LEARNING URBAN PERCEPTION OF CITIES VIA STREET VIEW IMAGES

LIM XIN QI

SESSION 2019/2020

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY

FEBRUARY 2020

LEARNING URBAN PERCEPTION OF CITIES VIA STREET VIEW IMAGES

BY

LIM XIN QI

SESSION 2019/2020

THIS PROJECT REPORT IS PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT

FOR

BACHELOR OF COMPUTER SCIENCE

B.C.S. (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY

FEBRUARY 2020

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

_____

Lim Xin Qi

Faculty of Computing and Informatics

Multimedia University

Date: 23:03:2020

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my project supervisor, Dr. John See for his wise guidance and advice throughout the semester.

I would also like to thank my family and friends for offering me emotional support and my fellow project takers for sharing their advice and knowledge in the domain.

To everyone who has supported me in my life

ABSTRACT

For many years the urban planners, sociologists and policymakers have been trying to quantify the human perception of a city. However, the effort was tremendously tedious since the studies were mostly carried out manually. As computer vision technology and big data technology advance, there is a rise on making use of the technology to solve the issue. As a result, many researchers in the industry tried to solve the problem by building different machine learning models to predict the human perception for a certain city. In this paper, a deep learning model is built to predict the perceptual scores of 6 different perceptual attributes, namely: beautiful, boring, depressing, lively, safety and wealthy of neighbourhoods in Kuala Lumpur, Malaysia, whereby, a multi-labeling task is described. The model is trained on Google Street View (GSV) images as they contain a lot of visual information on city streetscapes. As a result of the predictions, the neighbourhoods in Kuala Lumpur achieved high scores for both beautiful perception and lively perception but relatively low safety perceptual score. Using the results obtained, an interactive map visualisation is presented on a web-based platform. The visualisation includes 6 choropleth maps to visualise the perceptual score for each perceptual attribute as well as a predominance map to visualise the predominant attribute for each neighbourhood.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Preface

Every city in the world is shaped by different elements which give the city its character and uniqueness. However, what is defined as the "character" and "uniqueness" of a city? As a matter of fact, humans are the one who bestow these descriptions upon the cities based on their perception of these cities. The perception of humans, particularly of a city, commonly coined "Urban Perception" is influenced by a large amount of factors. The factors mostly lie in the behaviour and appearance of the city in which they speak to humans in different human senses. A city's appearance and behaviour affect humans in various psychological ways, including well-being, behaviour as well as their sense of belonging and security. As humans perceive the city in a certain way, the character of the city would eventually be shaped according to the perception. Hence, understanding the urban perception of a city enables a more sustainable development of the city in various aspects such as property development, designing a safer city, creating a more vibrant area etc.

1.2 Motivation

For many years, the urban planners, sociologists and policymakers have been carried out researches on understanding urban perception of cities. However, most of the studies to quantify the urban perception have been carried out manually such as using surveys and questionnaires. Thus, studies have to be conducted from time to time since human perception varies to a certain extent over time. Manual social studies on urban perception are not only time consuming but also limiting in the sense of reachability and having low throughput. Luckily, in the past decade, we see a rise on the computed studies on urban perception. Moreover, nowadays with the breakthrough of big data technology on Computer Vision, we can carry out the studies in a more efficient and data-driven way using visually perceived information of city streetscapes since city streetscapes plays a big part in affecting the urban perception.

As an example, In 2013, Salesses, Schechtner, and Hidalgo from MIT Media Lab tried a novel approach in the urban perception studies by making use of Google Street View (GSV) images to quantify urban perception on Safety, Class and Uniqueness in 4 cities. In 2016, Dubey et al. expanded the project and crowdsourced a new GSV dataset which contains 110,988 images from 56 cities to quantify 6 perceptual attributes namely: Beautiful, Boring, Depressing, Lively, Safety and Wealthy. Dubey et al. (2016) mainly inspired this research since they gathered a huge dataset which is useful for carrying out the urban perception studies in a more data-driven and efficient way.

For the urban perception studies to be useful, there is a need to visualise the findings to reveal urban perception patterns so that investigation on the cause behind the patterns can be done by the expertise. Having a visualisation on the findings of the studies makes it easier for the authorities to visualise where and what to look into for the development of the city. Unfortunately, there is not many visualisations on urban perception which are available right now.

1.3 Research Objectives

• To identify and train a suitable deep learning model for a multilabelled classification task to predict the urban perception of a location given the GSV image of that certain location.

• To create an interactive neighbourhood-level visualisation on the predicted perceptual attributes.

1.4 Research Scope

In the first part, the research focuses on identifying and training a deep learning model for a multilabelled classification task for 6 perceptual attributes: Beautiful, Boring, Depressing, Lively, Safety and Wealthy. The deep learning model will be trained on GSV images to predict the labels of a given GSV image. The visualisation on the perceptual attributes will be done by averaging predicted scores of perceptual attributes in given coordinate-based locations in certain neighbourhoods in Kuala Lumpur, Malaysia. The covered neighbourhoods are visualised in Figure 5.2a.

CHAPTER 2

LITERATURE REVIEW

The literature review will be divided into two main parts: Urban Perception of Cities and Visualisation of Perceptual Attributes.

2.1 Urban Perception of Cities

In the first sub-section of the literature review, earlier work on urban perception will be discussed in Section 2.1.1 and computed and data-driven studies on urban perception will be discussed in Section 2.1.2.

2.1.1 Earlier Work on Urban Perception

A classic literature on urban studies, Lynch (1960) studied about human visual perception of cities in three American cities by getting research participants to draw mental images of the cities. The focus was on “legibility" of a city in which legibility is defined as how easily a particular city view can be perceived and categorised into a pattern. Lynch emphasized on the importance of visual sense of a city and induced that the mental images contain five elements: paths, edges, districts, nodes and landmarks. This study inspired the usage of GSV as it covers almost all the elements.

Nasar (1990) conducted a survey on the likability of areas in two American cities. The respondents were divided into two categories: the residents and the visitors. The respondents were asked to verbally describe the areas based on the likability. Nasar then came up with an evaluation map which describes the area based on likability. This approach is similar to extracting semantic information from an image for further processing. From the survey, it is found out that residents and visitors had different preferences in the areas. It is also suggested in the paper that maps and photograph could be utilised in future works to help the respondents to identify the areas. The shortcoming of the designed survey can be fixed by using a large dataset of GSV images since the different preferences would converge in a large set of data. Nasar also conducted various studies using different methodologies on urban perception in (Nasar, 1982), (Nasar, 1988) and (Nasar & Jones, 1997).

Along the years, researchers started studying urban perception via digital data. Tucker, Ostwald, and Chalup (2004) proposed a method to analyse streetscape using image segmentation and Hough Transform algorithm (Hough, 1962) which detects less obvious boundaries in images. The algorithm requires the user to manually specify a threshold hence it might need to be implemented in some higher level algorithms for it to be able to process bigger sets of data. In Ratti, Frenchman, Pulselli, and Williams's research (2006), they used cell phone usage data to study the intensity of mobile usage at a certain area. The authors suggested to make use of the data to induce the characteristic of an area based on the high intensity of usage on different hours of the

day. However, the induction might not be accurate due to the lack of an actual visual representation of an area. In this case, visual attribute of an area is crucial in telling the urban characteristics.

In an effort to gather a large dataset (Place Pulse 1.0) for urban perception studies, Salesses et al. (2013) collected Google Street View (GSV) images and self-captured street view images in 4 different cities. The collected street views were then rated based on collected human perception on preferred city for a particular perceptual attribute in a pairwise comparison. Computational urban perception of an image was then computed in win and loss ratio in pairwise comparison. This research is one of the initial approaches on collecting a large scale of street view data for urban perception studies. However, the street views collected only covers 4 Western cities, hence Place Pulse 2.0 which covers several Asian cities is used for this FYP.

2.1.2 Computed and Data-Driven Studies on Urban Perception

2.1.2 (a) Place Pulse 2.0

Place Pulse 2.0 (Dubey et al., 2016) is a dataset consisting of GSV images of 56 cities from all the continents except for Antartica. In the research, same methodology as in Salesses et al.'s work, pairwise comparison was carried out to collect the human perception of a certain perceptual attribute in a city street view. The paper discussed about the notable ranking methods - Streetscore (Naik, Philipoom, Raskar, & Hidalgo, 2014)

but the model was trained based on Place Pulse 1.0 (Salesses et al., 2013) which only consists of 4 cities. Thus the authors decided to train their own Streetscore model using Microsoft Trueskill (Herbrich, Minka, & Graepel, 2007). For the prediction of urban perception of street view images, Dubey et al. deployed a siamese-like CNN model which accepts an image pair as input and predicts winner in the pairwise comparison. Figure 2.1 shows the architecture of the proposed CNN model.

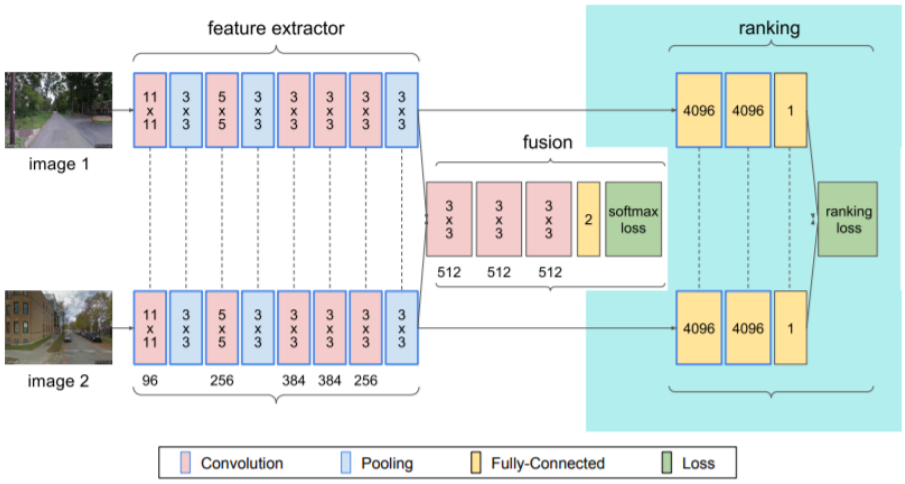


Figure 2.1: Architecture of Siamese-based CNN, reproduced from Dubey et al. (2016)

The model managed to achieve an accuracy of 73.5% for the prediction of urban perception. The limitation of this paper is on the rating methodology since it does not take semantic information of an image into account. Moreover, the GSV images collected by authors only focus on the default GSV angle which only displays the street. Hence information on neighbourhood in a close proximity might be missed out. The research can also be improved by measuring other data types such as audio information of a city, geo-tagged mobile data, social media information etc.

Place Pulse 2.0 has been referenced in numerous papers as the dataset to train machine learning models. Zhang et al. (2018) performed spatial mapping of perceptual attributes to two Chinese cities, namely Beijing and Shanghai. Each perceptual attribute was trained individually on the GSV images in such a way that the perceptual attributes were binarised in +1 as positive and -1 as negative in each image. A SVM classifier was then used to predict the perceptual distribution in the two cities. The authors also studied the relationship between a certain perceptual attribute and the visual elements in the certain image using multivariate regression.

