Integrating OpenStreetMap with Google Street View using image processing and machine learning

Ashutosh Choudhary
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01002
ashutoshchou@umass.edu

Abstract

OpenStreetMap is rapidly expanding crowd-sourced data base of world map. This data, if integrated with visual information from Google Street View would be an immense store of information without manual crowd collection. This paper proposes a method of probabilistically evaluating values of heading or camera angle of Google Street View which points directly to the entity of interest whose latitude and longitude information is provided by OpenStreetMap. A classifier is pre-trained on a set of images for a fixed number of classes present in OSM textual data. This classifier is then used to predict the probability of a particular heading value for a given location marked by its geolocation information. The final accuracy of prediction of heading is obtained as 27%

1 Introduction

OpenStreetMap is a collaborative project to create a free editable map of the world. The quality of any map is judged by three measures: depth, coverage and timeliness. Depth means the scale, or level of detail of the map. Coverage is the two spatial dimensions i.e whether the map covers the entire planet or city. And timeliness is how often is the map updated. It may not be feasible for any commercial mapping entity to maintain high standards in each of those measures, specially, timeliness. Crowd sourcing, hence, seems to be a very lucrative and achievable alternative to traditional mapping. OpenStreetMap becomes extremely important and though in its nascent stages, already has an immense amount of highly detailed open source data available for the entire planet. This data is present in the form of nodes for any tagged place and has geo-location co-ordinates (latitude, longitude), and a large amount of tagged data. It however does not have any visual information (images) for those nodes. Google Street View provides an API to get this visual information based on latitude and longitude. It also, requires the camera angle, called ıheading parameter, to give accurate pictures of the entity in question. This paper tries to estimate this value of heading, given geospatial information and the type of entity from the OpenStreetMap. This is achieved by training a classifier model on the images of the classes of data available in OpenStreetMap, for example, church, police station or a restaurant. This classifier is then used to predict the probability of any given image to be from a particular class. Given a set of equally spaced images in the 360 angle plane, the image with highest probability score of the given class from OpenSourceMap is the estimated value of heading.

2 Related work

While the classification part of the problem described is outdoor entity recognition, It is, however, a scene recognition problem, as the entire image is to be classified and not each object within the image annotated. Xiao et al. [1] from MIT research established bounds on performance of scene recognition problem by taking into consideration 899 classes with 130519 images. The database

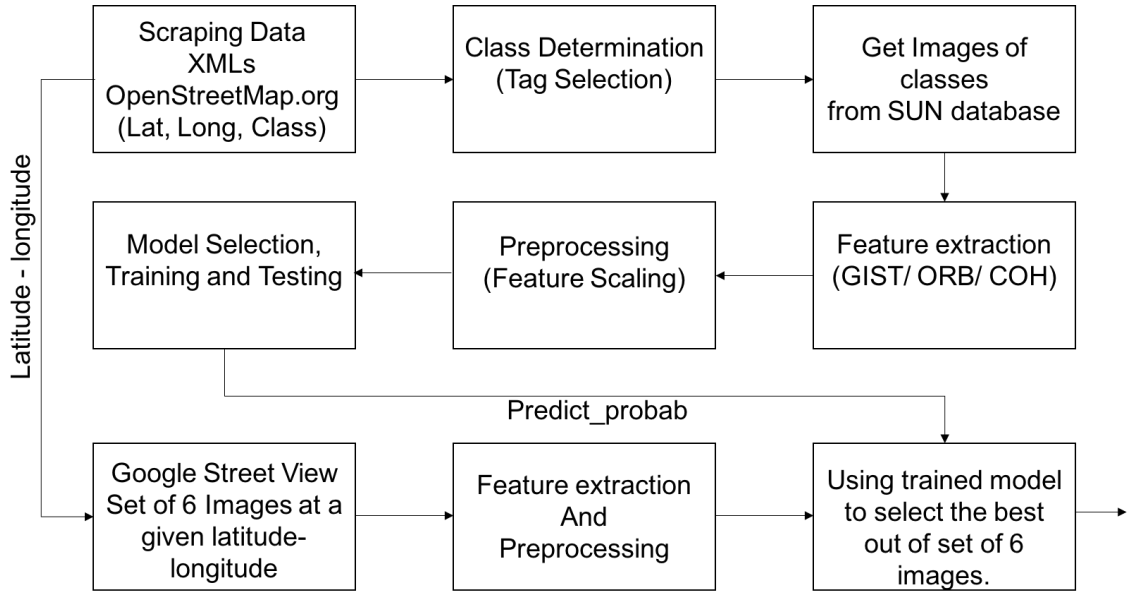

Figure 1: Flow of the model designed to estimate the heading value from the extracted data of OSM

