CSE6242 - Data and Visual Analytics

Progress Report

Team 64: Safe Street

Section 1: Introduction and Problem Definition

The trending bike movement in the United States is advocated by a large population who is seeking a healthy and greener lifestyle, especially in big cities like San Francisco. With its plan for net zero emissions in 2050, San Francisco is the second most bike friendly city in the United States [2, 20]. Meanwhile, as the bicycle becomes a more prevalent mode of transportation, the rate of bicyclist injury has increased drastically by 21% from 2013 to 2017[20]. Even though there are political will and cycling initiatives across the US to improve policy environment, cyclists are still extremely vulnerable when compared to other road users. To boost safety for cyclists and improve the civil infrastructure, designers need to know what features and characteristics of the biking environment directly relates to the accident rate. Civil projects like this are expensive in Bay Area. To make matter worse, there is a lack of a sufficient tool that transportation professionals can use to support their actions. There is an urgent need to identify and prioritize the area where changes would result the greatest improvement in safety. Our solution, Safe Street for Cyclist, will provide the key contributors of accidents and identify where the stakeholders should focus their resources and guide them with budget decisions. While the platform provides infinite support to planners, it serves as educational pathway and provides rich insight into San Francisco for rest of the world.

Section 2: Literature Survey

Today, many risk factors are assessed through high-level social studies such as vehicle counts, demographics, and social-economic status [14], but those studies rarely expose the root causes of accidents. One common problem of these approaches is the lack of associating transferable visual street features with accident data.

Other researches collect historical accident data and visualize the number of accidents on a map view. As a result, streets with higher accident frequency would be considered more dangerous [5, 15, 19]. The bias is that areas with higher traffic density are sampled and reported more often, and are thus more likely to be categorized as dangerous. Additionally, there is little visibility into the root cause of an accident.

Inspired by the work in [1, 5, 7], we aim to develop a scalable approach that leverages object detection algorithm [9, 10] to identify visual features of a street and combine these features to predict risk score. Our model will be trained with Microsoft's common objects dataset [11], which current practices [3, 7, 16, 18] do not have.

Moreover, we will perform clustering analysis on the detected objects and features extracted from street views. Similar to [9, 13], the significance level and effects of these clusters will be examined, and the significant ones will be the inputs for our final risk model. Our new approach has no sampling bias but covers all bike accident-related data, aiming to provide a unified method of collecting, storing, and transferring the features related to bike accidents. It is a one-stop shop for researchers and users.

Our approach is effective because Google Street Views API is a reliable and cost-effective data source for extracting infrastructure features and it has been used successfully in similar researches [4, 6]. San Francisco's route infrastructure differs from those in cities with high trip shares (like Amsterdam), so analyzing street features is a promising approach for investigating injury risks [18]. Our work will be presented in interactive visualization tools like mapbox.gl, rendering it user-friendly for users with non-technical background.

Section 3: Proposed methodology

The innovations are summarized here, and details can be found in a later section.

a. Computation:
    i. Use objective detection model to pull location characteristics from the images
    ii. Combine objective detection findings with public data in developing a Random Forest model for each location
    iii. Leverage Convolutional Neural Network to train the model
b. Data presentation:
    iv. Integrate bike collision data into one map whose roads are color-coded based on its composite risk score
    v. Craft user-friendly web page to display data for end users: provide toggles to further filter collision data to display meaningful statistics.

3.1 Data Collection & Data Cleaning

The data collection process begins with literature research and compiling a list of relevant datasets. The original datasets are then transformed to suit our needs. The bike collision datasets are from two pubic data resources: Transportation Injury Mapping System (TIMS) and TransBase. We gathered bike collision data of San Francisco from 2008 to 2012. We aggregated this collision dataset to street segment level, and then joined it to street infrastructure dataset.