Using the same dataset, Xu et al. (Xu et al., 2019) proposed a double column CNN architecture which takes both semantic features and generic features in the Google Street View (GSV) images as inputs. The authors evaluated the correlation between objects presented in an image and the perceptual attribute. Further studies can look into how the objects affect the perceptual attributes. Aside from these papers, (Ilic, Sawada, & Zarzelli, 2019), (Partridge, 2018) and (Min, Mei, Liu, Wang, & Jiang, 2019) also experimented with the Place Pulse 2.0 Dataset.

2.1.2 (b) Other Street View Imagery

There are also studies that performed data collection via mobile crowdsourcing with the aim of expanding the coverage of inaccessible streets by GSV (Santani et al., 2018). In this study, human urban perception was labelled automatically via low-level features and deep learning features which were extracted using GoogLeNet (Szegedy et

al., 2014). The researchers then built a Random Forest (Breiman, 2001) model for the prediction of urban perception. In this paper, the authors proposed a auto-inferred human perception of the streetscape using pre-trained CNN models. However, the maximum of R^2 value between the inferred human perception and predicted perception is only 0.49. Moreover, the reliability of ground truth which is the auto-inferred "human perception" is also questionable since computing it did not involve actual human perception.

Liu, Chen, Zhu, Xu, and Lin (2017) worked on predicting safety scores using multi-instance regression on street views along with crime records as the safety score for each place. Then, the safety scores which were derived from primary instance of each image were predicted using Expectation-Maximization (EM). In evaluating the result, the R^2 value reached 0.84 between predicted score and true score.

2.1.2 (c) Satellite Imagery

Urban perception studies using satellite imagery has also been looked into by the researchers in the domain. Wang et al. (2018) and He et al. (2018) both investigated commercial activeness using satellite images. In both the studies, patches of image regions were used to train the model. Wang et al. (2018) made use of Support Vector Regression (SVR) which accepts features extracted using Bag-of-Features (BOG) as input to predict the commercial activeness. He et al. (2018) implemented a CNN model to extract the feature vectors of the image patches and predicted the commer-

cial activeness using regression. The results were validated using the amount of online reviews on an area to denote the popularity of an area and thus representing the commercial activeness of the area. The accuracy of Wang et al. (2018) and He et al. (2018) achieved an accuracy of 62.66% and 74.3% respectively.

Piaggesi et al. (2019) looked into city poverty prediction using satellite imagery. The ground truth used was the household income obtained from surveys. Features of each image were extracted using CNN and the prediction was done using regression, which is similar to all the score predictions done in some papers discussed in this section.

Based on the result of these researches, satellite imagery is also a good alternative to be considered to train the model since it covers some secluded areas which are not reachable by GSV.

2.1.2 (d) Multimodal Approach

As a foreword, multimodality in this section is defined as the combination between different types of data or methodologies.

In a neuroscience research, it is proven that human perception is highly influenced by multisensory interaction (Watkins, Shams, Josephs, & Rees, 2007). Thus, multimodal approach which involves other forms of sensory data such as audio and smell data is also worth investigating. As being stressed by Lynch (1960), visual sense plays a

crucial part for the legibility of a city. Hence, for a more thorough investigation, visual data cannot be omitted while training a predictive model. In other words, on top of visual data, different forms of data can also be added as additional measurements.

Verma, Jana, and Ramamritham (2019) collected a time series of visual and audio data. The research focused on classifying the visual and audio data based on their semantic information for further urban perception studies. In the research, CNN was used for objects detection and semantic segmentation. Long Short-Term Memory (LSTM) network as the RNN was used for audio classification.

As another form of multimodal approach which involves visual data and mobile data, De Nadai et al. (2016) investigated the relationship between safe-looking and liveliness in a neighbourhood by using GSV for safe-looking prediction and mobile phone data as a measurement for liveliness. CNN was used to predict the safety score and the population density which denotes the liveliness was derived using the mobile phone data.

2.1.2 (e) Summary of Computed and Data-Driven Urban Perception Studies on Street View Imagery

The methodologies for computed and data-driven urban perception studies on street view images are summarised in Table 2.1. The table includes summarised researches which implemented multimodal approach but the emphasis is given on the the urban

perception prediction on the street view imagery in the research papers.

2.2 Visualisation of Perceptual Attributes Score in Cities

As proposed in the second objective, a neighbourhood-level visualisation will be constructed. In this sub-section, ways of visualisation are looked into. In scrutinising the research papers that did urban perception in the visual aspect, it is found that most of the findings are visualised using the following three main methods: (i) Geospatial representation, (ii) Graph and (iii) Correlation matrix.

2.2.1 Geospatial representation

In the Place Pulse 2.0 paper, Dubey et al. (2016) visualised the urban perception using geospatial representation. In Figure 2.2, we can see the safety scores represented geospatially in discrete values. On first look, information is hard to be gathered due to the discrete safety score of all categories being scattered evenly across the maps.

Ordonez and Berg (2014) visualised the predicted safety score using geospatial representation as well. The authors visualised three different results so they can be compared with each other. Here, the comparisons between predicted scores and ground truth scores can be seen very clearly since all of them were being visualised using the same technique. Another example of clear visualisation is that the safety scores are being presented in gradient form which is more intuitive to the human eyes. Figure 2.3 shows the visualisation result by Ordonez and Berg.

Table 2.1: Computed and Data-Driven Urban Perception Studies on Street View Imagery

Authors	Task	Methodology for Urban Perception Studies
Salesses et al. (2013) (Place Pulse 1.0)	Quantify urban perception on Safety, Class and Uniqueness in 4 cities	Win and loss ratio in pairwise comparison
Dubey et al. (2016) (Place Pulse 2.0)	Predict winner in a pairwise comparison given a perceptual attribute	• Crowdsourced large dataset • Siamese-like CNN model
De Nadai et al. (2016)	Investigate the relationship between social activeness and perception of safety	CNN for safety score prediction and measured it with mobile phone activity data as the social activeness metrics
Liu et al. (2017)	Develop a deep multi-instance regression method to predict weakly supervised GSV images.	Deep hierachical multi-instance regression which made use of Expectation-Maximization (EM)
Santani et al. (2018)	Automatically infer urban perception on outdoor scenes	• Mobile crowdsourced dataset • Random Forest for urban perception prediction
Wang et al. (2018)	Predict commercial activeness from satellite and street view images	• Bag-of-Features (BOG) for feature • Extraction and SVR for commercial activeness prediction
Zhang et al. (2018)	• Binary classification for each perceptual attributes (safe, lively, boring, wealthy, depressing, and beautiful) • Investigate the influence of visual elements on urban perception	• SVM classifier to classify binarised perceptual attributes on each image • Multivariate regression analysis
Xu et al. (2019)	Predict urban perceptual scores of images (Ranking Task)	Double column CNN architecture trained on semantic features and generic features
Verma et al. (2019)	Propose methodology for data collection on visual and audio data based on semantic information	CNN for objects detection and semantic segmentation

Figure 2.2: Safety scores for 6 cities, reproduced from Dubey et al. (2016)

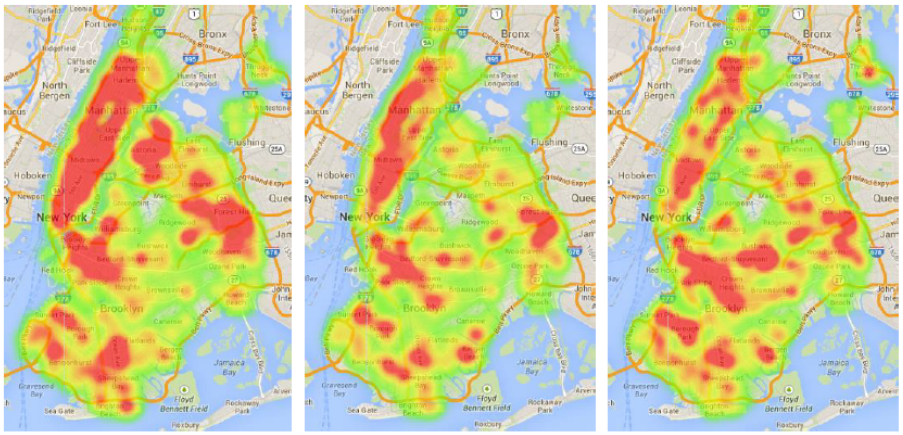


Figure 2.3: Safety scores for ground truth values, predicted values from a model trained from the same city as ground truth and predicted values from a model trained from a different city, reproduced from Ordonez and Berg (2014)

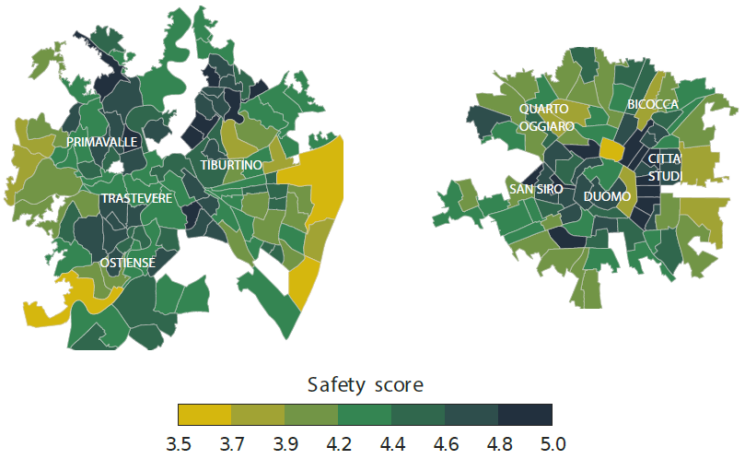


Figure 2.4: Safety score for each district for Rome and Milan, reproduced from De Nadai et al. (2016)

De Nadai et al. (2016) labelled each district with different discrete values which gives an overall representation for each district. This approach is clear in delivering the information but since each district is labelled only one single discrete value, some information might be lost through the representation. Figure 2.4 shows the visualisation of result by De Nadai et al.

2.2.2 Graph

In this domain, graphs are mainly used to compare the actual and predicted values to show the accuracy of the model.

Santani et al. (2018) plotted a comparative histogram to compare the actual and predicted score. While the difference between actual and predicted scores can be seen for each range, we can hardly measure the correlation of the two variables. The comparative histogram is shown in Figure 2.5.

