Image Classification to Determine the Level of Street Cleanliness: A Case Study

Abdullah Alfarrarjeh*, Seon Ho Kim†, Sumeet Agrawal‡, Meghana Ashok§, Su Young Kim¶, Cyrus Shahabi‖
Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089, USA
⟨ alfarrar*, seonkim†, sumeetag‡, meghanaa§, suyoungk¶, shahabi‖ ⟩@usc.edu

***Abstract*—The street cleanliness of big cities has a high impact on urban environment and health, so cities invest lots of effort to make their streets clean. With the recent advances in technology, it is feasible to develop smart systems for monitoring street cleanliness at scale such as automatic classification of street images to identify the cleanliness level by utilizing conventional classifiers (e.g., Naive Bayes and AdaBoost). However, these baseline methods do not provide a desired classification accuracy in practice. Thus, we propose a geo-spatial classification approach to enhance the classification accuracy. In particular, since the crowdsourced images are tagged with geo-location (i.e., GPS coordinates), we devise a novel framework with multiple local trained models exploiting the similarity of local images so that the proposed models learn better street image classification for each geographical region. This paper also presents a case study of street cleanliness classification using a large real-world geo-tagged image dataset obtained from Los Angeles Sanitation Department (LASAN). Experimental results showed that our framework was able to achieve an F1 score of around 0.9.**

***Index Terms*—street cleanliness, image-based classification, geo-aware classification, street-scene geo-tagged image**

I. Introduction

Urban streets are littered with waste deposits from natural sources and human activities, such as tree leaves, dumped items (e.g., furniture), and scattered trash. According to the "*Let's do it*" organization[1], there are 100 million tons of wastes dumped on streets around the world. Waste deposits negatively impact the public health, environment, economy, and tourism. Therefore, many cities devote a large annual budget and effort to enhance the cleanliness of their streets.

Different cities have adopted various approaches to enhance street cleanliness. Some are monitoring waste bins using various sensors [15] and image-based techniques [8]. Some others devise a street-rating system based on samples of geo-tagged images collected and evaluated by trained observers. An observer labels an image with a category (i.e., manual classification) which reflects the cleanliness level of a street. For example, New York adopted a rating system referred to as *ScoreCard* [21]. Los Angeles developed a smartphone app [13] which is used by city employees to collect street images and assess them manually based on a rating system. Then, street-rating systems enable identifying waste hotspots and help managing cleanup crews. As a result, it was reported that the percentage of unclean streets had been reduced by 82% in Los Angeles [13]. However, manual assessment of images has its limitations due to the cost of human labor and time. Hence, developing automatic classification of street scenes is essential for an efficient analysis of geo-tagged images collected by the city employees and, furthermore, scaling up the classification process of large-scale crowdsourced images by the public[2].

The focus of this study is automating the classification of street scenes based on their cleanliness level using a big real dataset of geo-tagged images from Los Angeles Sanitation Department (LASAN). A naive and straightforward approach is to generate one trained model with all images in a dataset. Towards this end, we investigate various image features and classifiers (e.g., SVM) to identify the best features and classifier to label an image based on predefined levels of street cleanliness. Furthermore, we observe that street scenes widely vary depending on their locations across a city (e.g., a street scene in downtown Los Angeles usually includes tall buildings with less vegetation while a street scene in Beverly Hills may include tall trees with no highrises) which might affect classification accuracy. Hence, we propose a classification scheme which aims at generating a set of efficient trained models to overcome the complexity of diverse street views associated with street scene locations that affect the accuracy of the classification mechanism. Our methodology relies on spatial partitioning techniques and utilizes existing machine learning classifiers to construct a local trained model per partition. Our experimental results show that the accuracy of local trained models generated using geo-spatial partitioning techniques outperforms that of one global trained model. These partitioning techniques have a set of parameters which affect the accuracy of the local trained models.

The remainder of this paper is organized as follows. In Section II, we review the related work. Section III introduces a set of definitions and provides background on image features and classifiers. Section IV describes our approaches for street cleanliness classification and Section V reports on our experimental results. Finally, in Section VI, we conclude.