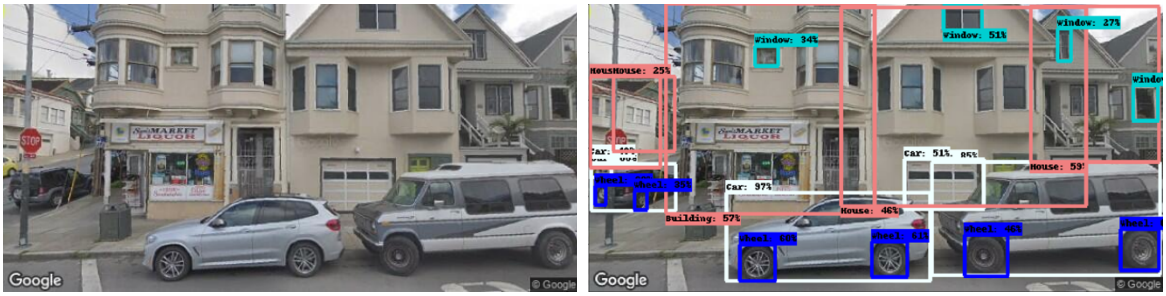
Since both datasets are public, they contain some redundant information and have large amount of missing data. The cleaning process is done with Python Pandas, NumPy and SQLite3 to minimize the chance of human error on our end. After a thorough review of the original datasets, we removed columns that are not meaningful to street demographics. We used KNN for imputation to replace the missing values with substitutes from KNN.

To get image data for object detection later, we extract a list of latitude and longitude pairs for each intersection in San Francisco and a list of coordinates for each collision. Using the Google Street Views API, we collected over 9,147 images from lat-long coordinates for intersection and over 3,852 images for collision locations.

3.2 Computation

To enrich street features, we leveraged a Tensorflow object detection model trained on Google's Open Image dataset to identify possibly relevant objects and street characteristics on all the Google Street Views images collected. For each image, we detect 20 objects, ranked by number

of occurrence and probability score. From there we were able to cluster streets using K-means into custom groups on those streets. The clusters obtained from the object detection process were combined into other existing street features in the TransBase dataset.



Latitude, Longitude: 37.7483528, -122.40871

We then developed a baseline Random Forest model using the features collected to identify street features that are highly correlated with a street segment being more dangerous than others. Random Forest is selected because some of the independent variables, such as number of parking space and speed limits, were highly skewed, so the assumptions for a linear model were not met.

Prior to fitting the model, we also needed to preprocess the TransBase dataset to handle missing values for speed limits and number of meters. The missing values of these variables are imputed using KNN on the street segment key. It is intuitive because nearby streets are more likely to have similar speed limits and number of meters. For features with significant missing values like number of parking space, we decided to add a binary column to capture the effects of missingness. Secondly, some of the categorical features are removed because more than 80% of the values belong to one category. The resulted Random Forest is able to achieve a 65% out-of-bag R-squared on our safety metrics, and the most significant feature is Estimated number of daily ridership per quarter mile from street segment. Going forward, we will explore more feature transformation and selection methods and compare to this baseline model.

Our objective of using CNN is different from Tensorflow. With CNN, we do not aim to identify the street object, but we try to identify the image similarity or dissimilarity. Convolutional neural networks (CNN) is a powerful tool for analyzing Google Street Views because of its ability to comprehensively capture visible features when analyzing image data. In our project, we want to do it in an automated process for feature extraction by learning the visual feature representation using CNN. We apply several layers of filters to transform the image data from pixel representations to the complimentary feature dataset that we need. In order to make predictions, we aim to use these complimentary features in conjunction with those from the TransBase dataset. The model needs a lot of training data with corresponding labels to give feedback to the network to optimize effective feature representation. We will try to use CNN as a learned representation of visual features which we can use to interactively explore the TransBase data.

3.3 Visualization

Data is meaningless to the public without visualization. Our challenge is to design an interactive visualization that depicts the hierarchical relationship between findings for each coordinate's

features and the street risk index, while preserving spatial relationships between each coordinate and each street. Specifically, we'll be addressing:

- How are we going to allow users to see each collision details?
- How are we going to express the relationship between a street and a coordinate?
- How are we going to allow users to select for their desired granularity and still see the bigger picture?
- What is most useful information that will be presented to attract interest from public and have them walk away with more than just ridership facts?

