Street Infrastructure: urban pedestrian crosswalks localization via Google satellite imagery using a deep learning model

Zheng Li

Adviosr: Janille Smith-Colin, Ph.D., P.E.

Fall 2018

1 Introduction

Pedestrian crosswalk plays an important role by connecting urban streets and providing proper access to pedestrian. Crosswalks are primarily painted on the road intersection and the zebra-crossing design has been widely used in most of modern cities. With the continuation of urbanization, more streets are going to be built, so does crosswalks. Hence it becomes necessary for city to establish a crosswalk inventory to better operate regular crosswalks maintenance and replacement. Because the data collecting process is repetitive, time-intensive and there is little benefits directly coming from knowing the crosswalks location, the crosswalk data collection has always been in a low-priority and neglected. However, due to crosswalk's unique marking pattern, it is possible nowadays to automate the collection process at low-cost within a short time period. With the prevalence of object detection in computer vision, it motivates interests to collect urban infrastructures including crosswalks using computer-aided techniques. Intrigued by this fact, this project aims to build a automatic framework to detect crosswalks via Google satellite imagery using a deep learning model. The report is organized in the following sections: section 2 presents related studies on crosswalks detection with various methods. Section 3 focuses on data acquisition and labeling. Section 4 covers image preprocessing and deep learning model and we tested the model and presented its performance in Section 5. Discussion and further development are include in Section 6.

2 Previous Studies

There are several studies that attempted to use computer vision to detect crosswalks using satellite images. Ghilardi et al.[3] developed a Support Vector Machine (SVM) model with 30x30 pixel window to detect crosswalks using low-resolution satellite images; Sun et al. [5] applied a joint-boosting model to successfully extract zebra crossings from high-resolution aerial image and reconstructed the crosswalks spatial geometry (dimension, orientation, etc.); Although many models have been developed in the field of object detection, those models are mostly using derived features from images though image augmentation such as Local Binary Pattern (LBP) features [3], Grey-Level Co-occurrence Matrix (GLCM) and Gabor features[5]. The expressivity of those derived features are limited because they are artificially created to better represent some aspect of an image (intensity, edge, texture, etc.). Recently the rising of convolutional neural network (CNN) has changed the entire image recognition field and outperformed majority of detection models. Berriel et al. [2] used a famous CNN architecture called VGG-16 to conduct a large-scale crosswalks classification from streetview images taken by camera-mounted vehicles. Ahmetovic et al.[1] combined two data sources and used a featured-driven detection model to detect crosswalks from satellite image and use streetview image as the verification with other image recognition model. This project is mostly motivated from above studies and presents a framework to automatically classify crosswalks from satellite images using a convolutional neural

network model(VGG-16).

3 Data Acquisition and Labeling

The entire project framework is presented in Fig 1. According to the overview, the framework uses street location as the input information, it first disaggregates streets into series of geo-points and retrieves corresponding street satellite images using Googe Static Map API[1] given points' latitude and longitude. Part of satellite images (100 images) are going to be manually labeled and used to generate training image patches for crosswalk recognition model and the rest of unseen satellite images go through image preprocess including extracting road segments, dilated road segments and masking streets. Each processed satellite image will be divided into 80x80 pixels patches. We use trained model to make classification on patches and output a binary response indicating whether or not the patch is recognized as a crosswalk. All predicted patches will be used to derive its coordinates and eventually assign to corresponding street as final crosswalk information inventory.

