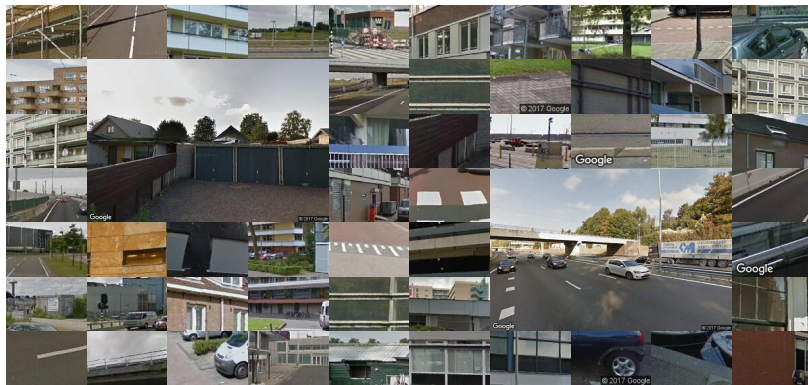


Quantifying and Predicting Urban Attractiveness with Street-View Data and Convolutional Neural Networks

Master's Thesis



Hendra Hadhil Choiri

Quantifying and Predicting Urban Attractiveness with Street-View Data and Convolutional Neural Networks

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK DATA SCIENCE AND TECHNOLOGY

by

Hendra Hadhil Choiri
born in Sragen, Indonesia



Web Information Systems
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
<http://wis.ewi.tudelft.nl>

Quantifying and Predicting Urban Attractiveness with Street-View Data and Convolutional Neural Networks

Author: Hendra Hadhil Choiri
Student id: 4468457
Email: HendraHadhilChoiri@student.tudelft.nl

Abstract

Analysing attractiveness of places in a region is beneficial to support urban planning and policy making. However, the attractiveness of a place is a subjective and high-level concept which is difficult to quantify. The existing methods rely on traditional surveys which may require high cost and have low scalability. This thesis attempts to quantify attractiveness of a place in a more efficient way and develop a model which can automatically predict attractiveness based on Street-View data (i.e. from Google Street View).

As a study case, 800 Google Street View images from 200 locations in Amsterdam have been extracted, and their attractiveness perceptions have been evaluated via crowd-sourcing to get the ground-truth information. The other attributes which are presumed to have a relationship with attractiveness are also assessed, such as familiarity, uniqueness, friendliness, pleasure, arousal, and dominance. The research and analysis revealed several insights related to the attractiveness of places. Attractive perception when seeing a place is positively correlated with perception of uniqueness, friendliness, pleasure, and dominance. Moreover, pleasure is possibly multi-collinear with attractiveness. It was also found that attractiveness perception has low spatial auto-correlation, which means that nearby places do not necessarily have similar attractiveness. Some visual features related to attractiveness were also investigated. The result indicated that scenes related to roads and residential buildings are less attractive, meanwhile, scenes related to greenery, blue sky, and water environment are more attractive.

A Convolutional Neural Network (CNN) model has been developed via machine learning approach which could automatically predict attractiveness perception of a place based on its representing Google Street View image. The developed model achieved 55.9% accuracy and RMSE of 0.70 to predict attractiveness in 5 ordinal values.

Thesis Committee:

Chair: Prof. Geert-Jan Houben, Faculty EEMCS, TUDelft
University supervisor: Dr. Alessandro Bozzon, Faculty EEMCS, TUDelft
Daily Supervisor: Dr. Achilleas Psyllidis, Faculty EEMCS, TUDelft
Committee Member: Dr. Cynthia Liem, Faculty EEMCS, TUDelft

Preface

All praises and thanks to Allah SWT for His blessings so that I can finish my study in TU Delft.

I would like to thank **Dr. Alessandro Bozzon** and **Dr. Achilleas Psyllidis** for their motivation, guidance, and feedback throughout my thesis work. They are always open whenever I have issues and questions about my research or my writing. We had some insightful discussions that really helped to improve the quality of my research.

I would also like to thank the other members of my supervision team, **Dr. Judith Redi**, **Dr. Pavel Kucherbaev**, and **Jie Yang** for the given suggestions as well as for providing me the resources that were needed during my thesis work. Moreover, I give huge appreciation to the participants of the survey conducted for this thesis. Without them all, this thesis could not be done.

Last, but not the least, I want to give my best thanks to my dear parents (**Hardani** and **Siti Aminah**), brothers, and sisters who always support me and pray for my success. To my fellow Indonesian guys in Computer Science, **Arkka**, **Hesa**, and **Sukma**, thanks for the great moments we shared during the past two years and the supports in the studies and projects. Big thanks to all members of **Keluarga Muslim Delft (KMD)** and **Perhimpunan Pelajar Indonesia Delft (PPI Delft)**, which became my second family and made Delft become like home. Finally, special thanks to the scholarship from **Lembaga Pengelola Dana Pendidikan (LPDP)** for providing me the financial support and opportunity to study at this amazing university.

Hendra Hadhil Choiri
Delft, the Netherlands
20th September 2017

Contents

Preface	iii
Contents	v
List of Figures	vii
1 Introduction	1
1.1 Background	1
1.2 Research Questions	3
1.3 Methods	4
1.4 Contributions	5
1.5 Outline	6
2 Related Works	9
2.1 Quantifying Urban Attractiveness	9
2.2 Quantifying Urban Perception by Means of Street-View Images	10
2.3 Development of Urban Perception Prediction System	11
2.4 Application of Urban Perception Quantification	13
2.5 Chapter Summary	14
3 Urban Attractiveness Quantification and Dataset Generation	17
3.1 Data Acquisition	17
3.2 Data Labelling	18
3.3 Attractiveness Quantification	22
3.4 Public Crowd-sourcing	29
4 Urban Attractiveness Prediction System	35
4.1 Convolutional Neural Network (CNN)	35
4.2 Dataset Expansion	41
5 System Evaluation and Understanding Urban Attractiveness	45
5.1 Urban Attractiveness Model Training	45
5.2 Final Trained Attractiveness Model	49
5.3 Visual Aspects Related to Urban Attractiveness	51

6 Discussion	57
6.1 Discussion	57
6.2 Threats to Validity	59
7 Conclusions	63
7.1 Conclusions	63
7.2 Outlook	65
Bibliography	67
A Source Codes for Development and Analysis	70
A.1 GitHub Repository for Development and Analysis	70
A.2 GitHub Repository for Crowd-sourcing Interface Website	71
B Google Street View API	72
C Crowd-sourcing Website	73
C.1 Data Model	73
C.2 Procedure	73
D Test The Qualification of Applying Exploratory Factor Analysis	75
E Crowd-sourcing Data	77
F Image Pre-processing	78

List of Figures

1.1	Diagram to Explain Research Questions and Methods in This Thesis . . .	7
2.1	Architecture of The 3-block rCNN Designed by [19]	12
3.1	Example of Street View data, four images represent a location	18
3.2	Questions in The Crowd-sourcing Task	21
3.3	Aggregated Attractiveness Label Distribution From Pilot Crowd-sourcing in Image Level (Left) and Location Level (Right)	23
3.4	Accuracy Distribution of Object Annotations on Golden Questions (by Using Exact Match Criteria)	24
3.5	Distributions of Answers Correctness of Object Annotation Tasks for each Golden Image	24
3.6	Variances of Attractiveness Label Judgments per Image (Top) and per Lo- cation (Bottom), Ordered from Low to High	25
3.7	Example of anomalous images. In the crowd-sourcing, each image has attractiveness value of 4. When all of them is shown together, the overall attractiveness is only 2.	26
3.8	Examples of location which the attractiveness is labelled 1 or 2 by Amer- ican/European Workers and labelled 4 or 5 by Asian Workers in AMT . .	32
3.9	Aggregated Attractiveness Label Distribution from Public Crowd-sourcing in AMT	32
3.10	Variances of Attractiveness Label Judgments per Location from Public Crowd-sourcing, Ordered from Low to High	32
3.11	Aggregated Attractiveness Label Distribution from Public Crowd-sourcing in AMT Grouped Based on Nationality	33
4.1	Architecture of CNN for The Attractiveness Prediction System	39
5.1	Attractiveness Distribution in Amsterdam Based on Assessed Dataset . .	50
5.2	Attractiveness Distribution in Amsterdam Based on Prediction System . .	51
5.3	Image Examples of Top Scenes in Each Attractiveness Category	53
5.4	Top 10 of The Most and The Least Attractive Places Based on The Pilot Crowd-sourcing	55
5.5	The Most Attractive Visual Patterns Based on The Developed CNN	55

5.6	The Least Attractive Visual Patterns Based on The Developed CNN	56
B.1	Example of HTTP request in Google Street View API and the returned image	72
C.1	Data Model Diagram of The Developed Crowd-sourcing Website	74
D.1	Scree Plot of Eigenvalue for each Factor Loading	76
E.1	Golden images and object annotation answer option list used in the crowd-sourcing. Options printed in bold indicate that the object appears in the image	77
F.1	Original Street-View Image to be Pre-Processed	78
F.2	Example of Re-sizing and Cropping Result of A Sample Street-View Image in Figure F.1	79

Chapter 1

Introduction

Urban planning and policy making require an understanding of city regions, in terms of both their physical and social structure, as well as in relation to high-level attributes (e.g., attractiveness, safety, walk-ability, etc.) of the urban environment. As cities become more complex, there is an increasing need to develop methods and implement tools that help to give insights into the structure and attributes of the urban environment. High-level concepts, such as the attractiveness of a place, are difficult to quantify and existing methods rely on traditional labor-intensive techniques (e.g. surveys) that cannot scale. From a computational perspective, it is also challenging to quantify such a high-level concept that relies on the subjectiveness of human perception.

Accordingly, it is essential to find an alternative approach to quantify attractiveness of a place, with less cost, better scalability, yet still reliably. This thesis proposes the quantification of the urban attractiveness with Street-View data, which is a kind of data containing coordinate of a location and a street-level image observed from that location (Street-View image). The attractiveness is assessed remotely by using crowd-sourcing and processed with statistical analyses. The research also aims to build a system that can automatically predict attractiveness of a place from its Street-View data, which is implemented through machine learning. The urban attractiveness dataset and prediction system generated from this thesis can be used for future research. This thesis also investigates the visual features and attributes that may be correlated to attractiveness.

This chapter introduces the background and motivation of the research. It also addresses the main research question and accompanying research sub-questions. Subsequently, the used methods and contributions of this research are elaborated. Finally, the structure of this document is outlined.

1.1 Background

Understanding the conditions of city regions is essential for various stakeholders, such as urban planners and governments. Numerous research have attempted to study the relationship between the physical structure of the regions to miscellaneous high-level attributes. One of the influential attributes is attractiveness.

Analysing urban attractiveness in a region can be useful to support the development of the city. Attractiveness of a city is a catalyst for sustainable economic growth [4]. Thus, a lot of cities try to make themselves more attractive by investing in provi-

sions related to various elements. The stakeholders can manage the accommodation, transportation, and other facilities to make the city more attractive and can attract more people to come and maintain social and economic contribution to the city.

Attractiveness of a place can be observed based on various perspectives. In the business point of view, a place is considered attractive if it can attract people to come [24]. As a foundation to that aspect, the other fundamental concept of attractiveness is from the environmental psychology perspective, which is related to human perception. Attractiveness of a place is determined based on how people perceive when being there and observe the views. Places which are perceived as attractive will attract more people to visit, either to live, to have a business, or to find pleasure, which leads to the economical benefit.

Some methods have been applied to quantify attractiveness perception. The conventional way is by managing on-site survey. The surveyors should go to the target location and assess the attractiveness based on some pre-defined criteria. This kind of method may be reliable, however, it is not scalable. To assess a new region, it requires a high cost to conduct more survey. Thus, more efficient way is required to conduct the assessment.

An alternative option is by remotely retrieving data from the city and assess their attractiveness without necessarily being there. Street-View image (street-level image observed from a location) is appropriate to represent a place. Street-View data of a location is defined as a Street-View image and the information of spatial data where that image is captured, such as its location coordinate (latitude and longitude), heading, and pitch (vertical angle). One of data sources that can provide Street-View data is Google Street View (GSV) ¹, which has a feature to remotely navigate places through interactive 360° panoramic images at the street level. This data source has been utilised for various urban analysis [21, 6].

Indeed, Google Street View has several advantages. This data source provides standard street-level view images with good resolution and has already covered extensive areas throughout the world. The specified location coordinates are also more precise because it is part of the features in Google Maps, which is largely used for navigation and location tracking. Another advantage is that GSV tries to normalise the captured images, which may minimise the effect of external factors (e.g. weather, amount of people, etc) on the perception of a place. Moreover, it is also simple to collect images and spatial data from Google Street View because it provides an API to facilitate image crawling based on coordinates and heading.

The attractiveness of a place can be quantified via Google Street View images representing it. The assessment is based on how people perceive a place when seeing its view. There are some challenges in using this approach. Observing image representation of a place may cause bias because the image is static and the observer does not have complete information of the area. Moreover, attractiveness is a complex and subjective attribute. So, the usage of the new source and assessment approach may produce noisy data. It is necessary to find a method that may minimise the noises and yield reliable data.

Despite the simpleness in collecting the Street-View data, the manual assessment is also not scalable. This issue can be solved by developing a system that can au-

¹<https://www.google.com/streetview/> (accessed 2016-10-15)

tomatically predict the attractiveness of a place from Street-View data. This system can replace human effort to assess the attractiveness. Nevertheless, considerate experiments and analyses are required to develop well-performed prediction system.

1.2 Research Questions

Motivated by previously elaborated background and challenges, this thesis aims to propose an alternative method to quantify and predict attractiveness of places to be more efficient and require lower cost, yet the result is still reliable. This objective is formulated into the following main research question.

MRQ How to implement a computational system that quantifies and predicts the attractiveness of places in city regions, based on Street-View data?

This main research question can be broken down into following research sub-questions.

RQ1 How to quantify the attractiveness of places in city regions by using Street-View data?

This research aims to provide a method to assess and quantify attractiveness of a place by using Street-View data. Attractiveness is a complex and subjective perception. So, the method should consider how the Street-View data can represent the actual place and be able to review the reliability of the assessment result. The proper analyses are also required to get more insights on attractiveness perception.

RQ2 How to develop a model that can automatically predict the attractiveness of places from Street-View data in city regions?

A thorough search of the relevant research yielded no machine-learning based model which attempts to automatically predict the attractiveness of a place based on its physical appearance. This research tries to develop that model, which can help to assess attractiveness of places in city regions in more efficient way. Considerate experiments and evaluations are required to develop an accurate model.

RQ3 How does the spatial dimension of the collected data affect the predictive performance of the machine learning model?

Besides the street-level images, Street-View data also contain spatial information, such as coordinates and headings of the observer when the image was captured. These information may contribute to improve the performance of the attractiveness prediction model. This research attempts to utilise them through semi-supervised approach in machine learning, and observe its impact on the model performance.

RQ4 Which visual features of the urban environment contribute to the attractiveness of a place in city regions?

Attractiveness of a place is related to people's perception when seeing it. Thus, there may be some visual patterns that make a place looks more or less attractive. This research tries to investigate the visual aspects of an environment that may relate to its attractiveness.

1.3 Methods

There are several activities done in this research to answer the research questions and tackle the challenges, which are outlined in the following ways. The summary of the research questions and the methods used in this thesis can be seen in Figure 1.1.

1.3.1 Dataset Generation

Street-View data are extracted from 200 locations in Amsterdam city (as a pilot study case) via Google Street View API. Each location is represented by four street-level images from four perpendicular headings to capture entire surrounding views and avoid bias. The overall attractiveness of the location is determined as the mean of attractiveness values from the four representing images.

The attractiveness assessment is done by means of crowd-sourcing. It requires several processes, such as defining the task questions, designing the protocol, and developing a crowd-sourcing interface website. Each image is assessed by multiple people, and their judgements are analysed to determine the attractiveness of the location. To get more reliable assessment, the crowd-sourcing is held internally in monitored lab sessions. After it is done, the judgments from the participants are validated, aggregated, and analysed. The output is an urban attractiveness dataset which contains collection of Street-View data from 200 locations in Amsterdam with attractiveness information as the label. In this thesis, these processes are called the dataset generation.

Attractiveness is a complex perception. It may be hard to be simply explained by using a single metric. There are various possible factors that may be associated to attractiveness. To help understanding it, other attributes that are suspected to be related should also be analysed, such as familiarity, uniqueness, friendliness, and emotion. These attributes can help to understand attractiveness by observing their correlations.

The main dataset in this thesis is assessed via internal crowd-sourcing, which the participants are limited and the process is guided and monitored. However, it may not be suitable for the assessment of huge amount of data. Instead, public crowd-sourcing is typically used (i.e. via Amazon Mechanical Turk). On the other hand, the results from public crowd-sourcing may contain more noises and be less reliable. To study its feasibility, a small sample from the dataset is assessed via Amazon Mechanical Turk and the result is compared to the result from the internal crowd-sourcing.

1.3.2 CNN Training

To deal with scalability issue, the automated attractiveness prediction system is developed by using machine learning approach. Machine learning is able to internally adapt on how to predict attractiveness based on the given training dataset. This trained model could automatically predict the attractiveness of new locations. It takes an image of location as the input, and predicts its attractiveness with a value in 5 Likert-scale. By

using this model, the attractive assessment of any new data can be performed without needing any more crowd-source.

The previously assessed dataset can be learnt by a machine learning algorithm to develop the attractiveness prediction model. Some research [19, 23] suggest that Convolutional Neural Network is one of the best models to handle image analytics, including classification task related to human perception. Some experiments are required to find suitable image processing, neural network design, and hyper-parameters that lead to an accurate trained model. The original dataset is split into training dataset and testing dataset. The model performance is measured based on its root mean square error (RMSE) towards the testing dataset.

1.3.3 Semi-Supervised Learning

The amount of labelled data obtained from crowd-sourcing is limited. Whereas, generally a huge size of training data is required to develop a machine learning model with high accuracy. One solution to handle this issue is by expanding the training dataset, which is generating new training data and estimate its label. When using image as the input, typical way to expand the dataset is by transforming the image (e.g. flipping, rotating, zooming, etc.). Google Street View data contain spatial information, such as location coordinate and heading. These information could be utilised to expand the dataset and boost the model's performance.

This research proposes semi-supervised learning methods by expanding the training dataset based on spatial information. The expansion works by adding Street-View data, and estimate their labels based on spatial proximity to existing locations in the original dataset. Some of the possible methods require an assumption that nearby locations are having correlated attractiveness, which can be tested by using spatial auto-correlation analysis.

1.3.4 Pattern Observation

The dataset and prediction model are analysed further to get visual features that may influence attractiveness of a place. It can be done by comparing images of attractive, neutral, and unattractive places. The relevant patterns can be obtained from various techniques, such as extracting scene categories and assessing attractiveness of image patches by using the developed CNN model.

1.4 Contributions

This thesis delivers several contributions which are explained as following.

1. Generate a new urban attractiveness dataset based on Street-View data

Various urban perceptions have been studied based on the physical appearance, which uses Street-View images to represent a location. For instances are perception of safety, liveliness, and wealth. However, research specific to attractiveness perception is still limited. On the other perspective, there are some existing approaches to quantify urban attractiveness. However, less studies measure it

based on images. This thesis combines those two aspects to gain a new perspective on the possibility of quantifying urban attractiveness based on images representing locations.

In this thesis, a collection of Street-View data from 200 places in Amsterdam has been retrieved and the attractiveness information (as well as other related attributes, such as familiarity, uniqueness, friendliness, and emotion) of each location has been assessed through controlled crowd-sourcing. Each place is represented by four Street-View images and each image is judged by 5 TU Delft students and staffs who have lived in Netherlands for less than 5 years. This urban attractiveness dataset can be used for further research in this subject. It can help to gain more insights on urban attractiveness, and can also be used to train and test another attractiveness prediction model.

2. Provide more insights on urban attractiveness related to physical appearance

This thesis shows that assessing attractiveness of a place by representing it with a single image from only one side may cause bias. Instead, images from the other headings should also be considered to determine the overall attractiveness of the location. Moreover, a spatial auto-correlation analysis indicated that neighbouring locations do not necessarily have similar attractiveness perception. The other investigation provides the information of visual features which may influence attractiveness. The analysis also revealed some attributes that strongly correlate to attractiveness, especially pleasure. These insights can help in urban planning and construction to monitor and improve attractiveness of the city.