II. Related Work

Systems for Street Cleanliness: The efforts for maintaining streets clean have followed two directions; waste bin monitoring systems and street-rating systems. Several street-rating systems were designed in various cities (e.g., New

[1] http://test.letsdoitworld.org/about

[2] The public can participate in collecting geo-tagged images using the spatial crowdsourcing mechanism [2], [11].

978-1-5386-5321-0/18/$31.00 ©2018 IEEE

York [21] and Los Angeles [13]). Some of these systems [13] are equipped with smartphone apps for the task of image collection and manual rating. Furthermore, Begur *et al.* [4] developed a smartphone app, deployed in the city of San Jose, for not only collecting but also analyzing images using object detection techniques to observe dumped wastes on streets. Regarding monitoring waste bins, Mahajan *et al.* [15] proposed an integrated sensor-based system (i.e., RFID and ultrasonic sensors) and Hannan *et al.* [8] classified the collected images by trash vehicles to identify the level of solid waste in bins.

Image-Based Classification: Recently, the advances in computer vision enable learning images more accurately; hence images become a reliable source of information in different domains. Various image-based applications are developed such as searching, object detection, and classification. In particular, image-based classification is adopted in serious problems related to human life and environment such as perceiving the safety level of a street view [16], land-use patterns (e.g., water body) from satellite images [17], disease diagnosis [6] and observing disaster situations from social networks [1]. Our approach is different from existing ones since we investigate the image-based classification with the geo-spatial information of images.

III. Preliminaries

A. Image Dataset

Definition 1 (Street-scene Geo-tagged Image)**.** An image I captures a street scene positioned at a geo-location (i.e., longitude and latitude). Hence, I is represented by two attributes: a visual scene of the image I_s and a geo-location I_g.

Based on the visual scene of an image I_s, each image is labeled with one of the five predefined image categories in this study: *bulky item*, *illegal dumping*, *encampment*, *overgrown vegetation*, and *clean*. LASAN uses these image categories in practice where each category implies the required resources and equipment to clean the area captured by the image[3]. Table I contains the description of each category in details and Figure 1 shows image examples of these categories.

Definition 2 (Street-scene Geo-tagged Image Dataset)**.** The dataset D is composed of n street-scene geo-tagged images ($D = \{I_0, I_1, \ldots I_{n-1}\}$).

B. Image Features

I_s can be represented by various feature descriptors. We selected three types of features which have been widely used in content-based image retrieval and image-based classification.

Color Histogram: Color is a visual cue which primarily helps humans to distinguish objects. The color histogram is produced by first discretizing colors into m bins, then counting the number of image pixels belonging to each bin color. A histogram based on pixels' color is a basic technique to represent the nature of an image. We adopted a histogram of color in the HSV color space for each image.

SIFT-based Bag of Words (SIFT-BoW): Scale-invariant feature transform (SIFT) [14] is a technique for detecting local interesting feature points in an image. The SIFT points usually lie on high-contrast regions of an image, such as object edges and corners. Street-scene images usually contain many objects, thus detecting their boundaries is essential to differentiate among the image categories. Following the success of the bag of word technique in text retrieval, Sivic and Zisserman [24] proposed the SIFT-based bag of word representation for an image which aggregates the distribution of local SIFT points.

Convolutional Neural Network (CNN) based Features: CNN is a hierarchical architecture consisting of a sequence of convolutional and sub-sampling layers followed by fully connected layers. The CNN architecture provides a rich image feature vector consisting of 4096 dimensions. It has been successfully used for various tasks in computer vision and multimedia (e.g., image classification [12] and retrieval [3]). Training a CNN from scratch often needs to learn a large number of parameters (e.g., 60 million), and subsequently requires both a large dataset and a long computational time. Alternatively, transfer learning is adopted which utilizes a pre-trained network on a different dataset and adapts it to another small dataset. This method is referred to as fine-tuning which has demonstrated satisfactory results in previous works (e.g., [18]). In literature, there are several available CNN architectures including Caffe [10] which we adopted. In our work, we customized the ImageNet pre-trained architecture provided with Caffe and fine-tuned the last three original fully-connected layers; fc6, fc7, and fc8. In particular, the fc8 layer is modified to represent a five-neuron layer (representing our image categories). The weights in these three layers are initialized from a zero-mean Gaussian distribution with standard deviation 0.01 and zero bias. The rest of layers are initialized using weights from the pre-trained model. The network is trained using stochastic gradient descent with the momentum of 0.9, gamma of 0.1 and a starting learning rate of 0.001 which we decay by a factor of 5e-4 every 10 epochs.