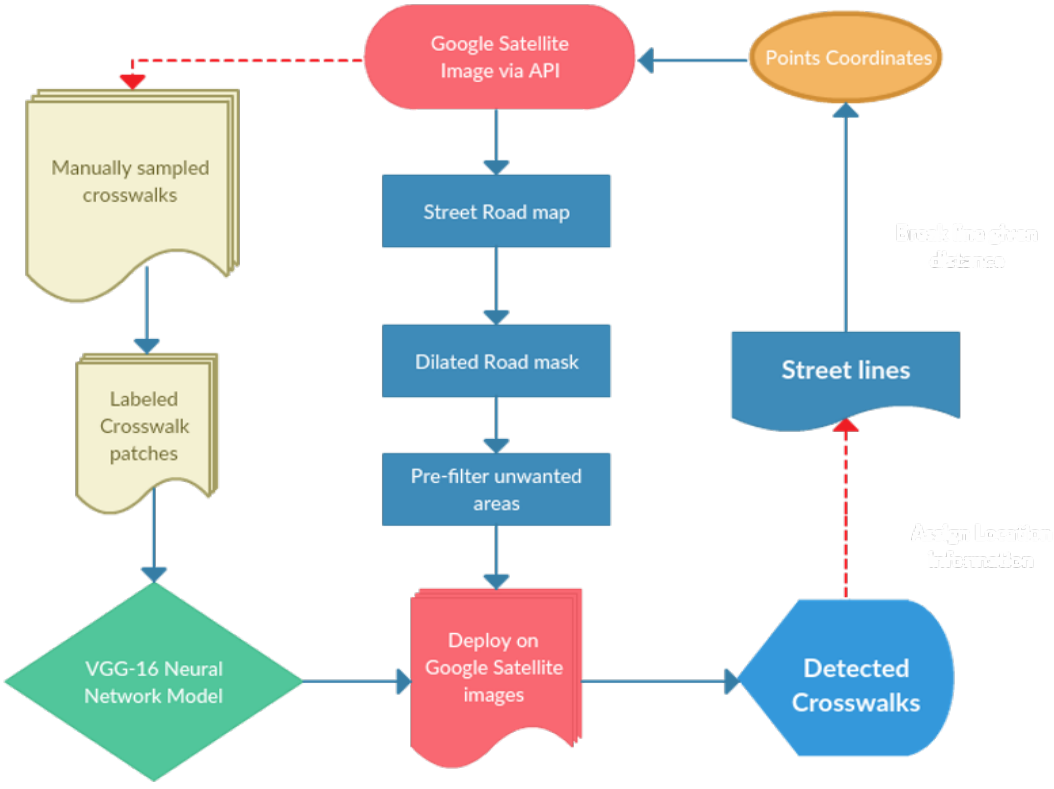

Figure 1: Automatic crosswalk localization framework

[1]Google Static Map API: https://developers.google.com/maps/documentation/maps-static/intro

3.1 Street satellite image retrieval

ArcGIS *shapefile* is one of the most popular file format developed and regulated by Esri that stores geospatial information. It has several geometric shapes including *point*, *polyline* and *polygon*. The street satellite image retrieval first evenly break street ArcGIS polylines into short line segments given a certain distance (60m in this project), and both start and end points of each line segment can be generated using ArcGIS toolbox. After removing the duplicate points, we get a series of points representing streets. And the coordinates of each point (latitude/longitude) are feed into Google Static Map API to download the satellite images of corresponding geo-location. The schematic acquiring process is shown in Fig 2 and executed in Python.

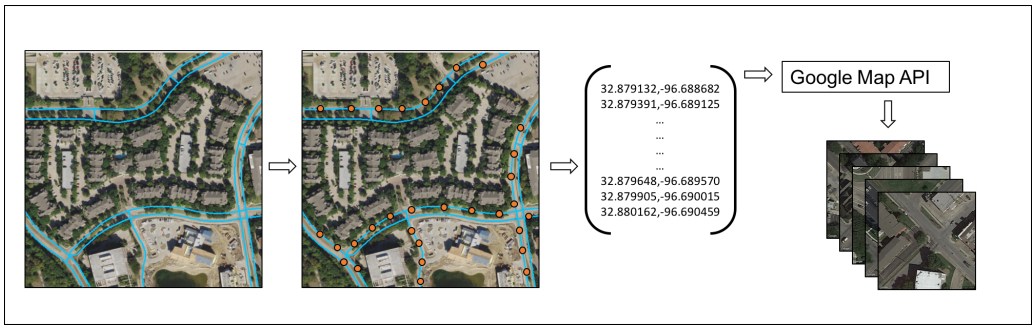

Figure 2: Acquiring Satellite images given ArcGIS Shapefile using Google Static Map API service