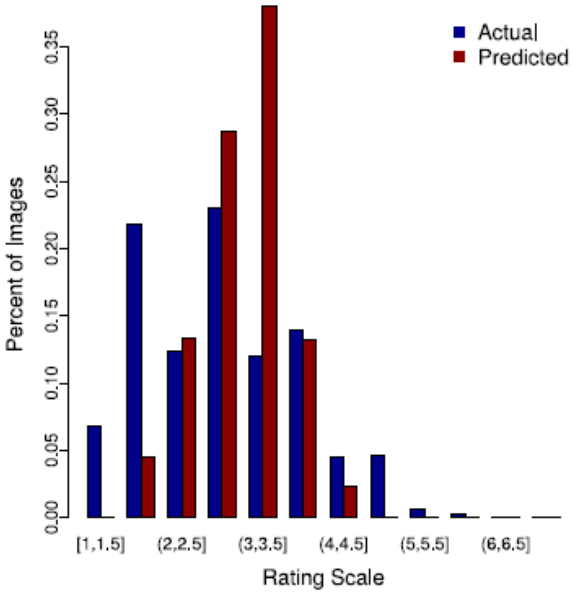


Figure 2.5: Comparative histogram of actual and predicted values for the degree of dangerous in a city, reproduced from Santani et al. (2018)

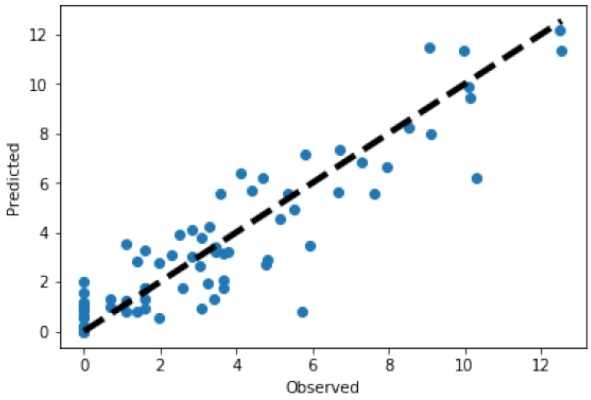


Figure 2.6: Scatter plot of predicted values against observed (actual) values of commercial activeness of a city, reproduced from He et al. (2018)

He et al. (2018) plotted a scatter plot as shown in Figure 2.6 to investigate the correlation between the actual and predicted value.

It is worth noting that while scatter plot is good at showing correlation between the actual and predicted values, it is hard to pinpoint the exact values to show the difference between the actual and predicted values using scatter plot. Hence, both comparative histogram and scatter plots can be used to analyse the accuracy of the model.

2.2.3 Correlation matrix

Correlation matrix is another way to visualise the correlation between variables. Zhang et al. (2018) and Santani et al. (2018) used correlation matrix to visualise the correlation of different urban perceptions. In an attempt to ensure that semantic information in images is closely related to urban perceptions, Xu et al. (2019) used correlation matrix to visualise correlation between semantic information in images and urban perceptions.

By comparing the presentation of the correlation matrix in all three papers, we can see that Santani et al. (2018)'s visualisation in Figure 2.7 shows the clearest patterns due to arranged rows and columns while in Xu et al. (2019)'s correlation matrix in Figure 2.8, it is relatively harder to find the patterns. However, this can also be caused by how a single semantic element can be correlated to different urban perceptions.

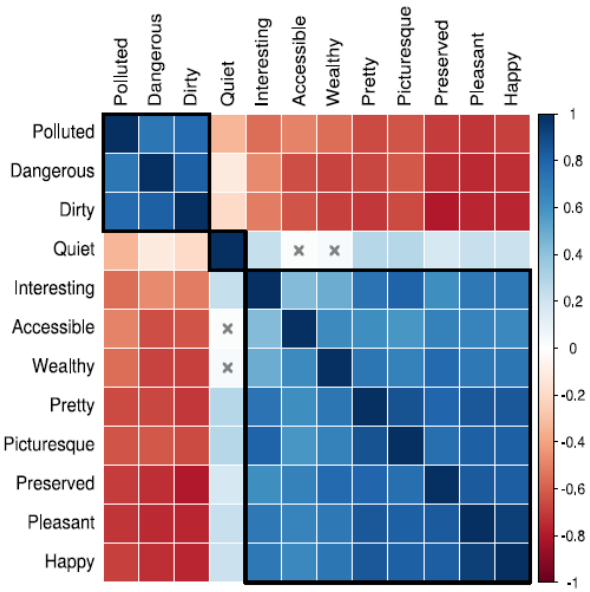


Figure 2.7: Correlation matrix between urban perceptions, reproduced from Santani et al. (2018)

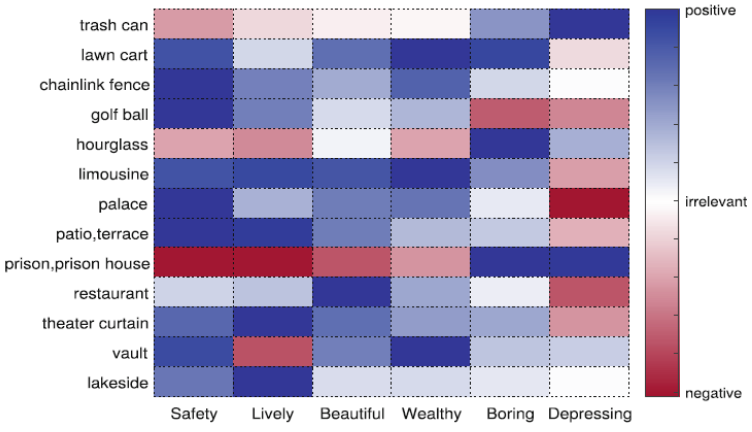


Figure 2.8: Correlation matrix between semantic information and urban perceptions, reproduced from Xu et al. (2019)

CHAPTER 3

THEORETICAL FRAMEWORK

3.1 Making Use of Place Pulse 2.0 Dataset

Place Pulse 2.0 (Dubey et al., 2016) is a dataset gathered by the MIT Media Lab. The dataset consists of the result of the pairwise comparison between 2 GSV images on different perceptual attributes. The motivation behind using the Place Pulse 2.0 dataset is that it is a large set of data and it consists of data for 7 Asian cities. Moreover, the study that built the dataset is based on Google Street View (GSV) in which, will be discussed in Section 3.2, it is suitable in this research.

3.2 Street View Images as Dataset

Human visual perception of cities, as described by Lynch (1960), are built on 5 main elements, namely: paths, edges, districts, nodes and landmarks. Built on top of this foundation, urban perception is largely influenced by these 5 elements. Hence, to study about urban perception in the visual way, it is important that our dataset consists images that contain these 5 elements.

3.2.1 Google Street View Images

Google Street View (GSV) is a service provided by Google for its users to visualise streets on Google Maps. It consists of street-level images captured by Google Street View cars from time to time. Since the images are street-level images, they could capture the city streetscapes very well.

Figure 3.1 shows a few examples of GSV images. As shown in the figure, paths (channels where observers move along (Lynch, 1960), e.g, pathways), edges (boundaries that set apart continuity (Lynch, 1960), e.g., buildings), nodes (strategic meeting points (Lynch, 1960), e.g., junctions) and landmarks (public reference points relevant to the city (Lynch, 1960), e.g., National Monument) are all present in the images.

Figure 3.1: Examples of Google Street View Image

3.3 Deep Learning Using CNN

In this section, Convolutional Neural network (CNN) will be discussed. CNN is good at recognising spatial patterns. Thus, it outperforms many models in image classification. Similar to the regular Artificial Neural Network, it has hidden layers where a list

of weighted inputs are used to produce a set of outputs via an activation function. In addition to that, CNN consists of Convolutional Layers and Pooling Layers which are responsible for the feature extraction part of CNN.

3.3.1 Architecture of CNN

In general, CNN is made up of 3 mian layers which consist of Convolutional Layer, Pooling Layer and the Fully-connected Layer.

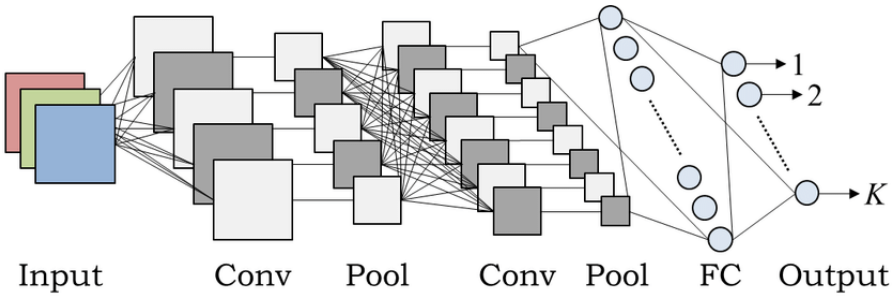


Figure 3.2: An example of CNN architecture, reproduced from Hidaka and Kurita (2017)

3.3.1 (a) Convolutional Layer

In the Convolutional layer, a filter will slide through the input image and a dot product will be produced for each pixel (*CS231n Convolutional Neural Networks for Visual Recognition*, 2017). As the filter successfully slides through the whole input image, a convolved output which is also known as the activation map will be produced. The convolutional layer is used to extract the features from the images. Figure 3.3 shows an example of an activation map.

Figure 3.3: An example of an Activation Map, reproduced from Xu et al. (2019)

3.3.1 (b) Pooling Layer

The pooling layer is used to for spatial size reduction to reduce computation and also reduce overfitting (*CS231n Convolutional Neural Networks for Visual Recognition*, 2017). Usually pooling is carried out by Max Pooling or Average Pooling. Max Pooling is found to perform better than Average Pooling (*CS231n Convolutional Neural Networks for Visual Recognition*, 2017).

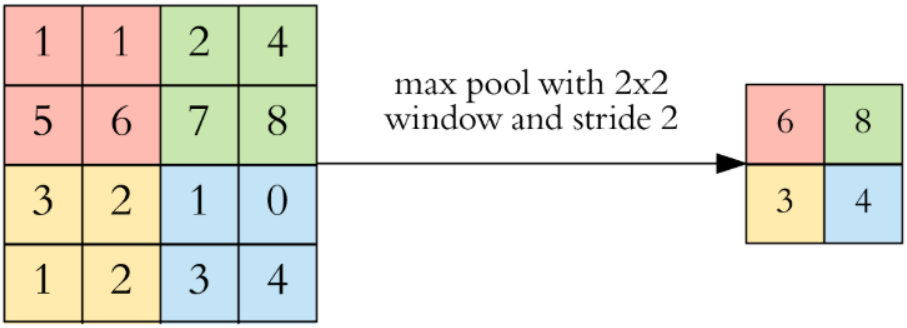


Figure 3.4: An example of Max Pooling, reproduced from Dertat (2017)

3.3.1 (c) Fully-connected Layer

The fully-connected layer is the part where classification is done. This layer holds all the activation information from each dimension in the feature extraction part. Flattening of the multidimensional information will be performed before the information is used for the classification task.

CHAPTER 4

RESEARCH METHODOLOGY

The research methodology is divided into 4 main stages, namely: Data Gathering, Data Pre-Processing, Model Building and Data Visualisation. The task will be subdivided into smaller tasks in each stage.

Figure 4.1 and 4.2 shows the Gannt chart and the tasks that were completed throughout the semester for both FYP 1 and FYP 2.