3. Develop a trained machine learning model that can automatically predict attractiveness of a place from Street-View data

This thesis developed a Convolutional Neural Network model which can predict attractiveness of a place from its Street-View data. The output is an ordinal value from 1 to 5 with higher value indicates more attractive location. The trained model achieved an estimated accuracy of 55.9% and RMSE of 0.70 based on a validation dataset, which gives significant improvement from random prediction. Predicting attractiveness is relatively more challenging because human perception is subjective and has deeper representation than other image analytic tasks, such as object recognition and scene classification.

1.5 Outline

The organisation of this thesis is as follows. In chapter 2, the literature related to urban attractiveness and implementation of machine learning in urban perceptions are presented to understand the state of the art and related works. Chapter 3 elaborates the generation of urban attractiveness dataset and attractiveness quantification, including urban images acquisition, labelling via crowd-sourcing, and initial data analysis. The generated dataset is then used to develop urban attractiveness prediction system that is elaborated in Chapter 4. Chapter 5 shows the results and evaluations experiments, as well as the investigation to find visual features that may be related to attractiveness.

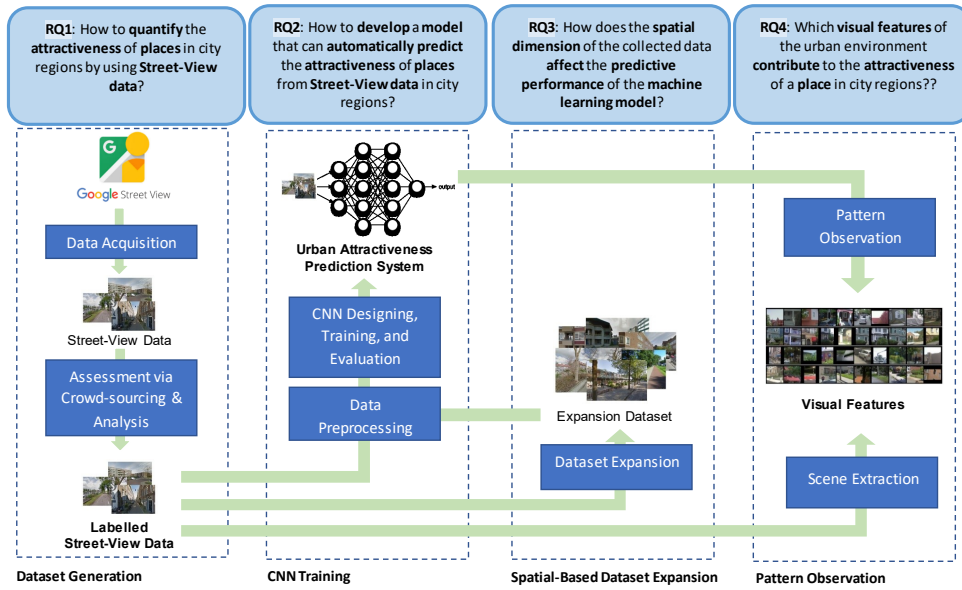


Figure 1.1: Diagram to Explain Research Questions and Methods in This Thesis

Moreover, the results and limitations are discussed in chapter 6. Finally in chapter 7, the studies are concluded and some future works are proposed.

Chapter 2

Related Works

This chapter elaborates various literature related to the thesis, which are broken down into four parts. The first part identifies the definition of urban attractiveness and some existing techniques to quantify it. The next part elaborates existing studies in quantifying urban perception specifically by using Street-View data. The third part presents literature related to the development of a system that can automatically predict urban perception, which may be adapted to develop the prediction system for urban attractiveness. The last part exposes some applications of urban perception quantification. After that, those previous works are summarised, and the contributions of this thesis related to them are stated.

2.1 Quantifying Urban Attractiveness

Urban attractiveness still has a vague definition and criteria. There are some studies to get further insights on quantifying attractiveness of places.

Hidalgo, et al. [9] conducted a survey to evaluate the attractiveness of places in Seville and Malaga, Spain. The participants are 58 residents of Malaga. There are several tasks in the survey. The first task is to list the most visually attractive and unattractive places of the city. The places are categorised into five main categories: cultural-historical places/landscapes, recreational places for leisure, panoramic places, housing areas, and industrial places. The result showed that the first three categories were considered as very attractive and the rest were very unattractive. Next, the participants are asked to fill the Perceived Restorativeness Scale (PRS). The result suggests that the most attractive places are more restorative than the most unattractive places. The third task is to assess the most attractive and the most unattractive places on a five-point scale for an 11-item battery with these aesthetic attributes: Vegetation, Visual richness, Congruence, Openness, Luminosity, Historic place, Cleanliness, Maintenance, Leisure, Meeting place, and Novel place. The result shows that the mean scores for each of them are significantly higher in attractive places than in unattractive ones.

Karmanov and Hamel [10] investigated the perception of attractiveness and interestingness of both the natural and the urban environment. The approach is by surveying participants having mildly stress and observing how the restorative potential of places can reduce their stress. The data are collected through Profile of Mood States (POMS)-

questionnaire. By using factor loading, they introduced two factors. The first factor is called 'attractiveness', which includes five scales: unfriendly-friendly, unpleasant-pleasant, unenjoyable-enjoyable, repulsive-inviting, unpersonal-personal. The other factor is called 'novelty' that contains four scales: simple-complex, dull-exciting, uninteresting-interesting, and average-exceptional. Further analysis also stated that the natural environment was significantly more attractive than the urban environment, while the urban environment was valued as more interesting.

Lankhorst, et al. [12] used GIS-based Landscape Appreciation Model (GLAM) to predict the attractiveness of the landscape by scoring physical aspects for each 250 x 250 meter cell in a grid map of Netherlands. The GLAM model consists of three positive indicators: Naturalness, Relief, Historical Distinctiveness, and three negative indicators: Skyline Disturbance, Urbanity and Noise Level. The result shows that the correlations (Pearson's) of each indicator pairs are below 0.30, except correlation of Naturalness-Relief which is 0.34. The predicted landscape attractiveness was formulated as a linear combination of these indicators by using regression weights.

Another approach of predicting attractiveness of a place is via digital footprints as proposed by Girardin, et al. [8]. The authors defined attractive places as places that have beneficial features for work, social interaction, or sightseeing purposes. Based on that definition, they assumed that attractive places will be visited by more people and thus leave higher density of digital footprints. They analysed two types of digital footprints generated by phones. The first one is cellular network activity, such as the number of calls, text messages, and network traffic. The other digital footprints type is photo activity, which is the number of photos and photographers in each location based on shared photos in Flickr.

Beside those approaches, there is a possibility of assessing urban attractiveness by means of Street-View data. This approach has been used to assess various urban perceptions which are elaborated in next section.

2.2 Quantifying Urban Perception by Means of Street-View Images

There have already been several research on quantifying and predicting urban perception by using images as the data.

Place Pulse 1.0 [21] is a pilot project for creating a dataset of urban images and using human perceptions as the labels. This dataset consists of 4,136 geo-tagged images from four cities in the US and Austria. The images from New York City (1,706) and Boston (1,236) were crawled from Google Street View, while the images from Linz (650) and Salzburg (544) were taken manually on site. They collected safety perception data through crowd-sourcing via a website. The given task is to select one of two given images to answer one of three questions: "Which place looks safer?" to assess safety, "Which place looks more upper-class?" to assess social class, and "Which place looks more unique" to assess uniqueness. The crowd-sourcing received 208,738 judgments cast by 7,872 individual participants from 91 countries. Next, each image is scored based on the win and loss ratio.

To develop a global dataset of urban appearance, Dubey, et al. [6] introduced a new crowdsourced dataset which is called Place Pulse 2.0. This dataset contains

1.17 million pairwise comparisons of 110,988 images from 56 cities in 28 different countries. There are six perceptual attributes that were annotated: safety, liveliness, boringness, wealth, depression, and beauty. The labeling method is the same as that of Place Pulse 1.0.

These research demonstrated that assessment of urban perception can be done by crowd-sourcing, and each location can be represented by Street-View image. Which means, this kind of approach may also be appropriate to assess other perceptions, such as attractiveness. The task of comparing image pairs may be more robust because it considers the relative perception difference between two images. It can also be used to order and rank the images. However, abundant image pairs are required to label a significant number of images.

Another option is to use rating task to assess the perception (e.g. rate the perception of each image with Likert scale or binary option). This approach is straightforward and easy to aggregate. It is also simpler to assign the score/label of an image.

This kind of approach is used in WeSense [27], which provides a procedure to collect urban images and assess some perceptions via a mobile application. That app lets the users take pictures around them and directly judge the photos. The attributes that are assessed including beauty, satisfaction, cleanliness, and tenseness. The pilot stage of this project is run in Amsterdam, as part of the AMS Institute (Amsterdam Institute for Advanced Metropolitan Solutions) initiative.

2.3 Development of Urban Perception Prediction System

Many research tried to use machine learning to develop a system that can predict urban perception of a place from its image. Some of the models that are already proven to perform well are the SVM and the Convolutional Neural Network (CNN).

Arietta, et al. [1] used Support Vector Regression (SVR) approach to building the predictor of seven city attributes (violent crime rate, theft rate, housing prices, population density, tree presence, graffiti presence, and perception of danger) based on visual appearance. The training dataset is created by extracting 10,000 Google Street View images and label them by interpolating the known data over the entire city with radial basis function (RBF). The extracted features are HOG + colour descriptor which was inspired by [5]. The experiments with the seven attribute data from each of six cities in US show vary results. For intra-city predictor, with the best result is achieved for prediction of housing prices in Boston with 82% accuracy, and the worst accuracy is 56% for graffiti presence prediction in Chicago. Meanwhile, cross-city predictions give more than 60% accuracy for most of the predictor.

This approach gives relatively low performance may because of the inaccuracy of the labels in the training data. The labelling via interpolation from the ground-truth data relies on a strong assumption and may cause high amount of noises. The other possible cause is that the extracted features may not representative enough to discriminate the predicted attributes.

Streetscore [16] trained an urban safety predictor by using v-Support Vector Regression (v-SVR). The training dataset is Place Pulse 1.0 with the safety label is ranked and scored by using Trueskill Algorithm. Streetscore considers these features: GIST, Geometric Classification Map, Texton Histograms, Geometric Texton Histograms,

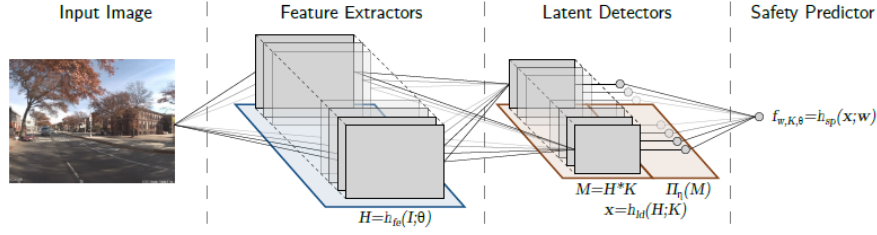


Figure 2.1: Architecture of The 3-block rCNN Designed by [19]

Color Histograms, Geometric Color Histograms, HOG2x2, Dense SIFT, LBP, Sparse SIFT histograms, and SSIM. By using an individual feature set, Geometric Texton Histograms shows the best performance, followed by GIST. By using forward selection to reduce the dimensionality, it was found that Geometric Texton Histograms, GIST, and Geometric Color Histograms are the top three features.

Porzi, et al. [19] has developed some machine learning models that try to predict urban safety from Street-View images, which was trained and tested by using Place Pulse dataset. They experimented with RankingSVM and Ranking Convolutional Neural Network (rCNN) to perform a ranking task which the objective is to automatically score Street-View images based on their safety perception.

They implemented RankingSVM for various types of features: GIST, HOG, SSIM (Self-similarity descriptors), features extracted from the sixth layer of Caffe reference network trained with ImageNet, features extracted from the sixth layer of Caffe reference network trained with PLACES dataset, and features derived from SUN Attribute dataset. The experiment result shows that the best performance is obtained by the Caffe reference network trained with PLACES dataset.

For the rCNN, they proposed an architecture which is described by Figure 2.1. This design consists of three compositional blocks as seen below.

1. Feature extractor
This block represents each input image into $r \times s \times t$ -dimensional features set. It uses the first 2 or 3 layers of the Alex Net.
2. Latent detector
This block receives the extracted features and uses m detectors of latent visual concepts. Each detector consists of a convolution with 3×3 linear kernel filters, a ReLU non-linearity, and a pooling operator. The pooling operator has a parameter η which combines average-pooling and max-pooling. Each detector will output a single scalar value.
3. Safety predictor
This block is a linear decision function with m weight parameters applied to each output of the latent detectors. This network uses the logistic loss function and the SGD solver with a momentum (μ) of 0.9 and a learning rate (α) of 0.1 (or $\alpha = 0.01$ for fine-tuning).

The experiment result shows that the best performance with accuracy of 70.25% is achieved with 2-layers feature extractor, 24 latent detectors split into four groups

with pooling factors 0, 0.01, 0.05, and 0.1. This experiment also confirms that CNN approach has better performance than SVM-based ones. However, this proposed CNN design is slightly outperformed by AlexNet-PLACES, which is an AlexNet model pre-trained by using PLACES dataset.

Based on this result, it seems that AlexNet-PLACES is a good feature extractor to predict safety from Street-View image. The research argued that deeper network typically guarantees better performance when the training data is sufficient. The experiment also revealed that AlexNet-PLACES has better performance than AlexNet-ImageNet, which means scene recognition is closer to perception prediction task, compared to object recognition. This study can be used as a starting point in developing a model to predict other urban attributes, in this case, is urban attractiveness for this thesis.

2.4 Application of Urban Perception Quantification

There are several useful applications of analysing and quantifying urban perception, which are explained as following.

2.4.1 Analysing The Impact of Changes in Neighbourhood's Physical Appearance

Naik, et al. [15] introduced a method to measure changes in the physical appearances of neighbourhoods from time-series Street-View images. They used in total of 1,645,760 Street-View images from Baltimore, Boston, Detroit, New York, and Washington DC captured in 2007 and 2014. They compared the images from the same locations, the same point of view (heading and pitch), but with a different time (2007 vs 2014).

For the comparison technique, each image is segmented into four geometric classes (ground, buildings, trees, and sky). Next, feature vectors (i.e. GIST and Texton Map) are extracted from each geometric class image and the features of streets and buildings are used to predict the safety of the place by using Streetscore [16]. The change of Streetscore value from 2007 to 2014 is computed which is called Streetchange, with positive value indicates upgrade in physical appearance, and vice versa. The Streetchange values are validated by both using human assessments and data from Boston's Planning and Development Authority (BPDA). The results showed a positive correlation to Streetchange.

Based on their study, there are three factors that relate physical appearance to economic and geographic data. First, population density and education of neighbourhoods correlate positively to the physical environments. Second, better initial appearance is also more likely to provide positive improvement. Third, physical proximity to other physically attractive neighbourhoods also gives positive correlation to neighbourhood improvement.

This research showed the benefit of using machine learning and Street-View data to understand the dynamics of a physical environment in the city. This method can also be applied in the context of urban attractiveness. The changes of attractiveness value (instead of Streetscore value which represents safety) from year to year can be

analysed and compared to the change of the other data, such as population density, education, and other data essential to city development.

2.4.2 Correlating to Other Data

Quantifying urban perception can be used to correlate it to other data. Salesses, et al. [21] studied the correlation between urban perception and crime rate in New York City. They found that the regression coefficients of the safety from Place Pulse 1.0 are negatively correlated to the crime rate, which means locations with safer looking have less crime. Meanwhile, social class has positive correlation to the crime rate, which means classier looking may lead to more crime.

Ordonez & Berg [18] also did a similar study for other cities. For Baltimore, they found a negative correlation between safety and crime rate ($\rho = -0.47$), and positive correlation between wealth and income data ($\rho = 0.61$). For Chicago, the correlations are weaker with $\rho = -0.21$ and $\rho = -0.32$ for safety-crime and wealth-income respectively.

These kinds of analysis can also be applied to attractiveness perception. By observing it with other data (e.g. crime rate, income, population density, etc.), more information can be gained related to the impact of attractiveness. If the correlation is strong, those data can be estimated by using attractiveness perception.

2.5 Chapter Summary

The previous sections elaborated previous research related to this thesis. Table 2.1 summarises the criteria and methods used in those works and their relevance to this thesis. The following are more explanations to emphasise the contribution of this research.

In the aspect of Quantifying Urban Attractiveness, some approaches to quantify attractiveness have been studied in past research with various media and parameters. This thesis provides an alternative by using Google Street View data as the media, which is easy to be extracted and covers a lot of places in the world. To get more insights on urban attractiveness, this thesis uses different attributes to be linked to attractiveness, which are familiarity, uniqueness, friendliness, and emotion. In addition, it also uses scene analysis and spatial auto-correlation analysis to obtain some patterns related to the attractiveness of places.

In the aspect of Quantifying Urban Perception by Means of Street-View Images, this thesis uses Street-View data to assess attractiveness, which has not been found in other research. The other novel approach in this thesis is to use 4 Street-View images to represent a single location, which may help to reduce bias. Moreover, the crowd-sourcing task used in this thesis is more straight-forward (i.e. directly rate a location) and requires relatively fewer judgments to assess an image. Lastly, the main crowd-sourcing implemented in this thesis was conducted in a controlled setting which is guided and monitored by the surveyor, so that the collected judgments are more reliable.

In the aspect of Development of Urban Perception Prediction System, this thesis tries to develop a new machine learning-based model to predict attractiveness of a

Table 2.1: Comparison Between Works in This Thesis to Various Past Relevant Research

Criteria	Used in the past relevant works	Used in this thesis
Quantifying Urban Attractiveness		
Assessment media	<ul style="list-style-type: none"> • Places categorised into five main categories (e.g. historical places, housing areas, etc.) • Video of places • Actual places (on-site survey) • Cellular network activity • Shared photos in Flickr 	Street-View data of places from Google Street View
Parameters to describe attractiveness	<ul style="list-style-type: none"> • Perceived Restorativeness Scale (PRS) • 11 aesthetic attributes • 9 attributes in Profile of Mood,States (POMS) • 6 indicators in Landscape,Appreciation Model (GLAM) • Density of digital footprints 	<ul style="list-style-type: none"> • Assessed attributes: familiarity, uniqueness, friendliness, and PAD triplet to represent emotion (pleasure, arousal, dominance) • Scenes and visual patterns • Spatial auto-correlation
Quantifying Urban Perception by Means of Street-View Images		
Assessed attributes	Safety, social class, uniqueness, liveliness, boringness, wealth, depression, beauty	Mainly attractiveness The other attributes are used to get insights on attractiveness: familiarity, uniqueness, friendliness, pleasure, arousal, dominance
Number of Street-View image representing a location	1 Street-View image	4 Street-View images with perpendicular heading
Crowd-sourcing task	Choose one from given two images which is more relevant to the asked attribute (i.e. answering a question "Which place looks more <an adjective attribute>?")	Rate a given image (or images) of location based on a perception
Crowd-sourcing setting	People do the task via a provided website	Internal crowd-sourcing: People are invited to do the task in controlled crowd-sourcing survey sessions Public crowd-sourcing: People do the task via Amazon Mechanical Turk
Development of Urban Perception Prediction System		
Machine learning model	SVM or Convolutional Neural Network (CNN)	Convolutional Neural Network (CNN)
Prediction output	A continuous value (regression)	An ordinal value (classification)
Performance metric	Rank accuracy	Root mean square error (RMSE)

place from Street-View data. The design of this model uses different architecture and output type.

The aspect of Application of Urban Perception Quantification is not compared because these past research are used as references for applications which can be applied in the context of urban attractiveness after this thesis is finished.