3.2 Training Patch generation

Once street satellite images are obtained, we use a moving window (80x80 pixels) with a 40 pixels stride to generate image patches for model prediction. In order to generate labeled image patches for model training, we manually collected 100 coordinates of satellite images that contains crosswalks on site using a web map application[2]. Each satellite image was manually labeled using Matlab computer vision image Labeler[3] and labels were constructed using the bounding box method with the concept of Intersection Over Union(IOU): An equal-sized pixel window parses over the image and count the area of crosswalk and non-crosswalks, then it computes the ratio between area of overlap and area of union as the criteria to assign positive/false label to the patch. An upper threshold 0.7 was chosen for positive labels (crosswalk exists) and lower threshold 0.4 was chosen for negative labels (crosswalk not exists). After labeling process, we are able to construct 10,000 labeled patches with 5000 positive patches and 5000 negative patches. An example of labeled patches is also presented in Fig 3.

[2]Satellite map of the world: https://satellites.pro/

[3]Matlab Computer Vision Image Labeler: https://www.mathworks.com/help/vision/ref/imagelabeler-app.html

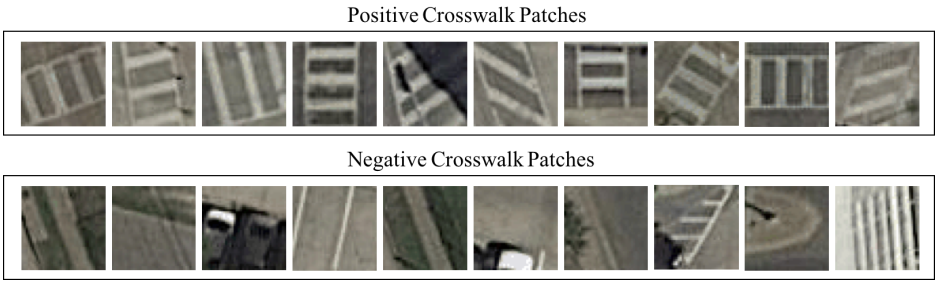

Figure 3: Examples of labeled positive and negative image patches at the size of 80x80 pixel window

4 Image Preprocessing and VGG-16 model

4.1 Image preprocessing

Since most of the crosswalks are designed to be near or at the road intersection, patches coming from non-road features (buildings, parking lots, grass, etc) have very little chance to have crosswalks. So it is more computationally efficient to use only road-covered patches and only conduct detection on those patches. Fig 4 shows the process of mask road-covered area within image and extract them for recognition. Besides the satellite images, we used the road map features provided by Google Static Map API to highlighted the road segments and performed several dilatation filters so that the mask covers the entire road. Then only patches within masked road area will be used for crosswalk detection. According to rough estimation shown in the example in Fig 4, the example image was taken in Dallas and approximately 50% of image was masked and therefore recognition process could potentially be speed up by 2 times.

Figure 4: Street segments extraction process. Steps (from left to right) include: Satellite image → Road map → Road mask → Dilated road mask → Masked streets

4.2 VGG-16

VGG-16 [4] (also called OxfordNet) is a convolutional neural network architecture named after the Visual Geometry Group[4] from Oxford, who developed it. It was used to win the ILSVR(ImageNet