created for this work, Scene UNderstanding or SUN is used in the current paper. This work evaluates a variety of image features with a 1 vs. all support vector machine. Among the features studied are GIST, Histogram of Gradient (HOG), dense SIFT, sparse SIFT, Local Binary Patterns (LBP), Color Histograms and line features. Human scene recognition performance was recorded for each of the scenes as 68% accurate. The best scene classification performance with all 899 features was obtained to be 38%.

Oliva and Torralba [2] in 2000 changed the approach to solving scene recognition. Previously, the approaches involved detection and recognition of each or most of the objects within a scene and then assembling them to get a holistic recognition of the scene. Features such as SIFT, FAST, BRIEF and ORB were predominantly used for this purpose. In their work on a holistic representation of the spatial envelope, they based their procedure on a low dimensional representation of the scene decomposed into perceptual dimensions like naturalness. openness, roughness, expansion and ruggedness. This low dimensional representation of a scene is conducive for classifiers like SVM and hence the GIST features as they were named by [2] became extremely popular for scene recognition along with SVMs. The current work paper evaluates GIST and SVM approach for scene recognition and finds it to have a better performance than all other model-feature set combination except neural networks.

The latest study on this topic by Zhou et al [5] in 2014 involves training Convolutional Neural Networks (CNNs) for learning high-level features. The task of feature extraction is left to the CNN and the features thus used are 'deep features'. After having trained CNN on Places database which is 60 times larger than the SUN database, accuracies of 50.0% and 66.2% were obtained on Places 205 and SUN 205 which are essentially test images of the corresponding databases. Comparing it with SVM trained on ImageNet + CNN features, accuracies were 40.8% and 49.6%. Hence, CNN shows a considerable improvement in scene recognition.

3 Data set

A total of 8766 tags with key as amenity were available in OSM database. Of these 223 were found to be wiki verified. Image availability of the top twenty of these label sorted by number of references in the world map was done and on the basis of conjoined score of the importance(number of references) and availability of their images in SUN[1] database, four classes were taken as a sample to represent the model proposed: parking lot, church, gas station, shops. Figure 3 For the second part of the model, which requires images from the same location taken at different angles, OSM data for Amherst, MA

was extracted and for each amenity or class of interest six pictures were extracted from Google Street View API by keeping latitude and longitude fixed and taking a picture every 60°from 0°to 300°.

Each image is represented in the form of a flattened three dimensional color histogram (COH), Oriented FAST and Rotated BRIEF (ORB) and GIST feature vectors. The three types of feature representation is solely for the purpose of experimentation with each of them used individually with no feature merging. This is due to the fact that each type of representation is chosen to depict different characteristics of image. While color histogram essentially provides no spatial information[3], ORB and GIST representations do. ORB is feature extraction by processing individual objects or regions[4] GIST descriptor[2] computes the output energy of a bank of 24 filters. The square output of each filter is then averaged on a 4 x 4 grid. It is a low dimensional holistic representation of scene based on a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness).