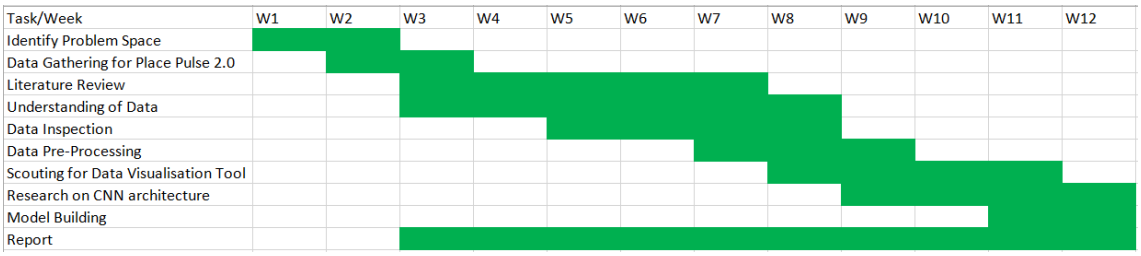


Figure 4.1: Gannt chart for FYP 1

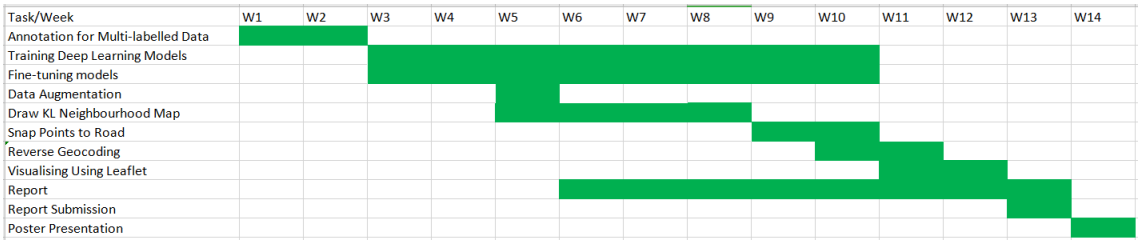


Figure 4.2: Gannt chart for FYP 2

4.1 Overall Research Methodology Process

Figure 4.3 shows the overall process framework of the research methodology. Firstly, a CSV file from Place Pulse 2.0 dataset which contains the result of the pairwise comparison between GSV images on different perceptual attributes was downloaded. The CSV file was then inspected to find out the amount of records of Asian cities contained in the dataset.

After inspecting the data, since the amount of records of Asian cities were found out to be high, data selection was done on the CSV file by selecting the records of Asian cities. Then, the selected dataset was checked to find out if the data are balance. Undersampling was carried out after finding out that the dataset is imbalance. Annotation of perceptual attributes were then done for the selected records after undersampling.

GSV images were then downloaded according to the coordinates of the selected data to be used as the training data. After the GSV images were downloaded and annotated with its perceptual attributes, it was used to train a deep transfer learning model. In the model building process, several parameters were fine-tuned and data augmentation was done to try to improve the training outcome.

After the model had been trained, the GSV images for Kuala Lumpur Neighbourhood were downloaded as the data to be predicted. Map visualisation was used to visualise

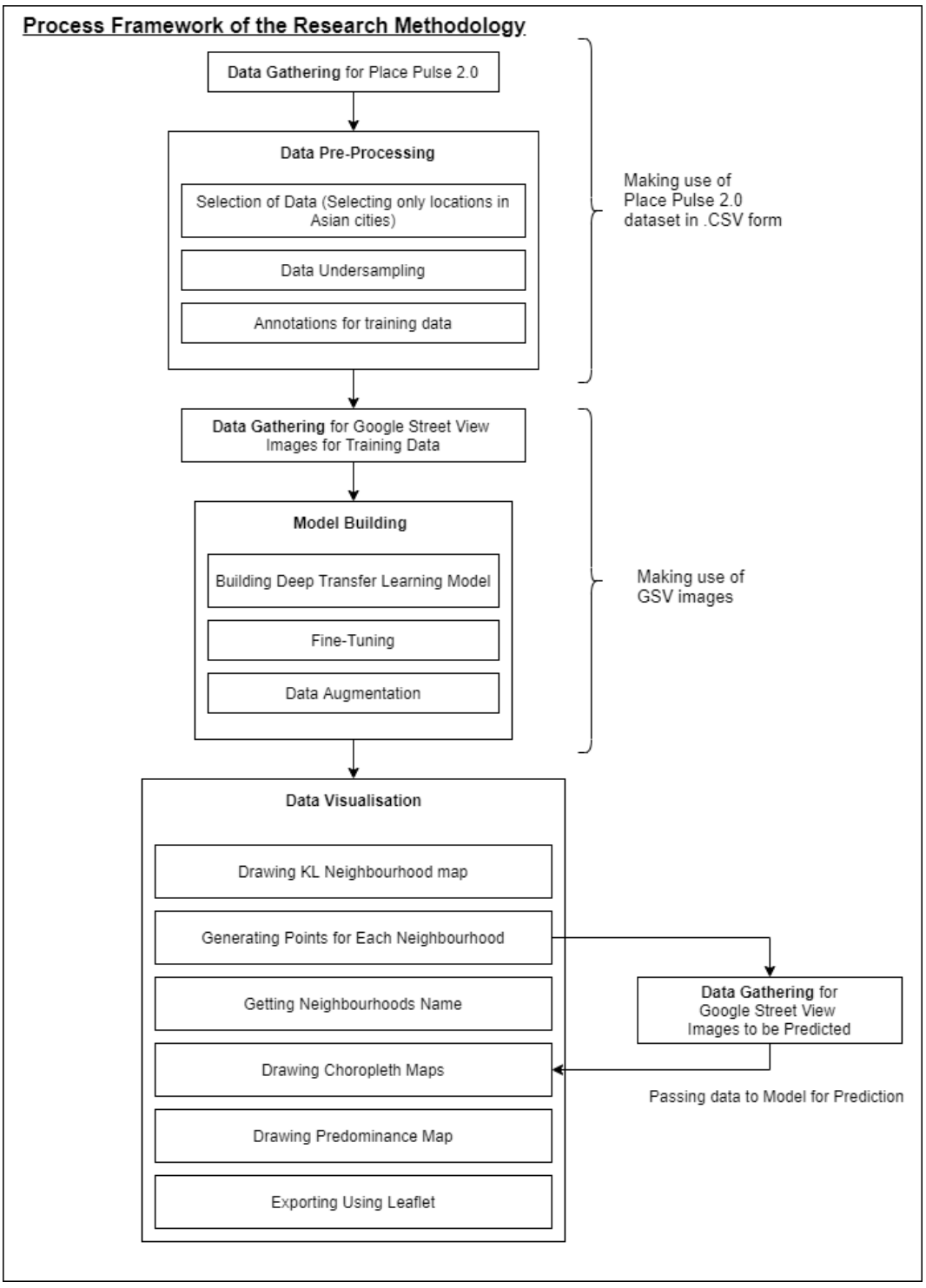


Figure 4.3: Process Framework of the Research Methodology

the perceptual attributes of each neighbourhood. Several steps were involved in the the Data visualisation stage. Data visualisation part will be discussed in details in Section 4.6.

4.2 Data Gathering

Data Gathering was carried out in 3 different steps to collect the (i) Place Pulse 2.0 pairwise comparison result which is available in a CSV file, (ii) GSV images in different Asian cities for training data and (iii) GSV images in KL neighbourhoods to be predicted.

4.2.1 Place Pulse 2.0 Pairwise Comparison Result (CSV)

The Place Pulse 2.0 dataset is the result of the pairwise comparison between 2 GSV images on a given perceptual attribute. The dataset is in .CSV form which consists of the coordinates of the 2 GSV images which were used in the pairwise comparison as well as the winner in the pairwise comparison for the given perceptual attributes (i.e.: the city which was perceived to suit the given perceptual attribute more).

The Place Pulse 2.0 dataset was downloaded from the Place Pulse website which is run by the MIT Media Lab. Data Pre-Processing (Section 4.3) was done on the Place Pulse 2.0 dataset to extract relevant data which are records of Asian cities.

4.2.2 Google Street View Images

Subsequently, GSV images were downloaded using the coordinates provided by the Place Pulse 2.0 dataset. The GSV images for training data were downloaded by querying using the Street View Static API.

GSV images to be predicted were downloaded after the KL neighbourhood map was drawn in the visualisation part (Section 5.2) using Street View Static API as well. The GSV images downloaded for each neighbourhood were used to predict the perceptual attributes for each neighbourhood.

The downloaded image size for the GSV images are 224x224. To preserve the similarity between GSV images in the Place Pulse 2.0 experiment and the GSV images to be downloaded, the heading, pitch and field-of-view (FOV) of GSV were not specified as the authors of Place Pulse 2.0 did not specify them for the experiment.

The GSV images were annotated (will be further discussed in Section 4.3.3) according to their coordinates and if they emerge as a winner in the Place Pulse 2.0 dataset. The shortcoming of querying using the Street View Static API is that there is a limitation on the cost-free query amount which is 28,000 queries per month.

4.3 Data Pre-Processing

The data pre-preprocessing are divided into 3 stages, namely data selection, undersampling and annotations for training data.

4.3.1 Data Selection

Data selection to was carried out so that the training data are more relevant to the GSV images to be predicted in Kuala Lumpur neighbourhoods. From the Place Pulse 2.0 dataset, there are 7 Asian cities that are present in the records, namely: Singapore, Tel Aviv, Hong Kong, Tokyo, Taipei, Bangkok and Kyoto. Since the city streetscapes of Asian cities are more similar to each other than city streetscapes of other continents, the records of the 7 Asian cities were extracted to be the training data.

3 bounding boxes were used to get the records which consist of the coordinates of the GSV images of the mentioned Asian cities. 3 smaller boxes were used instead of one big bounding box to ensure that irrelevant cities would not be included in the rows selected. Figure 4.4 illustrates the bounding boxes drawn to get the coordinates of the GSV images in the 7 different Asian cities.

The following state the cities being enclosed in each box.

- Box 1: Tel Aviv
- Box 2: Bangkok & Singapore
- Box 3: Tokyo, Kyoto, Hong Kong & Taipei

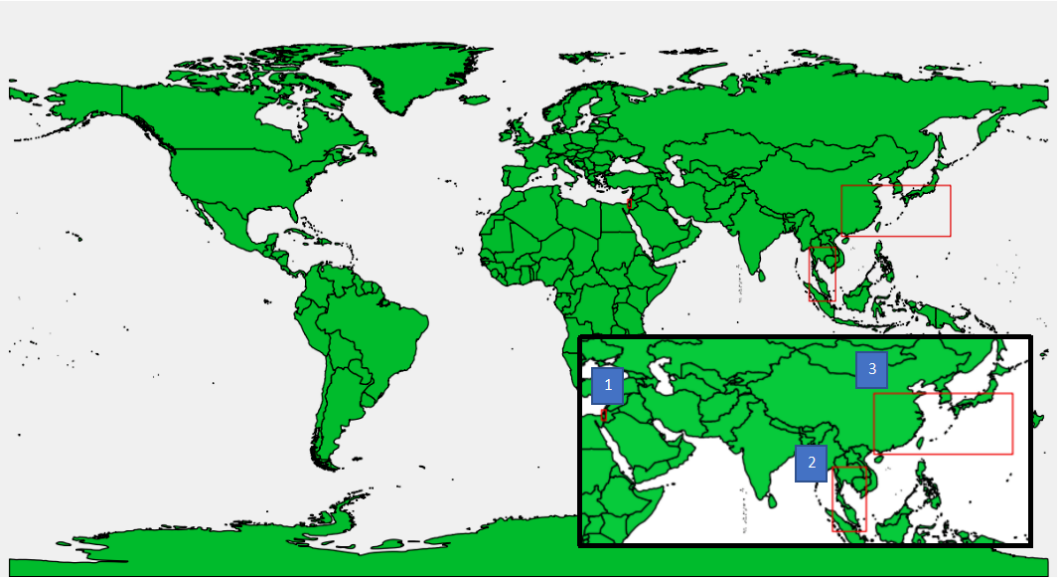


Figure 4.4: Bounding Boxes for Asian Cities

In addition to setting the coordinate boundaries to define the Asian cities, the coordinates of the GSV images have to be the winners for the pairwise comparison to be extracted. In other words, the data have to have coordinate which is (i) in one of the bounding boxes and (ii) a winner in the pairwise comparison in order to be extracted. As a result, the total amount of rows extracted is 108,570.