[4]Visual Geometry Group: http://www.robots.ox.ac.uk/∼ vgg/

[5]) competition in 2014. It contains 16 convolutional layers (Fig 5) and only uses 3x3 convolutional layers stacked on top of each other in increasing depth. It uses max pooling to reduce volume size and ends with softmax classifier. It is currently the most preferred choice in the community for extracting image features and simple pattern classification. The model contains up to 140 million parameters and was already pre-trained using ImageNet dataset. To transfer the model into this project, we took the pre-trained model and used the labeled patches constructed in section 3 to tune the parameters in the last three layers. The predicting patches(80x80) will be up-scaled to size of 224x224 which is the standard image input for VGG-16 and the model outputs a probability value for two classes (crosswalk, non-crosswalk) which turns into a binary response by applying a pre-defined threshold(e.g 0.5 in this project).

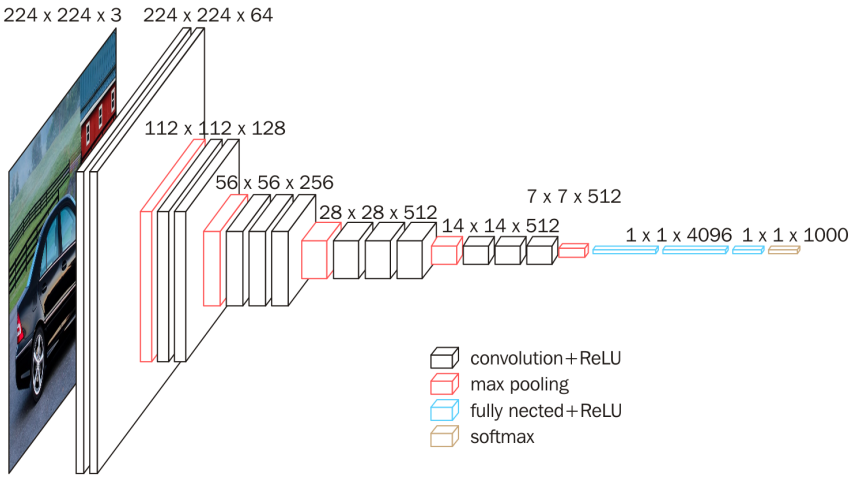

Figure 5: VGG-16 convolutional neural network architecture

5 Results

The model has been trained using Python package called *Keras*[6] with 10,000 labeled patches and tested on 1000 unseen patches. The model's best performance is 97.5% of classification accuracy[7]. We deployed the model and applied the framework to the neighborhood (Vickery Meadow[8]) to test its localization performance on crosswalk and the result is shown in Fig 6. We manually to find out all the ground-truth within Vickery Meadow as comparison and concluded that the framework found 16 crosswalks out of 17. And all five locations of crosswalks have been correctly identified.

[5]ImageNet Challenge: http://www.robots.ox.ac.uk/∼ vgg/research/very_deep/

[6]Keras: https://gist.github.com/baraldilorenzo/07d7802847aaad0a35d3

[7]Accuracy = Number of corrected positives instances/Total number of positive instances

[8]Vickery Meadow: Vickery Meadow is an ethnically-diverse neighborhood located at north Dallas, it has the population of 41,623 in an area of 2.86 square miles