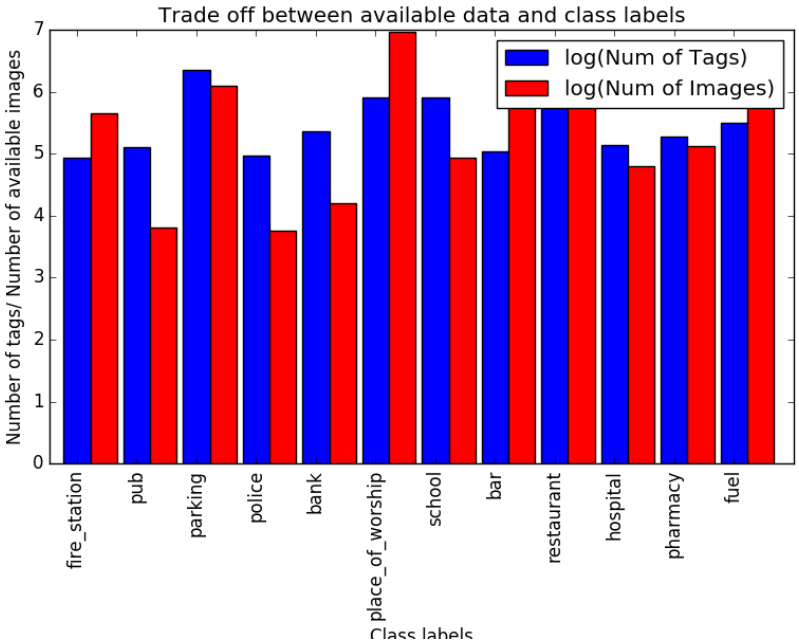

Figure 2: Class selection based on the popular tags and availability of image data

3.1 Data preparation

3.1.1 Data extraction from OSM, SUN and Google Street View

OpenStreetMaps metadata API is used to get class or label information which helps in selecting class label for the scope of this paper, as represented in Figure 1. SUN database is used to get the outdoor images of church, gas station, shopfront and parking lot. For testing the prepared and trained model in this heading evaluation problem, OSM data for the city of Amherst is used with 30 points of interest. For scaling it to bigger problems it is tried on the map of North America 185GB data by writing a fast iterating XML parser. Once, the information about entities is available with its location, Google Street View API is used to get six sliced panoramic images while labelling them with their heading angles in the process.

3.1.2 Computing COH, ORB and GIST feature vectors

The 3-D color histogram in RGB color space of the images is computed such that interactive information between the R, G, B planes is preserved. The bin size is taken to be (8x8x8) and therefore 512 features are available in COH feature vector. For ORB, number of features is taken as 10. Since, number of windows in each image is 32, the total number of ORB features are 320. GIST transformation gives a fixed number of features which is 960.

3.1.3 Preparing Training and Test data

Since the data obtained is streamed in a sequential manner without any inherent class information, during data preparation two key points are dealt with. One is to attach class information during scrapping and second is to randomly shuffle data before test and training data split. The shuffled data was then split into training data and test data. 80% of the images were randomly sub-sampled and marked as training data. The remaining 20% were marked as test data. The final data statistics found after data preparation is listed in Table 1

Table 1: Data set attributes

Amenity	No. of training images	No. of test images	GIST	COH	ORB
Church	847	211	960	512	320
Parking	356	89	960	512	320
Gas Station	266	66	960	512	320
Shop	587	147	960	512	320

4 Proposed solution

4.1 Pre-Processing

4.1.1 Normalization

Since the variance in the data values is extremely low, normalization of the feature vectors would help spread the data to extract more information. Also, ORB feature scales are different and hence feature vectors of zero mean and unit variance are obtained by normalization. The experiment with normalization clearly shows an improvement in cross validation scores particularly for SVMs.

4.2 Model Training

A set of eleven classifiers with default hyper-parameters are evaluated on the three feature sets. The eleven classifiers are: k-Nearest Neighbors, Linear Discriminative Analysis, Adaboost classifier, Multi-Level Perceptron, Quadratic Discriminative Analysis , decision tree classifiers, Naive Bayes, RBF-kernel SVC, Random Forest and Logistic Regression. This, is done in order to understand the inherent nature of feature space owing to lack of any semantic information about these features. In case of GIST features, due to high-dimensionality and relative small data set size, SVM seems to be good candidate. For color histogram (CoH), since data is linearly separable and also the classes intuitively would have quite varied histogram feature values, it is expected to have better cross validation scores than GIST or ORB. For example, grey and green colors will dominate in ṕarking lot classẃhile ćhurchćlass would have color in red, blue end of the RGB space. ORB on the other hand, depends on the object/ feature identification in an image to create an idea of whole scene. It may have accuracies lower than other two features spaces. These ideas were verified by the cross-validation scores obtained by the experiment as shown in figure 4.2.