C. Background on Classifiers

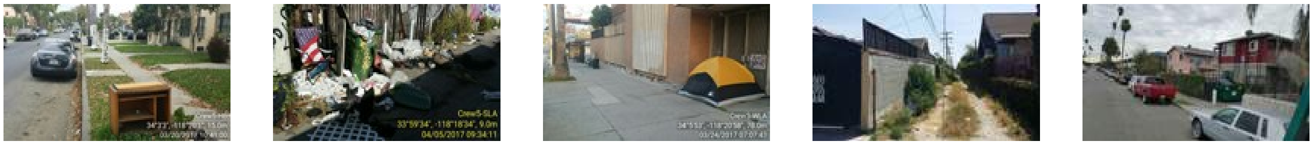
Extracted image features are fed into classifiers to learn models for categorizing street-scene images. Here, we review a set of well-known relevant classifiers.

First, the k-Nearest Neighbors (kNN) classifier is the simplest one which relies on the kNN search on the training dataset. The image class is identified based on the majority voting of its kNN. Second, Naive Bayes is a probabilistic classifier based on Bayes theorem which enables calculating a posterior probability for each class at prediction. Third, Support Vector Machine (SVM) is designed for binary classification where it constructs a hyperplane which divides the two classes with the largest margin. For multi-label classification, SVM is generalized in two schemes; one-versus-all and one-versus-one. Our work considers the one-versus-one scheme in which each pairwise class group is chosen as a two-class SVM each time. Fourth, the SoftMax classifier is a generalization of the binary form of Logistic Regression classifier [7]. SoftMax

[3] These image categories are defined by experts in the field (i.e., LASAN); thus can be used in other cities for developing their own frameworks.

TABLE I: LASAN's Categories of Street Scenes based on the Cleanliness Level

Image Category	Description
Bulky Item	There are some big items (e.g., couch, desk, mattress, and tire) thrown on a street.
Illegal Dumping	There is a pile of littered waste.
Encampment	There is a tent inhabited by people to live on the street.
Overgrown Vegetation	There is extra vegetation on a street and sidewalk.
Clean	The street is clean.



(a) Bulky Item (b) Illegal Dumping (c) Encampment (d) Overgrown Vegetation (e) Clean

Fig. 1: Image Examples for the Categories of Street Scenes based on the Cleanliness Level

converts the unnormalized values at a linear regression to normalized probabilities for classification.

There are known tree-based classifiers such as Decision Tree (DT). DT organizes a series of test conditions in a tree structure [20] where the internal nodes represent criteria for the attributes of each class, and the leaf nodes are associated with class labels. Random Forest includes a set of DTs, and the output of classification is defined by the leaf node that receives the majority of votes [22]. Another extension of DT is AdaBoost which builds a set of DTs adaptively. In the training phase of AdaBoost, a mis-classification of an instance is used to build a new optimized tree.

IV. Approaches

This section describes two proposed approaches for street scene classification; the first considers only the visual scenes of the image dataset (i.e., I_s) while the second considers both visual scenes (I_s) and geographical properties (I_g) since street scenes vary across different regions in a city.

A. Global Classification Scheme (GCS)

Among the classifiers discussed earlier, GCS approach constructs one single trained model using one of the well-known classifiers. The classifier learns the image features throughout the overall geographical region in a dataset. Thus, this approach does not consider the geo-properties of images and forms the baseline approach in this study. In the experiment section, we discuss the impact of the choice of both image features and the classifier on the street scene classification problem.

B. Geo-spatial Local Classification Scheme (LCS)

GCS suffers from data noise caused by the variety of street scenes. Hence, constructing a local trained model per sub-region enhances learning the visual characteristics of the surrounding areas in a sub-region. In particular, each local trained model focuses on learning the features of the categories without distraction caused by the features of different street views. Hence, the probability of correct image classification increases.

