techniques, including neural network interpretation and network neurons' selections, to link the deep feature representations to the psychological characteristic. Thus, these representations are useful, e.g., generating hypotheses that can be verified in experimental studies. This study found that the new golden ratios impact female faces in images more than those of males in FADNN. However, there is no empirical evidence for this in human behavior, even with the new golden ratios employed for both female and male face images (Bóo et al., 2013; Yoo et al., 2013).

Conversely, psychological theories also provide tools (such as psychophysical experimental settings) for exposing implicit representations of DNNs, which remain poorly understood. To our knowledge, work applying theories of cognition to DNNs for this purpose has been relatively limited. We are only aware of studies (Parde et al., 2019; Ritter et al., 2017; P. Wang & Cottrell, 2017; Zoran et al., 2015) used psychophysical-like experiments to understand DNN models better. In addition, the human-relevant representations are also valuable for improving DNNs and extending the applications of DNNs. For example, the human-relevant neurons learned by FADNN can be used to advance facial attractiveness prediction and generation, making the models or the generated images in line with human cognition. In concrete, it is possible to use human-relevant neurons as the attributes to constrain the DNN training (e.g., attribute-aware DNN: Lin et al., 2019) or guiding the generative adversarial network (e.g., conditional GAN: Y. Lu et al., 2018). We will investigate these extensions in the future.

The case study using the second framework showed the photo-shooting behaviors (wide and narrow-view) of tourists related to the subjective rating of travel experience. Tourists with positive emotions tend to share photos of wide-view, while with negative emotions, they are more likely to narrow-view photos that captured elements. This preference partially confirmed the broaden-and-build theory, in which positive emotions broaden attentional scope and trigger to see a big picture. This finding is difficult to be investigated through classical psychological experiments because of the complex environments of tourism. In addition, informatics research does not focus on the cause of human action but rather explores the human (mind), thus also ignore such insights. Therefore, the proposed framework is critical for expanding the current psychological theories through constructing NODS.

Oppositely, the theories of cognition extend the applications of online data analysis. Understanding human cognition can become a new task for online data analysis. Specifically, Griffiths (2015) pointed out that computer scientists hold with most data about human behavior, exceeding psychologists while analyzing behavior data for only behavior. For example, recommendation systems often use collaborative filtering to predict what users will purchase based purely on their behavior's similarity to others' behavior. Based on the theories of cognition, it is possible to demonstrate the value of postulating a mind based on NODS.

In summary, our works provide valuable examples to bridge the paradigm gap between classical psychological experiments and informatics researches. Both of the complementary frameworks extend the scope of both theories of cognition and applications of informatics. The strategies of DNN extracted from complex stimuli and natural human behavior data provide new perspectives to study human cognition. Simultaneously, theories about human cognition provide tremendous knowledge to understanding DNNs' representations and novel tasks for online data analysis.

6.3 Integration of DNN and NODS

In our studies, we applied DNN and NODS for understanding human cognition, respectively. However, there may be another way of the integration of DNN and NODS in psychological studies. For this purpose, firstly, we discuss the possibility of extensions of our studies. In the first part, the face image dataset was collected from the Internet. Notably, the name list of the highly attractive face was collected from beauty ranking websites, which intended to inform public opinion for the modern ideal of beauty and listened to millions of suggestions from social media users. This is a NODS to a certain extent. In the second part, we trained a DNN model to represent the field-of-view of a photographer. The reason why this model shows the efficiency and what representation learned by this model is also interesting. We showed that this model learned some representations different from the focus cue and scale cue related to human visual systems, which may be uncovered by psychologists. Thus, the DNNs can be used to learn some rules that represent the NODS and as a tool to clean the NODS as meaningful concepts. However, exploring the feature domain of natural behavior and DNN representation is a choice from many factors. Therefore, understanding the current psychological theories will help us narrow the scope of the analysis when facing them.

Before using DNN, there is some success in applying informatics approaches with psychological findings for psychological studies. Said and Todorov (2011) applied machine learning algorithms to reveal unreported components of attractiveness perception by comparing with averageness and sexual dimorphism. In addition, Bainbridge (2007) suggested that human behavior in online games (such as World of Warcraft) can be used to predict a person's personality and decision-making tendency. Stafford and Dewar (2014) used user data (n=854064) of online games to confirm the results of a psychological finding regarded to the effects of practice quantity and quality on performance.

Moustafa et al. (2018) pointed out that the integration of big data analysis and DNN can solve more complex psychological problems. However, most attempts in previous studies focused on data-driven approaches, yielding insights on interesting phenomena empirically depended on the available data. For example, DNNs are utilizing to improve the classification of complex datasets using abstract feature representations instead of the linear classification models. Thus, DNNs can be used to improve the predict performance of some specific disorders (Choi & Jin, 2018). However, these usages of DNNs are data-driven methods, which treated the DNNs as more powerful machine learning techniques than before (e.g., linear classification, statistical learning, and SVM). Maass et al. (2018) claimed that the researchers engaged in data-driven research "face the challenge of building a cohesive body of knowledge about phenomena because relationships are determined by the available data". Similarly, the "data-driven" of the analytic techniques is also a common criticism of many big data approaches (Harlow & Oswald, 2016). Landers et al. (2016) proposed a primer on theory-driven data collection for psychological research. In concrete, researchers should be precisely accounted for "why the data they found exist and test the hypotheses implied by that theory with additional analyses" (Harlow & Oswald, 2016). Analogously, for DNN representation, it is also required to account for why the representations showed better performance and test them connecting to theory. Therefore, the combination of strengths of psychological research (e.g., rich theory and high-quality measurement) and those of informatics approaches (e.g., power and flexibility) is important for understanding human cognition.

6.4 Limitations and Future Directions

This section discusses how to expand the suggestions in this dissertation to bridge the paradigm gap between classical psychological experiments and informatics research.

6.4.1 From machine representation to psychological representation

Psychologists are perhaps the most aware of how to use the learned representations of DNNs to study human cognition. Our proposed framework (as shown in Figure 6.1a) is essential because it provided a way to observe the representations and link them to the theory domain. Psychologists perhaps get a source of fruit in the next few years with the improvements in the interpretation of the representations learned by DNNs.

For example, we visualized the implicit representation learned by the DNN model in Chapter 2. The visualized images depicted a meaningful spatial configuration and then hypothesized that the implicit representation might include putative ratios for facial attractiveness. However, the visualized images cannot depict meaningful 3-d configuration and color information, which is significant for facial attractiveness (O'Toole et al., 1999; Said & Todorov, 2011) and face perception (H. Hill et al., 1995; Walker & Vetter, 2009). Emerging studies (M. Q. Hill et al., 2019; H. Hong et al., 2016; Parde et al., 2017) showed that the DNNs retain a significant amount of information about the original images. For example, M. Q. Hill et al. (2019) demonstrated a DNN trained for face identification retained the illumination and viewpoint information in the deep feature space. On the side of informatics, how to reconstruct the better quality of color and 3D information from DNNs is also a hot topic (Antipov et al., 2017; H.-Y. Chen & Lu, 2019; Jackson et al., 2017). However, these techniques cannot directly interpret the representations because of different research purposes.

Supposing the representations of DNNs has been well interpreted with the advancement of techniques. These intuitive representations would become rich resources for psychologists, although we still need to consider how these representations deviate from humans. On the one hand, we can construct stimuli through the visualized high-level features of DNNs, just like manipulating the simple features to construct distorted and average faces. On the other hand, these representations would enrich our exploration of the unknown fields of human cognition. For example, Lake et al. (2017) pointed out that the learned representations of DNNs may potentially apply to understand the human intuition, which has significant. For instance, how to let an autonomous agent understand the scene to navigate and perform tasks in the complex environment effectively or how to allow autonomous cars predict pedestrian behavior and infer the mental states (e.g., Where do they want to go?).