We decided the platform will be a scrollable, story-telling presentation. When the end user enters the site, the page will first present the impact of bicycle ridership. As the user scrolls through the webpage, we will present more facts about San Francisco's bicycle collisions and how everyone can contribute to the community. The reading will be short, clean and right to the point. Once the user has scrolled to the end of the story, a geo map of San Francisco will appear. The map will use D3.js as a data representation tool and overlay with MapBox-gl to display the city data. On the map, the streets will be color-coded based on the risk index level. For each street, historical collision points will be represented by a dot. If the end user hovers over the desired street, a pop-up will display more details about the street. If the end user hovers over the dots, a pop-up will display details about the collision, and will also provide an object detection map regarding the collision scene. The end user would also be able to use additional toggles to filter the visualization down by collision year, risk score and other significant features.

Ultimately, the goal is to provide value to the end user from the interaction with this platform. For example, an urban planner should be able to learn visual features for civil project analysis and combine his/her professional knowledge to come to a decision that maximizes safety improvement.

Section 4: Evaluation

There will be two ways of evaluating the impact of our project.

4.1. Accuracy: Random Forest

We implemented a random forest and calculated the score on the train set. In order to make sure that the model is not overfitting, a validation set was created.

4.2 Accuracy: Baseline & CNN

Our dataset consists of N rows where N is the number of streets segments in the training set. Each row contains a street id and our prediction of street clusters.

Scored on Accuracy and Macro-F1:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)$$

$$Macro - F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c}$$

where:

N is the number of streets segments in the training set,
ŷi and yi are the predicted and true label of the i-th street, respectively,
𝟙 denotes the indicator function,
C is the number of possible labels,
and F1c are the precision, recall and F1 score of class c, respectively.

4. 3 User Study:

4.3.1 Cyclist feedback & GUI User Friendly

Safe Street for Cyclists aims to assist cyclist with their daily biking. We are conducting a user study to assess if users who use our "safe street map" to route their biking path had better riding experiences, i.e., feel safer on the street. we asked 10 cyclists to test our map. We asked the users to first report their daily biking route and have them decide their own "improved" bike route using our dangerous vs safe heatmap. Later they try the "improved' map for a week and give us their subjective results.