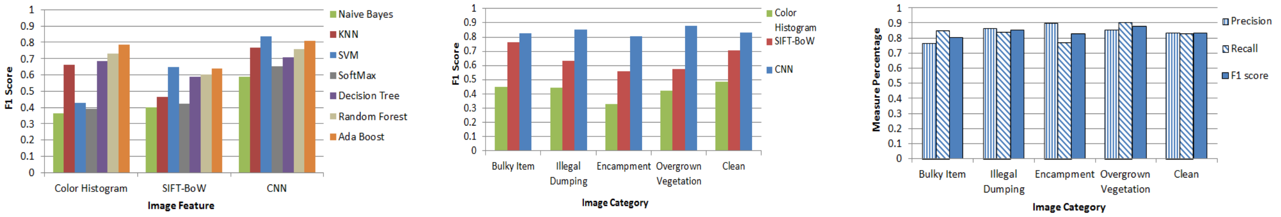
Constructing an efficient local trained model for a sub-region requires addressing two issues; having homogeneous views of streets in a sub-region and assuring a sufficient number of images in each sub-region for training purpose. Here, we explored two partitioning techniques to recognize sub-regions using the geospatial properties of images: Grid and Bucket Quadtreee [23]. Using Grid, the entire region in a datset is divided into fixed equal-sized cells ($Size$). This mechanism is simple but may result in imbalanced data distribution. Meanwhile, Quadtree partitions the region adaptively into four sub-regions until no sub-region contains more than a certain fixed number of images (Cap_{Max}). This mechanism produces sub-regions with varying area sizse and a large one potentially contains heterogeneous street scenes.

Design Approach: Optimally, we can create one local trained model per each cell (i.e., sub-region) generated by Grid or Quadtree. However, Grid varies in the number of images per cell and Quadtree varies in the size of cells. Moreover, to assign a cell to a classifier, the cell should contain a sufficient number of images, and the size of the cell should not be too large to avoid containing heterogeneous street scenes. Thus, we add constraints in creating a local model. A Grid cell should have at least a certain number of images (Cap_{Min}) and contain images from all street-cleanliness categories. The size of a selected Quadtree cell should not exceed the threshold value ($Size_{Max}$) while satisfying the other constraints on Grid cells. The cells which do not satisfy these criteria are assigned to a separate unified trained model.

V. Experiments

A. Dataset and Classifiers

We received around 22K geo-tagged street-scene images (D) from LASAN. These images were correctly labeled by LASAN experts based on the image categories for street cleanliness described in Table I and the distribution of images among categories is shown in Table II. Due to the imbalanced distribution across categories, a trained model can be biased to one of the categories. To overcome the problem of an imbalanced dataset in machine learning, researchers

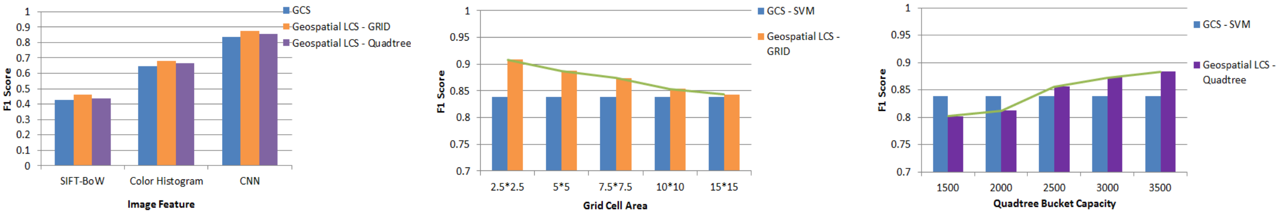


(a) W/ Various Classifiers and Image Features

(b) W/ SVM and Various Features w.r.t. Image Categories

(c) W/ Various Classifiers and CNN Features w.r.t. Various Measures

Fig. 2: The Efficiency of the Global Classification Scheme (GCS) Approach



(a) W/ Various Approaches and Image Features

(b) W/ Varying the Cell Size of Geo-spatial LCS - Grid

(c) W/ Varying the Cell Capacity of Geo-spatial LCS - Quadtree

Fig. 3: The Efficiency of the Geo-spatial Local Classification Scheme (LCS) using SVM