6.4.2 From human daily lives to psychological studies

NODS allow psychological scientists to test and expand theories by analyzing human behavior in the real world (Paxton, 2020). However, the purpose of NODS is targeting to understand human cognition, which is different from the general scientific purposes of informatics research. Our proposed framework (as shown in Figure 6.1b) is crucial because it provided a way to complement the existing theory with natural behaviors and experiences. The framework is applicable to many experience domains if there is a suitable hypothesis related to current theory and rich data. For example, with the smartphones and wearable devices, it is relatively easy to capture various types of data that describe daily actions, e.g., experiences of shopping, cooking, working, driving, and learning. Therefore, it is possible to expand the framework to transmit human daily activities and experiences to study human cognition through more and more data resources in the next few years.

Our study detected the photo-shooting behaviors from the travel images, which partially reflect tourists' attentional scopes. We were motivated to explore the connection between tourist satisfaction and their attentional scopes reflected from shooting behaviors. In reality, TripAdvisor also provides experience reviews, which provide meaningful information for sentiment analysis (Taecharungroj & Mathayomchan, 2019; Valdivia et al., 2019) and accurately refer to their experience rating (Y. Liu et al., 2019). Thus, we consider expanding our study to combine natural language processing with image processing for a comprehensive research about the tourism experience feelings in the future.

Additionally, the wearable cameras record daily activities captured by the egocentric vision. The recorded movie data provide the potential opportunity to study daily activities and cognition. They allow the researchers to capture the participants' gaze that they record the moments in their interaction with environments. The data collected from wearable cameras are used to investigate experience summarization (L. Chen et al., 2019), memory recall (Sellen et al., 2007), and personality traits (Blake et al., 2020). For example, L. Chen et al. (2019) focused on employing wearable cameras to assist the manipulation of daily machines. Sellen et al. (2007) focused on exploring how to use the images collected from wearable cameras to increase experiential recall. However, the current insights from these data are based on the data-driven approaches, which cannot complement the existing psychological theories.

To use our proposed frameworks for psychological studies is required to search the domain of theory. We hope our research would encourage more psychologists to conduct the studies using daily activities to expand their theory of human cognition. For example, what human activities are related to the decision-making in the shopping experience, and what human activities are related to the teachers' workload in the educational experience.

6.5 Concluding Remarks

There is a long history of exploring human cognition in philosophy and psychology. However, recent advancements in informatics begins to provide a new direction in the research of human cognition. At present, DNN provides a framework to predict responses of human brain and behaviors. Social network platforms gather large scales of human behavior data (e.g., mobile trajectory and human preferences). These informatics approaches provide computational and data resources for understanding human cognition. In this thesis, we applied these resources to connect real-world stimuli and natural behavior data to psychological theories. But the current works still have some limitations. For example, real-world stimuli and naturally occurring data have more confounding factors than data collected from laboratory experiments. However, we believe that these presented works or approaches could supplement psychological research and improve the external validity of existing theories or even discover new theories in the future.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917.

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research*, *58*(2), 175–191.

Altwaijry, H., & Belongie, S. (2013). Relative ranking of facial attractiveness. In *Proceedings of IEEE workshop on applications of computer vision* (pp. 117–124).

Antipov, G., Baccouche, M., & Dugelay, J.-L. (2017). Face aging with conditional generative adversarial networks. In *Proceedings of the IEEE international conference on image processing* (pp. 2089–2093).

Arase, Y., Xie, X., Hara, T., & Nishio, S. (2010). Mining people's trips from large scale geo-tagged photos. In *Proceedings of the ACM international conference on multimedia* (pp. 133–142).

Arnett, J. J. (2008). The neglected 95%: Why american psychology needs to become less american. *American Psychologist*, *63*(7), 602–614.

Atiyeh, B., & Hayek, S. (2008). Numeric expression of aesthetics and beauty. *Aesthetic Plastic Surgery*, *32*(2), 209–216.

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science*, *317*(5837), 472–476.

Bandura, A. (2001). Social cognitive theory of mass communication. *Media Psychology*, *3*(3), 265–299.

Bardou, D., Zhang, K., & Ahmad, S. M. (2018). Classification of breast cancer based on histology images using convolutional neural networks. *IEEE Access*, *6*, 24680–24693.

Barshan, E., Ghodsi, A., Azimifar, Z., & Jahromi, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, *44*(7), 1357–1371.

Bartie, P., & Mackaness, W. (2016). Mapping the visual magnitude of popular tourist sites in edinburgh city. *Journal of Maps*, *12*(2), 203–210.

Bashour, M. (2006). An objective system for measuring facial attractiveness. *Plastic and Reconstructive Surgery*, *118*(3), 757–774.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1), 1–14.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541–6549).

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. In *Proceedings of the European conference on computer vision* (pp. 404–417).

Bjørkelund, E., Burnett, T. H., & Nørvåg, K. (2012). A study of opinion mining and visualization of hotel reviews. In *Proceedings of the international conference on information integration and web-based applications & services* (pp. 229–238).

Blake, A. B., Lee, D. I., De La Rosa, R., & Sherman, R. A. (2020). Wearable cameras, machine vision, and big data analytics: Insights into people and the places they go. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 347–372). American Psychological Association.

Bóo, F. L., Rossi, M. A., & Urzúa, S. S. (2013). The labor market return to an attractive face: Evidence from a field experiment. *Economics Letters*, *118*(1), 170–172.

Borissavliévitch, M., & Hautecœr, L. (1958). *The golden number and the scientific aesthetics of architecture*. Alec Tiranti.

Broß, J. (2013). *Aspect-oriented sentiment analysis of customer reviews using distant supervision techniques* (Doctoral dissertation).

Burleson, M. H., Hall, D. L., & Gutierres, S. E. (2016). Age moderates contrast effects in women's judgments of facial attractiveness. *Evolutionary Behavioral Sciences*, *10*(3), 179–187.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963: 1–18.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cao, Y., & O'Halloran, K. (2015). Learning human photo shooting patterns from large-scale community photo collections. *Multimedia Tools and Applications*, *74*(24), 11499–11516.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Chen, F., Xiao, X., & Zhang, D. (2016). Data-driven facial beauty analysis: Prediction, retrieval and manipulation. *IEEE Transactions on Affective Computing*, *9*(2), 205–216.

Chen, F., & Zhang, D. (2014). Evaluation of the putative ratio rules for facial beauty indexing. In *Proceedings of the international conference on medical biometrics* (pp. 181–188).

Chen, H.-Y., & Lu, C.-J. (2019). Nested variance estimating VAE/GAN for face generation. In *Proceedings of the international joint conference on artificial intelligence* (pp. 1–8).

Chen, L., Nakamura, Y., Kondo, K., & Mayol-Cuevas, W. (2019). Hotspot modeling of hand-machine interaction experiences from a head-mounted RGB-D camera. *IEICE Transactions on Information and Systems*, *102*(2), 319–330.

Chen, Y., Mao, H., & Jin, L. (2010). A novel method for evaluating facial attractiveness. In *Proceedings of the international conference on audio, language and image processing* (pp. 1382–1386).

Cheng, M.-M., Liu, Y., Lin, W.-Y., Zhang, Z., Rosin, P. L., & Torr, P. H. (2019). BING: Binarized normed gradients for objectness estimation at 300 frames per second. *Computational Visual Media*, *5*(1), 3–20.

Choi, H., & Jin, K. H. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural Brain Research*, *344*, 103–109.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1–13.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346–358.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755:1–13.

Coetzee, V., Greeff, J. M., Stephen, I. D., & Perrett, D. I. (2014). Cross-cultural agreement in facial attractiveness preferences: The role of ethnicity and gender. *PloS One*, *9*(7), e99629:1–8.

Colón, Y. I., Castillo, C. D., & O'Toole, A. J. (2020). Facial expression is retained in deep networks trained for face identification. *https://doi.org/10.31234/osf.io/dphsv.*

Conway, B. R. (2003). Colour vision: A clue to hue in V2. *Current Biology*, *13*(8), R308–R310.