Based on the assessment from figure 4.2, GIST features seem to perform well with SVM and and a multi-layer perceptron. COH performs well with Random Forest ensemble and AdaBoost classifier. ORB on the other hand does not perform well and is dropped from the rest of the experiment pipeline. Thus, SVM and Random Forest ensemble are the two core classification model which are considered for experimentation along with MLP perceptron and adaboost classifier for GIST and Color histogram features respectively.

4.2.1 SVC

A support vector machine is a discriminative classifier which uses sigmoid based parameter function. Despite having linear combination of weights and feature vector, SVM can efficiently perform non-linear classification using kernel techniques. The classification function of SVM:

$$f_{SVM}(x) = \text{sign}(w^T x + b)$$

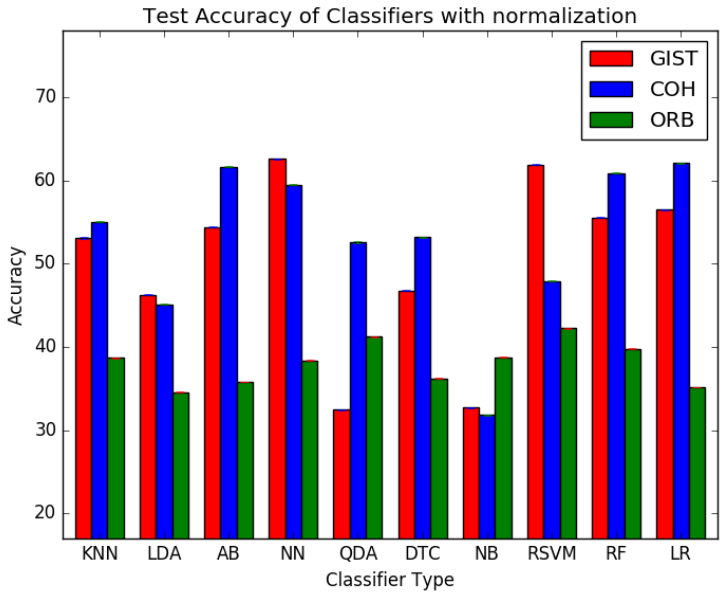

Figure 3: Evaluation of models and feature sets based on cross validation accuracy obtained

The class of data point x is determined by a linear combination of weight vector and the feature vector along with the threshold.

Hyper-parameters searched over: kernel type: 'rbf', 'linear', 'C' : 0.01, 0.1, 1, 10, 100, γ : $0.01, 0.1, 1, 10, 100$

4.2.2 Random Forest

Random Forest is an ensemble learning method that trains a number of decision trees and outputs the most frequent or mode of the individual outputs from each of these trained decision trees. Random Forest reduces the variance while keeping the bias low as well, and therefore correct the overfitting tendency of decision trees.

Hyper-parameters: No. of estimators: range(200, 300, 20); Max. depth: range(10, 40, 10) ;

4.3 Model selection

A 5-fold cross validation is used on the models described and their cross validation scores are used to determine the best set of hyper-parameters. Accuracy score is used as a metric for comparison between two classifiers or between two different set of features. Accuracy is chosen as a metric because the amount of data is almost evenly distributed in each class and hence a general metric like accuracy gives a good estimate of the model quality. Once a coarse grid search is done over the hyper-parameter ranges described in section 4, fine tuning of hyper-parameters is achieved by a grid search over the best parameters obtained +- 2 in logspace if appropriate. Coarse tuning of hyper-parameter for SVM is shown in figure 4.3

5 Experiments and results

Grid search is performed over a range of hyper-paramters with a 5-fold cross validation for chosen subset of 11 models. This is done for their corresponding feature types. After fine tuning hyper-parameters and choosing the set yielding best cross-validation scores, for SVM it is obtained as 'kernel': 'linear', 'C': 0.01, 'gamma': 0.01. Using such parameter values, accuracies are determined on test data and provided inTable 2. Also, a confusion matrix of the obtained results on test data is provided in Table 3