4.3.2 Undersampling

After extracting the data which focus on Asian cities, statistical analysis was done on the extracted data. As shown in Figure 4.5, the classes are imbalanced since the amount of data in "Lively" class and "Safety" class are exceptionally high.

In the statistical analysis of the extracted data, it was found out that the data are im-

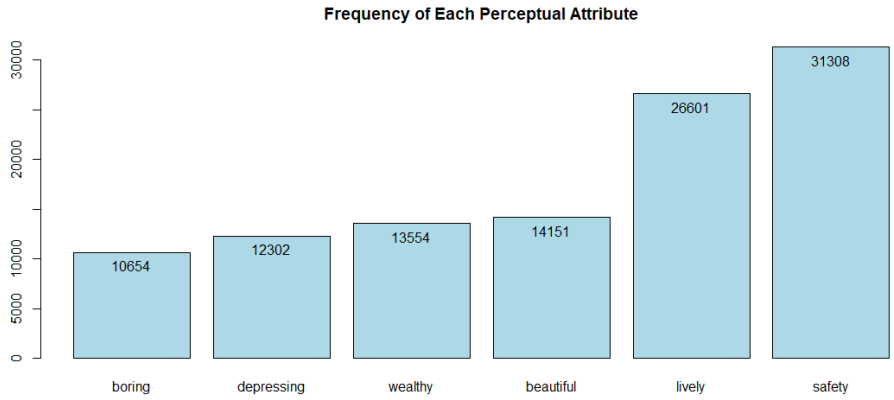


Figure 4.5: Frequency of Each Perceptual Attributes

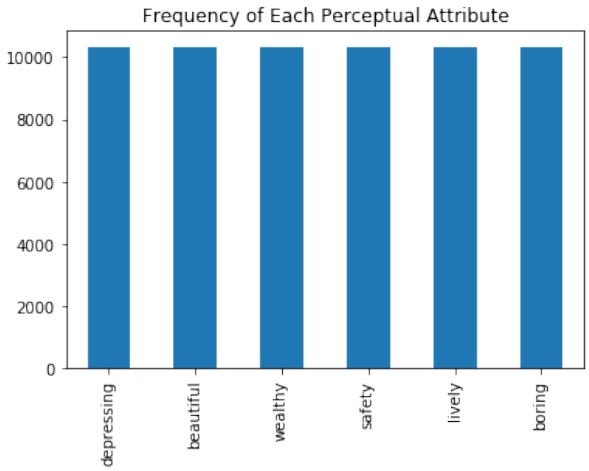


Figure 4.6: Frequency of Each Perceptual Attributes After Undersampling

balanced with the "Lively" class and the "Safety" class having exceptionally higher amount than the other classes. Hence, random undersampling was used to balance the data. Undersampling was performed on each of the classes to achieve balance. After undersampling, each class contains 10,344 records. Figure 4.6 shows the barplot of the frequency of each perceptual attribute after undersampling.

4.3.3 Annotation for Training Data

Attributes annotation was done by multilabelling each unique location according to the winner of the pairwise comparison for the targetted perceptual attributes.

Firstly, from the Place Pulse 2.0 dataset, the winners for the pairwise comparisons for one targetted perceptual attribute was extracted out together with the winning perceptual attribute. A list of winners with their own winning perceptual attribute were then created from the extraction. Then, unique winners which are defined by their own unique location and panorama ID (`panoID`) were extracted from the list of winners.

Each of the `panoID` of the list of unique winnner was then compared with the `panoID` of the list of winners. If the `panoID` from both files match, the winning perceptual attribute from the list of winners were then attached to its own unique location. Figure 4.7 shows an illustration of the whole annotation process.

4.4 Model Building

After the data were gathered and pre-processed, it was ready to be used to train a deep learning model. The task was initially described as a multiclass, single-labelled classification task. However, the training outcome (Section 5.1.1) obtained for this task was not satisfying. Hence another approach was used to tackle the problem statement. The new approach was to do multi-labelled classification on a given GSV image. Since the

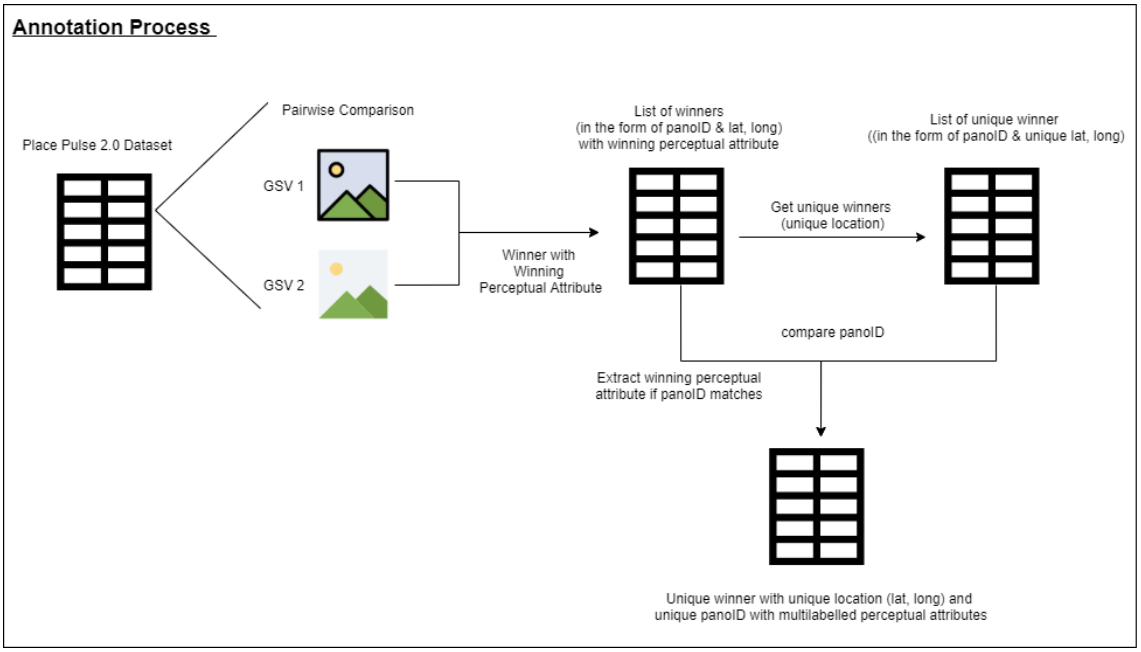


Figure 4.7: Annotation Process

GSV dataset collected was relatively small, transfer learning was opted to be applied. Tensorflow Keras (Chollet et al., 2015) was used to build the model and scikit-learn (Pedregosa et al., 2011) was made use of to log the mAP (will be discusssed in Section 4.5).

4.4.1 Model Architecture

The deep transfer learning model used was a model with VGG 16 ((Simonyan & Zisserman, 2014)) architecture, pretrained on Places 365 weight (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). Places365 dataset was used as the pretrained weight because it consists of images that includes indoor scenes, nature scenes, and urban scenes in which nature scenes and urban scenes are particularly similar to the GSV images to be predicted in the multilabelled classification. After building the feature

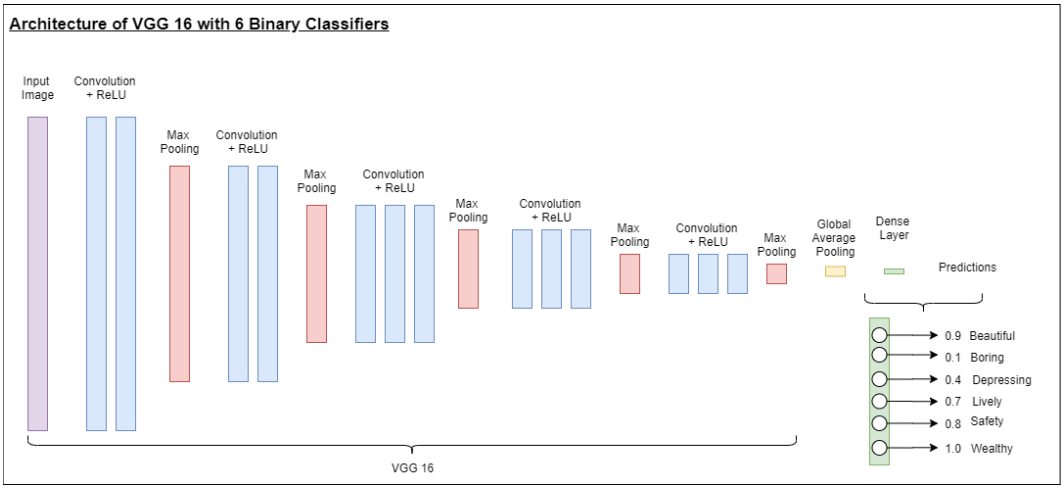


Figure 4.8: Multi-labelled Classifier with 6 Binary Classifiers

extraction part of the deep learning model, 6 binary classifiers were added at the fully connected layer to make predictions on the perceptual attributes. The model was compiled using binary cross-entropy loss since the loss computed for each output is not affected by each other.

Figure 4.8 shows an illustration of the model architecture. At the last layer, we can see that there are 6 binary classifiers which are used to predict 6 different perceptual attributes. The score for each perceptual attribute is not affected by each other.

4.4.2 Fine-Tuning and Data Augmentation

After the whole model was built, the model was fine-tuned on the following parameters: learning rate, batch size and optimiser. Different weight namely ImageNet was also used to try to achieve a better training outcome. Data augmentation was also tried to improve the training outcome. Data augmentation was performed by downloading

the GSV images of another angle from the same location. The angle was set to be 90° horizontally rotated.

4.5 Evaluation Metrics of the Model

The dataset of GSV images were split into 90% training data and 10% test data in which training data were further split into 80% training data and 20% validation data.

Since the task was described as a multi-labelled classification task, mean average precision (mAP) was used as the evaluation metrics instead of accuracy. The reason of using mAP rather than accuracy is because in multi-label classification, the model's performance is measured by getting the right prediction for each class for each sample instead of getting the whole prediction right for each sample.

The mAP was calculated at the end of each training epoch by averaging the average_precision_score metrics provided by scikit-learn (Pedregosa et al., 2011). In the scikit-learn documentation, it is stated that the formula summarises a precision-recall curve by using the weighted mean of the precisions obtained at a defined threshold, in which the weight is defined as the increase in recall from a previous threshold. The threshold is assumed to be calculated by scikit-learn internally.

$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1}) P_n \tag{4.1}$$

Equation 4.1 is the formula for the average_precision_score by scikit-learn in which P_n and R_n are the precision and recalll at the n^{th} threshold.

4.6 Data Visualisation

The visualisation of the predicted perceptual attribute score was presented using a map. One of the objectives of the project is to create a neighbourhood-level visualisation. Kuala Lumpur, Malaysia was chosen as the targeted city since it consists of various kinds of areas with different streetscapes ranging from the very famous Petronas Twin Towers building which intuitively signifies a wealthy area in the human perception to the Kampung Baru area which is a huge contrast to the streetscape of the area where Petronas Twin Towers resides at.