Cooper, P. A., Geldart, S. S., Mondloch, C. J., & Maurer, D. (2006). Developmental changes in perceptions of attractiveness: A role of experience? *Developmental Science*, *9*(5), 530–543.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *Computer Vision and Image Understanding*, *61*(1), 38–59.

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology*, *50*(5), 925.

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). "their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, *68*(2), 261–279.

Danel, D., & Pawlowski, B. (2007). Eye-mouth-eye angle as a good indicator of face masculinization, asymmetry, and attractiveness (Homo sapiens). *Journal of Comparative Psychology*, *121*(2), 221–225.

Data, I. B., & Hub, A. (2013). Infographics & animations: The four V's of big data. *IBM. https://www. ibmbigdatahub. com/infographic/four-vs-big-data.*

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the international conference on machine learning* (pp. 647–655).

Easley, G., Labate, D., & Lim, W.-Q. (2008). Sparse directional image representations using the discrete shearlet transform. *Applied and Computational Harmonic Analysis*, *25*(1), 25–46.

Edensor, T. (2000). Staging tourism: Tourists as performers. *Annals of Tourism Research*, *27*(2), 322–344.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119–142.

Elsaesser, T., & Buckland, W. (2002). *Studying contemporary american film: A guide to movie analysis*. London: Hodder Arnold.

Endel, F., & Piringer, H. (2015). Data wrangling: Making data useful again. *IFAC-PapersOnLine*, *48*(1), 111–112.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1–13.

Eriksen, C. W., & James, J. D. S. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*(4), 225–240.

Fang, Z., Cao, Z., Xiao, Y., Zhu, L., & Yuan, J. (2016). Adobe boxes: Locating object proposals using object adobes. *IEEE Transactions on Image Processing*, *25*(9), 4116–4128.

Farkas, L. G., & Kolar, J. C. (1987). Anthropometrics and art in the aesthetics of women's faces. *Clinics in Plastic Surgery*, *14*(4), 599–616.

Farkas, L. G., & Schendel, S. A. (1995). Anthropometry of the head and face. *American Journal of Orthodontics and Dentofacial Orthopedics*, *107*(1), 112–112.

Fenske, M. J., & Raymond, J. E. (2006). Affective influences of selective attention. *Current Directions in Psychological Science*, *15*(6), 312–316.

Feser, D. K., Gründl, M., Eisenmann-Klein, M., & Prantl, L. (2007). Attractiveness of eyebrow position and shape in females depends on the age of the beholder. *Aesthetic Plastic Surgery*, *31*(2), 154–160.

Fink, B., Grammer, K., & Thornhill, R. (2001). Human (Homo sapiens) facial attractiveness in relation to skin texture and color. *Journal of Comparative Psychology*, *115*(1), 92–99.

Förster, J. (2012). Glomosys: The how and why of global and local processing. *Current Directions in Psychological Science*, *21*(1), 15–19.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, *56*(3), 218–226.

Fredrickson, B. L. (2004). The broaden–and–build theory of positive emotions. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1449), 1367–1377.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion*, *19*(3), 313–332.

Frieze, I. H., Olson, J. E., & Russell, J. (1991). Attractiveness and income for men and women in management 1. *Journal of Applied Social Psychology*, *21*(13), 1039–1057.

Gan, J., Li, L., Zhai, Y., & Liu, Y. (2014). Deep self-taught learning for facial beauty prediction. *Neurocomputing*, *144*(20), 295–303.

Gao, Z., Flevaris, A. V., Robertson, L. C., & Bentin, S. (2011). Priming global and local processing of composite faces: Revisiting the processing-bias effect on face perception. *Attention, Perception, & Psychophysics*, *73*(5), 1477–1486.

Garcia, M. A., & Solanas, A. (2004). 3d simultaneous localization and modeling from stereo vision. In *Proceedings of IEEE international conference on robotics and automation, 2004* (pp. 847–853).

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological Science*, *13*(1), 34–40.

Ge, M., Zhuang, C., & Ma, Q. (2019). Robust visual object clustering and its application to sightseeing spot assessment. *Multimedia Tools and Applications*, *78*(12), 17135–17164.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780.*

Geldart, S., Maurer, D., & Henderson, H. (1999). Effects of the height of the internal features of faces on adults' aesthetic ratings and 5-month-olds' looking times. *Perception*, *28*(7), 839–850.

Gindl, S., Weichselbraun, A., & Scharl, A. (2010). Cross-domain contextualisation of sentiment lexicons. In *Proceedings of the European conference on artificial intelligence* (pp. 771–776).

Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 270–279).

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*(3), 548–568.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Goodhew, S. C. (2020). *The breadth of visual attention.* Cambridge University Press.

Goodhew, S. C., Lawrence, R. K., & Edwards, M. (2017). Testing the generality of the zoom-lens model: Evidence for visual-pathway specific effects of attended-region size on perception. *Attention, Perception, & Psychophysics*, *79*(4), 1147–1164.

Grant, E., Sahm, S., Zabihi, M., & van Gerven, M. (2016). Predicting and visualizing psychological attributions with a deep neural network. In *Proceedings of the international conference on pattern recognition* (pp. 1–6).

Gray, D., Yu, K., Xu, W., & Gong, Y. (2010). Predicting facial beauty without landmarks. In *Proceedings of the European conference on computer vision* (pp. 434–447).

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Gu, L., Yang, X., Li, L. M. W., Zhou, X., & Gao, D.-G. (2017). Seeing the big picture: Broadening attention relieves sadness and depressed mood. *Scandinavian Journal of Psychology*, *58*(4), 324–332.

Güçlü, U., Thielen, J., Hanke, M., van Gerven, M., & van Gerven, M. A. (2016). Brains on beats. In *Advances in neural information processing systems* (pp. 2101–2109).

Gunes, H., & Piccardi, M. (2006). Assessing facial beauty through proportion analysis by image processing and supervised learning. *International Journal of Human-computer Studies*, *64*(12), 1184–1199.

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In *Proceedings of the IEEE international conference on computer vision* (pp. 1633–1640).

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447–457.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29.

Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: The role of shape and colour. *Proceedings of the Royal Society of London B: Biological Sciences*, *261*(1362), 367–373.

Hill, M. Q., Parde, C. J., Castillo, C. D., Colon, Y. I., Ranjan, R., Chen, J.-C., Blanz, V., & O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, *1*(11), 522–529.

Holland, E. (2008). Marquardt's Phi mask: Pitfalls of relying on fashion models and the golden ratio to describe a beautiful face. *Aesthetic Plastic Surgery*, *32*(2), 200–208.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622.

Hong, W., Chen, T.-S., & Chen, J. (2015). Reversible data hiding using delaunay triangulation and selective embedment. *Information Sciences*, *308*, 140–154.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 15037:1–15.

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. In *Advances in neural information processing systems* (pp. 2429–2437).

Ittelson, W. H. (1976). Environment perception and contemporary perceptual theory. *Environmental Psychology: People and Their Physical Settings*, 141–154.

Jackson, A. S., Bulat, A., Argyriou, V., & Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE international conference on computer vision* (pp. 1031–1039).

Jacobsen, J. K. (2007). Use of landscape perception methods in tourism studies: A review of photo-based research approaches. *Tourism Geographies*, *9*(3), 234–253.

Jayaratne, Y. S., Deutsch, C. K., McGrath, C. P., & Zwahlen, R. A. (2012). Are neoclassical canons valid for southern Chinese faces? *PloS One*, *7*(12), e52593:1–7.

Jefferson, Y. (2004). Facial beauty-establishing a universal standard. *International Journal of Orthodontics-Milwaukee*, *15*(1), 9–26.

Ji, L.-J., Yap, S., Best, M. W., & McGeorge, K. (2019). Global processing makes people happier than local processing. *Frontiers in Psychology*, *10*, 670:1–10.

Jia, J., Wu, S., Wang, X., Hu, P., Cai, L., & Tang, J. (2012). Can we understand van gogh's mood? learning to infer affects from images in social networks. In *Proceedings of the ACM international conference on multimedia* (pp. 857–860).