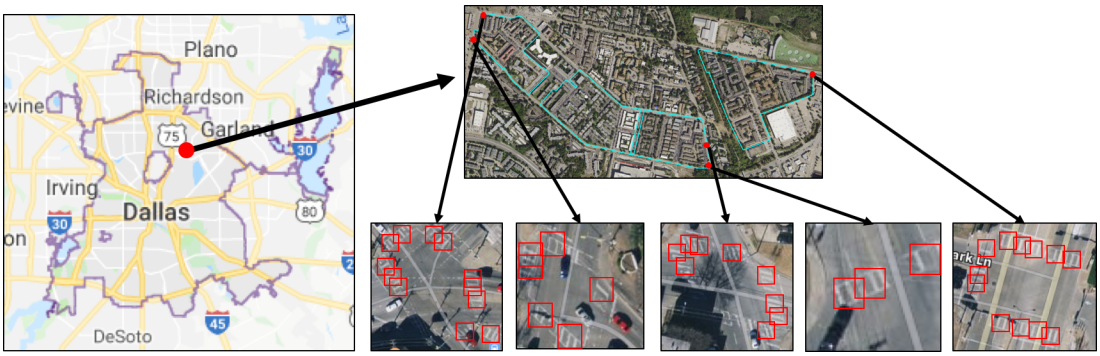

Figure 6: Graphical representation of crosswalk localization framework performance results. The proposed found 16/17 crosswalks in Vickery Meadow and 5/5 road intersections containing crosswalks

6 Conclusion and Future Direction

6.1 Performance Justification

- Model Accuracy

  Due to the time limit of project, the model has only been trained once without applying any cross-validation which could leads to some bias. Although it achieved 97.5% accuracy in this project, a CNN model is expected to have a higher accuracy in this context since this model contains $\sim$ 140 million parameters and the model should be complex enough to correctly recognize crosswalks with high confidence ($\geq$ 98%). A full performance evaluation and rigorous parameter optimization need to be conducted to obtain optimal model. Under the scope of this project, the VGG-16 model serves as core function in the framework to make prediction on individual image patch and a model with high recall should be preferred. During the localization phase, a low recall model would produce more false positives (classifies true crosswalk labeled as non-crosswalk), it reduces the number of true crosswalks, and thus disobey the essence of this framework which is to replace human labors and automatically detect crosswalks. However, uses a model with high recall, it produces less false positives and it is acceptable for model to generate false negative(classifies true non-crosswalk as crosswalk). Nevertheless, it still captured all the ground-truth without losing truth information. Those false negatives can easily be identified/filtered by human eye during the later evaluation phase.

- Performance Metric

  A simple accuracy metric was used in this project to measure the goodness of the model as well as framework. However, we should recognize that the difference between model prediction and framework prediction. Model prediction shows the predictability of simple image patch which do not provide any spatial coordinates of predicted result. framework prediction is set to describe

how accurate the framework is able to find the crosswalks and assign proper geo-coordinates. The procedure of assigning geo-coordinates was barely discussed in this project due to the time limit, but it should have a clear-defined method to convert VGG-16 model binary prediction into a spatial vertices related to crosswalks. This task will be left as future work to extend the project.

6.2 Future Direction

- Localization Method

  This project attempted to build a framework that retrieves Google satellite images and localize the crosswalks using VGG-16 model. The development and application of VGG-16 model was the main focus in this project and the crosswalk localization (get geo-coordinates) based on model's binary response is lightly discussed and needs to be fully developed for practice use. As discussed in the previous section, the high recall model is expected to be used in the framework. To accommodate the model, Ghilardi et al[3] proposed a method to localize crosswalks based on binary response from recognition model. Besides the binary response indicating the existence of the crosswalk in the patch, he also used neighbor patches as a second verification: it patch A has been recognized as existed crosswalk, then all the patches around patch A will be examined, if at least one of the surrounding patches is also recognized as crosswalk by model, patch A is a crosswalk, otherwise patch A is non-crosswalk. This method can be used as the base to complete the framework.

- Crosswalk Quality

  One of the most interesting questions we want to ask once we have all the crosswalks locations would be: can we assess the quality of crosswalk? Unfortunately, there is no existing/on-going literature that answers this question. It currently faces two main challenges: 1) The quality of crosswalks has not been widely defined and used in any of infrastructure assessment projects. The lack of quantitative definition of crosswalk quality hinders the efforts to proceed crosswalks quality study. For example, some might think the clearness of stripe pattern could be used as an index to measure crosswalk quality, but others may argue the completeness of crosswalk markings across the road should be considered because while crossing the street, pedestrian would pay more attention to the overall markings pattern rather other the texture of the stripes. 2) The uncertainty of the model performance on quality assessment. Assume that we eventually come up with a measure for crosswalk quality, it still needs to be defined in a way that is detectable in the context of computer vision techniques. If the pattern exhibited at different crosswalk quality is not distinguishable via human eyes, it is also very difficult for a computer to differentiate them. Given those challenges, it is recommended to carefully examine the definition