Chapter 3

Urban Attractiveness Quantification and Dataset Generation

This chapter shows the method of assessing attractiveness of urban locations that are captured by Google Street View data, which answers RQ1. The assessment is done by means of crowd-sourcing and the output is an urban attractiveness dataset. The dataset is used as the ground-truth to evaluate attractiveness of places in general. The generated dataset is also essential to develop an urban attractiveness prediction system. The dataset generation is divided into three main steps: the data acquisition, the data labelling, and the attractiveness quantification. Besides, the assessment via public crowd-sourcing is also presented.

3.1 Data Acquisition

This research uses Google Street View data to represent each location that will be assessed. Google Street View provides panoramic views from locations along many streets in the world. It enables people to observe a place remotely without physically being on the spot. To observe a location, the Street-View data are crawled by using Google Street View Image API ¹, which can be accessed via URL parameters sent through a standard HTTP request. As a result, a static street-level view image will be returned based on the parameters. If the image is not available, the API returns a default grey image containing text *"Sorry, we have no imagery here"*. The parameters to be inputted for this research are location (latitude and longitude) and heading. Meanwhile, the other parameters are set as default (size = 600x400, pitch=0, and fov=20). An example of a HTTP request for Google Street API and the returned image is shown in Appendix B.

In the existing research of quantifying attribute of location via Street-View data (e.g. Place Pulse [21, 6] and UrbanGems [20]), each location is usually represented by a single Street-View image. However, observing a location from only one heading may causes bias due to its visible view limit. There are areas in the same location which are not visible in the image, e.g. the area behind the observer.

¹<https://developers.google.com/maps/documentation/streetview/intro> (accessed 2016-10-15)

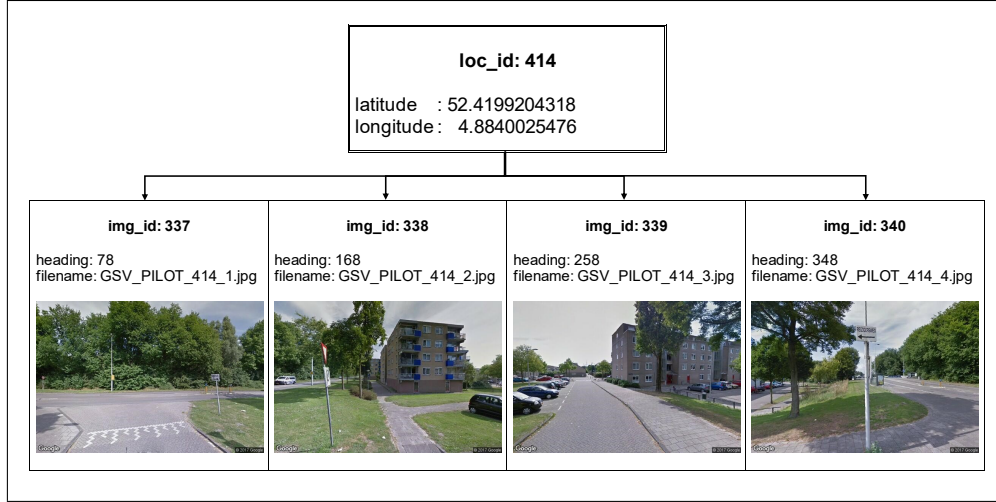


Figure 3.1: Example of Street View data, four images represent a location

To reduce this bias, this research uses four images observed from four perpendicular headings (each direction is separated by 90°) to represent each location. The initial heading is selected randomly. By using this approach, all of surrounding views observed from the target location is covered. The overall attractiveness of a location is assumed to be the mean of the attractiveness values assessed from each of representing images. This assumption will be evaluated in 3.3.3. To get the idea, figure 3.1 shows example of Street-View data extracted from a sample location. It can be seen that each location has attributes of *loc_id* (location id), *latitude*, and *longitude*. Each location is linked to four representing Street-View images. Each image has *img_id* (image id), *heading*, and *filename* attributes.

During the crawling process, the coordinates (*latitude* and *longitude*) are picked randomly inside a defined boundary. For Amsterdam area, the boundary is a rectangular with latitude between 52.29 and 52.42, and longitude between 4.73 and 4.98. After picking a coordinate, an initial heading is also selected randomly as an integer in a range of 0 to 360. The other headings are simply computed as $(h + 90k) \bmod 360$ with h is initial heading and $k = 1, 2, 3$. Then, for each tuple of coordinate and heading, an HTTP request is sent to Google Street View API, and an image is returned. Note that each image has the same size of $600px \times 400px$. If the crawled image is empty (Street-View image in the specified coordinate is unavailable) or shows indoor place, then it is rejected and removed from the dataset.

3.2 Data Labelling

Each target location to be assessed is represented by Street-View images which are crawled in the data acquisition step. Next, the attractiveness assessment can be outsourced to crowd of workers, which the act is usually called as crowd-sourcing. By using crowd-sourcing, the assessment can be done without any help from experts. The assessment value of a target object is decided based on the judgments from multiple people (crowd-sourcing participants). However, attractiveness perception is subjective.

tive, so each participant may give different judgments for the same object and their answers should be analysed and aggregated to get the final attractiveness information of the target location. The result from crowd-sourcing will be used as the ground-truth of attractiveness information of each image and location.

3.2.1 Crowd-sourcing Task

Each crowd-sourcing participant is asked to do the specified task to assess the Street-View data. The task is conducted via a web-based crowd-sourcing tool developed in Ruby on Rails modified from [22] (the details of this tool is explained in Appendix C). To participate in the crowd-sourcing task, a user should go to the website and fill in their identity information (name, email, gender, age, and nationality). Next, the user will be given an instruction on how to do the task, and given one example page to explain how to answer the questions in the task. Next, the crowd-sourcing task is begun.

In one task, a participant is assigned with a set of locations from the dataset. As stated in 3.1, each location in the dataset is represented by four images. Related to this setting, the crowd-sourcing task is divided into two parts.

The first part is the image-level assessment. Individual Street-View images are shown to the participants, one image at a time. For each image, they are asked to answer each given question. For each location, all four representing images are required to be judged in the same task. The order of the images to be shown are shuffled, because if several images from the same location are shown consecutively, the judgments may be influenced by the previous view.

The second part is intended to evaluate the overall attractiveness of the location, which is called location-level assessment. Instead of one image, all four images representing the same location are shown simultaneously at a time. The given questions remain the same. The judgment results from this part is designated to observe how location attractiveness judged from each partial view image contributes to the overall attractiveness of the location.

The crowd-sourcing relies on the answers given by each participant. However, there is no guarantee that all of the participants give proper answers. One way to detect fraud or incompetent participant is by giving golden questions. In the pilot crowd-sourcing, five golden images are given to the users. Four images are from the same location, and the fifth image is from a distinct location. The same golden images are asked to every participants.

When a golden image is shown, an additional question is given to the user, which is object annotation. The user should annotate which objects that appear in the queried image. There are three given options, and the correct answer can be one or more. Figure E.1 in Appendix E shows the golden images and the options in object annotation question. This image annotation is objective, which there exists a correct answer that can be used to verify users' competence in answering the questions.

After a crowd-sourcing task is done, each location gets five new judgments. Four judgments of each representing image in part 1, and one judgment of the overall location in part 2.

3.2.2 Crowd-sourcing Questions

For each image (in task part 1) or location (in task part 2), the participants are requested to answer several questions based on how they perceive the location by looking at the given representing images.

The main label to be evaluated is the attractiveness. Other than attractiveness, other attributes are also asked to be judged, such as familiarity, uniqueness, friendliness, and user's emotion. These attributes are presumed to have relationship with attractiveness, which will be observed after the crowd-sourcing is done. Analysing them can help to understand attractiveness. Intuitively, more unique and friendly place is more attractive, and if a place is already familiar, its attractiveness may decrease. Attractive place also probably gives positive emotion.

To ensure that no question is skipped, the next question will not be shown before the user answer the current question. After the whole questions in one page is answered, the user can submit the answers to move to next page. After it is submitted, the judgment data is stored into the database and the users cannot go back or change their answers.

The details of each question is elaborated as following. Example of the interface containing the questions is shown in Figure 3.2.

1. Attractiveness

The main question in the task is to judge attractiveness of a location by answering the question "Would you like to visit this place?". If a place is more likely to be visited, then it can attract more people and considered as more attractive. The answer of this question is 5-points Likert scale, which is encoded into values 1 to 5.

2. Familiarity

This question is to observe the influence of people's familiarity to a location towards their perception of attractiveness. The answer of this binary (yes=1 or no=0). If the user has seen a place with a view similar to the queried image, then it is considered as familiar. Generally, the answer is more dependent to the user and cannot be generalised over the image/location.

3. Uniqueness

The uniqueness information is asked in the context of its occurrence in Netherlands. This question is to check the correlation between uniqueness and attractiveness, which can help to understand the urban attractiveness. The answer of this question is 5-points Likert scale, which is encoded into values 1 to 5.

4. Friendliness

This question is to check the correlation between friendliness and attractiveness, which also can help to understand the urban attractiveness. The answer of this binary (yes=1 or no=0). The place is considered friendly if the user feels good being in there.

5. Emotion

The affect button [3] is used to provide emotion feedback, which is represented by three affective dimensions: Pleasure, Arousal and Dominance (PAD). Based

Would you **like to visit** this place?

Extremely unlikely Unlikely Neutral Likely Extremely likely

Are you **familiar** with this place? Yes No

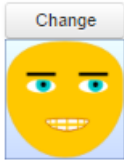
How **unique** is this place?

I frequently see something like this Ordinary Neutral Unique I've never seen anything like this

Do you consider this place **friendly**? Yes No

How does this place make you **feel**?

Change



Which of the following **objects** appear in this image?

Sky
 Blue Car
 Canal/River

Figure 3.2: Questions in The Crowd-sourcing Task

on PAD factor-based theory, every emotion can be mapped into a combination of PAD values [13]. Each dimension has a continuous value in a range between -1 and 1.

6. Object Annotation

This question is special for golden image. The user is asked to annotate which objects that appear in the queried image. There are three given options, and the correct answer can be one or more. This question is asked 5 times in one task and only available in image level judgment (part 1).

3.2.3 Pilot Crowd-sourcing

The quality of the judgments from crowd-source varies. Some people may provide random answers, or do not fully understand the task. To obtain more consistent evaluations, a pilot crowd-sourcing is set in controlled lab sessions, which has been done on 28-30 March 2017 and 2-4 May 2017. Each session is attended by up to four participants, so the surveyor can conveniently guide and monitor each participant during

the process. In this type of crowd-sourcing, the participants' actions are monitored so their competence on the judgments are more trustworthy. Moreover, the participants can ask and interact to the surveyor.

The participants in this pilot survey are 50 students and staffs of TU Delft, which volunteered after the survey information was announced. The participants consists of 86% males and 14% females, with ages in range of 17-35, age average is 25.5, and $\sigma_{age} = 2.4$. Based on nationality, there are British, Chinese, Greek, Indian, Indonesian, and Russian. All of them had been staying in Netherlands for less than 5 years. People who have been living in Netherlands for a long time (e.g. Dutch people) are not included in the survey because they have been exposed with places and scenes in Netherlands (especially Amsterdam) which may interfere with their perception judgments compared to the people who are not too familiar with Netherlands. Each location in the dataset is judged by five distinct participants. The dataset is randomly grouped into 10 task sets with each set consists of 20 locations. Each participant judges one task set, which consists of 20 locations and 80 Street-View images to be judged.

Afterwards, 200 locations are assessed in this pilot with 800×5 judgments for image-level assessment and 200×5 judgments for location-level assessment. There are seven attributes that are extracted from the judgments based on 3.2.2 : attractiveness, familiarity, uniqueness, friendliness, pleasure, arousal, and dominance.

In addition, five Street-View images and one location in golden questions are also assessed. These data can support some analyses with higher reliability because each object is judged by all of the participants.

3.2.4 Data Label Aggregation

To determine the ground-truth label of each image and location, the votes from the participants are aggregated. For Likert-scale type question (i.e. attractiveness and uniqueness), the values are encoded into ordinal values of 1 to 5. It is assumed that each participant has equivalent expertise. So, the overall label can be presented as the median of the votes which will still yield ordinal labels. For binary type question (no=0, yes=1), the majority of the votes is used. Values from affect button (pleasure, arousal, and dominance) have continuous type and be rounded to 3 decimal values. The labels for continuous values are computed based on the mean.

Focusing on the attractiveness (as the main attribute), the distribution of the aggregated labels is shown in Figure 3.3 for both image level judgments (crowd-sourcing part 1) and location level judgments (crowd-sourcing part 2). The expected distribution is normal distribution. Both graphs show small amount of locations with extreme value of 1 and 5, and interestingly displays that the number of unattractive locations are more than attractive ones. Label 3 has the highest frequency, which is reasonable because 3 is the middle value.

3.3 Attractiveness Quantification

After the crowd-sourcing is done, the judgments data from the participants are analysed. The initial analysis is to validate the reliability of the judgments, and then some statistical analyses are applied to the data to get insights on attractiveness.

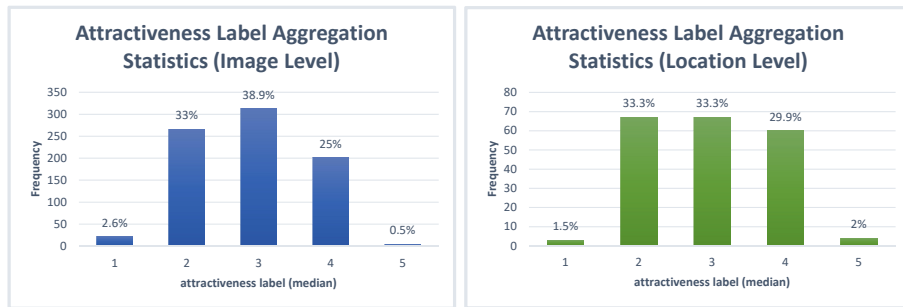


Figure 3.3: Aggregated Attractiveness Label Distribution From Pilot Crowd-sourcing in Image Level (Left) and Location Level (Right)

3.3.1 Answers of Golden Questions

For each golden image, there is a question to annotate objects that appear in the image. There are two perspectives to verify the correctness of the answer.

1. The first one is the exact match criteria. The user's answer is considered as correct if all of the objects that appear in the image are selected, and the objects that are not in the image are not selected. Users with high accuracy by using this criteria is considered competent. If not, their answers should be audited.

By using this criteria, the accuracy distribution is shown in Figure 3.4. The orange bars reveal that there are 6 participants who have low accuracy of 40%. From an investigation, it is found that for each object annotation question, they only selected exactly one answer. Based on their testimonial, they admitted that they assumed only one option can be selected.

The annotation correctness rate of each image is shown in Figure 3.5. It seems that most of participants failed to annotate objects in image with `img_id 10003`. There is a tricky option in here. By looking at the image, actually there is a canal in the image, but most of them did not realise that. The blue bars in 3.4 display the accuracy if "canal" option in image with `img_id 10003` is omitted. Now, all of the participants have accuracy above 50%.

2. Another less strict criteria is that only one correct selected option is enough to validate that the object annotation is correct. However, it is still considered as wrong if they annotate an object which does not appear in the image. By using this criteria, all of the participants has 100% accuracy.

Based on this observation, all of the judgments are accepted and can be used for further analysis.

3.3.2 Judgments Variance

The other analysis to check the reliability of the crowd-sourcing result is by observing the judgments variance. Since the task itself is subjective, it cannot be expected that all of the users will have mutual votes on the same asked object. However, by observing

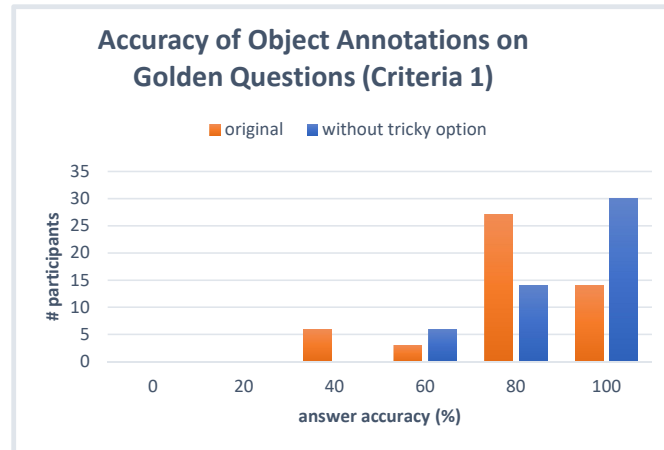


Figure 3.4: Accuracy Distribution of Object Annotations on Golden Questions (by Using Exact Match Criteria)

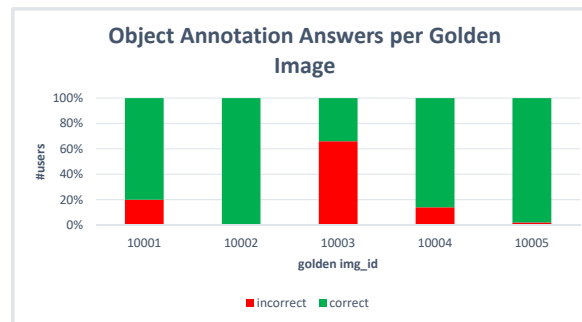


Figure 3.5: Distributions of Answers Correctness of Object Annotation Tasks for each Golden Image

the votes distribution for each object (image/location), the confidence level of each labels can be analysed. For each object, the votes variance of attractiveness label is computed, and then ordered among all objects from low to high. The graph is shown in Figure 3.6 for both image and location judgments. Higher variance means that the votes are more spread from the average value.

The votes are considered as good (having high confidence) if the variance is lower than a threshold. It means that the users' votes are converged into a value. This research uses variance of 1 as the threshold, which statistically gives a tolerance of 1 value deviation to the given label (e.g. when the label is 4, then the actual attractiveness may be 3, 4, or 5, but is less likely to be 2). By taking 1 as the variance threshold, it is found that 14.4% of the image judgments and 14.4% of the location judgments have variance exceed the threshold. Which means the crowd-sourcing produces around 85% judgments with high confidence.



Figure 3.6: Variances of Attractiveness Label Judgments per Image (Top) and per Location (Bottom), Ordered from Low to High

3.3.3 Estimating Overall Attractiveness of a Location Based on Representing Street-View Images

As stated in 3.1, one location is represented by 4 Street-View images. The overall attractiveness of the location can be computed based on the attractiveness values of those 4 images, in this case simple mean function is used. By comparing the predicted location attractiveness (the mean of 4 attractiveness values from each images is calculated, then rounded) and the actual attractiveness (based on location-level assessment in crowd-sourcing), it gives an accuracy of 68.2% and RMSE of 0.54.

To get more detail observation, Table 3.1 shows the confusion matrix. The combination of 4 attractiveness labels from images is mapped into a labels set. For example, if the attractiveness values from representing images of a location are 3,4,3,2 (or any of its permutation), then it belongs to labels set $\{2,3,3,4\}$. It can be seen that the predicted value mostly gives the correct value, or deviated by 1 rank. Except for one case with attractiveness values $\{4,4,4,4\}$, which the mean is 4, but the actual attractiveness is only 2. It is considered as an anomaly because most of the other locations with the same labels set give the correct actual attractiveness of 4. An investigation revealed that this location label has a high variance (i.e. 2.16), so it may be the case that the assessed label itself is less reliable.

The Street-view images of the location with this anomalous case is shown in Figure 3.7. When only viewed from one side, it seems that it shows a scene with greenery, road, or field, with beautiful sky. However, when all of the images are combined, it seems the scenery of empty roads and field make it less attractive for some participants.