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM international conference on multimedia* (pp. 675–678).

Jones, B. C., DeBruine, L. M., & Little, A. C. (2007). The role of symmetry in attraction to average faces. *Perception & Psychophysics*, *69*(8), 1273–1277.

Jones, M. N. (2016). Developing cognitive theory by mining large-scale naturalistic data. *Big Data in Cognitive Science*, 1–12.

Kagian, A., Dror, G., Leyvand, T., Cohen-Or, D., & Ruppin, E. (2007). A humanlike predictor of facial attractiveness. In *Advances in neural information processing systems* (pp. 649–656).

Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., & Ruppin, E. (2008). A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, *48*(2), 235–243.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kimchi, R., & Palmer, S. E. (1982). Form and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(4), 521.

Kinchla, R. A., & Wolfe, J. M. (1979). The order of visual processing: Top-down, bottom-up, or middle-out. *Perception & Psychophysics*, *25*(3), 225–231.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Kuhbandner, C., Lichtenfeld, S., & Pekrun, R. (2011). Always look on the broad side of life: Happiness increases the breadth of sensory memory. *Emotion*, *11*(4), 958–964.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the IEEE international conference on 3D vision* (pp. 239–248).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253:1–72.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*, *21*(4), 475.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121.

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental Psychology*, *23*(3), 363–369.

Lawson, F., & Baud-Bovy, M. (1977). *Tourism and recreation development, a handbook of physical planning.* Architectural Press.

Le, V. N. T., Apopei, B., & Alameh, K. (2019). Effective plant discrimination based on the combination of local binary pattern operators and multiclass support vector machine methods. *Information Processing in Agriculture*, *6*(1), 116–131.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, S. H., Chan, C. S., Mayo, S. J., & Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, *71*, 1–13.

Lewin, K. (1951). *Field theory in social science.* New York, NY: Harpers.

Leyvand, T., Cohen-Or, D., Dror, G., & Lischinski, D. (2008). Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics*, *27*(3), 38.

Li, G., Zhu, T., Hua, J., Yuan, T., Niu, Z., Li, T., & Zhang, H. (2019). Asking images: Hybrid recommendation system for tourist spots by hierarchical sampling statistics and multimodal visual bayesian personalized ranking. *IEEE Access*, *7*, 126539–126560.

Li, J., Xiong, C., Liu, L., Shu, X., & Yan, S. (2015). Deep face beautification. In *Proceedings of the ACM international conference on multimedia* (pp. 793–794).

Liang, L., Lin, L., Jin, L., Xie, D., & Li, M. (2018). SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *Proceedings of the international conference on pattern recognition* (pp. 1598–1603).

Liang, X., Tong, S., Kumada, T., & Iwaki, S. (2019). Golden ratio: The attributes of facial attractiveness learned by CNN. In *Proceedings of the IEEE international conference on image processing* (pp. 2124–2128).

Lin, L., Liang, L., Jin, L., & Chen, W. (2019). Attribute-aware convolutional neural networks for facial beauty prediction. In *Proceedings of the international joint conference on artificial intelligence* (pp. 847–853).

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1638–1659.

Liu, L., Zhou, B., Zhao, J., & Ryan, B. D. (2016). C-image: City cognitive mapping through geo-tagged photos. *GeoJournal*, *81*(6), 817–861.

Liu, Y., Huang, K., Bao, J., & Chen, K. (2019). Listen to the voices from home: An analysis of Chinese tourists' sentiments regarding australian destinations. *Tourism Management*, *71*, 337–347.

Loh, Y. P., & Chan, C. S. (2019). Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, *178*, 30–42.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Lu, X., Suryanarayan, P., Adams Jr, R. B., Li, J., Newman, M. G., & Wang, J. Z. (2012). On shape and the computability of emotions. In *Proceedings of the ACM international conference on multimedia* (pp. 229–238).

Lu, Y., Tai, Y.-W., & Tang, C.-K. (2018). Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European conference on computer vision* (pp. 282–297).

Ma, Y., Zhang, Z., Chen, S., Yu, Y., & Tang, K. (2019). A comparative study of aggressive driving behavior recognition algorithms based on vehicle motion data. *IEEE Access*, *7*, 8028–8038.

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, *19*(12), 1253–1273.

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the ACM international conference on multimedia* (pp. 83–92).

MacLin, O. H., & MacLin, M. K. (2004). The effect of criminality on face attractiveness, typicality, memorability and recognition. *North American Journal of Psychology*, *6*(1), 145–154.

Macrae, C. N., & Lewis, H. L. (2002). Do I know you? processing orientation and face recognition. *Psychological Science*, *13*(2), 194–196.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5188–5196).

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, *120*(3), 233–255.

Mao, H., Jin, L., & Du, M. (2009). Automatic classification of Chinese female facial beauty using support vector machine. In *Proceedings of the IEEE international conference on systems, man and cybernetics* (pp. 4842–4846).

Marcinkowska, U. M., Kozlov, M. V., Cai, H., Contreras-Garduño, J., Dixson, B. J., Oana, G. A., Kaminski, G., Li, N. P., Lyons, M. T., Onyishi, I. E., Prasai, K., Pazhoohi, F., Prokop, P., Cardozo, S. L. R., Sydney, N., Yong, J. C., & Rantala, M. J. (2014). Cross-cultural variation in men's preference for sexual dimorphism in women's faces. *Biology Letters*, *10*(4), 20130850.

Marquardt, S., & Stephen, R. (2002). Marquardt on the golden decagon and human facial beauty. interview by dr. gottlieb. *Journal of Clinical Orthodontics*, *36*(6), 339–47.

Mathias, M., Benenson, R., Pedersoli, M., & Van Gool, L. (2014). Face detection without bells and whistles. In *Proceedings of the European conference on computer vision* (pp. 720–735).

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4040–4048).

McCurrie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S., & Scheirer, W. J. (2018). Convolutional neural networks for subjective face attributes. *Image and Vision Computing*, *78*, 14–25.

Moonon, A.-U., & Hu, J. (2015). Multi-focus image fusion based on nsct and nsst. *Sensing and Imaging*, *16*(1), 401–416.

Moustafa, A. A., Diallo, T. M., Amoroso, N., Zaki, N., Hassan, M., & Alashwal, H. (2018). Applying big data methods to understanding human behavior and health. *Frontiers in computational neuroscience*, *12*, 84:1–4.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383.

Nguyen, A., Yosinski, J., & Clune, J. (2016). Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*.

Nguyen, T. V., & Liu, L. (2017). Smart mirror: Intelligent makeup recommendation and synthesis. In *Proceedings of the ACM international conference on multimedia* (pp. 1253–1254).

Niedenthal, P. M., & Kitayama, S. (2013). *The heart's eye: Emotional influences in perception and attention*. Academic Press.

Ning, Z., Zhou, G., Chen, Z., & Li, Q. (2018). Integration of image feature and word relevance: Toward automatic image annotation in cyber-physical-social systems. *IEEE Access*, *6*, 44190–44198.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (7), 971–987.

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

Osborne, P. (2000). *Traveling light: Photography, travel and visual culture*. Manchester University Press.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, *22*(9), 794–809.

O'Toole, A. J., Price, T., Vetter, T., Bartlett, J. C., & Blanz, V. (1999). 3D shape and 2D surface textures of human faces: The role of averages in attractiveness and age. *Image and Vision Computing*, *18*(1), 9–19.

Pajares, G., & De La Cruz, J. M. (2004). A wavelet-based image fusion tutorial. *Pattern Recognition*, *37*(9), 1855–1872.

Pallett, P. M., Link, S., & Lee, K. (2010). New "golden" ratios for facial beauty. *Vision Research*, *50*(2), 149–154.

Pan, S., Lee, J., & Tsai, H. (2014). Travel photos: Motivations, image dimensions, and affective qualities of places. *Tourism Management*, *40*, 59–69.

