LEARNING URBAN PERCEPTION VIA STREET VIEW IMAGES

Lim Xin Qi

Multimedia University
Faculty of Computing and Informatics
Persiaran Multimedia
63100 Cyberjaya
Selangor

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.

Index Terms— urban perception, Google Street View, perceptual attribute, map visualisation

1. INTRODUCTION

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 factor.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.

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. However, manual studies on urban perception are inefficient in which they are time consuming, limiting in the sense of reachability and have low throughout. 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 way using visually perceived information of city streetscapes.

As an example, In 2013, researchers 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 [1]. In 2016, The project was expanded 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 [2]. The researchers mainly inspired this paper since they gathered a huge dataset which is useful for carrying out the urban perception studies in a more data-driven and effective way.

As a result of the computed studies, visualisation is also easier to be created. 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.

Hence, the research objectives for this paper are: 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 and to create an interactive neighbourhood-level on the predicted perceptual attributes.

2. RELATED WORK

2.1. Earlier Work on Urban Perception

A classic literature on urban studies, [3] studied about human visual perception of cities in three American cities by getting research participants to draw mental images of the cities. 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 all the elements presented by the mental images.[4] [5], [6] and [7] also conducted urban perception studies in a manual way.

Along the years, researchers started studying urban perception via digital data. [8] proposed a method to analyse streetscape using image segmentation and Hough Transform algorithm [9] which detects less obvious boundaries in images. In [10], the researchers used cell phone usage data to study the intensity of mobile usage at a certain area.

In an effort to gather a large dataset (Place Pulse 1.0) for urban perception studies, [1] collected Google Street View (GSV) images and self-captured street view images in 4 different cities. 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.

2.2. Computed and Data-Driven Studies on Urban Perception

2.2.1. Place Pulse 2.0

Place Pulse 2.0 [2] is a dataset consisting of GSV images of 56 cities from all the continents except for Antartica. In the research, pairwise comparison was carried out to collect the human perception of a certain perceptual attribute in a city street view. A siamese-like CNN model which accepts an image pair as input and predicts winner in the pairwise comparison. Place Pulse 2.0 has been referenced in numerous papers as the dataset to train machine learning models including [11], [12], [13], [14] and [15].

2.2.2. Street View and Satellite Imagery

There are also studies that performed data collection via mobile crowdsourcing with the aim of expanding the coverage of inaccessible streets by GSV [16]. In this study, human urban perception was labelled automatically via low-level features and deep learning features which were extracted using GoogLeNet. [17] worked on predicting safety scores using multi-instance regression on street views along with crime records as the safety score for each place. [18] and [19] both investigated commercial activeness using satellite images. In both the studies, patches of image regions were used to train the model. [18] made use of Support Vector Regression (SVR) which accepts features extracted using Bag-of-Features (BOG) as input to predict the commercial activeness. whereas [19] implemented a CNN model to extract the feature vectors of the image patches and predicted the commercial activeness using regression. [20] looked into city poverty prediction using satellite imagery.

[21], [22], [23] made use of multimodal approach which combines different types of data or methodologies in the urban perception study.

2.3. Visualisation of Perceptual Attributes Score in Cities

As proposed in the second objective, a neighbourhood-level visualisation will be constructed. 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 geospatial representation.

In the Place Pulse 2.0 paper, [2]visualised the urban perception using geospatial representation. [24] 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.

[23] 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.

3. THEORETICAL FRAMEWORK

3.1. Making Use of Place Pulse 2.0 Dataset

Place Pulse 2.0 [2] is dataset gathered by the MIT Media Lab. The dataset consists of the result of the pairwise comparison between 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.

3.2. Street View Images as Dataset

Human visual perception of cities, as described by Lynch [3], 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 of images that contain these 5 elements.

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.

4. RESEARCH METHODOLOGY

4.1. 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.

4.2. Model Building

4.2.1. Model Architecture

The deep transfer learning model used was a model with VGG 16 architecture, pretrained on Places 365 weight [25]. 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 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.