Based on the evaluation, the accuracy and RMSE are considered as acceptable because locations with the same labels sets may have different overall attractiveness, and the prediction by using mean value gives the best estimation almost all the time (e.g. For labels set 2,3,3,4, estimating location attractiveness as 3 is the best choice. If

Table 3.1: Confusion Matrix of Location Attractiveness Prediction Based on The Mean of Attractiveness from 4 Representing Street-View Images.

labels set	labels mean	prediction	location attractiveness					
			1	2	3	4	5	
{1,1,2,2}	1.5	2		2				
{1,1,2,3}	1.75	2		1				
{1,2,2,2}	1.75	2	2	3				
{1,2,2,3}	2	2		5				
{2,2,2,2}	2	2	1	14				
{1,2,3,3}	2.25	2		1				
{2,2,2,3}	2.25	2		12	2			
{1,2,3,4}	2.5	3			1	1		
{2,2,2,4}	2.5	3		4	1			
{2,2,3,3}	2.5	3		8	6			
{1,3,3,4}	2.75	3					2	
{2,2,3,4}	2.75	3		4	10			
{2,3,3,3}	2.75	3		4	12			
{2,2,4,4}	3	3					3	
{2,3,3,4}	3	3		5	8	2		
{3,3,3,3}	3	3		2	10	6		
{2,3,4,4}	3.25	3		1	2	5		
{3,3,3,4}	3.25	3			7	5		
{1,4,4,5}	3.5	4					1	
{2,4,4,4}	3.5	4			1	1		
{3,3,4,4}	3.5	4			6	9		
{2,4,4,5}	3.75	4				2	1	
{3,4,4,4}	3.75	4			1	11	1	
{4,4,4,4}	4	4		1		11	1	
{4,4,4,5}	4.25	4				1	1	



Figure 3.7: Example of anomalous images. In the crowd-sourcing, each image has attractiveness value of 4. When all of them is shown together, the overall attractiveness is only 2.

it is predicted as 2 or 4, then it will give higher error).

3.3.4 Correlations Among Assessed Attributes

Knowing the correlations between each judged attributes can provide more insight about the attractiveness perception. Due to the difference of data type and range of

Table 3.2: Spearman's Correlation Matrix of The Assessed Attributes

	attr	fami	uniq	frie	plea	arou	domi
attr	1	0.115	0.579	0.549	0.776	0.460	0.502
fami	0.115	1	-0.127	0.076	0.122	-0.022	0.126
uniq	0.579	-0.127	1	0.249	0.514	0.445	0.297
frie	0.549	0.076	0.249	1	0.597	0.197	0.374
plea	0.776	0.122	0.514	0.597	1	0.484	0.481
arou	0.460	-0.022	0.445	0.197	0.484	1	0.211
domi	0.502	0.126	0.297	0.374	0.481	0.211	1

each attribute, Spearman's rank correlation coefficient is used to observe the correlation (i.e. instead of Pearson correlation, although both of them actually show similar result). Table 3.2 shows the correlation matrix of merged data from image level and location level aggregated labels. It can be observed that several attributes are correlated (having correlation > 0.5). Attractiveness as the main label is correlated to uniqueness, friendliness, pleasure, and dominance. The other correlated attributes is friendliness-pleasure and uniqueness-pleasure. Meanwhile, familiarity shows relatively small correlation to all of other attributes. This matrix also reveals that there is a possibility of multi-collinearity between attractiveness and pleasure.

3.3.5 Exploratory Factor Analysis

Several variables are assessed in the data to help explaining attractiveness. There is a possibility that there are actually several types of attractiveness (i.e. factors that can explain attractiveness) which are not observable in the data. To find these hidden factors, the exploratory factor analysis [7] is applied. This analysis assumes that there are m underlying factors (F_1, F_2, \dots, F_m) whereby each observed variables (X_1, X_2, \dots, X_p) is a linear function of these factors together plus a residual variate and reproduce the maximum correlation as formulated in 3.1.

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \dots + a_{jm}F_m + e_j \quad (3.1)$$

where $j = 1, 2, \dots, p$.

The procedure is done based on a guide from [29]. Based on the guide, several statistical analyses have to be done to check that explanatory factor analysis can be applied to the generated dataset. Appendix D shows the results (by using SPSS) and it confirms that the analysis can be applied and familiarity variable should be omitted. Data from Extraction Sums of Squared Loadings and scree plot suggest that there is only 1 factor. The generated factor matrix is shown in Table 3.3.

The generated factor has very high correlation to pleasure, and high correlation in range 0.5 – 0.6 to other variables. If this factor is called "attractiveness", then it confirms the observation in correlation matrix that indicated that attractiveness has multi-collinearity with pleasure.

Table 3.3: Factor Matrix Extracted from The Dataset (There is Only 1 Factor)

variable	Factor 1
pleasure	0.931
uniqueness	0.591
friendliness	0.575
arousal	0.544
dominance	0.508

3.3.6 Spatial Analysis

The generated dataset has spatial dimension, which are latitude and longitude. Based on Tobler's first law of geography [28], nearby things are more likely to be correlated. So, there is a possibility that nearby locations will have similar value of attractiveness (or other attributes as well). If this similarity behaviour is confirmed, then it can influence the design of attractiveness prediction model. For example, attractiveness of a new location can be estimated based on the attractiveness of neighbouring locations. To verify this hypothesis, a spatial auto-correlation is performed.

The spatial auto-correlation can be analysed from two perspectives. Global auto-correlation can detect the existence of patterns which display spatial clustering. Meanwhile, local auto-correlation is usually used to identify the clusters or hot spots which reflect the global pattern. This research only focuses on the global auto-correlation because the interest is to find out if there is any pattern in the spatial distribution.

One of the metrics commonly used for auto-correlation is Moran's I coefficient [4], which was first proposed by P.A.P. Moran [14]. The Global Moran's I, which measures the global auto-correlation, is formulated as Equation 3.2. Supposed that there are N observed locations from 1 to N , then x_i represent the observed value in i 'th location, and \bar{x} is the mean. The other required variable is $w_{i,j}$, which is spatial weight between i 'th and j 'th location. Generally, $w_{i,j}$ is defined as 1 if the two locations are nearby (i.e. neighbours, which mean that they have potential interaction), and 0 otherwise (note that $w_{i,i}$ is also defined as 0). W is the sum of all weights ($W = \sum_i \sum_{j \neq i} w_{i,j}$).

$$I = \frac{N \sum_i \sum_j w_{i,j} (x_i - \bar{x})(x_j - \bar{x})}{W \sum_i (x_i - \bar{x})^2} \quad (3.2)$$

Based on the formula, the Global Moran's I value will range from +1 which indicate that the observed values are spatially more clustered, to -1 which means the values has high heterogeneity. A value of 0 shows random pattern. The expected value of Moran's I under no spatial auto-correlation is $E(I) = -1/(N-1)$, which in the case of 201 locations, it is -0.005 .

To obtain the values of spatial weights, the neighbours of each location should be defined. Two of the commonly used criteria are explained as following.

1. Distance threshold

The neighbours are defined as locations within a distance threshold. To compute the estimated distance in metric units, the coordinates in the dataset are mapped

into Cartesian coordinates. Based on latitude/longitude distance calculator ², in Amsterdam distance between the minimum and maximum latitude boundary is around 14 km and its of longitude boundary is around 17 km (the boundaries are mentioned in 3.1). Thus, the mapped coordinates have x -axis value between 0 and 14, and y -axis value between 0 and 17.

Table 3.4 shows the Moran's I calculation for various distance thresholds. p -rand is the p -value under randomisation assumption and p -norm is the p -value under normality assumption (both of them are checked because there is no confirmed distribution, the higher value will be used as the measure). A result is statistically significant when the p -Value is below 0.01, which in this case is achieved for a threshold of 3 to 4 km with Moran's I around 0.05. For threshold below 2 km, there are locations without any neighbour due to the absence of any other location within the threshold boundary. In this case, this approach is not suitable to be used.

2. k-Nearest Neighbour (k-NN)

An alternative to define the neighbours of a location is based on k nearest observed locations. Unlike distance threshold based, this approach ensures that all of the location has neighbours. However, if the k is set high, then some locations may have neighbours with far distance. Table 3.5 shows the Moran's I calculation for $k = 2$ to 10. The significant result is achieved with $k \geq 5$ with Moran's I values are around 0.1.

These results show positive values of Global Moran's I, which means that there is possibility of clustered patterns in the data over the map. However, the value is small which may be caused by the random selection of locations during the data acquisition. But, most probably it is because the observed locations are scattered with relatively big distance among each others. The p -value shows that significant result is achieved with distance threshold above 3 km, or above 5 nearest neighbours, which are a large area. It is reasonable that places separated more than 3 km are having weak attractiveness correlation. However, there is limited information when the locations have smaller distance. Hence, 4.2.2 attempts an alternative approach to estimate attractiveness from very near location. Moreover, the auto-correlation analysis will be updated with more data based on the prediction system in 5.2.3.

3.4 Public Crowd-sourcing

Performing crowd-sourcing in controlled lab sessions is generally more reliable. However, the volunteered participants are usually limited and each session is ideally only attended by small number of participants, so it is not really applicable to assess a lot of locations. Other way to do the crowd-sourcing is by using public platforms such as Amazon Mechanical Turk (AMT) ³, which enable people from around the world to assess the locations via online, with small payments.

²URL: <http://www.nhc.noaa.gov/gccalc.shtml> (accessed on 2017-08-31)

³<https://www.mturk.com> (accessed 2017-08-15)

Table 3.4: Moran’s I and p-Values (Under Random and Normality Assumption) with Weights Based on Distance Threshold for Various Thresholds

distance threshold (km)	Moran’s I	p-rand	p-norm	# locs with no neighbour
5	0.014	0.0555	0.1148	0
4	0.038	0.0003	0.0046	0
3	0.092	0.0000	0.0000	0
2	0.057	0.0108	0.0660	1
1	0.085	0.0817	0.2167	25
0.9	0.094	0.0936	0.2344	40
0.8	0.038	0.5137	0.6434	55
0.7	0.053	0.4511	0.5933	76
0.6	0.191	0.0381	0.1418	110
0.5	0.114	0.3194	0.4806	138
0.4	0.109	0.4965	0.6299	168
0.3	-0.439	0.1726	0.3336	192

Table 3.5: Moran’s I and p-Values (Under Random and Normality Assumption) with Weights Based on k-NN for Various k

k	Moran’s I	p-rand	p-norm
2	0.090	0.0362	0.1370
3	0.098	0.0064	0.0524
4	0.101	0.0013	0.0221
5	0.116	0.0000	0.0036
6	0.131	0.0000	0.0003
7	0.108	0.0000	0.0013
8	0.117	0.0000	0.0002
9	0.118	0.0000	0.0001
10	0.105	0.0000	0.0002

3.4.1 Implementation of Public Crowd-sourcing

A crowd-sourcing task has been released in AMT on 1-5 September 2017. The task is only for location-level assessment (each location is directly represented by 4 Street-View images). The same crowd-sourcing interface in pilot was used, so that the workers have the same type of questions and user experience as the participants in pilot crowd-sourcing. Via AMT, crowd-sourcing workers will get a link to the crowd-sourcing interface website and do the task. After it is finished, they will receive a voucher code to be inputted back to the AMT to receive the reward.

The workers can be anyone outside Netherlands, with the same reason as the internal crowd-sourcing. In this task, each worker will judge a task set containing 10 locations plus 1 golden location. Each task set is set to be judged by 13 workers.

For the first trial, 2 task sets are assessed (i.e., task set 11 and 12). Each task set

contains 5 locations from the generated dataset with each attractiveness value of 1 to 5. The other 5 locations are new randomly picked locations.

3.4.2 Result of Public Crowd-sourcing

The collected data show that there are 4 workers who failed to answer the golden question correctly (based on 2nd criteria in 3.3.1). The judgments from these workers are rejected. The other rejected judgments are from workers who always judge the attractiveness and uniqueness with value 4 or 5 (coincidentally, all of them are from India). This kind of judgments may be because those workers do not have clear experience about European places and consider every places is attractive and unique, or the worse case is because they were not serious during the assessment. There is no other case which the worker always gives the same values for all images. The other validation is based on the assessment time. All of the remaining users completed the task with in more than 4 minutes which is normal.

After the filtering, finally there are 7 valid workers for each task set. They consists of 7 males and 7 females, with an average age of 34 and $\sigma_{age} = 10.2$. Their nationalities are Indian, American, Filipino, and Irish. From these accepted judgments, the same analysis as the pilot crowd-sourcing can be performed.

The judgments of each location are aggregated. The aggregated label for attractiveness is shown in Figure 3.9. It shows imbalance labels which is dominated with unattractive locations. There is no location labelled with class 1 or 5. The ordered variance per locations is shown in Figure 3.10. Generally, they have high variance with 47.6% of them are above the threshold of 1. Thus, the judgments from public crowd-sourcing are rarely well-converged and having high uncertainty.

To investigate the cause of high variances, the judgments are separated based on the worker's nationality. American and Irish workers are grouped as "American-European" (there are 8 workers, with 4 per task set), and Indian and Filipino workers belong to "Asian" group (there are 6 workers, with 3 per task set). There is interesting distribution difference between them, which can be seen in Figure 3.11. For American-European, most of the locations are perceived as unattractive. Even, there is a location with an attractiveness of 1. Note that some of the locations have median of 1.5, 2.5, or 3.5, but all of them are rounded up. On the other hand, the judgments aggregation from Asian workers dominate the attractive locations. As a result, some places are perceived as unattractive for American/Europeans, but attractive for Asians, and the aggregated label has a high variance. Figure 3.8 shows two location examples which labelled with 1 or 2 by the American/European workers, but labelled 4 or 5 by the Asian workers.

As an important note, this finding may only valid in these limited samples and cannot be generalised for all of the workers. However, this analysis result suggests that public crowd-sourcing may not be reliable to assess attractiveness from Street-View data. The difference of living environment may influence the standard of people's perception of attractiveness.

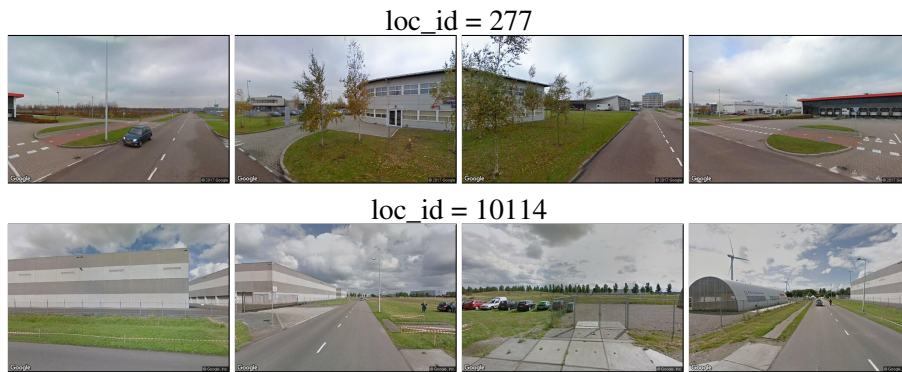


Figure 3.8: Examples of location which the attractiveness is labelled 1 or 2 by American/European Workers and labelled 4 or 5 by Asian Workers in AMT

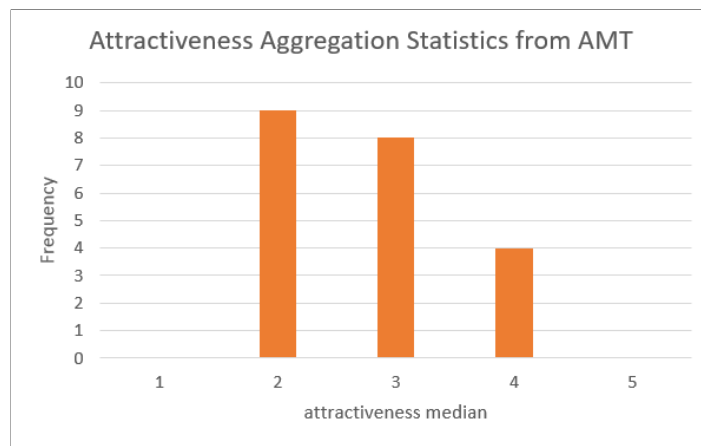


Figure 3.9: Aggregated Attractiveness Label Distribution from Public Crowd-sourcing in AMT

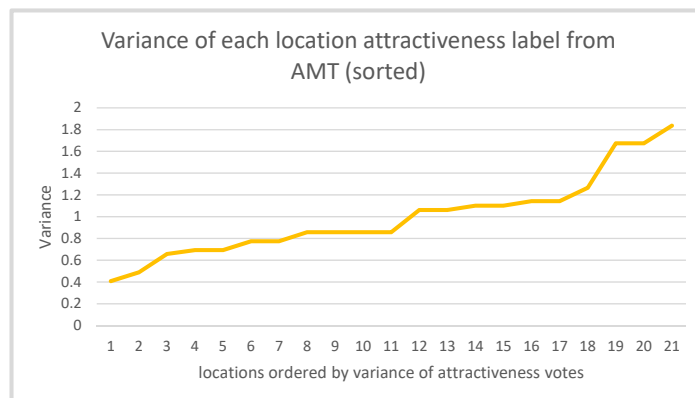


Figure 3.10: Variances of Attractiveness Label Judgments per Location from Public Crowd-sourcing, Ordered from Low to High

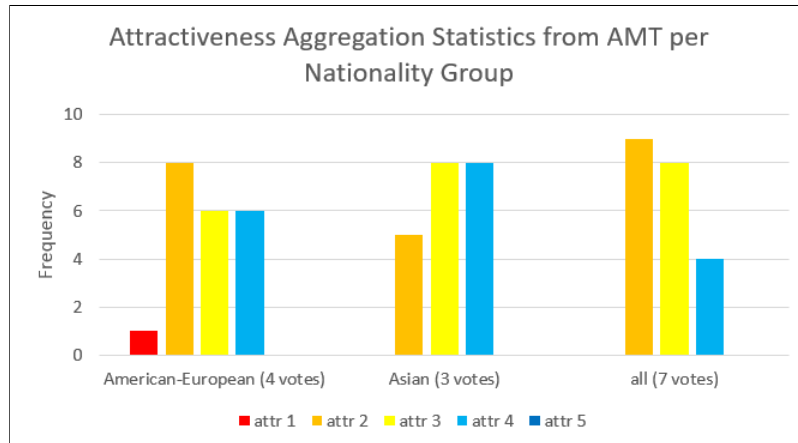


Figure 3.11: Aggregated Attractiveness Label Distribution from Public Crowd-sourcing in AMT Grouped Based on Nationality

Table 3.6: Confusion Matrix of Locations Attractiveness Label Between Judgments via Internal (Pilot) and Public (AMT) Crowd-sourcing

		label from pilot				
		1	2	3	4	5
label from AMT	2	2	1	1		
	3		1	1	1	1
	4				2	1

3.4.3 Comparison to Pilot Crowd-sourcing

For the next analysis, the aggregated attractiveness labels from AMT are compared to the result from the pilot crowd-sourcing. There are 11 locations from pilot which are also assessed in AMT. Table 3.6 shows the confusion matrix of attractiveness labels between judgments from internal (pilot) and public (AMT) crowd-sourcing. It indicates a positive correlation, even though there are still many discrepancies (with accuracy = 63% and RMSE = 0.95). There is a pattern that when location a is labelled with 2 in public crowd-sourcing, then its label in pilot is ≤ 3 , and locations with label 4 in public crowd-sourcing has a label > 3 in public crowd-sourcing.

Table 3.7 shows the detail which compares the label (judgments median) and variance for each location. It seems that small variance in pilot crowd-sourcing does not assure small variance in public crowd-sourcing, and vice versa. Even, small variances in both crowd-sourcing also do not always lead to the same aggregated label. Location with id 26 is the only one with a big error (5 vs 2), even though both of them have variance < 1 .

Based on these analyses, it is not recommended to use public crowd-sourcing to assess attractiveness based on Street-View data, except with a lot of judgments per object, which should be studied with more supporting data. Moreover, in public crowd-sourcing, the workers cannot be monitored and cannot ask during the task execution. So, for the next analyses and machine learning model development, the results from public crowd-sourcing will not be considered.

Table 3.7: Attractiveness Label and Variance Comparison Between Location Judgments from Internal (Pilot) and Public (AMT) Crowd-sourcing

loc_id	#votes	labels from pilot		labels from AMT		error
		median	variance	median	variance	
564	7	1	1.36	2	1.14	1
640	7	1	0.24	2	1.14	1
602	7	2	1.04	2	1.67	0
448	7	2	1.36	3	0.86	1
277	7	3	1.36	2	1.67	1
224	7	3	0.56	3	1.10	0
315	7	4	1.36	4	0.41	0
514	7	4	0.56	3	0.78	1
26	7	5	0.24	3	0.86	2
666	7	5	0.24	4	0.49	1
9999 (golden)	14	4	0.59	4	0.66	0

Chapter 4

Urban Attractiveness Prediction System

Manual assessment, either via on-site survey or controlled crowd-sourcing does not scale. For example, to assess attractiveness of places in Netherlands with each adjacent places have an interval of 1 km, then more than 40,000 assessments have to be done. If the interval is less than 1 km, then even more assessments are required. On the other hand, using public crowd-sourcing to assess attractiveness is less reliable based on the previous analysis. Thus, it is essential to develop a system that can automatically predict the attractiveness of a location from Street-View data. The system can be developed by using machine learning, such that the model can learn the attractiveness from an assessed dataset. Some experiments have to be conducted to achieve a model with high performance. This machine learning approach answers the RQ2 of this thesis. Part 4.2 shows a semi-supervised learning approach which utilise spatial information to expand the dataset which can answer RQ3.