4.6.1 Drawing KL Neighbourhood Map

The whole visualisation is mostly map visualisation, this is because locations are better understood and easier to be visualised using a map. Hence, the first and foremost step would be to draw a KL neighbourhood map for visualisation. The tool to draw KL Neighbourhood map was QGIS (QGIS Development Team, 2009), an open sourced geographical information system software. Instead of programmatically drawing square grids to represent each area of Kuala Lumpur, irregular polygons were drawn to represent each neighbourhood in Kuala Lumpur. The reason for drawing irregular polygons is that it is more intuitive such that the whole neighbourhood can be represented by a certain perceptual score hence it gives a simplified idea on how is the neighbourhood

perceived as.

4.6.2 Generating Points and Getting Neighbourhoods Name

After the KL neighbbourhood map was drawn, random points were generated for each neighbourhood so that GSV images of that particular point can be downloaded based on the coordinate of the point. The random points were then snapped to the nearest road by using the Google Maps "Snap to Roads" API so that GSV images for the random points are available to be downloaded since GSV are mostly available on roads.

After the points had been snapped to the nearest road, the points were used to query for its own neighbourhood name. This was done by doing reverse geocoding of the given address using Google Maps "Geocoding" API. The coordinate of the points were used to find out about the neighbourhood where the points belong to.

4.6.3 Choropleth Map

There are several ways in visualising the data. One very direct way is to geospatially label the predicted perceptual scores for each perceptual attribute for each neighbourhood. Labelling the scores for each neighbourhood enables us to visualise clearly on which neighbourhood has a higher perceptual score. It also quantifies the answer for questions such as: "How safe is this place?" Hence, a choropleth map for each perceptual attribute was drawn to visualise the perceptual scores.

4.6.4 Predominance Map

On top of the choropleth maps, a predominance map was also drawn to visualise the predominant perceptual attribute of the neighbourhoods. A predominance map is useful in showing patterns across different attributes with the same measurement on a map. In this project, since there are 6 perceptual attributes to be looked into, a predominance map was drawn to show the predominant attribute for each neighbourhood.

After the predominant attribute was found out for each neighbourhood, the weight of the predominant attribute among the 6 perceptual attributes was annotated. The weight is calculated as in Equation 4.2. In the equation, p represents the predominant attribute. The weight is then used to denote the strength of predominance of the predominant attribute for each neighbourhood.

$$W_p = \frac{PerceptualScore_p}{\sum_{i=1}^{n} PerceptualScore_i} \tag{4.2}$$

4.6.5 Exporting the Maps

After the maps were drawn, Leaflet (Agafonkin, 2011), a Javascript library for interactive map was used to visualise the maps onto a web platform. Leaflet was chosen as the visualisation tool because it is light-weighted and it provides built-in interactive features with the maps.

CHAPTER 5

IMPLEMENTATION

The implementation part of the project involves two main parts in order to achieve the objectives of the project, namely to: (1) identify and train a suitable deep learning model to predict the labels of a location given the GSV image of that certain location and (2) create a neighbourhood-level visualisation on the predicted perceptual attributes.

5.1 Training a Suitable Deep Learning Model

Building a suitable deep learning model was absolutely important for the classification task. Before the task was described as a multi-labelled task, it was described as a multi-class, single-labelled classification task. However the training outcome was not up to expectation. Thus, the task was then described as a multi-labelled classification task.

5.1.1 Training Details and Outcome for Single-Labelled Classification

A few models which include VGG 16 (Simonyan & Zisserman, 2014), AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) and InceptionNet (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2015) were trained on ImageNet (Deng et al., 2009) to complete the task.

However, the training outcomes were very unsatisfying in which the highest accuracy reached was only 0.2 and the loss fluctuated between 1.7 and 1.8.

5.1.2 Training Details and Outcome for Multilabelled Classification

Several deep learning models were built using Tensorflow Keras (Chollet et al., 2015) according to the architecture of VGG 16 and trained on both ImageNet and Places365 (Kalliatakis, 2017) weights. Training outcome for these two were compared and it was found out that models trained on ImageNet could not learn well since it has low variance and high bias on the wrong predictions.

Focus were then put on training models using Places 365 weights. Initially, Adam optimiser was used as the optimiser algorithm. However, the mean average precision (mAP) and the loss fluctuated a lot during the training and could not converge. Hence SGD was used to replace Adam as the optimiser. The training result using SGD as the optimiser indicates that the model learns better.

The learning rate of the model was also fine-tuned to get better result. It is found out that the model learns better with 0.001 learning rate compared to 0.01 learning rate. In a few early model trainings, the number of epochs was set to be 100 but with the implementation with SGD as the optimiser, Keras's Early Stopping was made use of by monitoring the validation loss. The patience of the Early Stopping which denotes the delay of epochs before stopping the training was set to be 100 epochs. In other

Table 5.1: Comparison of Training Outcomes Using Different Model Settings

Model	Optimiser	Learning Rate	Data Augmentation	Validation Loss	Validation mAP
1	Adam	0.001	No	**0.84**	0.69
2	Adam	0.001	Yes	0.85	0.67
3	SGD	0.01	No	0.85	0.66
4	SGD	0.001	No	0.85	0.70
5	SGD	0.001	Yes	0.86	**0.72**

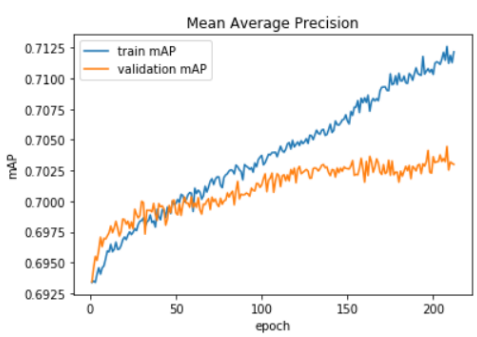
words, if there is no decrease for the validation loss in 100 epochs, the training would be stopped.

In an effort to increase the model performance, data augmentation was also performed. The newly created GSV images were rotated 90° horizontally on Google Street View. They were downloaded by setting the `heading` parameter for the Google Maps "Static Street View" API to be 90 (i.e., `heading=90`).

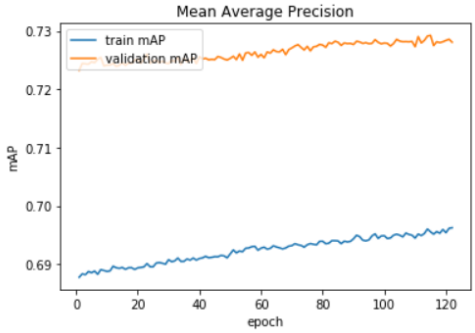
Table 5.1 shows the comparison of training outcomes using different model settings while the architecture and weights used were all respectively VGG16 and Places365. The batch size specified was also constantly 16. The validation loss and validation mAP are both taken at the point of the lowest validation loss.

From Table 5.1, Model 4 and Model 5 both performed similarly well but Model 4 was picked because the training mAP of Model 5 was generally lower than the validation mAP of Model 5. Figure 5.1 shows a comparison of training mAP and validation mAP

of Model 4 and Model 5. Model 4 was picked because it behaved more normally.



(a) Training mAP and Validation mAP of Model 4



(b) Training mAP and Validation mAP of Model 5

Figure 5.1: Comparison of Training mAP and Validation mAP of Model 4 & Model 5

5.2 Visualising the Predicted Perceptual Attributes

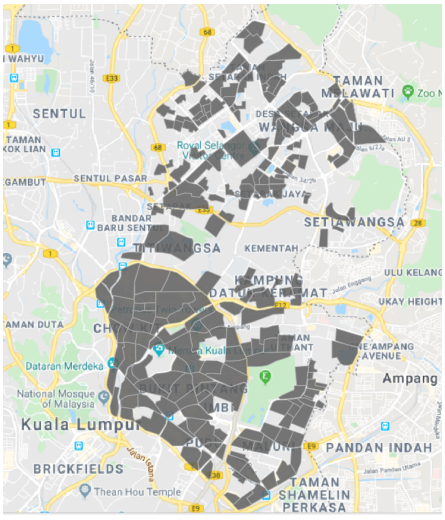
There are several steps involved in visualising the predicted perceptual attributes. The first step was to draw a KL neighbourhood map since there is no neighbourhood map of Kuala Lumpur currently available online. Next, random points for each neighbourhood were generated. The generated points were used to label the neighbourhood name of each neighbourhood.

After downloading the GSV images for each point, the images were passed to the model created earlier to get predicted perceptual scores for each attribute. The perceptual scores for each neighbourhood were calculated by averaging the perceptual scores of all the points in a particular neighbourhood. The perceptual score for each attribute for each neighbourhood were then visualised using choropleth maps. A predominant map was also drawn to illustrate the predominant attribute for each neighbourhood.

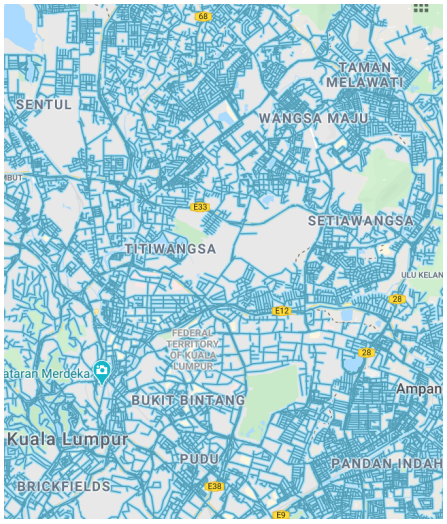
The map visualisation layers were then exported to be an interactive visualisation map using Leaflet (Agafonkin, 2011), a Javascript library for interactive maps.

5.2.1 KL neighbourhood map

Firstly, a neighbourhood map that covers a part of Kuala Lumpur was drawn using an open-sourced geographical information system software, QGIS (QGIS Development Team, 2009). Polygons which cover a certain neighbourhood area would be drawn if GSV is available at that particular area. The process was done by checking if GSV is available while drawing the neighbourhood map manually.



(a) Drawn Kuala Lumpur Neighbourhood Map



(b) Available Google Street View

Figure 5.2: Plotting of Kuala Lumpur Neighbourhood Map by checking if GSV is Available at a Particular Area

Figure 5.2a shows the KL neighbourhood map that was drawn. In Figure 5.2b, the blue lines denote that GSV is available.