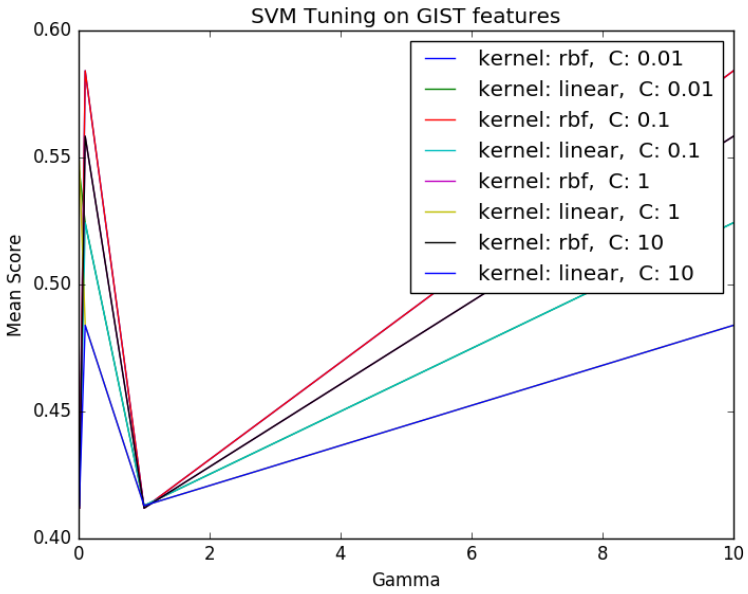

Figure 4: Scores obtained for grid search over hyper-parameters of SVM

Table 2: Accuracy scores on test data

Classification		
Model	Feature Type	Score
SVC	GIST features	0.59649122807
ML Perceptron	GIST features	0.615984405458
Random Forest	Color Hist	0.462890625
AdaBoost	Color Hist	0.46484375

Table 3: Confusion matrix for Test data results

	Confusion Matrices			
SVM + GIST	175	29	39	43
	9	9	3	5
	6	6	31	8
	21	22	16	91
MLP + GIST	149	12	20	18
	21	24	19	15
	12	1	33	4
	29	29	17	110
RF + CoH	204	49	79	114
	0	0	0	0
	0	0	0	0
	6	17	10	33
AB + CoH	202	35	63	86
	1	12	9	37
	5	14	16	16
	2	5	1	8

6 Discussion and Conclusion

SVM is used to then obtain the heading value of 120 datapoints, with each incorrect heading being produced as a 0 score and each correct one as +1. Based on this measure of score 27% times a correct heading value gets predicted. A case of correct and incorrect evaluation is shown in the figure 3 With the rise of Convolutional Neural Networks (CNNs) and huge databases, heading values can be predicted with higher accuracies using a similar pipeline and then can be used to create a supplemented data extension to OpenStreetMap until OpenStreetView reaches the potential of Google Street View.

Figure 5: Correct estimation of heading value as 180°. A church is clearly in frame

Figure 6: Incorrect estimation of heading value as 240°rather than 300°for a shopfront

References

[1] Xiao, Jianxiong, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. "Sun database: Large-scale scene recognition from abbey to zoo." In *Computer vision and pattern recognition (CVPR)*, 2010 IEEE conference on, pp. 3485-3492. IEEE, 2010.

[2] Oliva, Aude, and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope." *International journal of computer vision* 42, no. 3 (2001): 145-175.

[3] Chapelle, Olivier, Patrick Haffner, and Vladimir N. Vapnik. "Support vector machines for histogram-based image classification." *IEEE transactions on Neural Networks* 10, no. 5 (1999): 1055-1064.

[4] Rublee, Ethan, Vincent Rabaud, Kurt Konolige, and Gary Bradski. "ORB: An efficient alternative to SIFT or SURF." In 2011 *International conference on computer vision*, pp. 2564-2571. IEEE, 2011.

[5] Zhou, Bolei, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. "Learning deep features for scene recognition using places database." In *Advances in neural information processing systems*, pp. 487-495. 2014.