Parde, C. J., Castillo, C., Hill, M. Q., Colon, Y. I., Sankaranarayanan, S., Chen, J.-C., & O'Toole, A. J. (2017). Face and image representation in deep CNN features. In *Proceedings of the IEEE international conference on automatic face and gesture recognition* (pp. 673–680).

Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S., & O'Toole, A. J. (2019). Social trait information in deep convolutional neural networks trained for face identification. *Cognitive Science*, *43*(6), e12729:1–19.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British machine vision conference* (pp. 1–12).

Pashos, A., & Niemitz, C. (2003). Results of an explorative empirical study on human mating in germany: Handsome men, not high-status men, succeed in courtship. *Anthropologischer Anzeiger*, 331–341.

Patterson, G., Xu, C., Su, H., & Hays, J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, *108*(1-2), 59–81.

Paxton, A. (2020). The belmont report in the age of big data: Ethics at the intersection of psychological science and data science. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 347–372). American Psychological Association.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods*, *49*(5), 1630–1638.

Penton-Voak, I. S., Jacobson, A., & Trivers, R. (2004). Populational differences in attractiveness judgements of male and female faces: Comparing british and jamaican samples. *Evolution and Human Behavior*, *25*(6), 355–370.

Perrett, D. I., Lee, K. J., Penton-Voak, I. S., Rowland, D., Yoshikawa, S., Burt, D. M., Henzi, S., Castles, D. L., & Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature*, *394*(6696), 884–887.

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature*, *368*(6468), 239–242.

Perrett, D. I., Penton-Voak, I. S., Little, A. C., Tiddeman, B. P., Burt, D. M., Schmidt, N., Oxley, R., Kinloch, N., & Barrett, L. (2002). Facial attractiveness judgements reflect learning of parental age characteristics. *Proceedings of the Royal Society of London B: Biological Sciences*, *269*(1494), 873–880.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2017). Leveraging deep neural networks to capture psychological representations. *arXiv preprint arXiv:1706.02417*.

Proctor, R. W., & Xiong, A. (2020). From small-scale experiments to big data: Challenges and opportunities for experimental psychologists. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 35–58). American Psychological Association.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107.

Rafegas, I., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision Research*, *151*, 7–17.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, *8*(4), 364–382.

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1), 1–12.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, *57*, 199–226.

Rhodes, G. (2013). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception*, *42*(11), 1179–1199.

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science*, *10*(1), 52–58.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the international conference on machine learning* (pp. 2940–2949).

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358.

Rothe, R., Timofte, R., & Van Gool, L. (2016). Some like it hot-visual guidance for preference prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5553–5561).

Rowe, G., Hirsh, J. B., & Anderson, A. K. (2007). Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences*, *104*(1), 383–388.

Rubenstein, A. J., Langlois, J. H., & Roggman, L. A. (2002). What makes a face attractive and why: The role of averageness in defining facial beauty. In G. Rhodes & L. A. Zebrowitz (Eds.), *Advances in visual cognition, vol. 1. facial attractiveness: Evolutionary, cognitive, and social perspectives* (pp. 1–33). Westport, CT, US, Ablex Publishing.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Li, F.-F. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Russell, R., Porcheron, A., Sweda, J. R., Jones, A. L., Mauger, E., & Morizot, F. (2016). Facial contrast is a cue for perceiving health from the face. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(9), 1354–1362.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Workshop* *on proceedings of the IEEE international conference on computer vision* (pp. 397–403).

Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science*, *22*(9), 1183–1190.

Samuels, C. A., & Ewy, R. (1985). Aesthetic perception of faces during infancy. *British Journal of Developmental Psychology*, *3*(3), 221–228.

Schmid, K., Marx, D., & Samal, A. (2008). Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition*, *41*(8), 2710–2717.

Schmitz, T. W., De Rosa, E., & Anderson, A. K. (2009). Opposing influences of affective state valence on visual cortical encoding. *Journal of Neuroscience*, *29*(22), 7199–7207.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., & van Gerven, M. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.

Sellen, A. J., Fogg, A., Aitken, M., Hodges, S., Rother, C., & Wood, K. (2007). Do life-logging technologies support memory for the past? an experimental study using sensecam. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 81–90).

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13.

Shapley, R., & Hawken, M. J. (2011). Color in the cortex: Single-and double-opponent cells. *Vision Research*, *51*(7), 701–717.

Shen, H., Chau, D. K., Su, J., Zeng, L.-L., Jiang, W., He, J., Fan, J., & Hu, D. (2016). Brain responses to facial attractiveness induced by facial proportions: Evidence from an fMRI study. *Scientific Reports*, *6*, 35905:1–13.

Shen, L., & Bai, L. (2006). A review on Gabor wavelets for face recognition. *Pattern Analysis and Applications*, *9*, 273–292.

Shen, Y., Ge, M., Zhuang, C., & Ma, Q. (2018). Sightseeing value estimation by analysing geosocial images. *International Journal of Big Data Intelligence*, *5*(1-2), 31–48.

Shi, S., Gao, F., Meng, X., Xu, X., & Zhu, J. (2019). Improving facial attractiveness prediction via co-attention learning. In *2019 IEEE international conference on acoustics, speech and signal processing* (pp. 4045–4049).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srinivasan, N., & Hanif, A. (2010). Global-happy and local-sad: Perceptual processing affects emotion identification. *Cognition and Emotion*, *24*(6), 1062–1069.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological Science*, *25*(2), 511–518.

Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in phuket, thailand. *Tourism Management*, *75*, 550–568.

Tamir, M., & Robinson, M. D. (2007). The happy spotlight: Positive mood and selective attention to rewarding information. *Personality and Social Psychology Bulletin*, *33*(8), 1124–1136.

Tarrés, F. (2012). Gtav face database. *http://gps-tsc. upc. es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase. html.*

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). YFCC100M: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64–73.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(9), 1226–1238.

Urry, J. (1992). The tourist gaze “revisited”. *American Behavioral Scientist*, *36*(2), 172–186.

Valdivia, A., Hrabova, E., Chaturvedi, I., Luzón, M. V., Troiano, L., Cambria, E., & Herrera, F. (2019). Inconsistencies on TripAdvisor reviews: A unified index between users and sentiment analysis methods. *Neurocomputing*, *353*, 3–16.

Valenzano, D. R., Mennucci, A., Tartarelli, G., & Cellerino, A. (2006). Shape analysis of female facial attractiveness. *Vision Research*, *46*(8-9), 1282–1291.

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142:1–6.

Vedaldi, A., & Lenc, K. (2014). Matconvnet-convolutional neural networks for matlab. *arXiv preprint arXiv:1412.4564*.

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361.

Vinson, D. W., Dale, R., & Jones, M. N. (2016). Decision contamination in the wild: Sequential dependencies in yelp review ratings. In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 1433–1438).

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*(11), 12–12.

Wang, P., & Cottrell, G. W. (2017). Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of Vision*, *17*(4), 9–9.

Wang, S., Shao, M., & Fu, Y. (2014). Attractive or not?: Beauty prediction with attractiveness-aware encoders and robust late fusion. In *Proceedings of the ACM international conference on multimedia* (pp. 805–808).

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257.

Whitehill, J., & Movellan, J. R. (2008). Personalized facial attractiveness prediction. In *Proceedings of the IEEE international conference on automatic face and gesture recognition* (pp. 1–7).

Winter, S. (2003). Route adaptive selection of salient features. In *International conference on spatial information theory* (pp. 349–361).

Workman, S., Souvenir, R., & Jacobs, N. (2017). Understanding and mapping natural beauty. In *Proceedings of the IEEE international conference on computer vision* (pp. 5589–5598).

Xu, L., Fan, H., & Xiang, J. (2019). Hierarchical multi-task network for race, gender and facial attractiveness recognition. In *Proceedings of the IEEE international conference on image processing* (pp. 3861–3865).

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yang, M., Zhang, L., Shiu, S. C., & Zhang, D. (2013). Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary. *Pattern Recognition*, *46*(7), 1865–1878.