4.2.2. 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.

4.3. 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 multilabelled classification task, mean average precision (mAP) was used as the evaluation metrics instead of accuracy.

The mAP was calculated at the end of each training epoch by averaging the average_precision_score metrics provided by scikit-learn [26]. 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.

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n$$

The equation above 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.

4.4. Data Visualisation

The visualisation of the predicted perceptual attribute score was presented using a map. one of the objectives of the project is to present a neighbourhood-level visualisation. Kuala Lumpur, Malaysia was chosen as the targeted city.

Several Google Maps API were made use of in the visualisation process including Static Street View API, Snap to Road API and Geocoding API.

To visualise the predicted perceptual scores, a choropleth map for each perceptual attribute was drawn to visualise the perceptual scores. 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?"

On top of the choropleth maps, a predominance map was also drawn. 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 follows:

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

where p represents the predominant attribute. The weight is then used to denote the strength of predominance of the predominant attribute for each neighbourhood.

5. IMPLEMENTATION

5.1. Evaluation of Model Performance

Several deep learning models were built according to the architecture of VGG 16 and trained on both ImageNet and Places365 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. SGD was used as the optimiser and as compared to Adam, the model learned better with SGD in this task.

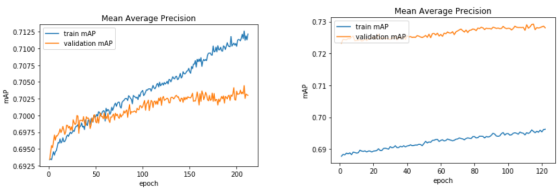
In an effort to increase the model performance, data augmentation was also performed. The newly created GSV images were rotated 90° horizontally on Google Street View. They were downloaded by setting the "heading" parameter for the Google Maps "Static Street View" API to be 90 (i.e., heading=90).

Table 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.

Table 1: Comparison of Training Outcomes Using Different Model Settings

Model	Optimiser	Data Augmentation	Validation Loss	Validation mAP
1	Adam	No	**0.84**	0.69
2	Adam	Yes	0.85	0.67
4	SGD	No	0.85	0.70
5	SGD	Yes	0.86	**0.72**

From Table 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 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

Fig. 1: Comparison of Training mAP and Validation mAP of Model 4 & Model 5

5.2. Visualising the Predicted Perceptual Attributes

Firsly, KL neighbourhood map was drawn to represent the KL neighbourhoods. Then points were generated for each neighbourhood using the Google Snap to Roads API. As the points were generated, the GSV image for each point was downloaded using the Google Static Street View API. After that, the neighbourhood names were queried using the coordinate of the points by using the Google Geocoding API. GSV Images were then passed to the model for predictions. Predicted perceptual attribute scores were visualised using choropleth maps and predominance map. Lastly, the map visualisations were exported using Leaflet.

6. CONCLUSION

In this paper, 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 multilabelled 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 crowdsourced 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.

7. REFERENCES

[1] Philip Salesses, Katja Schechtner, and César A. Hidalgo, "The collaborative image of the city: Mapping the inequality of urban perception," *PLOS ONE*, vol. 8, no. 7, pp. 1–12, 07 2013.

[2] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo, "Deep learning the city : Quantifying urban perception at A global scale," *CoRR*, vol. abs/1608.01769, 2016.

[3] K. Lynch, *The Image of the City*, Harvard-MIT Joint Center for Urban Studies Series. Harvard University Press, 1960.

[4] Jack L. Nasar, "The evaluative image of the city," *Journal of the American Planning Association*, vol. 56, no. 1, pp. 41–53, 1990.

[5] Jack Leon Nasar, "A model relating visual attributes in the residential environment to fear of crime,," *Journal of environmental systems*, vol. 11, no. 3, pp. 247–255, 1982.

[6] Jack L. Nasar, *Perception and evaluation of residential street scenes*, p. 275–289, Cambridge University Press, 1988.