investigated various solutions [9]. One solution is a data-level approach by generating extra synthesized dataset through sampling mechanisms, which we adopted. In particular, we balanced only the training subset[4] by generating synthesized images using the Python Augmentor library [5][5] which applies image processing techniques (e.g., cropping and rotating) on a subset of images per category to obtain a balanced distribution. Meanwhile, we did not balance the testing subset of the dataset which was used for reporting the F1 score of the classification approaches. Then, all images are processed for feature extraction. For color feature extraction, images were processed in the HSV color space, and the color histogram was divided into 20, 20, and 10 bins in H, S, and V, respectively. For SIFT-BoW, to generate the dictionary of visual words, SIFT key points were extracted from 80% of D and clustered into 1000 clusters (using kMeans). For CNN, the Caffe architecture was fine-tuned using 80% of D. For the classifiers adopted in our approaches, we used the Python scikit-learn [19] library. All classifiers were trained on 80% of the dataset using 10-fold cross-validation. Table III shows the parameters and constraints of the *Geo-spatial* LCS approach, where the default values are underlined.

TABLE II: Dataset Distribution among Image Labels

Image Label	# of images
Bulky Item	12,315
Illegal Dumping	1,007
Encampment	886
Extra Vegetation	932
Clean	6,815

TABLE III: Parameter Values in *Geo-spatial* LCS

Parameter	Values
$Size$	2.5*2.5, 5*5, 7.5*7.5, 10*10, 15*15 mi^2
Cap_{Max}	1500, 2000, 2500, 3000, 3500 images
Cap_{Min}	1000 images
$Size_{Max}$	10*10 mi^2

B. Results

1) Impact of the Choice of Classifier and Image Features: Fig. 2a shows the F1 Scores of GCS for various combinations of classifier and image features. Classification with GCS achieved the best F1 score using the CNN image features due to the rich feature representation provided by CNN. Among the used classifiers, SVM achieved the best F1 score with both SIFT-BoW and CNN obtaining scores of 0.64 and 0.83, respectively, while AdaBoost achieved the best F1 score using color histogram with a score of 0.78. Since GCS shows the best F1 score with SVM using two types of image features, we further studied its F1 score among all image categories for street cleanliness. As shown in Fig.2b, GCS with SVM achieved an F1 score higher than 0.80 in all image categories obtaining the highest score with the “Overgrown Vegetation” category and the lowest score with the “Encampment” category. Across all image features SVM obtained the lowest F1 score with the “Encampment” category because “Encampment” images may be confused with the “Bulky Item” images. Fig. 2c provides the detailed performance of SVM classifier using various metrics (precision, recall, and F1 score) based on the CNN image features. Across all categories, both precision and recall showed good values. Hence, we chose to use F1 score since it depicts the harmonic average of precision and

[4] Balancing only the training dataset for generating non-biased trained model does not distort the nature and the reality of the problem.

[5] Our approaches are not restricted to the currently used augmentation technique and it can be easily replaced with other techniques.

recall values.

2) The Impact of Geo-spatial Local Classification Scheme: Fig. 3a shows the F1 Score of various mechanisms of Geo-spatial LCS compared with GCS using the SVM classifier while varying image features. Overall, both variants of *Geo-spatial* LCS achieved a better F1 score than that of GCS demonstrating that *Geo-spatial* LCS generates a set of regions where each region contains street images which share similar visual characteristics of streets; thus overcoming the scenes heterogeneity impact on street cleanliness classification.

3) The Impact of Varying the Cell Size of Grid : Fig. 3b shows the F1 scores of the *Geo-spatial* LCS *- Grid* approach using the SVM classifier while varying the cell size (i.e., $Size$) of the Grid compared with GCS. In general, *Geo-spatial* LCS *- Grid* achieved a better F1 score than that of GCS when a grid cell is small. When creating a local trained model for a small size cell (which has sufficient data based on the Cap_{Min} constraint), the trained model can distinguish the homogeneous visual characteristics of a small region; thus increasing the certainty of detecting waste objects. Meanwhile, a grid with large cells generates sub-datasets with heterogeneous street views; hence the F1 score of the *Geo-spatial* LCS approach decreases. For example, a cell of size 15×15 mi^2 covers the area of both downtown Los Angeles and its neighborhood; thus potentially contains street scenes with various visual characteristics. For the dataset, our approach obtained the best F1 score reaching to 0.90 when $Size$ is 2.5×2.5 mi^2.