Yang, Y., Tong, S., Huang, S., & Lin, P. (2015). Multifocus image fusion based on nsct and focused area detection. *IEEE Sensors Journal*, *15*(5), 2824–2838.

Yoo, J.-Y., Kim, J.-N., Shin, K.-J., Kim, S.-H., Choi, H.-G., Jeon, H.-S., Koh, K.-S., & Song, W.-C. (2013). Centralization or decentralization of facial structures in Korean young adults. *Journal of Craniofacial Surgery*, *24*(3), 1007–1010.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

You, Q., Luo, J., Jin, H., & Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence.*

Yu, J., Sharpe, S. M., Schumann, A. W., & Boyd, N. S. (2019). Deep learning for image-based weed detection in turfgrass. *European Journal of Agronomy*, *104*, 78–84.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European conference on computer vision* (pp. 818–833).

Zhang, L., Zhang, D., Sun, M.-M., & Chen, F.-M. (2017). Facial beauty analysis based on geometric feature: Toward attractiveness assessment application. *Expert Systems with Applications*, *82*, 252–265.

Zhao, J., Zhang, M., He, C., & Zuo, K. (2019). Data-driven research on the matching degree of eyes, eyebrows and face shapes. *Frontiers in Psychology*, *10*, 1466:1–11.

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., & Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. In *Proceedings of the ACM international conference on multimedia* (pp. 47–56).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.

Zhou, B., Liu, L., Oliva, A., & Torralba, A. (2014). Recognizing city identity via attribute analysis of geo-tagged images. In *Proceedings of the European conference on computer vision* (pp. 519–534).

Zhuang, C., Ma, Q., Liang, X., & Yoshikawa, M. (2014). Anaba: An obscure sightseeing spots discovering system. In *Proceedings of the IEEE international conference on multimedia and expo* (pp. 1–6).

Zhuang, C., Ma, Q., & Yoshikawa, M. (2017). SNS user classification and its application to obscure poi discovery. *Multimedia Tools and Applications*, *76*(4), 5461–5487.

Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *Proceedings of the European conference on computer vision* (pp. 391–405).

Zoran, D., Isola, P., Krishnan, D., & Freeman, W. T. (2015). Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE international conference on computer vision* (pp. 388–396).

Zou, J., Ji, Q., & Nagy, G. (2007). A comparative study of local matching approach for face recognition. *IEEE Transactions on Image Processing*, *16*(10), 2617–2628.

Appendix A

Selected List of Publications

Journal Article

- **Tong, S.**, Liang, X., Kumada, T., & Iwaki, S. (2020). Putative ratios of facial attractiveness in a deep neural network. *Vision Research*, 178, 86-99.
- **Tong, S.**, Loh, Y. P., Liang, X., & Kumada, T. (2019). Wide or narrow? A visual attention inspired model for view-type classification. *IEEE Access*, 7, 48725-48738.
- Yang, Y., **Tong, S.**, Huang, S., Lin, P., & Fang, Y. (2017). A hybrid method for multi-focus image fusion based on fast discrete curvelet transform. *IEEE Access*, 5, 14898-14913.

International Conference

- Liang, X., **Tong, S.**, Kumada, T., & Iwaki, S. (2019). Golden ratio: The attributes of facial attractiveness learned by CNN. In *Proceedings of the IEEE international conference on image processing* (pp. 2124-2128).
- Liu, Y., Liang, X., **Tong, S.**, & Kumada, T. (2019). Photo shot-type disambiguation by multi-classifier semi-supervised learning. In *Proceedings of the IEEE international conference on image processing* (pp. 2466-2470).
- Liang, X., **Tong, S.**, Kumada, T., & Loh, Y. P. (2019). Understanding the photo-shooting patterns of sightseeing. In *Proceedings of the international conference on data science and information technology* (pp. 36-41).

- **Tong, S.**, Liang, X., Kumada, T., Iwaki, S., & Tosa, N. (2017). Learning the cultural consistent facial aesthetics by convolutional neural network. In *Proceedings of the international conference on culture and computing* (pp. 97-103). (Best paper honorable mention award)
- Peng Loh, Y., **Tong, S.**, Liang, X., Kumada, T., & Chan, S. C. (2017). Understanding scenery quality: A visual attention measure and its computational model. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 289-297).
- **Tong, S.**, Loh, Y. P., Liang, X., & Kumada, T. (2016). Visual attention inspired distant view and close-up view classification. In *Proceedings of the IEEE international conference on image processing* (pp. 2787-2791).

Domestic Conference

- **Tong, S.**, & Liang, X. (2015). The distant view and close-up view classification using DWT and PCANet, In *Proceedings of 18th meeting on image recognition and understanding.*

Preprint

- **Tong, S.**, Liang, X., Kumada, T., Loh, Y. P., & Nakashima R. (2021). Happy travelers take big picture: The relationship between tourist satisfaction and photo-shooting behaviors, In *Proceedings of the ACM international conference on multimedia.* (Under Review)

- Liang, X., Fan, L., Loh, Y. P., Liu, Y., & **Tong, S.** (2017). Happy travelers take big pictures: A psychological study with machine learning and big data. arXiv preprint arXiv:1709.07584.

Appendix B

Datasets

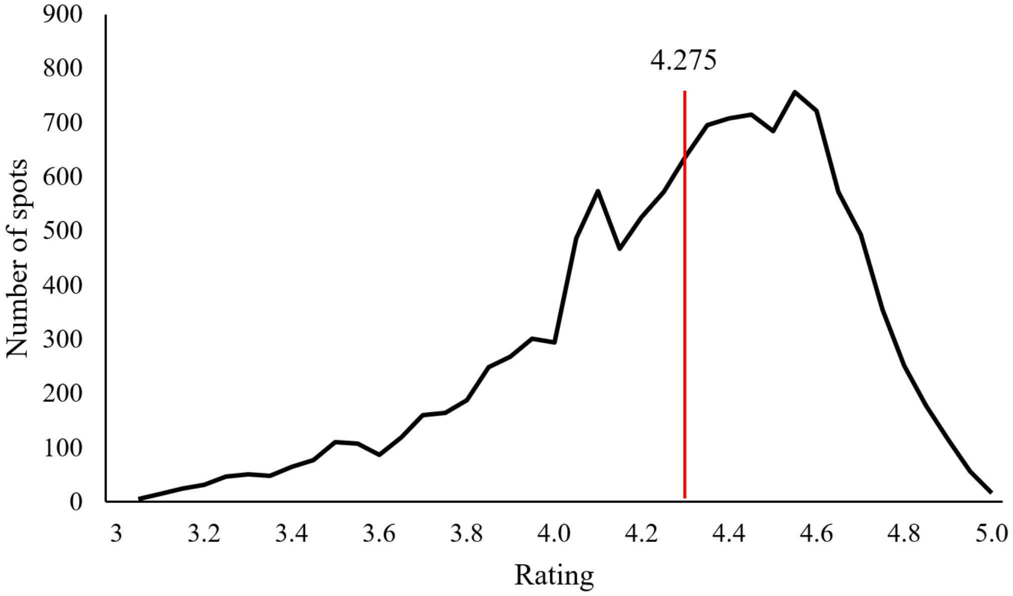


Figure B.1: The experience rating distribution of 12,482 scenic spots collected from *TripAdvisor*. The median of the experience ratings among these spots is 4.275.

Table B.1: The downloaded Photos from *Flickr*. N_p^s and N_{wp}^s represent the number of photos and wide-view photos in s-th scenic spots.