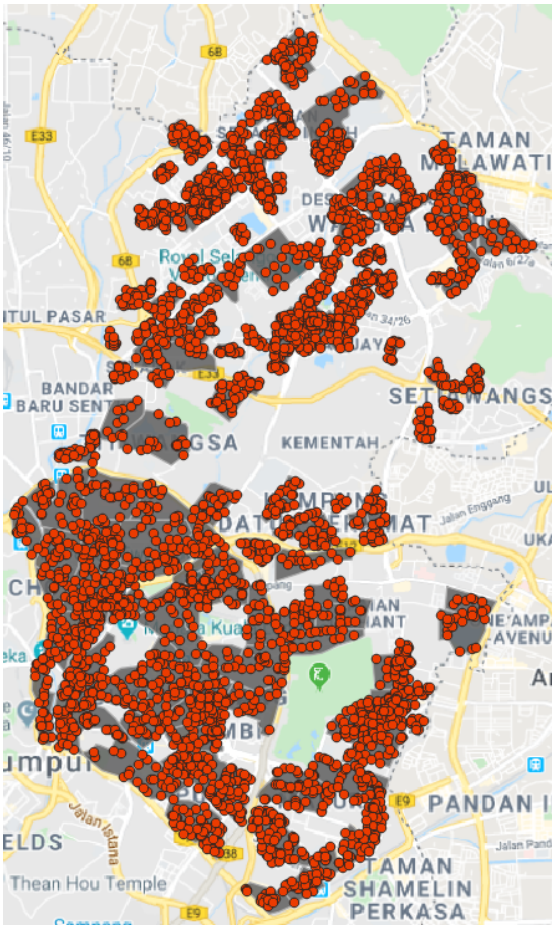


Figure 5.3: Snapped Points on the Neighbourhood Map

5.2.2 Generating New Points for Each Neighbourhood

To generate new points for each neighbourhood, random points were first generated for each neighbourhood by checking if the points fall within the neighbourhood polygon. Then, by making use of the Google Maps "Snap to Roads" API, the generated random points for each neighbourhood were snapped to roads so that GSV images can be downloaded since GSV is mostly available on the road. The process were repeated until there are suffice points to represent the whole neighbourhood.

Figure 5.3 shows the points being plotted on the neighbourhood map. Initially, the number of points were set to be 20 for each neighbourhood, however, due to having fewer roads available for GSV, some areas only have as few as 5 points. Aside from that, most of the areas have 15-20 points.

5.2.3 Getting Neighbourhood Name for Each Neighbourhood

The points were then used to find out the name of the neighbourhood (named as "sublocality" by Google Maps) each of them belongs to. In this part, Google Maps "Geocoding" API were used. The "Geocoding" API is usually used to query for the coordinate of a given location name. Here, a reverse geocoding was done by getting the neighbourhood name given the coordinate of the points.

Some of the points despite being in the same neighbourhood, the query returned a different neighbourhood name. To solve this problem, The neighbourhood name for the points that appear the most frequent for each neighbourhood were used to be the label as the neighbourhood name for all the points that belong to the particular neighbourhood.

5.2.4 Choropleth Maps to Visualise Each Attribute

The perceptual scores for each neighbourhood were calculated by averaging all the scores of the points for a particular perceptual attribute in a particular neighbourhood. The labelled perceptual scores for each neighbourhood were then used to plot choro-

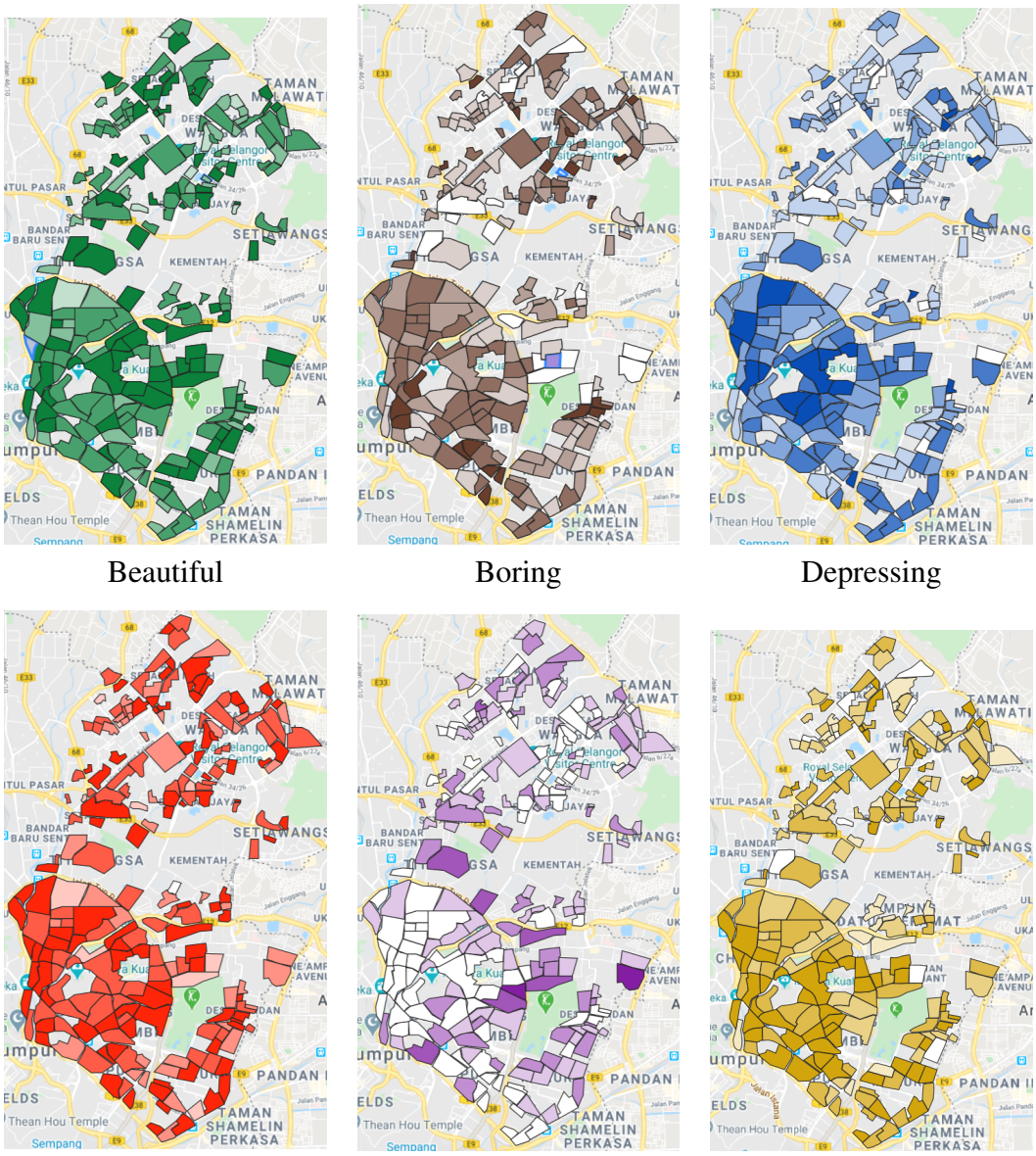


Figure 5.4: Choropleth Maps of the 6 Perceptual Attributes

pleth maps for each perceptual attribute. Figure 5.4 shows the choropleth map to visualise the perceptual scores in the neighbourhoods. The color encoding of the choropleth map denotes the perceptual score of each area. The perceptual score for each attribute were divided into 5 ranges.

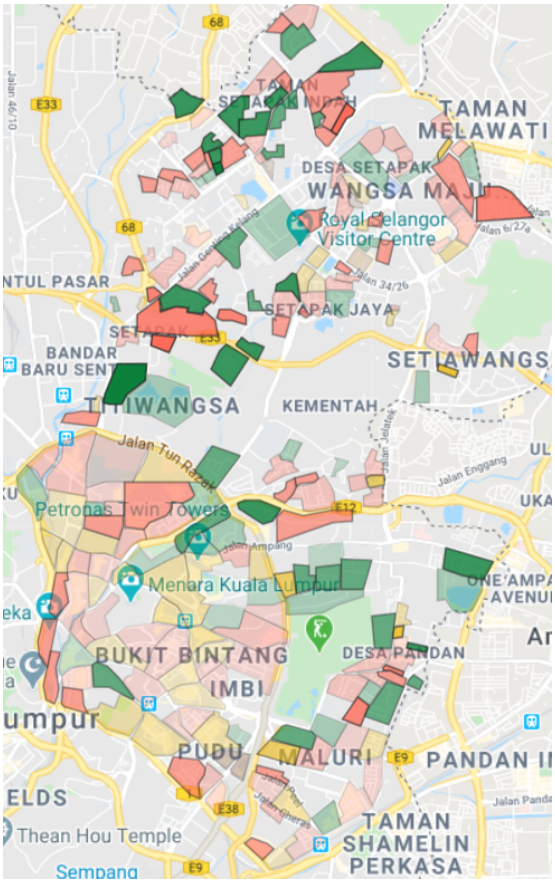


Figure 5.5: Predominance Map for Urban Perception in KL Neighbourhood

5.2.5 Predominance Map

In visualising the predominant perceptual attribute for each neighbourhood, a predominance map was drawn.

Figure 5.5 shows the predominance map that was created. The predominant attribute for each neighbourhood was calculated by getting the attribute that has the highest perceptual score across all 6 perceptual attributes. In the map, the transparency encoding of the map represents predominance strength of the predominant attribute among the

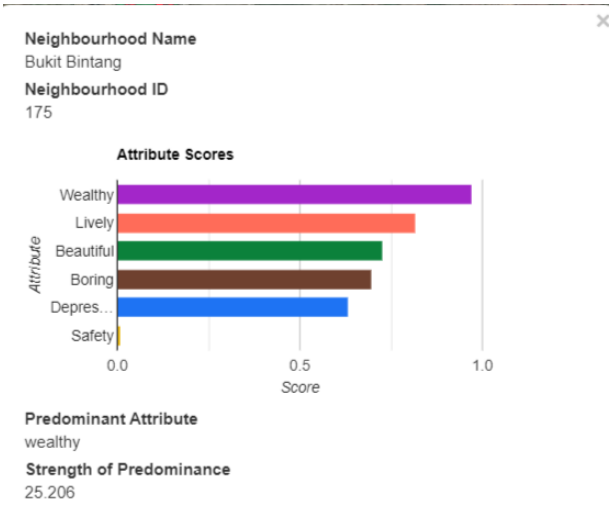


Figure 5.6: Example of a Pop Up from a Neighbourhood

6 perceptual attributes. The predominance strength calculated by getting the weight of the predominant attribute (as discussed in Section 4.6.4). It was divided into 5 levels of transparency.

5.2.6 Exporting to Web-Based Visualisation

The exporting was done using a plugin, qgis2web (Chadwin, Klinger, Olaya, & Dawson, 2015) on QGIS (QGIS Development Team, 2009) which exports map layers into Leaflet (Agafonkin, 2011) compatible form.

On the web-based visualisation, Google charts were embedded in each neighbourhood to visualise each perceptual score. Figure 5.6 shows an example of a pop up when mouse is hovered at a particular neighbourhood area.