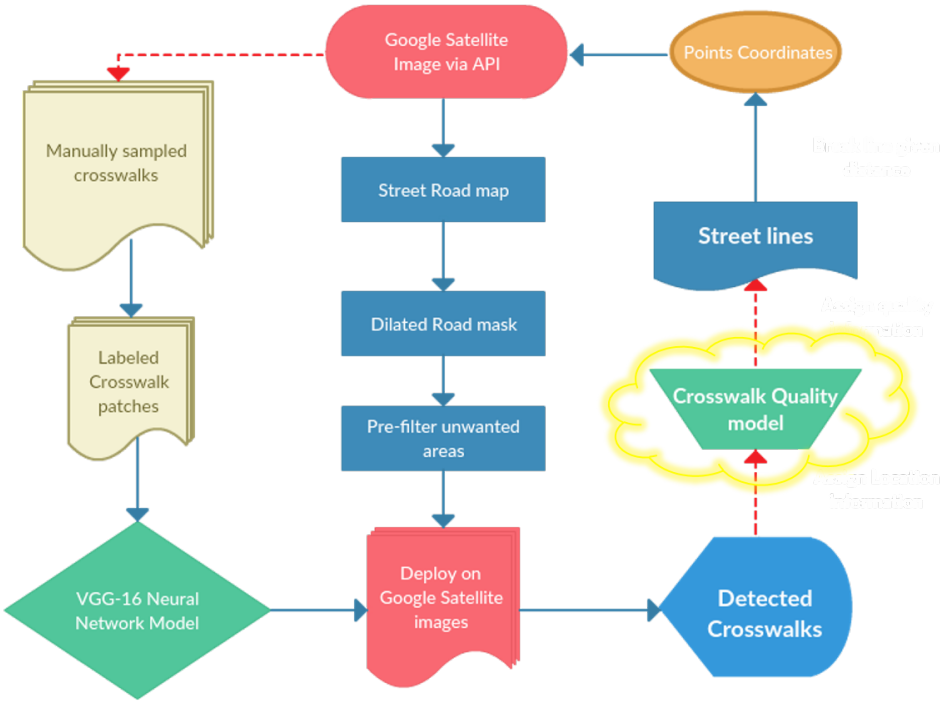

Figure 7: Full crosswalk localization framework with quality assessment model

of crosswalk quality and after that choose the proper model for quality assessment. Fig 7 shows a final framework including crosswalk quality assessment model (highlighted in yellow). Once developed, it will be a cost-effective alternatives for both researchers and city planners to collect crosswalks information.

References

[1] Dragan Ahmetovic, Roberto Manduchi, James M Coughlan, and Sergio Mascetti. Mind your crossings: Mining gis imagery for crosswalk localization. *ACM Transactions on Accessible Computing (TACCESS)*, 9(4):11, 2017.

[2] Rodrigo F Berriel, Franco Schmidt Rossi, Alberto F de Souza, and Thiago Oliveira-Santos. Automatic large-scale data acquisition via crowdsourcing for crosswalk classification: A deep learning approach. *Computers & Graphics*, 68:32–42, 2017.

[3] Marcelo Cabral Ghilardi, Julio Jacques Junior, and Isabel Manssour. Crosswalk localization from low resolution satellite images to assist visually impaired people. *IEEE computer graphics and applications*, 38(1):30–46, 2018.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] Yanbiao Sun, Fan Zhang, Yunlong Gao, and Xianfeng Huang. Extraction and reconstruction of zebra crossings from high resolution aerial images. *ISPRS International Journal of Geo-Information*, 5(8):127, 2016.