No. (s)	Spot name	N_p^s	N_{wp}^s (CFCNN)	N_{wp}^s (Human)
1	Hollywood Walk of Fame	999	209	182
2	Haight-Ashbury	999	312	291
3	Cannery Row	994	416	377
4	Fisherman's Wharf	1,002	432	390
5	Venice Beach	986	475	479
6	Torrey Pines State Natural Reserve	1002	644	628
7	Fort Rosecrans Cemetery	918	579	595
8	Carmel City Beach	1003	729	737
9	Yosemite Valley	975	732	842
10	Dantas View	1,045	918	911
11	Manneken Pis	998	196	-
12	The Little Mermaid Den Lille Havfrue	1,997	919	-
13	Jungle Island	982	259	-
14	Marina Beach	640	334	-
15	Khaosan Road	1,006	174	-
16	Las Ramblas	990	330	-
17	Ho Chi Minh Mausoleum	1,015	559	-
18	Television Tower	844	513	-
19	Tiananmen Square	997	599	-
20	National Monument	1,033	379	-
21	Kiyomizu dera Temple	988	598	-
22	Tower Bridge	1,998	1,480	-
23	Chichen Itza	1,753	1,279	-
24	Eiffel Tower	1,982	1,529	-
25	Colosseum	1,970	1,578	-
26	Sydney Opera House	1,994	1,589	-
27	Parthenon	1,998	1,437	-
28	Golden Gate Bridge	1,993	1,553	-
29	Taj Mahal	2,000	1,638	-
30	Bryce Canyon National Park	1,998	1,671	-

Table B.2: Regions about the downloaded scenic spots from *TripAdvisor*.

No. (i)	Spot name	Country	Continent
1	Manneken Pis	Belgium	Europe
2	Pattaya Floating Market	Thailand	Asia
3	Chessington World of Adventures Resort	United Kingdom	Europe
4	The Little Mermaid Den Lille Havfrue	Denmark	Europe
5	Beijing Zoo	China	Asia
6	Gold and Silver Pawn Shop	Colombia	South America
7	Hollywood Walk of Fame	United States	North America
8	Hin Ta Hin Yai Rocks	Thailand	Asia
9	Yogyakarta Palace	Indonesia	Asia
10	Central Market Mercado Central	Chile	South America
11	Bourbon Street	United States	North America
12	Casa di Giulietta	Italy	Europe
13	Marineland	France	Europe
14	Brindavan Garden	India	Asia
15	Chiang Mai Zoo	Thailand	Asia
16	Pinnawala Elephant Orphanage	Sri Lanka	Asia
17	Poble Espanyol	Spain	Europe
18	Chinatown Kuala Lumpur	Malaysia	Asia
19	Christiania	Denmark	Europe
20	Temple Street Night Market	China	Asia
21	Parque Isla Magica	Spain	Europe
22	Fuente de los Candados	Peru	South America
23	Little Havana	United States	North America
24	Santa Justa Lift	Portugal	Europe
25	Little India	Singapore	Asia
26	John Lennon Wall	Czech Republic	Europe
27	Khao San Road	Thailand	Asia
28	Holy Land Experience	United States	North America
29	Torre del Oro	Spain	Europe
30	La Boca	Argentina	South America
31	Vaci Street	Hungary	Europe
32	Atomium	Belgium	Europe
33	Charminar	Italy	Europe
34	Haight Ashbury	United States	North America
35	Cannery Row	United States	North America
36	Las Ramblas	Spain	Europe
37	Hachiko	Italy	Europe
38	Dr Sun Yat Sen Classical Chinese Garden	Canada	North America
39	Medina of Tunis	Iraq	Asia
40	Tram 28	Portugal	Europe
41	Prater	Austria	Europe
42	Arab Street	Singapore	Asia
43	Taipei 101	Taiwan	Asia
44	Fisherman s Wharf	United States	North America
45	Bayside Marketplace	United States	North America
46	Parque Kennedy Parque Central de Miraflores	Peru	South America

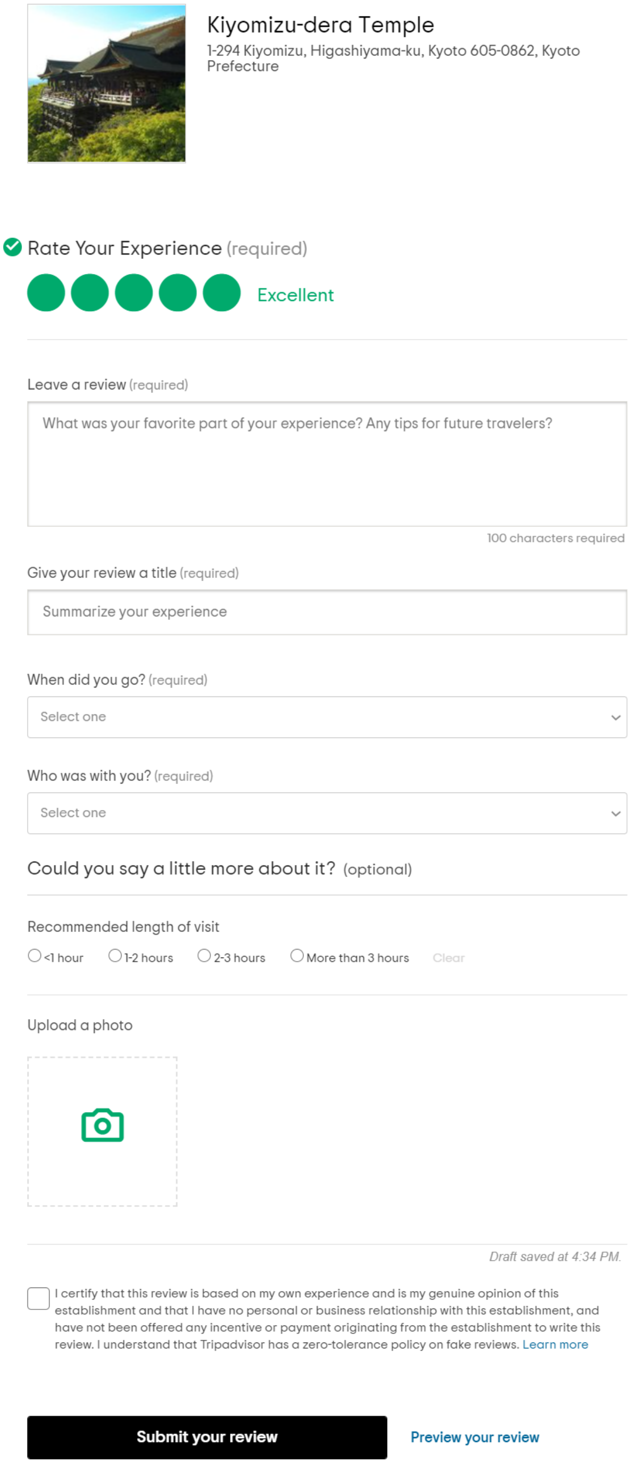
No. (i)	Spot name	Country	Continent
47	Merlion Park	Singapore	Asia
48	Santa Monica Pier	United States	North America
49	Gateway of India	India	Asia
50	Galata Bridge	Turkey	Asia
51	Insadong	South Korea	Asia
52	Torre de Belem	Portugal	Europe
53	Castelo de S Jorge	Portugal	Europe
54	Lombard Street	United States	North America
55	Downtown Disney	United States	North America
56	Eureka Skydeck 88	Australia	Oceania
57	Trevi Fountain	Italy	Europe
58	Trinity College Dublin	Ireland	Europe
59	17 Mile Drive	United States	North America
60	Edinburgh Castle	United Kingdom	Europe
61	Coronado Bridge	United States	North America
62	Kiyomizu dera Temple	Japan	Asia
63	Old Tel Aviv Port Area	Israel	Asia
64	Piazza Venezia	Italy	Europe
65	Castello Sforzesco	Italy	Europe
66	Victoria Harbour	China	Asia
67	Tower Bridge	United Kingdom	Europe
68	Griffith Park	United States	North America
69	La Jolla Cove	United States	North America
70	Chichen Itza	Italy	Europe
71	Eiffel Tower	France	Europe
72	Royal Botanic Garden Edinburgh	United Kingdom	Europe
73	Colosseum	Italy	Europe
74	Sydney Opera House	Australia	Oceania
75	Parthenon	United States	North America
76	Golden Gate Bridge	United States	North America
77	AT & T Park	United States	North America
78	Yosemite Valley	United States	North America
79	Carmel City Beach	United States	North America
80	Torrey Pines State Natural Reserve	United States	North America
81	Fort Rosecrans Cemetery	United States	North America
82	Centro Storico	Italy	Europe
83	Niagara Falls	Canada	North America
84	Taj Mahal	India	Asia
85	Old Quebec	Canada	North America
86	Dante s View	United States	North America
87	Savannah Historic District	United States	North America
88	Zona Arqueologica Teotihuacan	Mexico	North America
89	Killarney National Park	Ireland	Europe
90	Sassi di Matera	Italy	Europe
91	Grand Canyon South Rim	United States	North America
92	Gettysburg National Military Park	United States	North America
93	Iguazu Falls	Brazil	South America
94	Bryce Canyon National Park	United States	North America

Table B.3: The averaged experience rating (R^i) and the wide percentages (P^i_{wp}) of 94 scenic spots. Note that the ratings of scenic spots here are calculated by the photo providers only, thus, some of the scenic spots are different from the ratings as shown in Table 5.2, which are calculated by the review providers. N^i_p and N^i_{pp} represent the numbers of photos and photo providers of i-th scenic spots, respectively.