Results: **under construction**

4.3.2 Urban Designer Feedback

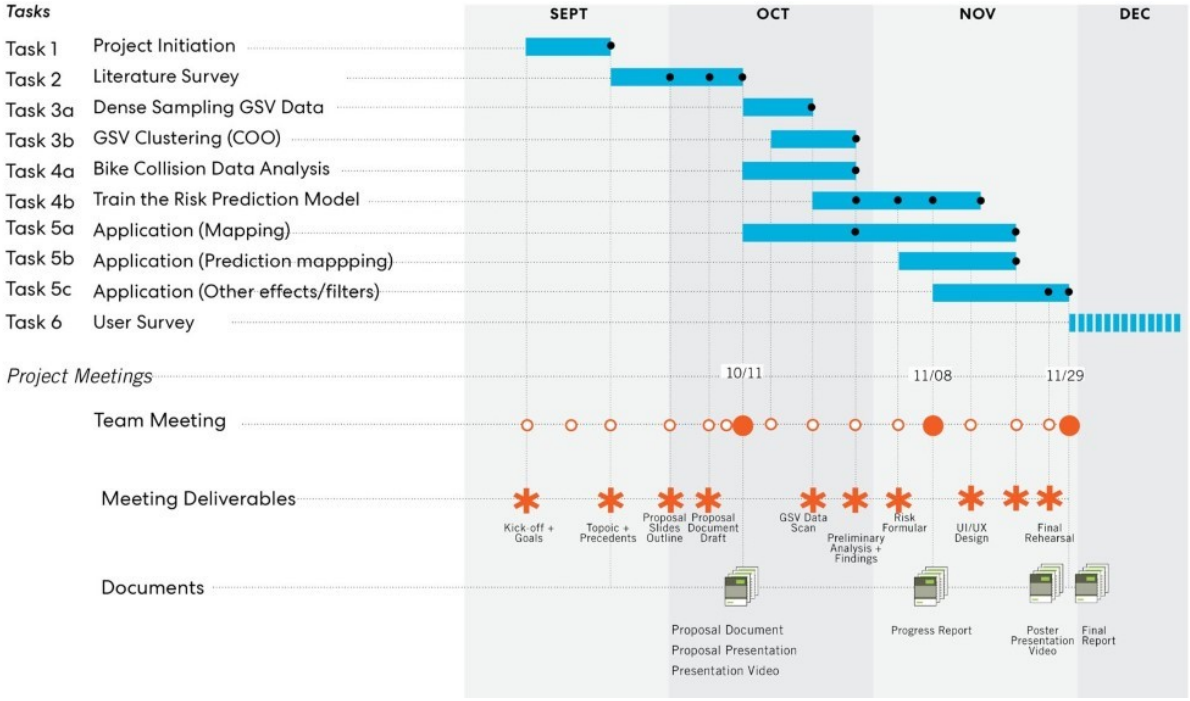
Several professional urban designers are also distributed with copies of our tool. They test our tool functionality to determine if the project is indeed useful.

Results: **under construction**

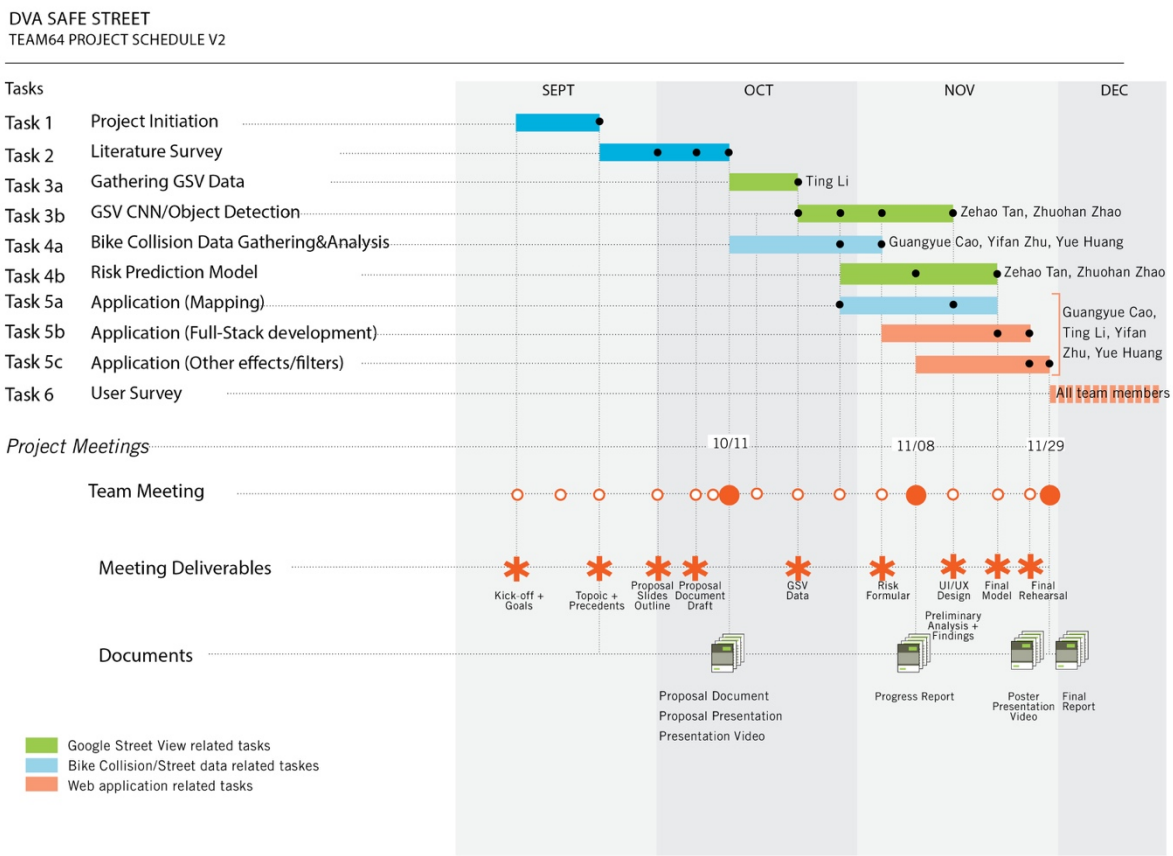
Section 5: Plan of Activities and Distribution of team effort