[7] Jack L. Nasar and Kym M. Jones, "Landscapes of fear and stress," *Environment and Behavior*, vol. 29, no. 3, pp. 291–323, 1997.

[8] Chris Tucker, Michael Ostwald, and Stephan Chalup, "A method for the visual analysis of streetscape character using digital image processing," 01 2004, pp. 134–140.

[9] Paul V. C. Hough, "General purpose visual input for a computer*," *Annals of the New York Academy of Sciences*, vol. 99, no. 2, pp. 323–334, 1962.

[10] Carlo Ratti, Dennis Frenchman, Riccardo Maria Pulselli, and Sarah Williams, "Mobile landscapes: Using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, 2006.

[11] Fan Zhang, Bolei Zhou, Liu Liu, Yu Liu, Helene H. Fung, Hui Lin, and Carlo Ratti, "Measuring human perceptions of a large-scale urban region using machine learning," *Landscape and Urban Planning*, vol. 180, pp. 148 – 160, 2018.

[12] Yongchao Xu, Qizheng Yang, Chaoran Cui, Cheng Shi, Guangle Song, Xiaohui Han, and Yilong Yin, *Visual Urban Perception with Deep Semantic-Aware Network: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part II*, 01 2019.

[13] Lazar Ilic, M. Sawada, and Amaury Zarzelli, "Deep mapping gentrification in a large canadian city using deep learning and google street view," *PLOS ONE*, vol. 14, no. 3, pp. 1–21, 03 2019.

[14] Simon Partridge, *Finding The Scenic Route*, Ph.D. thesis, University of Bristol, 2018.

[15] W. Min, S. Mei, L. Liu, Y. Wang, and S. Jiang, "Multi-task deep relative attribute learning for visual urban perception," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.

[16] Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez, "Looking south: Learning urban perception in developing cities," *Trans. Soc. Comput.*, vol. 1, no. 3, pp. 13:1–13:23, Dec. 2018.

[17] Xiaobai Liu, Qi Chen, Lei Zhu, Yuanlu Xu, and Liang Lin, "Place-centric visual urban perception with deep multi-instance regression," in *Proceedings of the 25th ACM International Conference on Multimedia*, New York, NY, USA, 2017, MM '17, pp. 19–27, ACM.

[18] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang, "Urban perception of commercial activeness from satellite images and streetscapes," in *Companion Proceedings of the The Web Conference 2018*, Republic and Canton of Geneva, Switzerland, 2018, WWW '18, pp. 647–654, International World Wide Web Conferences Steering Committee.

[19] Zhiyuan He, Su Yang, Weishan Zhang, and Jiulong Zhang, "Perceiving commerial activeness over satellite images," in *Companion Proceedings of the The Web Conference 2018*, Republic and Canton of Geneva, Switzerland, 2018, WWW '18, pp. 387–394, International World Wide Web Conferences Steering Committee.

[20] Simone Piaggesi, Laetitia Gauvin, Michele Tizzoni, Ciro Cattuto, Natalia Adler, Stefaan Verhulst, Andrew Young, Rhiannan Price, Leo Ferres, and Andre Panisson, "Predicting city poverty using satellite imagery," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[21] S. Watkins, L. Shams, O. Josephs, and G. Rees, "Activity in human v1 follows multisensory perception," *NeuroImage*, vol. 37, no. 2, pp. 572 – 578, 2007.

[22] Deepank Verma, Arnab Jana, and Krithi Ramamritham, "Machine-based understanding of manually collected visual and auditory datasets for urban perception studies," *Landscape and Urban Planning*, vol. 190, pp. 103604, 2019.

[23] Marco De Nadai, Radu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Hidalgo, Nicu Sebe, and Bruno Lepri, "Are safer looking neighborhoods more lively? a multimodal investigation into urban life," 08 2016.

[24] Vicente Ordonez and Tamara L. Berg, "Learning high-level judgments of urban perception," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 494–510, Springer International Publishing.

[25] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.