4.1 Convolutional Neural Network (CNN)

Image classification is the task of taking an image as the input, and classifying it into a class (or providing probability of classes) which suitably describes the image. This research tries to implement an image classification task, which is classifying a Street-View image into an attractiveness class (from class 1 to class 5). A Street-View image in this research is read by a computer as a matrix of pixel values with size of $3 \times 400 \times 600$. In machine learning, these data typically are converted into some features (which is called feature extraction), and then a machine learning algorithm can generate a model which fit those features into the expected class. The challenge is determining which features to be extracted, and which algorithm to be used so that the model can differentiate each class based on the pattern in these features.

Porzi,et al [19] showed that Convolutional Neural Network (CNN) has the best performance in the task of predicting human perception from image. Thus, this research will focus on CNN, including the network architecture and the hyper-parameters. Thoma [26] has summarised the concept of CNN in his thesis.

4.1.1 VGG-PLACES

Training the whole network from a random weight initialisation can take a lot of iterations and require a long time to learn and adapt the training dataset. Transfer learning is usually used to transfer knowledge from a source domain to a target domain [17]. Thus, a CNN for attractiveness prediction can be developed from a pre-trained CNN for other purposes, such as object recognition or scene classification. The pre-trained model will act as a feature extractor, which has already learned low level features, such as edges, curves, shapes, and colours.

By using this approach, the training phase can be focused for the more important layers, which are some of the last layers in the network. They should be replaced with some new layers designed for the required task.

The next challenge is to choose a suitable pre-trained network for the task. It has been found that AlexNet-PLACES is one of the best pre-trained model for safety perception prediction task[19]. The original AlexNet-PLACES is a CNN model used for scene classification, which uses the AlexNet [11] architecture and trained with PLACES205 dataset [30]. It can be used to extract scenes from an input image (there are 205 possible scenes). For the task of classifying safety perception, transfer learning from AlexNet-PLACES outperformed the model developed from pre-trained AlexNet-ImageNet (AlexNet trained with ImageNet for object recognition task) as well as AlexNet trained from scratch.

It is understandable that a CNN trained with PLACES dataset is more adaptable for human perception task, because scene classification has deeper representation than object recognition. In scene classification task, PLACES205 dataset has been used to train several convolutional neural networks, such as AlexNet, GoogLeNet, and VGG. Based on the experiment [30], VGG architecture has the best performance for scene classification trained with PLACES dataset. Thus, it is possible that VGG-PLACES will be a better pre-trained model for attractiveness prediction task. Thus, the CNN architecture and feature extractor to be used in this research is based on VGG-PLACES.

VGG-PLACES consists of 5 convolutional blocks, followed by 2 fully connected layers (fc6 and fc7), and the last is output layer with 205 output nodes represent 205 scenes. Totally, there are 16 main layers with the details can be seen in Table 4.1. Each convolutional layer uses kernel size of 3×3 and ReLU activation function. Meanwhile, each max-pooling layer uses pool size of 2×2 and strides of 2×2 .

4.1.2 Development by using Keras Framework

The implementation of the CNN is done by using Keras framework ¹. It provides various types of neural network layers, and various settings. It is also simpler to design and configure the network by using this framework. Moreover, it supports Tensorflow or Theano backend, which in this research the former is used.

A CNN model requires the layers architecture, and the nodes in each layer have weight values as the parameters. PLACES205-VGG architecture and weight values (stored in HDF5 format) are retrieved from the official project ², which is implemented in Caffe framework. Due to structure difference between HDF5 formatted

¹<https://keras.io/> (accessed 2017-01-15)

²<http://places.csail.mit.edu/downloadCNN.html>(accessed 2017-01-20)

Table 4.1: Layers Architecture in VGG

Block	Layer#	Layer type	Name in Caffe HDF5	Layer name in Keras HDF5	Size
input image					(3,224,224)
Block 1	1	Conv2D	/data/conv1_1/0 /data/conv1_1/1	/block1_conv1/block1_conv1_W_1:0 /block1_conv1/block1_conv1_b_1:0	(3, 3, 3, 64) (64,)
	2	Conv2D	/data/conv1_2/0 /data/conv1_2/1	/block1_conv2/block1_conv2_W_1:0 /block1_conv2/block1_conv2_b_1:0	(3, 3, 64, 64) (64,)
MaxPooling2D					
Block 2	3	Conv2D	/data/conv2_1/0 /data/conv2_1/1	/block2_conv1/block2_conv1_W_1:0 /block2_conv1/block2_conv1_b_1:0	(3, 3, 64, 128) (128,)
	4	Conv2D	/data/conv2_2/0 /data/conv2_2/1	/block2_conv2/block2_conv2_W_1:0 /block2_conv2/block2_conv2_b_1:0	(3, 3, 128, 128) (128,)
MaxPooling2D					
Block 3	5	Conv2D	/data/conv3_1/0 /data/conv3_1/1	/block3_conv1/block3_conv1_W_1:0 /block3_conv1/block3_conv1_b_1:0	(3, 3, 128, 256) (256,)
	6	Conv2D	/data/conv3_2/0 /data/conv3_2/1	/block3_conv2/block3_conv2_W_1:0 /block3_conv2/block3_conv2_b_1:0	(3, 3, 256, 256) (256,)
	7	Conv2D	/data/conv3_3/0 /data/conv3_3/1	/block3_conv3/block3_conv3_W_1:0 /block3_conv3/block3_conv3_b_1:0	(3, 3, 256, 256) (256,)
MaxPooling2D					
Block 4	8	Conv2D	/data/conv4_1/0 /data/conv4_1/1	/block4_conv1/block4_conv1_W_1:0 /block4_conv1/block4_conv1_b_1:0	(3, 3, 256, 512) (512,)
	9	Conv2D	/data/conv4_2/0 /data/conv4_2/1	/block4_conv2/block4_conv2_W_1:0 /block4_conv2/block4_conv2_b_1:0	(3, 3, 512, 512) (512,)
	10	Conv2D	/data/conv4_3/0 /data/conv4_3/1	/block4_conv3/block4_conv3_W_1:0 /block4_conv3/block4_conv3_b_1:0	(3, 3, 512, 512) (512,)
MaxPooling2D					
Block 5	11	Conv2D	/data/conv5_1/0 /data/conv5_1/1	/block5_conv1/block5_conv1_W_1:0 /block5_conv1/block5_conv1_b_1:0	(3, 3, 512, 512) (512,)
	12	Conv2D	/data/conv5_2/0 /data/conv5_2/1	/block5_conv2/block5_conv2_W_1:0 /block5_conv2/block5_conv2_b_1:0	(3, 3, 512, 512) (512,)
	13	Conv2D	/data/conv5_3/0 /data/conv5_3/1	/block5_conv3/block5_conv3_W_1:0 /block5_conv3/block5_conv3_b_1:0	(3, 3, 512, 512) (512,)
MaxPooling2D					
Flatten					
Fully Connected Layers	14	Dense	/data/fc6/0 /data/fc6/1	/fc6/fc6_W_1:0 /fc6/fc6_b_1:0	(25088, 4096) (4096,)
	15	Dense	/data/fc7/0 /data/fc7/1	/fc7/fc7_W_1:0 /fc7/fc7_b_1:0	(4096, 4096) (4096,)
Output Layer	16	Dense	/data/fc8/0 /data/fc8/1	/predictions/predictions_W_1:0 /predictions/predictions_b_1:0	(4096, 205) (205,)

weights data in Caffe and Keras, it should be converted. Each convolutional layer and fully connected layer generally consists of neuron weights and activation constants parameters. The complete layers name and shape conversions are also shown in Table 4.1.

4.1.3 Network Modification

After the pre-trained network is tested and works well to classify scene from any image, some layers have to be modified to be able to learn the attractiveness of a location from training images.

The most important layer to be modified is the output layer. It should be replaced to adjust the number of output nodes. This research uses 5-scale Likert scale as the attractiveness label, which is decoded into five ordinal classes (1 to 5). The issue with this class type is that the values are discrete and has an order, however the distance between each class is not necessarily equal and clearly determined. To handle it, the developed network uses 4 output binary nodes with the following rule.

- [0, 0, 0, 0] => class 1
- [1, 0, 0, 0] => class 2
- [1, 1, 0, 0] => class 3
- [1, 1, 1, 0] => class 4
- [1, 1, 1, 1] => class 5

By using this approach, each output node can search the appropriate boundary of each adjacent class (e.g. output 1 tries to separate class 1 from class 2 and above, output 2 tries to separate class 1,2 and class 3,4,5, etc). The output layer uses sigmoid activation (Equation 4.1), which makes the output nodes are independent and have a range from 0 to 1. For each output node, the output value should be rounded to get a binary value.

$$S(x) = \frac{1}{1 + e^{-x}} \quad (4.1)$$

The convolutional layers in block 1 - 5 and the first fully connected layer (fc6) in VGG-PLACES are used as deep feature extractors that can map an image into 4096 scalar features. Without fc6, the extracted features will be 25,088 which is too much to be trained by a small dataset. The weight values in feature extractor are frozen, so they will not be updated during training.

In this research, the feature extractor is followed with two fully connected layers (FC1 and FC2) with 4096 nodes each. All of the nodes in a fully connected layer are connected to nodes in previous layer. The weights in these layers are randomly initialised and will be tuned during training.

With a small size of training dataset, the trained network may encounter overfitting quickly, which is highly fitted to the learnt data, but lose the generality to predict new data. To avoid that, two dropout layers are added, each between FC1 and FC2, as well as between FC2 and the output layer. A dropout layer will randomly drop some nodes from previous layer (set their values to 0). The fraction amount of the dropped nodes are based on a dropout rate parameter. The dropout rate for each dropout layer will be determined based on experiments.

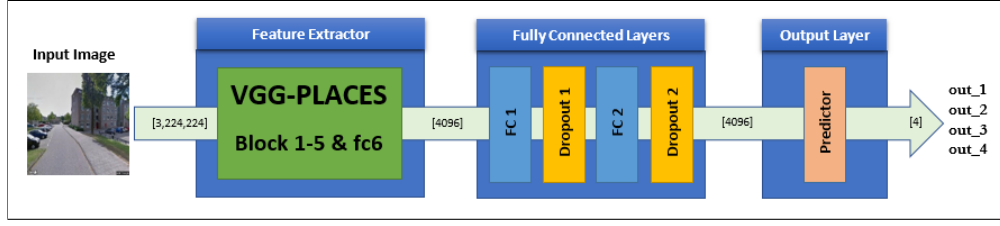


Figure 4.1: Architecture of CNN for The Attractiveness Prediction System

After the modification, the developed CNN has an architecture as seen in Figure 4.1.

4.1.4 Network Training Configuration

Training a neural network involves feed-forward and back-propagation stages. Feed-forward is a process which the input data is processed by the network, from input layer, intermediate layers, up to output layer and return output values. Back-propagation is done after feed-forward, which the output loss is computed, and the weights of previous layers are refined to minimise the loss. These processes are applied several times with various input data to adjust the weights until the loss converges to a desired threshold.

One of the standard formula to update the weights is Stochastic Gradient Descent (SGD) [2], as shown in Equation 4.2 which the weight is updated after each batch of N inputs. w_t is current weight of a node. $J(x_i, y_i; w_t)$ is the loss function and its gradient indicates that when the value is positive, then the weight is too big and needs to be reduced, and vice versa. α is a learning rate parameter which can determine how big the weight value will be updated.

$$w_{t+1} = w_t - \alpha \cdot \frac{1}{N} \cdot \sum_{i=1}^N \nabla_w J(x_i, y_i; w_t) \quad (4.2)$$

To evaluate the performance of each output node, the loss function used is binary_crossentropy (so that each output node tries to fit its binary value prediction).

In training a neural network, several hyper-parameters have to be configured. Following [19], the optimizer to be used to update the weights is standard Stochastic Gradient Descent (SGD) with learning rate (α) of 0.01, and Nesterov momentum of 0.9. The other hyper-parameters are set through experiment in 5.1.2. The parameters and their selected value options are as following.

1. Batch size

Batch size (N) determines the number of samples that is going to be propagated through the network. The weights in the network are updated after each batch. Small batch will update the weights more often but noisy data may cause high fluctuation on the weights update. Conversely, big batch size may reduce noise potential (because the gradient values are accumulated before updating the weights), but the weights are updated more rarely. The tested batch size values are 1, 5, 10, and 20.

2. Learning rate decay value

After several iterations, the weights in the network are updated and are expected to converge to the optimum values. Thus, the learning rate should be reduced to avoid big fluctuation. Small decay may causes the SGD to hardly finding the optimum weight values. Meanwhile, big decay may causes the learning rate to decrease too fast before reaching the optimum.

In Keras, the learning rate in SGD can be decayed by using a decay value which updates the learning rate after each iteration. Equation 4.3 shows the formula of learning rate update (α_{old} to α_{new}) per *iteration* with a *decay* value.

However, this research uses custom *decay_rate*, which will update the learning rate after each epoch based on Equation 4.4. Some decay values to be tested are 0 (no decay), 0.01, 0.05, and 0.1.

$$\alpha_{new} = \alpha_{old} * \frac{1}{(1 + decay * iteration)} \quad (4.3)$$

$$\alpha_{new} = \alpha_{old} * (1 - decay_rate) \quad (4.4)$$

3. Dropout rate

Dropout can help to avoid over-fitting. However, if the dropout rate is too big, it can be hard to find optimal model.

There are two dropout layers that are added in the network. A literature [25] suggests that the best dropout rate is between 0.2 to 0.5. Thus, the experiment will compare dropout rate of 0 (without dropout), 0.2, and 0.5.

4.1.5 Image Pre-processing

The developed CNN requires specific format of image as the input. Thus, several pre-processings need to be applied to the input image before fed into the CNN.

1. Pre-processing based on Caffe setting

The weight values in the pre-trained VGG-PLACES are designed for Caffe setting. Meanwhile, there are several differences between image format in Caffe and Keras network. To make it compatible, some pre-processings need to be applied to the input images for Keras.

a) Set each value to be in range 0-255.

In Caffe, each pixel in input image is represented by a value in range 0 to 255. In Keras, a loaded image may have values with range 0-1. In this case, they should be multiplied by 255.

b) Reverse the channels order (RGB into BGR).

Caffe uses OpenCV, which has BGR format for the image. Meanwhile, Keras uses RGB, so the channel order should be reversed to be BGR.

- c) Subtracts each channel value with the mean of PLACES205.

Mean subtraction is usually performed in Caffe, which involves subtracting the mean of each feature across every image in the data. This pre-processing is intended to centralise the data around the origin along each dimension. The mean value of each channel in PLACES205 dataset are: B = 105 , G = 114, and R = 116. Thus, each values in each image input should also be subtracted by these values.

2. Re-sizing and cropping

The input of the network is an image represented by $(3, 224, 224)$ matrix, which represents an image size of 224×224 , with 3 channels (B,G,R) per pixel. The original image has a size of 600×400 . Thus, the image has to be processed to get the desirable size. One obvious way to do it by re-sizing the original image directly into size 224×224 .

Other than re-sizing it, cropping method can also be used to generate image with size 224×224 . To get equivalent ratio, the original image is firstly re-sized to 400×400 , then a cropped image is generated. There are five area to be captured with the following positions.

- Center: (88,88) to (312,312)
- Top Left: (0,0) to (224,224)
- Top Right: (0, 176) to (224,400)
- Bottom Left: (176,0) to (400,224)
- Bottom Right: (176,176) to (400,400)

Example of the results after re-sizing or cropping are shown in Figure F.2 in Appendix F. By using both re-sizing and cropping, a dataset size will increase to be 6 times the original size.

4.2 Dataset Expansion

Ideally, huge training dataset is required to train a deep learning model. 800 images (which are currently available) are still considered as small amount to develop an accurate model. To generalise the model, more images should be added to the dataset. A common way to do this is by transforming the image, so that the image array will change and be read differently by computer, but the label stays the same.

4.2.1 Image Transformation

Keras has provided various types of image transformations to generate new images from existing images, which is called Image Data Generator. The transformations will add more variations to input images, and each epoch may have different images arrays. Some transformations to be applied are as following.

1. Flipping horizontally (vertical flipping is not applied because it is uncommon to use an image with ground at the top and sky at the bottom)

2. Shearing in counter-clockwise direction randomly with maximum angle of 0.1 radians
3. Random rotation with maximum angle of 5°
4. Shifting the values of each colour channel with maximum of 5
5. Zooming with random scale between 0.8 and 1.2

4.2.2 Spatial-Based Expansion

In case of using Street-View data as the input dataset, fortunately it is simple to add more Street-View data. The challenge is to estimate the attractiveness label of the new data. This research proposes a dataset expansion approach by adding unlabelled Street-View data and estimate their label based on some assumptions. These data may help to improve the performance of machine learning.

Formulation for Spatial-Based Expansion





To explain the next methods, there are some symbols that should be defined.

- att(X)** = Attractiveness value of location represented by Street-View image X . In the original dataset, the value is determined based on the result of crowd-sourcing. For the expansion dataset, the value is estimated based on a formula.
- lat(X)** = The latitude coordinate where Street-View image X is taken. This value is inputted as parameter in Google Street View API during the image crawling.
- long(X)** = The longitude coordinate where Street-View image X is taken. This value is inputted as parameter in Google Street View API during the image crawling.
- hdg(X)** = The heading which Street-View image X is taken. This information is inputted as parameter in Google Street View API during the image crawling.
- round(k)** = The rounded value (closest integer) of a number k

Heading Expansion

In the original dataset, each location is represented by 4 images in 4 perpendicular headings. Based on the analysis in 3.3.3, the attractiveness from Street-View images in the same location with different headings are correlated. It can be assumed that places viewed from the same location may have similar attractiveness, which is supported from an example in Figure 4.2. That figure displays Street-View images extracted from the same coordinate (52.3033,4.9292), but with different headings. The left-most

Table 4.2: Sample of Street-View Images From The Same Coordinate but with Different Headings

heading			
129	159	189	219
			

and right-most images (with heading 129 and 219) are images from the dataset, and both of them have attractiveness values of 4. It can be observed that the other images with heading between them contain similar view. Thus, headings from these middle headings can be added to the training dataset.

The concept can be formulated as following. Let A is an assessed location in the dataset. This location is represented by four images. A_1 and A_2 are two images of A with consecutive headings (i.e. $hdg(A_2) = hdg(A_1) + 90$). Between A_1 and A_2 , other representing image A_i can be extracted with $hdg(A_1) < hdg(A_i) < hdg(A_2)$. The challenge is to estimate the label of A_i . If $att(A_1) = att(A_2)$, then it is simple to assume that $att(A_i)$ will have the same label. If they are different, then the label can be formulated as a linear function as in Equation 4.5.

$$att(A_i) = \text{round} \left(\frac{\text{closeness}(A_i, A_2) * att(A_1) + \text{closeness}(A_i, A_1) * att(A_2)}{\text{closeness}(A_i, A_2) + \text{closeness}(A_i, A_1)} \right) \quad (4.5)$$

A function $\text{closeness}(X_1, X_2)$ is defined to express the similarity between X_1 and X_2 with smaller value means more similar. In this case, the closeness can be formulated as the heading difference (see Equation 4.6).


$$\text{closeness}(X_1, X_2) = \frac{|hdg(X_1) - hdg(X_2)|}{90} \quad (4.6)$$

Location Expansion

Another possibility to expand the dataset is by adding Street-View data from new locations, with an assumption that locations in a specific area may have similar attractiveness values. Based on spatial analysis in 3.3.6, the data has a small positive Moran's I coefficient, which indicates small existence of spatial correlation. However, it is suspected to be caused by the sparsity of the sample locations. Thus, attractiveness of a location cannot be estimated based on attractiveness of neighbouring locations if the neighbours have relatively far distance.