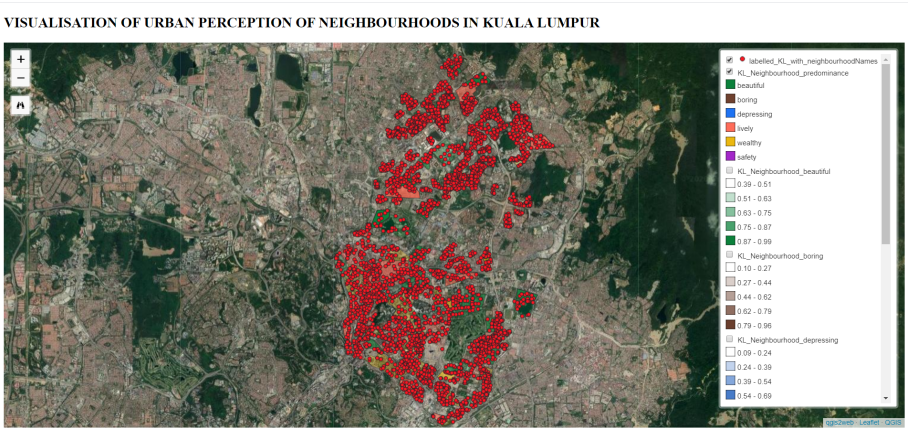


Figure 5.7: Screenshot of exported web-based visualisation

Figure 5.7 shows a screenshot of the exported web-based visualisation. There is a search button on the top left corner to help users locate a neighbourhood by its neighbourhood name. The map layers can also be hidden or shown as preferred by the user.

When clicked on the points, pop up such as Figure 5.8 would appear if the user is curious about the GSV images that are used for the prediction of the perceptual scores.

5.2.7 Analysis of Findings

From the choropleth maps in Figure 5.4, we can see that most of the neighbourhoods in Kuala Lumpur have high beautiful and lively scores. As for the boring attribute, depressing attribute and wealthy attribute, the scores are higher as it goes near the bottom left area where Kuala Lumpur city centre is located at. It is interesting to see how neighbourhoods with high wealthy score also have a high boring score and a high

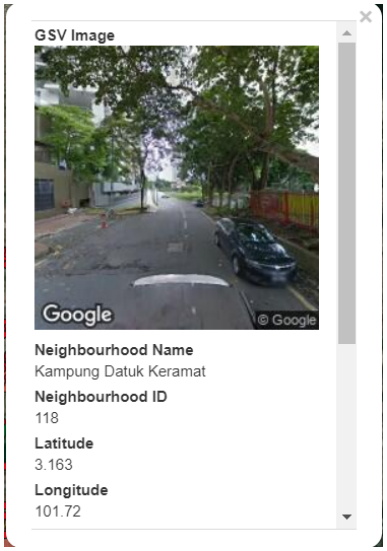


Figure 5.8: Example of a Pop Up from a Point

depressing score.

The safety score for all of the areas of in Kuala Lumpur is relatively low compared to all the other perceptual attributes. The highest safety score is only 0.54 whereas all the other attributes achieve at least 0.84 perceptual score. Moreover, Kuala Lumpur city centre surprisingly have a lower safety score. Furthermore, the depressing score is also much higher at the Kuala Lumpur city centre. This is probably caused by skyscrapers and towers were perceived as depressing despite being wealthy looking.

CHAPTER 6

CONCLUSION

In this research project, in order to solve the inefficiency of manual social studies on city streetscapes and the lack of visualisation on urban perception of cities, two research objectives were proposed. As an accomplishment, a suitable deep learning model was identified and trained for the multi-labelled classification task to predict the perceptual attributes for a location given the GSV image of that location. By having the model, the urban perception for a given city can be predicted without having to carry out time consuming surveys and questionnaires. The reachability limitation and low throughput issue will also be consequently eliminated since the deep learning model is able to predict the perceptual scores for a large amount of different locations. Using that model, predictions of perceptual attributes were done for some neighbourhoods in Kuala Lumpur. The predictions were then visualised on a web-based platform which enables user interactions.

For future enhancement, the performance of the deep learning model can still be improved by searching for better suited specifications for the model. If a model with better performance can be identified, the project can be further extended to be a crowd-sourced project in which users can expand the neighbourhood map and predictions can

be made on the newly expanded map. There is also a need for automating the pipeline of the whole process so that users can get the predictions immediately after drawing the neighbourhood map.

REFERENCES

Agafonkin, V. (2011). Leaflet: an open-source javascript library for mobile-friendly interactive maps [Computer software manual]. Retrieved from http://https://leafletjs.com

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. Retrieved from http://dx.doi.org/10.1023/A%3A1010933404324 doi: 10.1023/A:1010933404324

Chadwin, T., Klinger, R., Olaya, V., & Dawson, N. (2015). *qgis2web.* Retrieved from https://github.com/tomchadwin/qgis2web/wiki

Chollet, F., et al. (2015). *Keras.* https://github.com/fchollet/keras. GitHub.

Cs231n convolutional neural networks for visual recognition. (2017). http://cs231n.github.io/convolutional-networks/. (Accessed: 2019-06-19)

De Nadai, M., Vieriu, R., Zen, G., Dragicevic, S., Naik, N., Caraviello, M., ... Lepri, B. (2016, 08). Are safer looking neighborhoods more lively? a multimodal investigation into urban life.. doi: 10.1145/2964284.2964312

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09.*

Dertat, A. (2017, 11). *Applied deep learning - part 4: Convolutional neural networks.* https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2. Medium. (Accessed: 2019-06-19)

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city : Quantifying urban perception at A global scale. *CoRR*, *abs/1608.01769*. Retrieved from

http://arxiv.org/abs/1608.01769

He, Z., Yang, S., Zhang, W., & Zhang, J. (2018). Perceiving commerial activeness over satellite images. In *Companion proceedings of the the web conference 2018* (pp. 387–394). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from https://doi.org/10.1145/3184558.3186353 doi: 10.1145/3184558.3186353

Herbrich, R., Minka, T., & Graepel, T. (2007). Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 569–576). MIT Press. Retrieved from http://papers.nips.cc/paper/3079-trueskilltm-a-bayesian-skill-rating-system.pdf

Hidaka, A., & Kurita, T. (2017, 12). Consecutive dimensionality reduction by canonical correlation analysis for visualization of convolutional neural networks. In (Vol. 2017, p. 160-167). doi: 10.5687/sss.2017.160

Hough, P. V. C. (1962). General purpose visual input for a computer*. *Annals of the New York Academy of Sciences*, *99*(2), 323-334. Retrieved from https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1962.tb45317.x doi: 10.1111/j.1749-6632.1962.tb45317.x

Ilic, L., Sawada, M., & Zarzelli, A. (2019, 03). Deep mapping gentrification in a large canadian city using deep learning and google street view. *PLOS ONE*, *14*(3), 1-21. Retrieved from https://doi.org/10.1371/journal.pone.0212814 doi: 10.1371/journal.pone.0212814

Kalliatakis, G. (2017). *Keras-vgg16-places365*. https://github.com/GKalliatakis/Keras-VGG16-places365. GitHub.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.),

Advances in neural information processing systems 25 (pp. 1097–1105). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Liu, X., Chen, Q., Zhu, L., Xu, Y., & Lin, L. (2017). Place-centric visual urban perception with deep multi-instance regression. In *Proceedings of the 25th acm international conference on multimedia* (pp. 19–27). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3123266.3123271 doi: 10.1145/3123266.3123271

Lynch, K. (1960). *The image of the city*. Harvard University Press. Retrieved from https://books.google.com.my/books?id=_phRPWsSpAgC

Min, W., Mei, S., Liu, L., Wang, Y., & Jiang, S. (2019). Multi-task deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing*, 1-1. doi: 10.1109/TIP.2019.2932502

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014, June). Streetscore – predicting the perceived safety of one million streetscapes. In *2014 ieee conference on computer vision and pattern recognition workshops* (p. 793-799). doi: 10.1109/CVPRW.2014.121

Nasar, J. L. (1982). A model relating visual attributes in the residential environment to fear of crime,. *Journal of environmental systems*, *11*(3), 247-255. Retrieved from http://dx.doi.org/10.2190/4EEQ-C09R-M4MX-JGA0 doi: 10.2190/4EEQ-C09R-M4MX-JGA0

Nasar, J. L. (1988). Perception and evaluation of residential street scenes. In J. L. Nasar (Ed.), *Environmental aesthetics: Theory, research, and application* (p. 275–289). Cambridge University Press. doi: 10.1017/CBO9780511571213.026

Nasar, J. L. (1990). The evaluative image of the city. *Journal of the American Planning Association*,

56(1), 41-53. Retrieved from https://doi.org/10.1080/01944369008975742 doi: 10.1080/01944369008975742

Nasar, J. L., & Jones, K. M. (1997). Landscapes of fear and stress. *Environment and Behavior*, *29*(3), 291-323. Retrieved from https://doi.org/10.1177/001391659702900301 doi: 10.1177/001391659702900301

Ordonez, V., & Berg, T. L. (2014). Learning high-level judgments of urban perception. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 494–510). Cham: Springer International Publishing.

Partridge, S. (2018). *Finding the scenic route* (Unpublished doctoral dissertation). University of Bristol.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... et al. (2011, November). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.

Piaggesi, S., Gauvin, L., Tizzoni, M., Cattuto, C., Adler, N., Verhulst, S., ... Panisson, A. (2019, June). Predicting city poverty using satellite imagery. In *The ieee conference on computer vision and pattern recognition (cvpr) workshops.*

QGIS Development Team. (2009). Qgis geographic information system [Computer software manual]. Retrieved from http://qgis.org

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, *33*(5), 727-748. Retrieved from https://doi.org/10.1068/b32047 doi: 10.1068/b32047

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013, 07). The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*, *8*(7), 1-12. Retrieved from https://doi.org/

10.1371/journal.pone.0068400 doi: 10.1371/journal.pone.0068400

Santani, D., Ruiz-Correa, S., & Gatica-Perez, D. (2018, December). Looking south: Learning urban perception in developing cities. *Trans. Soc. Comput.*, *1*(3), 13:1–13:23. Retrieved from http://doi.acm.org/10.1145/3224182 doi: 10.1145/3224182

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, *abs/1409.1556*. Retrieved from http://arxiv.org/abs/1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). *Going deeper with convolutions.*

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, *abs/1512.00567*. Retrieved from http://arxiv.org/abs/1512.00567

Tucker, C., Ostwald, M., & Chalup, S. (2004, 01). A method for the visual analysis of streetscape character using digital image processing. In (p. 134-140).

Verma, D., Jana, A., & Ramamritham, K. (2019). Machine-based understanding of manually collected visual and auditory datasets for urban perception studies. *Landscape and Urban Planning*, *190*, 103604. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169204618313835 doi: https://doi.org/10.1016/j.landurbplan.2019.103604

Wang, W., Yang, S., He, Z., Wang, M., Zhang, J., & Zhang, W. (2018). Urban perception of commercial activeness from satellite images and streetscapes. In *Companion proceedings of the the web conference 2018* (pp. 647–654). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from https://doi.org/10.1145/3184558.3186581 doi: 10.1145/3184558.3186581

Watkins, S., Shams, L., Josephs, O., & Rees, G. (2007). Activity in human v1 follows multisensory perception. *NeuroImage*, *37*(2), 572 - 578. Retrieved from http://www.sciencedirect.com/science/article/pii/S1053811907004326 doi: https://doi.org/10.1016/j.neuroimage.2007.05.027

Xu, Y., Yang, Q., Cui, C., Shi, C., Song, G., Han, X., & Yin, Y. (2019). *Visual urban perception with deep semantic-aware network: 25th international conference, mmm 2019, thessaloniki, greece, january 8–11, 2019, proceedings, part ii*. doi: 10.1007/978-3-030-05716-9_3

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, *180*, 148 - 160. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169204618308545 doi: https://doi.org/10.1016/j.landurbplan.2018.08.020

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