No. (i)	N^i_p	N^i_{pp}	R^i	P^i_{wp}	Scenery type
1	4,082	2,324	3.519	0.067	Historic Sites
2	1,900	369	3.680	0.403	Landmarks
3	1,884	475	3.958	0.283	Parks
4	4,080	2,203	3.659	0.294	Landmarks
5	1,791	438	3.493	0.257	Parks
6	1,452	480	3.885	0.312	Neighborhoods
7	7,999	2,193	3.595	0.208	Landmarks
8	1,477	484	3.713	0.631	Landmarks
9	1,415	325	3.711	0.473	Historic Sites
10	1,828	589	3.616	0.313	Neighborhoods
11	1,764	571	3.930	0.497	Landmarks
12	5,587	1,744	3.725	0.329	Museums
13	1,737	442	3.982	0.310	Parks
14	850	176	3.869	0.808	Parks
15	1,647	391	3.693	0.194	Parks
16	3,446	801	3.869	0.399	Nature
17	1,996	321	3.966	0.641	Museums
18	2,593	709	3.788	0.358	Neighborhoods
19	2,125	575	3.744	0.501	Neighborhoods
20	677	205	3.629	0.405	Neighborhoods
21	515	140	3.729	0.564	Parks
22	457	237	3.827	0.335	Parks
23	1,149	254	4.181	0.242	Neighborhoods
24	4,224	1,348	3.775	0.819	Landmarks
25	1,992	388	4.003	0.477	Neighborhoods
26	1,471	530	3.708	0.187	Landmarks
27	4,101	1,311	4.063	0.380	Neighborhoods
28	637	166	4.337	0.418	Museums
29	1,343	376	3.923	0.726	Landmarks
30	2,601	546	4.168	0.414	Neighborhoods
31	1,120	126	4.063	0.554	Landmarks
32	7,727	2,407	4.098	0.377	Museums
33	1,258	352	4.045	0.649	Religious Sites
34	896	192	4.167	0.485	Neighborhoods
35	1,005	272	4.154	0.663	Historic Sites
36	7,996	2,368	4.182	0.606	Neighborhoods
37	1,012	593	4.148	0.311	Landmarks

No. (i)	N_p^i	N_{pp}^i	R^i	P_{wp}^i	Scenery type
38	836	227	4.137	0.597	Landmarks
39	1,310	125	4.208	0.433	Neighborhoods
40	4,253	1,914	4.197	0.542	walks
41	5,033	1,260	4.189	0.689	Parks
42	988	231	4.087	0.587	Neighborhoods
43	7,062	1,960	4.196	0.616	Landmarks
44	5,710	1,894	4.272	0.572	Neighborhoods
45	3,150	939	4.327	0.670	Landmarks
46	740	202	4.366	0.518	Parks
47	5,312	1,741	4.305	0.732	Historic Sites
48	8,889	2,488	4.282	0.729	Landmarks
49	2,877	1,039	4.286	0.789	Landmarks
50	1,842	533	4.330	0.817	Bridges
51	2,533	625	4.331	0.401	Neighborhoods
52	12,484	4,489	4.327	0.858	Landmarks
53	12,877	3,245	4.343	0.798	Historic Sites
54	5,462	1,938	4.282	0.828	Landmarks
55	992	291	4.354	0.374	Landmarks
56	2,722	857	4.450	0.759	Landmarks
57	24,816	12,023	4.632	0.638	Landmarks
58	3,421	998	4.416	0.709	Historic Sites
59	3,848	937	4.679	0.843	Nature
60	19,045	4,722	4.512	0.728	Historic Sites
61	651	269	4.487	0.935	Bridges
62	9,734	2,243	4.445	0.735	Historic Sites
63	582	131	4.527	0.755	walks
64	1,587	495	4.517	0.887	Landmarks
65	6,606	1,379	4.513	0.728	Historic Sites
66	2,505	779	4.502	0.876	Marinas
67	16,561	5,203	4.680	0.763	Landmarks
68	697	199	4.608	0.845	Parks
69	5,829	1,625	4.725	0.665	Beaches
70	18,805	4,697	4.675	0.819	Historic Sites
71	59,439	18,827	4.681	0.850	Landmarks
72	2,559	508	4.791	0.507	Art Galleries,
73	51,505	16,701	4.682	0.889	Historic Sites
74	9,589	2,863	4.729	0.791	Landmarks
75	4,901	1,490	4.703	0.897	Historic Sites
76	18,651	6,257	4.754	0.851	Historic Sites
77	2,344	723	4.714	0.663	Landmarks
78	4,588	1,094	4.793	0.879	Nature
79	880	271	4.727	0.884	Beaches
80	1,231	364	4.783	0.859	Parks
81	255	101	4.703	0.919	Cemeteries
82	516	63	4.810	0.323	Landmarks
83	16,515	4,382	4.840	0.882	Landmarks
84	15,876	4,355	4.817	0.811	Historic Sites
85	5,774	1,196	4.816	0.780	Neighborhoods
86	942	362	4.831	0.894	Landmarks

No. (i)	N_p^i	N_{pp}^i	R^i	P_{wp}^i	Scenery type
87	3,262	710	4.848	0.758	Historic Sites
88	5,067	1,211	4.822	0.785	Ancient Ruins,
89	2,424	631	4.875	0.830	Parks
90	8,301	1,837	4.876	0.871	Landmarks
91	16,923	4,271	4.893	0.862	nature
92	3,051	700	4.870	0.676	Parks
93	19,385	5,582	4.906	0.752	Nature
94	6,201	1,413	4.932	0.906	Parks



Kiyomizu-dera Temple
1-294 Kiyomizu, Higashiyama-ku, Kyoto 605-0862, Kyoto Prefecture

✓ Rate Your Experience (required)

● ● ● ● ● Excellent

Leave a review (required)

What was your favorite part of your experience? Any tips for future travelers?

100 characters required

Give your review a title (required)

Summarize your experience

When did you go? (required)

Select one

Who was with you? (required)

Select one

Could you say a little more about it? (optional)

Recommended length of visit

○ <1 hour ○ 1-2 hours ○ 2-3 hours ○ More than 3 hours Clear

Upload a photo

Draft saved at 4:34 PM.

□ I certify that this review is based on my own experience and is my genuine opinion of this establishment and that I have no personal or business relationship with this establishment, and have not been offered any incentive or payment originating from the establishment to write this review. I understand that Tripadvisor has a zero-tolerance policy on fake reviews. Learn more

Figure B.2: The questionnaire survey provided by *TripAdvisor*. Firstly, tourists make an overall rating of the travel experience — ‘Excellent’, ‘Very good’, ‘Average’, ‘Poor’, ‘Terrible’. Secondly, they will be asked to leave a review to describe — “What are your favorite part of your experience? Any tips for future travelers”. Thirdly, upload the photos.