Nevertheless, there is still a possibility that places with very near distance when observed from the same heading will have similar attractiveness. Figure 4.3 shows an intuitive example of this assumption. Image in the middle is a Street-View image from dataset with coordinate 52.3033,4.9292 and heading 129, based on the crowdsourcing, its attractiveness value is 4. The other images are observed from nearby

Table 4.3: Sample of Street-View Images in Nearby Coordinates with The Same Heading. The Images Show Similar Scenes

		Longitude		
		4.9260	4.9262	4.9264
Latitude	52.3031			
	52.3033			
	52.3035			

locations (see the latitudes and longitudes) with the same heading. Those images show identical scene (similar composition of road, trees, and sky). Thus, the attractiveness for those images can be assigned to be the same as its of the original image, which is 4. Obviously, some of them may be inaccurate and having low confidence. For example is an expansion image in bottom-left, which contains a white tractor which may affect the actual attractiveness. This kind of image will become a noise, but maybe also be covered by other images with correct assignment. The validity and efficacy of this approach will be tested via experiment.

The formulation of the location expansion is as following. Suppose A is a Street-View image which $att(A)$ is known. Then other Street-View images from nearby location (with distance threshold d_{max}) with the same heading as A (for example is A') is assumed to have similar attractiveness as A , as formulated in

$$\text{if } dist(A', A) \leq d_{max} \text{ and } hdg(A') = hdg(A), \text{ then } att(A') = att(A) \quad (4.7)$$

In this research, the additional Street-View images for A are extracted from locations with latitudes in $\{lat(A) + 0.0002, lat(A), lat(A) - 0.0002\}$ and longitudes in $\{long(A) + 0.0002, long(A), long(A) - 0.0002\}$, the same as used in example in Figure 4.3. The distance between the original location to the expansion locations is around 20 - 50 m). If a smaller distance is used, the Google Street View will return the same image.

Chapter 5

System Evaluation and Understanding Urban Attractiveness

In previous chapters, urban attractiveness dataset has been created and the design of the CNN training to develop attractiveness prediction system has been elaborated. This chapter presents the result of the experiments to evaluate the CNN design. It also analyses the visual aspects that may contribute to attractiveness which answers RQ4.

5.1 Urban Attractiveness Model Training

This section explains the experiments that have been done in the CNN training and presents the results.

5.1.1 Performance Evaluation Method

Following [19], the evaluation is done by splitting the dataset with 80% of the dataset is used as training dataset and 20% for validation dataset. 40 locations (which consist of 160 Street-view images) are randomly selected with the label distributions are preserved to match the distribution in the original dataset. Golden image with `img_id` 10005 is also included into the validation dataset as the 20% split from golden images. In total, 161 Street-View images are picked as the validation dataset. The remaining 644 Street-View images from 161 locations are used to train the CNN. The validation dataset is independent to the training dataset. So, the performance of the developed CNN can be estimated on the its evaluation when predicting the data in the validation dataset.

The performance is measured based on the root-mean-square error (RMSE) of the prediction on the validation dataset. RMSE can estimate how far the prediction is to the actual target value. Smaller RMSE means better performance because the predictions are more converged to the target labels. For the next explanation, terms *RMSE_train* and *RMSE_val* are used to represent the value of RMSE when predicting the original training dataset and the validation dataset respectively.

By using random guessing based on distribution (random predictor), the classifier will get an expected accuracy of 32.6% and RMSE of 2.04. Meanwhile, if a classifier

naively classifies every images into class 3 (which is the middle as well as the majority class), then the accuracy will be 39.8% and RMSE=0.82. This naive predictor may get better performance for the prediction of this specific validation dataset. However, if the predicted dataset has different distribution, then its performance can be poor. These values will be used as an initial standard for the performance. For the experiments to develop the machine learning model, the accuracy and RMSE should be better than this standard.

5.1.2 Determining hyper-parameters

The first experiment is to determine appropriate hyper-parameters for the CNN training. The CNN is trained by using re-sized original images (without cropping or using Keras Image Generator), so the training and evaluation dataset have the same pre-processing. The training is done in 10 epochs, and then the performance is observed.

As stated in 4.1.4, the pre-defined optimizer is the SGD with *learning rate* = 0.01 and *Nesterov momentum* = 0.9. The other parameters are determined based on grid search, which compares the performance of each usage of hyper-parameter combination. The parameters to be set are batch size, learning decay rate, and dropout rates.

Batch size and learning decay rate

The first grid search is to determine the *batch size* and *decay rate*, which is important to set the training pace. It should not be too slow which makes the model hardly learn the data, but should not be too fast which causes the model to easily over-fit. The *batch sizes* to be tested are 0, 5, 10, and 20. The decay values are tested for values of 0.1, 0.05, 0.01, and 0 (no decay). The dropout rates are set as 0.2 in both layers.

Table 5.1 shows the result of each parameter combination. During the experiment for 10 epoch, the smallest *RMSE_val* (RMSE to the validation dataset) is reported to this table. These best *RMSE_val* of each row are then sorted increasingly. The top rows show the parameter combinations with the smallest *RMSE_val*, which are assumed to be the best parameters that could train the CNN to successfully classify external data. The information of the *RMSE_train* (RMSE to the training dataset) are also provided, which can tells whether the model is general or too over-fitted to the training data. If the *RMSE_train* is too small compared to the *RMSE_val*, then it is considered as over-fitting.

Based on the result, the best performance is accomplished by using a batch size of 5 and a decay rate of 0.1 or 0. Both of them got the best *RMSE_val* of 0.72. The *RMSE_train* is around 0.3 smaller than the *RMSE_val* which indicates that the CNN is quite over-fit. Nonetheless, this over-fitting issue will be handled via dropout and dataset expansion. Hence, *batch_size* = 5 and *decay_rate* = 0.1 are selected to be used in the next experiments because it is experimentally proven to provide a better performance, even though the other top combinations actually can also be used.

Dropout rates

The next experiment is to determine the dropout rates for both fully connected layers. The values to be tested are 0, 0.2, and 0.5 for each dropout layer. The *batch size* and *decay rate* are set to be 5 and 0.1 respectively based on the previous experiment.

Table 5.1: Training Result of The First Grid Search. Each Row Shows The Best $RMSE_{val}$ (RMSE to The Validation Dataset) Achieved for A Combination of Batch Size and Decay Rate

rank	batch size	decay	RMSE to validation dataset	RMSE to training dataset
1	5	0.1	0.72	0.43
2	5	0	0.72	0.4
3	20	0.1	0.74	0.51
4	1	0.1	0.74	0.42
5	20	0.01	0.75	0.52
6	10	0	0.75	0.2
7	20	0	0.76	0.55
8	20	0.05	0.76	0.37
9	10	0.05	0.76	0.33
10	5	0.01	0.77	0.12
11	10	0.1	0.78	0.57
12	5	0.05	0.78	0.53
13	10	0.01	0.78	0.29
14	1	0.01	0.82	0.84
15	1	0	1.31	1.42
16	1	0.05	1.31	1.42

Table 5.2: Training Result of The Second Grid Search. Each Row Shows The Best $RMSE_{val}$ (RMSE to The Validation Dataset) Achieved for A Combination of Dropout Rates

rank	batch size	decay	RMSE to validation dataset	RMSE to training dataset
1	0.2	0.2	0.72	0.4
2	0.5	0.5	0.73	0.6
3	0	0.2	0.73	0.4
4	0.2	0	0.73	0.28
5	0.5	0	0.74	0.62
6	0.5	0.2	0.75	0.61
7	0.2	0.5	0.75	0.59
8	0	0	0.75	0.42
9	0	0.5	0.76	0.44

The result is shown in Table 5.2 with similar representation as before. The result shows that the best $RMSE_{val}$ is achieved with $dropout\ rate\ 1 = 0.2$ and $dropout\ rate\ 2 = 0.2$. Thus, this combination will be still used for the next experiments. Higher dropout rates may help to avoid over-fitting, but there is a risk that the developed CNN becomes too general and fail to learn attractiveness.

5.1.3 Learning Transformed Images

After the hyper-parameters are set ($learning\ rate = 0.01$, $batch\ size = 5$, $decay\ rate = 0.1$, $dropout\ rate\ 1 = 0.2$, $dropout\ rate\ 2 = 0.2$), the training can be begun with transformed training data. The first experiment is to use Keras Image Generator to add image variation. The training still uses 644 images in the training dataset per epoch, but each image will be different in each epoch due to the transformation. The best achieved $RMSE_val$ is 0.72, which is still the same as the result of training without image transformation. However, now the $RMSE_train$ is 0.52 which is less over-fitted. In this case, image transformation gives slight positive impact to the training.

The next experiment is to check the effect of data augmentation through cropping. Each original Street-View image is processed into 6 images (1 re-sized image and 5 cropped images explained in 4.1.5). The resulted 3,864 images are then used to train the model for each epoch. The hyper-parameter configurations are still the same as the previous experiment. During the training, the Keras Image Generator is still applied to add variations to the training dataset in each epoch. Based on the experiment, the best achieved $RMSE_val$ is 0.75, which is worse than the the training with only the original training dataset.

There is a possibility that when a Street-View image is cropped, its attractiveness perception may be altered. For example, the cropping caused some objects in the image to be removed. Thus, giving it the same label as the original leads to inaccuracy during training. This case is different from the data transformation by using Keras Image Generator. After an image is flipped, rotated, or sheared, visually it still looks similar to the original image and may not lose any significant features which may influence the perception. From this experiment, it has been found that image transformation could slightly reduce over-fitting during training, but image cropping did not improve the performance of the training.

5.1.4 Learning Spatial-Based Expansion Dataset

As elaborated in 4.2, dataset expansion methods based on spatial data are proposed. The first one is the heading expansion. For each location in the training dataset, 12 Street-View images are extracted with heading interval of 30° starts from its initial heading. Each image is then labelled based on Equation 4.5. The training is re-started by using the extracted 7,728 Street-View images. This approach improves the best $RMSE_val$ to be 0.70.

The next approach is location expansion. For each Street-View image in the training dataset, 8 additional images are extracted from the nearby coordinates stated in 4.2.2. Sometimes, the target coordinate shows no Street-View image, or shows an indoor place, which should be omitted. Finally, there are 2,281 Street-View images included in the location-based expansion dataset. The result of CNN training by using this dataset did not yield a better performance. The best achieved $RMSE_val$ is only 0.78.

The performance of each dataset expansion methods are summarised in Table 5.3. The dataset expansion method relies on the correctness of labels given to the additional input images. From these experiment, it was found that the heading-based expansion could slightly improve the performance of the CNN. It is possibly because each ad-

Table 5.3: The Performance of CNN Training by using Various Dataset Expansion Techniques

Dataset Expansion	RMSE to expanded training dataset	RMSE to original training dataset	RMSE to validation dataset
Image transformation	N/A	0.52	0.72
Image cropping	0.69	0.64	0.75
Heading-based expansion	0.62	0.67	0.70
Location-based expansion	0.51	0.58	0.78

Table 5.4: Accuracy and RMSE of Attractiveness Prediction of Images in Validation Dataset Based on Developed CNN, Random, and Naive Predictor

Predictor	Accuracy	RMSE
CNN	55.9%	0.70
Naive	39.8%	0.82
Random	32.6%	2.04

ditional Street-View image has intersection to some labelled images in the original dataset. They have common visual features and validate the assumption that adjacent heading leads to similar attractiveness in most cases. However, location-based expansion failed to develop a better model. A possible explanation is that the assumption that a very near place will have similar attractiveness perception is inappropriate, which is relevant to the analysis result that the attractiveness data have low auto-correlation.

5.2 Final Trained Attractiveness Model

After the experiments are done, the best attractiveness prediction system so far has been developed when learning the heading expansion dataset. This model will be used and reviewed for the next analyses and applications.

5.2.1 Model Performance

The developed model has an accuracy of 62.9% and RMSE of 0.67 to the training dataset, which is still not too over-fitted and is expected to generalise the characteristics of attractiveness perception. When tested to the validation dataset, this model got an error of 0.70, and accuracy of 55.9%. This is already an improvement from the random and naive prediction (Table 5.4 shows their comparison).

The confusion matrix of the prediction to the validation dataset is shown in Table 5.5. The table shows that there is no image classified into class 1 or 5. It is maybe caused by the fact that the number of training images from those classes are very small and did not give strong influence. So, the CNN classified them into class 2 or 4 instead.

Table 5.5: Confusion Matrix Between Predicted Attractiveness Label from Developed CNN and Actual Label from Pilot Crowd-Sourcing for Street-View Images in Validation Dataset

		Label from pilot crowd-sourcing				
		1	2	3	4	5
Label from CNN prediction	2	3	25	14	2	
	3		20	47	27	
	4		1	3	18	1

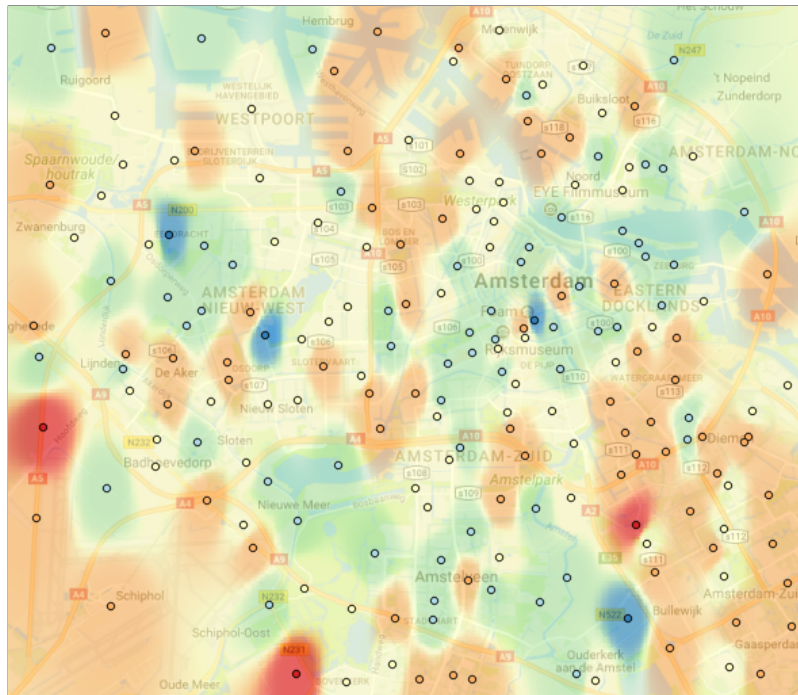


Figure 5.1: Attractiveness Distribution in Amsterdam Based on Assessed Dataset

5.2.2 Visualising Attractiveness Distribution in Amsterdam

Figure 5.1 shows the attractiveness distribution in Amsterdam by using 201 locations in the pilot dataset (shown as circle points). The heatmap visualisation is generated via QGIS¹ with inverted weight interpolation. The colour codes range from red for attractiveness=1, to orange for attractiveness=2, to yellow for attractiveness=3, to green for attractiveness=4, and to blue for attractiveness=5.

The developed CNN model can be used to predict new locations. More sample locations in Amsterdam are picked with interval of 0.003 latitude and 0.003 longitude. In each location, four Street-View images are extracted each with a heading of 0, 90, 270, and 360. The images which are invalid or show indoor scenes were removed. In total, 7,392 valid Street-View images from 1,848 locations were crawled. The attractiveness of each location is predicted by classifying the four representing Street-View

¹URL: <http://www.qgis.org/en/site/> (accessed: 2017-08-28)

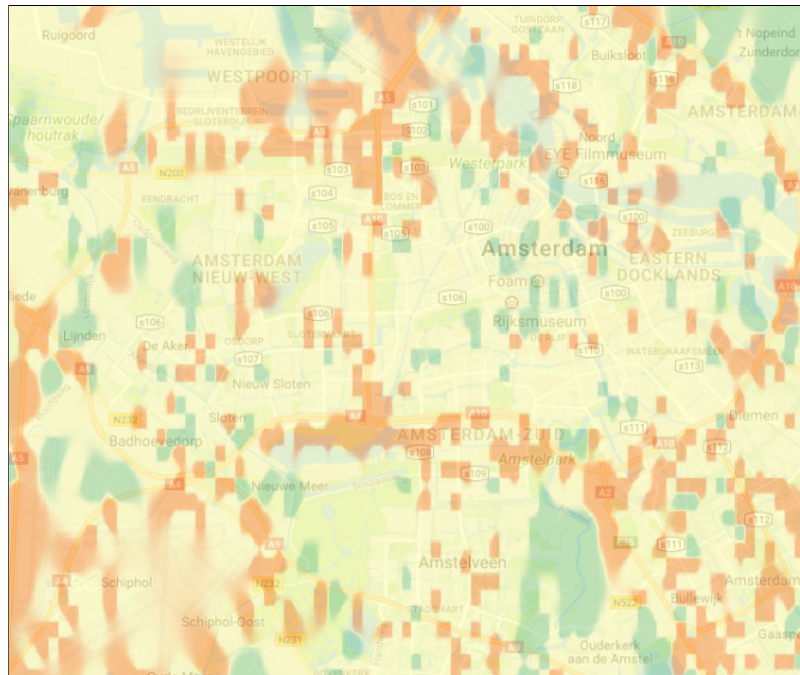


Figure 5.2: Attractiveness Distribution in Amsterdam Based on Prediction System

images with the CNN model and then compute their mean. Figure 5.2 shows the attractiveness distribution in Amsterdam based on those samples. Generally, it shows that most of the places has neutral (class 3) attractiveness, and there are some spots with scattered attractive (class 4) places as well as unattractive (class 2) places.

5.2.3 Spatial Analysis Based on Predicted Attractiveness

By using the 1,848 sample locations which are already extracted and predicted in 5.2.2, the Global Moran's I coefficient can be re-computed. The result with distance threshold weighting is shown in Table 5.6. The table shows that all of the results are significant. Moreover, the smaller the threshold, the Moran's I coefficient tends to increase. Even though, the maximum auto-correlation can be achieved is around 0.2, which is still considered as small. Similar result is also obtained by using k-NN based weighting, which is shown in Table 5.7. These results assure that the attractiveness of neighbouring places are generally weakly correlated, but the correlation is stronger when the distance is smaller.

5.3 Visual Aspects Related to Urban Attractiveness

Attractiveness is a complex perception. This thesis assesses attractiveness of places based on how people perceive them. So, there should be some visual characteristics which make a place looks more attractive or less attractive. Some analyses have been done to get these information.