plan of activities and task distribution.

Ver1(as of 10/11/2019)



Updated plan of activities:



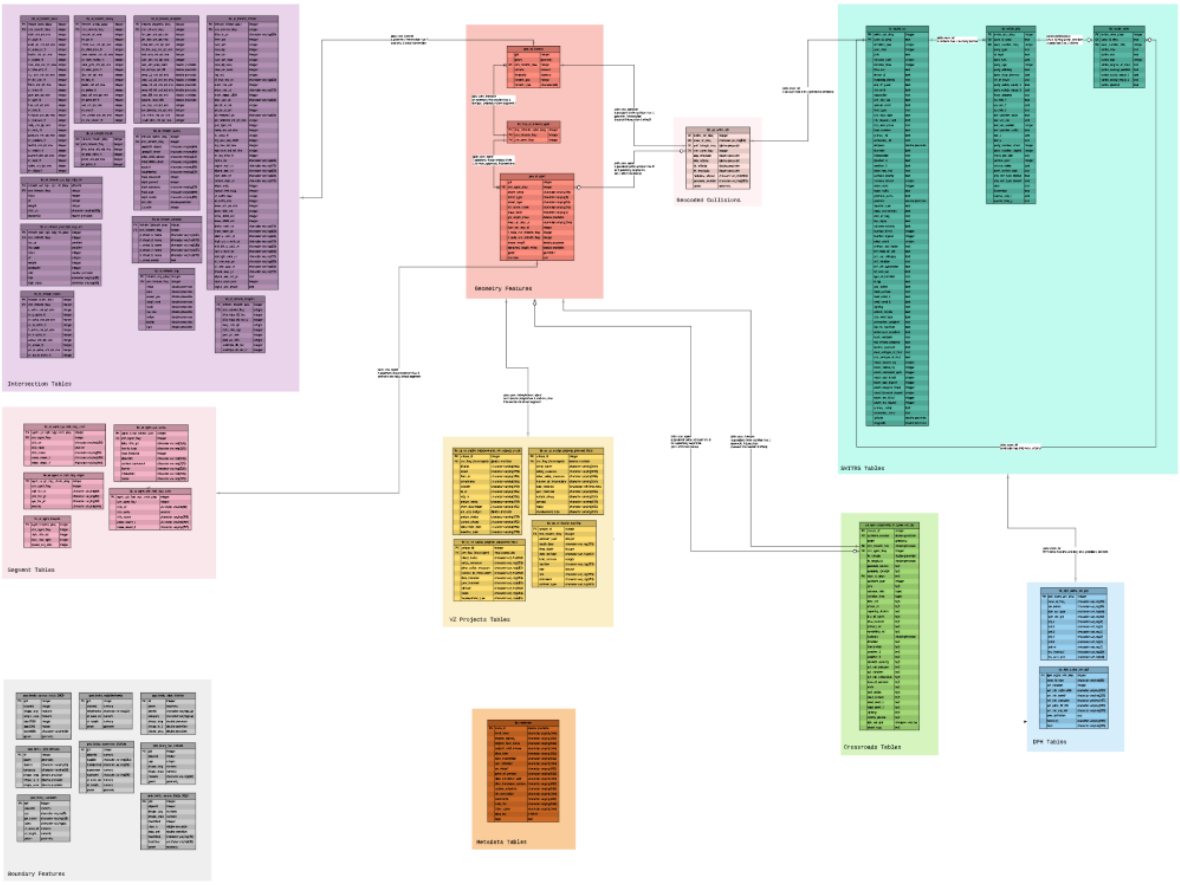
Section 6: Conclusion and Discussion

Future work:

Under the Construction

Appendix

TransBase EER Diagram



References

[1]V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[2]G. Bryden. Causal exploration of bike accidents in the bay area. *Open Journal of Safety Science and Technology*, 02:75–83, 01 2012.

[3]F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun. Anticipating accidents in dashcam videos. volume 10114, pages 136–153, 03 2017.

[4]P. Clarke, J. Ailshire, R. Melendez, M. Bader, and J. Morenoff. Using google earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health Place*, 16(6):1224 – 1229, 2010.

[5]J. DiGioia, K. E. Watkins, Y. Xu, M. Rodgers, and R. Guensler. Safety impacts of bicycle infrastructure: A critical review. *Journal of Safety Research*, 61:105 – 119, 2017.

[6]C. S. Hanson, R. B. Noland, and C. Brown. The severity of pedestrian crashes: an analysis using google street view imagery. *Journal of Transport Geography*, 33:42 – 53, 2013.

[7]K. Hara, V. Le, and J. Froehlich. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 631–640, New York, NY, USA, 2013. ACM.

[8]D. Jones and M. Jha. The effect of urban form on traffic accident incidence. pages 212–222, 01 2010.

[9]L. Li, J. Tompkin, P. Michalatos, and H. Pfister. Hierarchical visual feature analysis for city street view datasets. 2017.

[10]J. Liang and R. Urtasun. End-to-end deep structured models for drawing crosswalks.

In *ECCV*, 2018.

[11]T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2014.

[12]W. Lu, D. M. Scott, and R. Dalumpines. Understanding bike share cyclist route choice using gps data: Comparing dominant routes and shortest paths. *Journal of Transport Geography*, 71:172 – 181, 2018.

[13]W. E. Marshall and N. N. Ferenchak. Why cities with high bicycling rates are safer for all road users. *Journal of Transport Health*, 13:100539, 2019.

[14]D. Mohan, M. Khayesi, W. H. Organization, G. Tiwari, D. Indian Institute of Technology, and F. Nafukho. *Road Traffic Injury Prevention Training Manual*. Nonserial Publication. World Health Organization, 2006.

[15]W. E. Moritz. Survey of north american bicycle commuters: Design and aggregate results. *Transportation Research Record*, 1578(1):91–101, 1997.

[16]N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. Streetscore – predicting the perceived safety of one million streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 793–799, 2014.

[17]H. Summala, E. Pasanen, M. Räsänen, and J. Sievänen. Bicycle accidents and drivers' visual search at left and right turns. *Accident Analysis Prevention*, 28(2):147 – 153, 1996.

[18]K. Teschke, M. A. Harris, C. C. O. Reynolds, M. Winters, S. Babul, M. Chipman, M. D. Cusimano, J. R. Brubacher, G. Hunte, S. M. Friedman, M. Monro, H. Shen, L. Vernich, and

P. A. Cripton. Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. *American journal of public health*, 102(12):2336—2343, December 2012.

[19]M. Winters and K. Teschke. Route preferences among adults in the near market for bicycling: Findings of the cycling in cities study. *American Journal of Health Promotion*, 25(1):40–47, 2010. PMID: 2080

[20] SF Environment.org. Focus 2030: A Pathway to Net Zero Emissions. Climate Report. July 2019. https://sfenvironment.org/download/focus-2030-a-pathway-to-net-zero-emissions-climate-report-july-2019