Table 5.6: Moran's I and p-Values of Predicted Attractiveness with Weights Based on Distance Threshold

distance threshold (km)	Moran's I	p-rand	p-norm	# locs without neighbour
5	0.007	0.0000	0.0000	0
4	0.016	0.0000	0.0000	0
3	0.028	0.0000	0.0000	0
2	0.052	0.0000	0.0000	0
1	0.135	0.0000	0.0000	0
0.9	0.145	0.0000	0.0000	1
0.8	0.154	0.0000	0.0000	1
0.7	0.171	0.0000	0.0000	3
0.6	0.194	0.0000	0.0000	4
0.5	0.203	0.0000	0.0000	18
0.4	0.196	0.0000	0.0000	44
0.3	0.202	0.0000	0.0000	224

Table 5.7: Moran's I and p-Values of Predicted Attractiveness with Weights Based on k-NN

k	Moran's I	p-rand	p-norm
2	0.221	0.0000	0.0000
3	0.211	0.0000	0.0000
4	0.209	0.0000	0.0000
5	0.192	0.0000	0.0000
6	0.181	0.0000	0.0000
7	0.177	0.0000	0.0000
8	0.170	0.0000	0.0000
9	0.161	0.0000	0.0000
10	0.152	0.0000	0.0000

5.3.1 Scenes Related to Urban Attractiveness

The first analysis is to find out scenes which may correlate to attractiveness perception. PLACES-VGG is used to classify scenes (as texts) of each Street-View image in the dataset. Top 5 predicted scenes of each image are extracted, and any scene with score less than 0.99 is omitted because it may be inaccurate. Next, the frequency of each scene occurring in attractive places (Street-View images with attractiveness label 4 and 5) are recapitulated and treated as weight score of the scene. The same process is applied to unattractive places (Street-View images with attractiveness label 1 and 2) as well as neutral places (attractiveness label 3). Next, the score of each scene is normalised based on its ratio among these 3 groups. Finally, the score in each group is ranked. If the frequency is less than 5, then it is ignored to avoid bias due to lack of sample. Table 5.8 shows the top 10 result. In that table, "ratio" determines the ratio

Table 5.8: Scenes Correlated to Unattractiveness and Attractiveness

rank	unattractive			neutral			attractive		
	scene	ratio	portion	scene	ratio	portion	scene	ratio	portion
1	train_station	0.857	0.43%	restaurant_patio	0.714	0.33%	harbor	0.857	0.61%
2	valley	0.833	0.36%	fire_escape	0.625	0.33%	fairway	0.750	0.61%
3	viaduct	0.833	1.44%	mansion	0.565	1.74%	river	0.731	1.93%
4	train_railway	0.818	0.65%	residential_houses	0.525	10.48%	dock	0.727	0.81%
5	railroad_track	0.765	0.94%	motel	0.504	4.34%	formal_garden	0.706	1.22%
6	bridge	0.719	1.66%	inn	0.487	2.54%	pond	0.684	1.32%
7	dam	0.714	0.36%	apartment_building	0.480	7.34%	bayou	0.650	1.32%
8	skyscraper	0.714	0.36%	courthouse	0.455	0.33%	cottage_garden	0.647	1.11%
9	water_tower	0.700	0.50%	plaza	0.449	2.67%	marsh	0.565	1.32%
10	shed	0.667	0.58%	crosswalk	0.447	3.94%	schoolhouse	0.556	0.51%

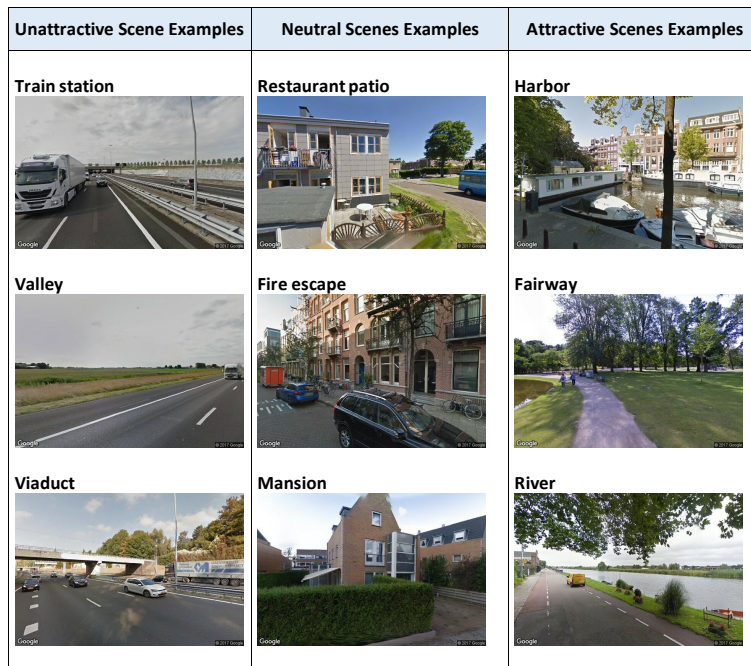


Figure 5.3: Image Examples of Top Scenes in Each Attractiveness Category

between the scene frequency in the group (attractive/neutral/unattractive) to the total frequency of a particular scene. Meanwhile, "portion" shows the scene frequency in the group relative to the total frequency in the same group.

Based on this result, the scenes correlated to attractiveness are obtained. Most of the unattractive scenes are related to roads (e.g. a checking in the dataset reveals that "train station", "valley", and "viaduct" are actually related to car streets). Meanwhile some scenes linked to attractiveness are related to water environment (e.g. harbour, river, dock, pond, bayou) and greenery field (e.g. fairway, formal garden, cottage garden, marsh). For the neutral places, the scenes are mostly related to buildings to stay (e.g. fire escape, mansion, residential houses, motel, inn, apartment building). Figure 5.3 shows the image examples of the top scenes in each of the category.

5.3.2 Visual Patterns Related to Urban Attractiveness

One of the simple way to investigate visual aspects that may relate to urban attractiveness is by directly observing Street-View images with top and bottom attractiveness values, then compare them and manually search the patterns. Figure 5.4 show top 10 of the most attractive and the least attractive Street-View image of places in the dataset based on pilot crowd-sourcing. However, there is no clear pattern which can differentiate them.

Another method is by feeding patches of Street-View images to the prediction system and check the output. Patches with high output values in the last output layer will likely contain patterns which contribute to attractiveness. This kind of technique is also used in [19, 5]. To create the patches, each Street-View image in the dataset is cut into 25 parts with size 120×80 (the size ratio is maintained). Next, the attractiveness of each patch is classified by using the developed CNN.

To check the most attractive patterns, the top 100 patches with the highest value in the third output node are observed. That output node represents the boundary between class 3 and 4, so its value will determine how likely that image is to be classified into class 4. By using the same approach, the least attractive patterns are observed from 100 patches with the lowest values in the second output node which represents the boundary between class 2 and 3. Figure 5.5 and 5.6 show those patches for the most and the least attractive patterns respectively.

It can be seen that they have different characteristics. The most attractive patches are dominated with images of trees and sky. Meanwhile, the least attractive patterns are mostly related to buildings and roads. These findings support the result from the scene analysis, that road is perceived as not attractive and greenery is attractive. Interestingly, scenes related to building which were mostly considered as neutral based on scene analysis, are classified into the least attractive based on the developed CNN.

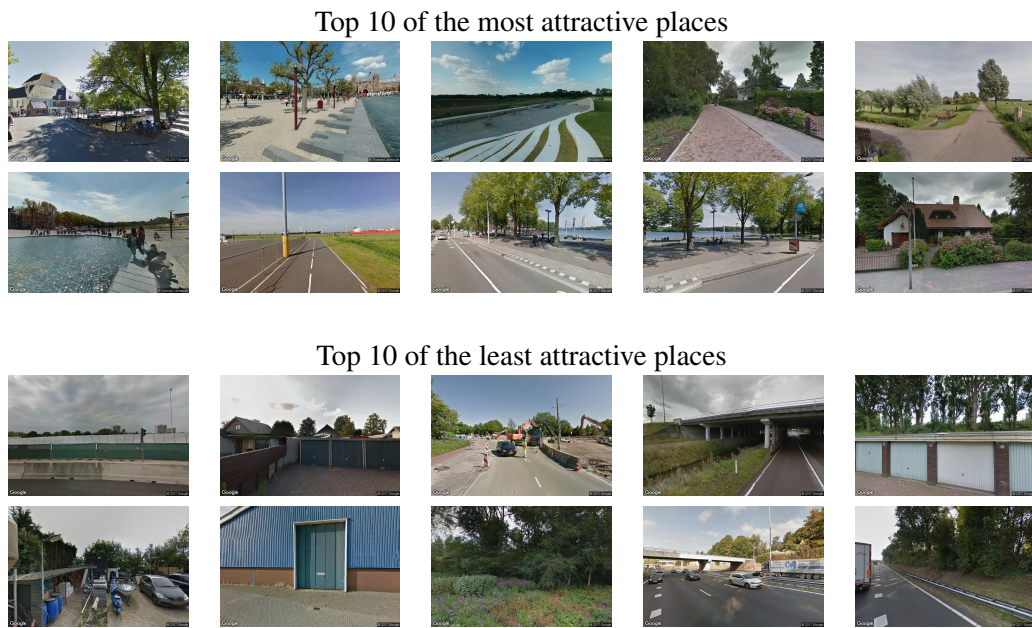


Figure 5.4: Top 10 of The Most and The Least Attractive Places Based on The Pilot Crowd-sourcing



Figure 5.5: The Most Attractive Visual Patterns Based on The Developed CNN



Figure 5.6: The Least Attractive Visual Patterns Based on The Developed CNN

Chapter 6

Discussion

This chapter reflects on the results of the experiments and relates them to existing literature. It also presents the possible threats to validity, either internal or external.

6.1 Discussion

The reflection and analysis of each process and result in the proposed method are discussed as follows.

6.1.1 Dataset Generation

The dataset generation consists of various processes, such as data acquisition, data labelling via crowd-sourcing (either internal or public), and data analysis. There are several topics that can be discussed, which are as seen below.

1. Representing a location with Street-View images

The generated dataset justifies that when assessing perception of a location, representing it with only a single Street-View image is not sufficient. The analysis has shown that when a location is represented by four Street-View images in perpendicular heading, its overall attractiveness can be estimated based on the mean of attractiveness from each image with an RMSE of 0.59. This result is considered as good, because the same combinations of image-level labels may lead to different location-level labels. Nevertheless, the confusion matrix in Table 3.1 shows that the predicted location-level label can accurately predict the best label for most of the labels set.

This property can be extended, when a location is represented by any number of Street-View images (and not necessarily separated by the same heading intervals), the overall attractiveness can be estimated as the weighted mean of their attractiveness labels. However, more studies are required to confirm this hypothesis.

2. Relating attractiveness to other perceptions

The analysis in this research found that attractiveness of a place has a relatively high correlation with pleasure ($\rho = 0.776$) and friendliness ($\rho = 0.549$). Based

on the factor loading result, pleasure and friendliness have the component value of 0.931 and 0.575 respectively to the only generated factor, which is assumed to be attractiveness. Some studies also comply with this inference, for example, Karmanov and Hamel [10] which showed high attractiveness correlation to pleasure and friendliness with $\rho = 0.904$ and $\rho = 0.756$ respectively based on the factor loading. Thus, the approach to assess the attractiveness of places based on Street-View data proposed in thesis could provide reliable result and comparable to the result from previous studies with different methods.

3. Internal crowd-sourcing vs public crowd-sourcing

The results from public crowd-sourcing showed relatively high variances and some of the aggregated labels are different from the results from the internal crowd-sourcing. It may be caused by the difference in participants' living environment. In internal crowd-sourcing, the participants have stayed in Netherlands and have been exposed to various scenes of Netherlands. Meanwhile, workers in CrowdFlower may not have visited Netherlands. Moreover, the workers in AMT were not monitored and guided. There is a probability that some of them did not fully understand the task, or just did not want to do the task seriously. Some of these fraudulent workers were detected based on answers to golden questions and judgments distributions. However, some of the fraudulent workers may still pass these qualifications (i.e. they judged randomly, but luckily answered golden questions correctly).

6.1.2 CNN Learning

In developing a CNN model, architecture design and training configuration are essential to successfully train the network. The final experiment shows that the model has an estimated RMSE of 0.70, which is an improvement from 0.82 achieved with a naive predictor which always classify any image into class 3. This performance can be significantly improved by adding more training data.

The usage of 4 binary outputs to predict 5-scale ordinal values seems to work well. Even though, the model never predicts any image into class 1 or 5 due to their size minority. So, the very unattractive and very attractive images were predicted into class 2 or 5 instead to minimise error. The confusion matrix in Table 5.5 also shows that the prediction rarely deviates to more than 1 class.

The selection of hyper-parameters is also important. In dealing with a training dataset with small size, the wrong choice of parameters may cause over-fitting, or even really slow learning. The grid search is a good strategy to find a good combination of parameters. From the grid search, only one combination is selected to be used in the CNN training. Other alternatives of combination can also be used, and it may lead to a similar result.

The experiments also showed that image pre-processing is essential in CNN training. Image transformations (i.e. horizontal flipping, shearing, rotating, channel shifting, and zooming) provides good impact to the developed CNN. It is reasonable that even though the image is modified, but visually there is no significant change to the view, and it is reasonable that the attractiveness perception stays the same.

6.1.3 Spatial-Based Dataset Expansion

Based on the spatial analysis, with only observing the 201 locations in the crowd-sourced dataset, the Global Morgan's I showed a small value (around 0.1) which indicates lack of pattern existence. After more locations are observed and the attractiveness labels were predicted by the CNN model, the Morgan's I coefficient became around 0.2, which is still considered as small. These results indicate that places in nearby locations are not necessarily having similar attractiveness. This finding is also supported with the experiment result that the location expansion approach (explained in 4.2.2) failed to improve the CNN performance. When people move from a location to another location in several meters distance, they may observe different scenes and objects, which lead to different attractiveness perception. When there are a lot of inaccurate labels in the expansion dataset, they will become noises and interfere with the accuracy of the trained model.

However, heading expansion approach could still work and improved the CNN performance. Any two Street-View images captured from the same location with adjacent heading (i.e. $< 45^\circ$) will have similar part of images. That intersection may influence the attractiveness of the location.

6.1.4 Pattern Observation

The analysis in 5.3 found the scenes and visual patterns correlated to the attractiveness of places. Water environments dominate the top ranks of attractive scenes with a relatively big portion. Amsterdam (as well as most cities in Netherlands) has a lot of canals and ponds, which are well-managed and successfully attract people to visit. Conversely, roads mostly show up in unattractive places. These kinds of scenes are usually used for transportation, so they are not really intended to attract people to visit. Buildings are mostly considered as having low attractiveness. However, some of them might be considered as neutral, especially if they are functioned to stay, such as mansion, residential houses, motel, inn, apartment building. Other interesting scenes are ones related to a large field. The field which contains greenery likes garden is more attractive than an empty field.

These observations match with the analysis result of some existing research. Hidalgo, et al. [9] found that recreational places and panoramic places are considered as attractive, which can be linked to greenery and blue sky. On the other hand, the same literature also considered housing areas and industrial places as unattractive. Another study also concluded that the natural environment was significantly more attractive than the urban environment [10]. In addition to those previously found patterns, this thesis found another information that water environments and blue sky are attractive, and transportation roads are unattractive.

6.2 Threats to Validity

Following are some limitations and threats to the validity of this research.

1. The size of assessed dataset

The generated dataset in this thesis contains 800 Street-View images from 200 locations. Compared to the other datasets on urban perception, this number is considered as small (e.g. Doersch, et al. [5] suggests that around 10.000 street-level images with 2.000 positive classes and 8.000 negative classes are sufficient to detect discriminative visual elements). There is a probability that it does not cover all of possible views and environments.

Despite that issue, the assessed locations are scattered over the city. It may not cover all kinds of city views, but each location may represent the sample view in its area. Some analysis of this small dataset also provided similar result to some past research. Even with small size, this dataset could develop a prediction model with better performance than random prediction. Even though, to develop a model with high accuracy, more assessed data are still required.

2. The number of crowd-sourcing participants

In data assessment via crowd-sourcing, more people to judge an object will lead to a more reliable result. In this thesis, each object is judged by only five people.

The crowd-sourcing in this research deals with an assessment of human perception, which is a subjective task. With any number of participants, there is always an uncertainty to the aggregated label. So, the variances were observed to check the confidence of the labels. For golden images which were each judged by 50 people, the variances were consistently less than 1, which is used as a standard threshold. The generated dataset already contains more than 80% images and locations with variance below this threshold, even with only 5 judgments.

3. The crowd-sourcing judgments validation

The competence of each crowd-sourcing participant is checked based on their ability to annotate objects in golden questions. However, object annotation is different from attractiveness assessment as the main task.

The object annotation task can test the attention of the participants during the survey. If they can answer the golden question correctly, then they are serious in doing the task and assumed to be able to do the attractiveness assessment well. In the internal crowd-sourcing, the process is guided and monitored by surveyor, so unreliable participant is unlikely to exist. Besides, the validity of the judgments is not only based on their competence, but also the variance of the judgments. If the variance is small, then the participants have an agreement to the assessment.

4. Dataset label aggregation

Attractiveness label used in this thesis is Likert-scale with 5 ordinal values. The judgments are aggregated based on simply their median. Meanwhile, currently, there are various advanced methods for the aggregation.

Label aggregation by taking the median of the judgments is already a valid method in statistics, especially for data with ordinal type. Each of the crowd-sources is treated as having the same competence. Some probabilistic approaches may also be used, however, it still will not guarantee better labels when each object is judged by 5 people.

5. Experimental Result

The performance of the developed CNN model in this thesis is estimated based on the experiment. When another dataset is used to train or test the CNN, it may give a different result.

The validation dataset used to estimate the performance was selected and set to have similar class distribution to the overall dataset. When the model is evaluated by using another dataset, the metric value may be different, but it usually does not have any significant discrepancy [19].

6. Ignored factors

In this research, there are some factors that were not considered. Temporal information such as time of the day and season of the year is assumed to be default in Google Street View. The existence of people in the images may also influence the attractiveness.

Most of Street-View data in the dataset is taken at noon and taken in the same season. Thus, all of the data are assumed to be in the same condition. It needs more study to deduce whether the attractiveness will be different if taken in low light (e.g. at night) or another season. For the existence of people, during the data acquisition, very crowded locations (e.g. tourism objects) were avoided to minimise the captured people. But, in fact, it is difficult to gather images with no people in all places. However, the pattern analysis showed that people existence did not really influence the attractiveness of places.

Chapter 7

Conclusions

This chapter summarises the works done in this thesis. The conclusions are drawn by answering the research questions. The last section outlooks some directions for future research.

7.1 Conclusions

The answers to each research question are concluded as the following.

RQ1 How to quantify the attractiveness of places in city regions by using Street-View data?

Even though attractiveness of a place is subjective, it can be quantified by extracting its Street-View data and use crowd-sourcing to assess the attractiveness. This is called dataset generation process, which consists of three steps: data acquisition, data labelling, and attractiveness quantification. In the data acquisition, Street-View data are extracted by using Google Street View API. Each location can be represented by four Street-View images from four perpendicular headings. The location attractiveness can be computed as the mean of assessed attractiveness from each representing image. In the data labelling, the attractiveness of each location can be assessed by looking at its representing image and answering question "Would you like to visit this place?". This thesis uses 5 Likert-scale as the answer, which can be mapped into an ordinal value from 1 to 5. However, it turned out that locations with attractiveness value of 1 or 5 are rare. So, using 3 ordinal values as the attractiveness label is already sufficient, which will differentiate unattractive, neutral, and attractive places. The assessment can be done via crowd-sourcing. It was experienced that the judgements are more reliable if the crowd-sourcing is monitored and done by participants who have lived in a similar environment with the area of the assessed location.

During the attractiveness quantification, there are some notable findings. The attractiveness of a place has a high positive correlation to uniqueness, friendliness, pleasure, and dominance. Moreover, there is a possibility of multi-collinearity between attractiveness and pleasure. It means that how people perceive attractive place is related to how much pleasure they feel when seeing that place. On

the other side, familiarity has a small correlation to any of the other assessed attributes.

RQ2 How to develop a model that can automatically predict the attractiveness of places from Street-View data in city regions?

A model which can automatically predict attractiveness of a place can be developed by using the Convolutional Neural Network (CNN). The model is trained by using the labelled dataset. The CNN used in this research consists of 5 convolutional blocks and the first fully connected layer in VGG-PLACES as feature extractor, followed by two fully connected layers with dropouts, and the last is output layer. To handle 5 ordinal values as the labels, the output layer uses 4 binary nodes which represent the boundary of each adjacent classes. The training of CNN with this architecture successfully developed an attractiveness prediction model with improved performance than the random and naive prediction (from an RMSE of 0.82 to 0.70).

A suitable combination of hyper-parameters can be selected by using grid search, which compares the performance of each hyper-parameter combinations. The experiment showed that with a small dataset, the model training is easily overfit, even with dropout layers. Thus, a lot of augmented data are required. Image transformations (i.e. horizontal flipping, shearing, rotating, channel shifting, and zooming) are effective to add variation to the dataset and may improve the performance of a CNN. However, image cropping did not improve its performance. The other methods to expand the dataset are by utilising spatial data which will answer the RQ3.

RQ3 How does the spatial dimension of the collected data affect the predictive performance of the machine learning model?

Spatial information can be used to observe the distribution pattern of attractiveness over the area, which can be analysed via auto-correlation (e.g. based on Global Moran's I coefficient). If the data show high positive spatial auto-correlation, then the attractiveness of a location can be estimated based on the attractiveness of neighbouring locations. Unfortunately, the analysis showed small spatial auto-correlation of attractiveness in the data. Thus, the attractiveness of a new place cannot be estimated based on the attractiveness of neighbouring locations because they do not necessarily have correlated attractiveness perception.

Another possibility to expand the dataset is by using an assumption that locations with very close distance (i.e. < 300 m) when observed from the same heading may have similar attractiveness. However, the experiment showed that this approach did not improve the performance of the machine learning model (with an increasing RMSE of 0.06 compared to the training without any dataset expansion), which may disproved the assumption or the specified distance is not small enough. Hence, it affirmed the spatial auto-correlation analysis that nearby places may not have similar attractiveness, even if they have small distance.

Another way to expand the dataset based on spatial data is by using heading expansion, which assumes that places viewed from the same location but slightly

different heading may have similar attractiveness. Thus, for each Street-View image in a training dataset, another image from other headings can be added, and its label is estimated based on the labelled images in the same location. Based on the experiment, this approach could slightly improve the performance of the machine learning model (from an RMSE of 0.72 to 0.70).

RQ4 Which visual features of the urban environment contribute to the attractiveness of a place in city regions?

The thesis has observed some visual patterns that may be correlated to the attractiveness of a place. Most of the unattractive scenes are related to roads and buildings. Meanwhile, some scenes linked to attractiveness are related to water environment, greenery field, and blue sky. These information can be considered for the planning and development of a city.

MRQ How to implement a computational system that quantifies and predicts the attractiveness of places in city regions, based on Street-View data?

Those answers of the research sub-questions can be summed up to answer this main research question. This thesis has confirmed that there is a relationship between the physical appearance of a place and its attractiveness, which is measured based on how people perceive it.

From RQ1, it was discovered that quantifying attractiveness of places can be done by means of Street-View data and the assessment via controlled crowdsourcing. Some information were gathered, such as attributes correlated to attractiveness, the influence of each observed view to the overall attractiveness of a place, and how attractiveness of places are spatially distributed. Moreover, through RQ4, some visual aspects which are related to attractiveness are also revealed. Some noises normally exist in the assessed data due to the subjectivity of attractiveness perception. Nevertheless, it was validated by the fact that the analysis results on the data in this thesis match to some existing research which used other methods.

RQ2 provided a solution to make a more scalable way to assess the attractiveness of places, which is by developing a CNN model which could automatically estimate the attractiveness of a place based on Street-View data. The accuracy of the developed CNN can be improved by learning larger dataset. In RQ3, some possibilities to take advantage of spatial information to improve the performance of machine learning were propose, even though some of them did not give significant impact.

7.2 Outlook

This thesis is a pilot research to study the attractive perception of places based on Street-View data. There are numerous rooms for improvement to continue this research, which is elaborated as the following.

1. Increasing the number of locations in the dataset to be assessed

This research has already generated an urban attractiveness dataset of 800 Street-View images from 200 locations. The crowd-sourcing can be continued to assess more Street-View images from more locations. By having more data, the applied analysis will be more accurate and the training of machine learning model may achieve better performance.

In another perspective, this research uses Amsterdam as a study case. Thus, the analyses in this thesis may only be relevant in the context of attractiveness perception in Netherlands. The research can be conducted for other countries to observe the attractiveness characteristic in there, as well as to get a global insight of urban attractiveness in general.

2. Improvement of the machine learning model

Besides increasing the training dataset size, various other approaches can be applied to improve the performance of the CNN model. Different extracted features, feature reduction methods, and CNN architecture may lead to better accuracy.

3. Implementation of the methods to quantify other urban attributes

The method to quantify urban attractiveness in this research can be adapted to other perceptions, especially which are related to the visual appearance. The approach of using multiple Street-View images to represent a single location can also be applied to the existing research on quantifying urban perception through Street-View data. The heading-based dataset expansion can also be applied to provide more training data in machine learning with Street-View images as the input. If the assessed attribute has high spatial auto-correlation, the location expansion technique can also be tested.

Bibliography

- [1] Sean Arietta, Alexei A. Efros, Ravi Ramamoorthi, and Maneesh Agrawala. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Transactions on Visualization and Computer Graphics*, pages 2624–2633, 2014.
- [2] Léon Bottou. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes 91*, Nîmes, France, 1991. EC2.
- [3] Joost Broekens and Willem-Paul Brinkman. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641 – 667, 2013.
- [4] A.D. Cliff and J.K. Ord. *Spatial autocorrelation*. Pion, London, 1973.
- [5] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *Communications of the ACM*, 58(12):103 – 110, December 2015.
- [6] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and Cesar A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. *European Conference on Computer Vision (ECCV)*, 2016.
- [7] Leandre R. Fabrigar and Duane T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, Oxford, 2012.
- [8] Fabien Girardin, Andrea Vaccari, Alexandre Gerber, Assaf Biderman, and Carlo Ratti. Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructures Research*, 4:175 – 200, 2009.
- [9] M. Carmen Hidalgo, Rita Berto, Maria Paz Galindo, and Anna Getrevi. Identifying attractive and unattractive urban places: Categories, restorativeness and aesthetic attributes. *Medio Ambiente y Comportamiento Humano*, 7(2):115 – 133, 2006.
- [10] Dmitri Karmanov and Ronald Hamel. Broken windows. *Assessing The Restorative Potential of Contemporary Urban Environment(s): Beyond the nature versus urban dichotomy*, 86:115 – 125, 2008.

- [11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [12] Janneke Roos-Klein Lankhorst, Sjerp De Vries, and Arjen Buijs. Mapping landscape attractiveness: A gis-based landscape appreciation model for the dutch countryside. *Research in Urbanism Series*, 2, 2011.
- [13] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.
- [14] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1):17–23, June 1950.
- [15] Nikhil Naik, Scott Duke, Ramesh Raskar, Edward L. Glaeser, and Cesar A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences of the United States of America*, 114(29):7571–7576, July 2017.
- [16] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. Streetscore - predicting the perceived safety of one million streetscapes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 779 – 785, 2014.
- [17] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. *Computer Vision and Pattern Recognition (CVPR)*, pages 1717–1724, 2014.
- [18] Vicente Ordonez and Tamara L. Berg. Learning high-level judgments of urban perception. *Computer Vision ECCV*, 2014.
- [19] Lorenzo Porzi, Samuel Rota Buló, Bruno Lepri, and Elisa Ricci. Predicting and understanding urban perception with convolutional neural networks. *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, 2015.
- [20] Daniele Quercia, Neil O’Hare, and Henriette Cramer. Aesthetic capital: What makes london look beautiful, quiet, and happy? *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, February 2014.
- [21] Philip Salesses, Katja Schechtner, and Cesar A. Hidalgo. The collaborative image of the city: Mapping the inequality of urban perception. *The PloS One*, 8(7), July 2013.
- [22] Ernestasia Siahaan, Judith A. Redi, and Alan Hanjalic. Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal. *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2014.

-
- [23] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, 2:958–963, 2003.
- [24] Jolita Sinkiene and Saulius Kromalcas. Concept, directions and practice of city attractiveness improvement. *Public Policy and Administration*, (31):147 – 154, 2010.
- [25] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, pages 1929–1958, 2014.
- [26] Martin Thoma. Analysis and optimization of convolutional neural network architectures. Master’s thesis, Karlsruhe Institute of Technology, July 2017.
- [27] Alexandra Tisma and Rene van der Velde. Wesense: Social sensing the quality of urban environments. 2016.
- [28] W.R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234 – 240, June 1970.
- [29] An Gie Yong and Sean Pearce. A beginner’s guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2):79 – 94, 2013.
- [30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv*, 2016.

Appendix A

Source Codes for Development and Analysis

The following are GitHub repositories that contain source codes of scripts and programs used during the work of this thesis.

A.1 GitHub Repository for Development and Analysis

The source codes for data acquisition, data analysis, and development of CNN model can be accessed in the following GitHub repository. Some of the main scripts and their purposes are explained in Table A.1. All of them are implemented in Python.

```
https://github.com/hendrahc/Quantifying-and-Predicting-Urban-Attractiveness-via-Street-View-Data
```

Table A.1: Scripts for Development and Analysis in This Thesis

No	Script file	Description
1	<code>crawl_image.py</code>	Crawling Street-View images based on given coordinate boundaries
2	<code>analysis.py</code>	Statistical analysis of the assessed dataset, such as label aggregation, variance analysis, and correlation analysis
3	<code>spatial_analysis.py</code>	Computing spatial auto-correlation (Moran's I)
4	<code>image_classifier.py</code>	Designing, training, and evaluating the CNN model, including image processing
5	<code>dataset_expansion.py</code>	Extracting and labelling additional Street-View images for dataset expansion, either heading-based or location-based
6	<code>scene_analysis.py</code>	Clustering scenes of Street-View images

A.2 GitHub Repository for Crowd-sourcing Interface Website

The source codes for the urban attractiveness crowd-sourcing tool used in this thesis can be accessed in the following GitHub repository. For more details on the system, please refer to Appendix C.

<https://github.com/hendrahc/Urban-Attractiveness-Survey-Website>

Appendix B

Google Street View API

Figure B.1 shows an example of HTTP request for Google Street View API with size, coordinate locations, and heading values as the parameters, followed by the returned Street-View image.

```
https://maps.googleapis.com/maps/api/streetview?  
size=600x400&location=52.4199204318,4.8840025476&heading=258
```



Figure B.1: Example of HTTP request in Google Street View API and the returned image

Appendix C

Crowd-sourcing Website

This appendix shows the overview of the developed website to be used as crowd-sourcing interface. The website is modified from [22] and developed with Ruby on Rails ¹.

C.1 Data Model

The data model diagram is shown in Figure C.1. As can be seen, there are six main tables. In campaigns, a task set contains several locations, and each location consists of 4 images. Some of the images are golden images with additional information of the golden questions and answers. All of these data have to be configured before the crowd-sourcing. During crowd-sourcing, after filling the profile a user is assigned with a task set. Next, locations and images linked to this task set will be judged by the user. The judgments are stored into scores table.

C.2 Procedure

Following is the step by step of developed web-pages presented to the users (crowd-sourcing participants).

1. As the opening, a welcome page is displayed that states that the users will be asked to evaluate the attractiveness, and that the test is for non-profit research purposes. In pilot crowd-sourcing, each user should input the user information, such as name, email, gender, age, and nationality.
2. Next page shows the instruction of the task part 1. In part 1, the user should evaluate the attractiveness of each one image. To prevent users from skipping the instructions, the web-page is set so that users can only move on to the next page after several seconds.
3. Users are then presented with a training image. The training has the purpose of letting users practising with the scoring interface. Afterwards, users could begin rating the test images.

¹The original source code is from <https://github.com/ernestacias/qualitytest> (accessed 2017-01-24)

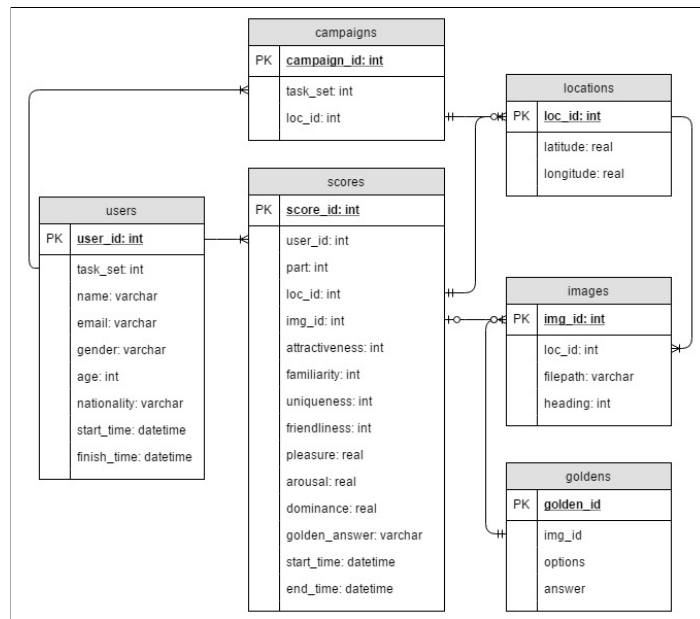


Figure C.1: Data Model Diagram of The Developed Crowd-sourcing Website

4. Next, the judgement of part 1 is started. The images to be judged are given to the user. One image per page. After judgments in a page is submitted, it cannot be changed.
5. During part 1, sometimes golden question is asked
6. After the Part 1 task is done, the part 2 will begin. As previously, an instruction page and trial phase will be provided. For part 2, instead of 1 image, 4 images that describe the surrounding views are given to be judged.
7. After all of the locations have been judged, the task is ended.

Appendix D

Test The Qualification of Applying Exploratory Factor Analysis

This appendix shows the statistical tests based on [29] to confirm that exploratory factor analysis is applicable to the generated dataset.

1. Correlation matrix

Variables that frequently have low correlation coefficient ($r < 0.30$) should be removed as they indicate a lack of patterned relationship. Table 3.2 shows that familiarity has low correlations to the other attributes, so this variable should be omitted for next analysis. The remaining variables to be used in factor analysis are uniqueness, friendliness, pleasure, arousal, and dominance which have correlation above 0.30 at least twice.

2. Bartlett's Test of Sphericity and KMO

Table D.1 shows the result of Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin measure (KMO) of Sampling Adequacy. If the significant level of p is less than 0.05, then the data are confirmed to have patterned relationship. Its value in the data is < 0.001 which is certainly less than the threshold. KMO should be above 0.5 to validate that it is suitable to perform factor analysis to the data, which is also complied by the data with $KMO = 0.720$. Thus, the data pass this test and next requirements can be checked.

3. Anti-Correlation Matrix

The diagonal element of the Anti-Correlation matrix should be above 0.50. If not, distinct and reliable factors cannot be produced. Table D.2 shows the matrix, and the diagonals (printed in bold) are above the threshold.

4. Determining Number of Factors

Table D.1: Result of Bartlett's Test and KMO on The Aggregated Data

Bartlett's Test of Sphericity	Approx. Chi-Square	1151.778
	df	10
	Sig.	<0.001
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.720

Table D.2: Anti-Image Correlation Matrix of The Aggregated Data

	uniq	frie	plea	arou	domi
uniq	0.769	0.062	-0.273	-0.327	-0.102
frie	0.062	0.662	-0.533	0.104	-0.125
plea	-0.273	-0.533	0.677	-0.313	-0.241
arou	-0.327	0.104	-0.313	0.728	0.020
domi	-0.102	-0.125	-0.241	0.020	0.851

Table D.3: Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.593	51.855	51.855	2.100	41.995	41.995
2	0.957	19.141	70.996			
3	0.664	13.288	84.285			
4	0.494	9.880	94.165			
5	0.292	5.835	100.000			

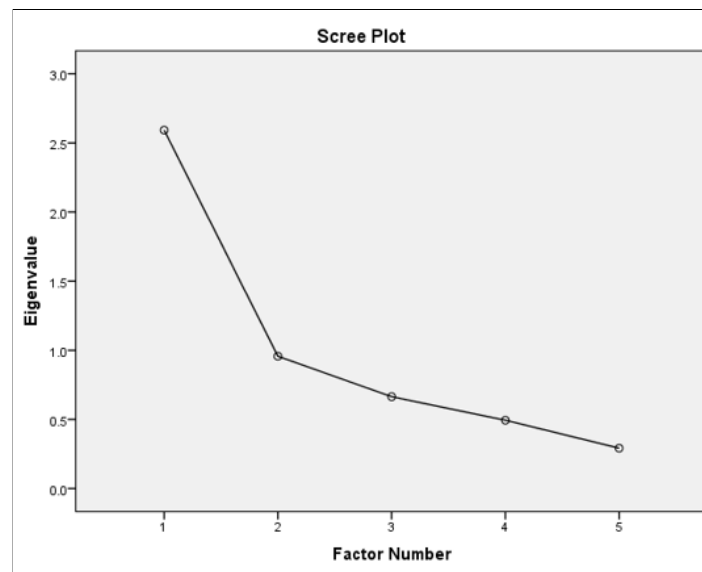


Figure D.1: Scree Plot of Eigenvalue for each Factor Loading

To determine the number of factors, there are two suggested approaches. The first one is based on the number of available rows in Extraction Sums of Squared Loadings in Table D.3. The table shows that there is only 1 row in the rightmost columns. The other way is by checking the scree plot (see Figure D.1) and looking at the factor number with eigenvalue below 1. It can be seen that the minimum factor number meeting this criteria is only 1. Thus, both approaches indicate that there is only 1 factor.

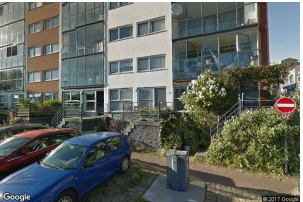
Appendix E

Crowd-sourcing Data

Figure E.1 shows the 5 golden Street-View images used in the internal crowd-sourcing.



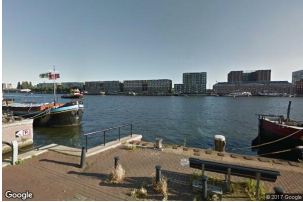
- img_id = 10001
- Canal/river
- Red car**
- Building**



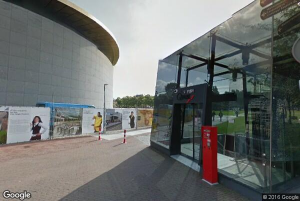
- img_id = 10002
- Canal/river
- Yellow car
- Building**



- img_id = 10003
- Sky**
- Blue car
- Canal/river**



- img_id = 10004
- Sky**
- Building**
- Blue car



- img_id = 10005
- Canal/river
- Building**
- Redcar

Figure E.1: Golden images and object annotation answer option list used in the crowd-sourcing. Options printed in bold indicate that the object appears in the image

Appendix F

Image Pre-processing

Figure F.1 shows an example of Street-View image, and Figure F.2 shows the image after re-sized or cropped.



Figure F.1: Original Street-View Image to be Pre-Processed

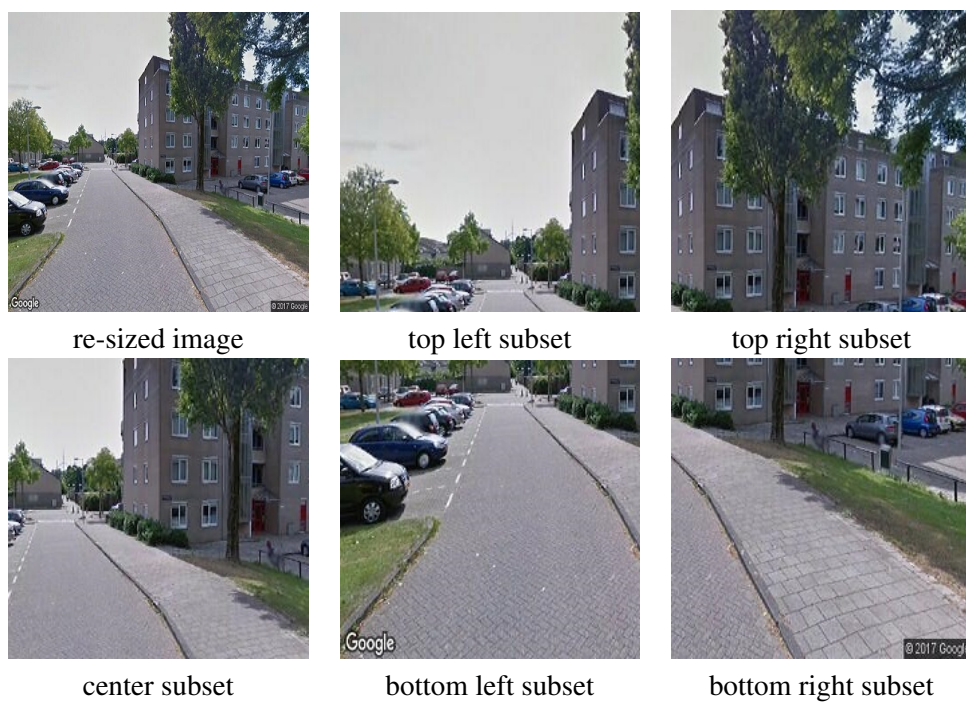


Figure F.2: Example of Re-sizing and Cropping Result of A Sample Street-View Image in Figure F.1