4) The Impact of Varying the Cell Capacity of Quadtree: Fig. 3c shows the F1 scores of the *Geo-spatial* LCS *- Quadtree* approach using the SVM classifier while varying bucket capacity of Quadtree compared with GCS. In general, the F1 score of *Geo-spatial* LCS *- Quadtree* increases when increasing Cap_{Max} because the number of images for learning increases for each trained model. In particular, it was not able to outperform GCS with a small bucket capacity (e.g., 1500). However, *Geo-spatial* LCS *- Quadtree* outperformed GCS with an F1 score of 0.88 when Cap_{Max} = 3500.

VI. Conclusions

In this paper, we have studied an automatic classification of street scenes to identify their cleanliness level using a big real dataset obtained from Los Angeles Sanitation (LASAN) Department. We first investigated the use of various image features and classifiers following the cleanliness labels defined by LASAN. Due to the visual differences in street scenes across geographical regions, we have proposed a classification scheme with multiple local trained models utilizing the geo-spatial characteristics associated with the images. The best variant of our approach achieved an F1 score of 0.9.

For future work, we plan to extend our approaches to a) estimate an aggregated cleanliness level of a region that contains images of various cleanliness levels, and b) generalize our solution for video streams collected from garbage collection trucks operated by LASAN.

Acknowledgment

This research has been supported in part by NSF grants IIS-1320149 and CNS-1461963, the USC Integrated Media Systems Center, and unrestricted cash gifts from Google, and Oracle. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors.

References

[1] A. Alfarrarjeh, S. Agrawal, S. H. Kim, and C. Shahabi, "Geo-spatial Multimedia Sentiment Analysis in Disasters," in *DSAA*. IEEE, 2017, pp. 193–202.

[2] A. Alfarrarjeh, T. Emrich, and C. Shahabi, "Scalable spatial crowdsourcing: A study of distributed algorithms," in *MDM*, vol. 1. IEEE, 2015, pp. 134–144.

[3] A. Alfarrarjeh, C. Shahabi, and S. H. Kim, "Hybrid indexes for spatial-visual search," in *ACM MM Thematic Workshops*. ACM, 2017, pp. 75–83.

[4] H. Begur, M. Dhawade, N. Gaur, P. Dureja, H. Jeon, and J. Gao, "An edge-based smart mobile service system for illegal dumping detection and monitoring in san jose," in *SCI*. IEEE, 2017.

[5] M. D. Bloice, "Augmentor," 2016. [Online]. Available: http://augmentor.readthedocs.io

[6] E. Chen, P. Chung, C. Chen, H. Tsai, and C. Chang, "An automatic diagnostic system for ct liver image classification," *IEEE Trans Biomed Eng*, vol. 45, no. 6, pp. 783–794, 1998.

[7] J. B. Copas, "Binary regression models for contaminated data," *J R Stat Soc Series B Stat Methodol*, pp. 225–265, 1988.

[8] M. Hannan, W. Zaila, M. Arebey, R. A. Begum, and H. Basri, "Feature extraction using hough transform for solid waste bin level detection and classification," *Environ Monit Assess*, vol. 186, pp. 5381–5391, 2014.

[9] H. He and E. A. Garcia, "Learning from imbalanced data," *TKDE*, vol. 21, no. 9, pp. 1263–1284, 2009.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*. ACM, 2014, pp. 675–678.

[11] L. Kazemi and C. Shahabi, "Geocrowd: enabling query answering with spatial crowdsourcing," in *SIGSPATIAL*. ACM, 2012, pp. 189–198.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[13] S. Ladin-Sienne, "Turning dirty streets clean through comprehensive open data mapping," 2017. [Online]. Available: http://datasmart.ash.harvard.edu/news/article/turning-dirty-streets-clean-through-comprehensive-open-data-mapping-1001

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *INT J COMPUT VISION*, vol. 60, no. 2, pp. 91–110, 2004.

[15] K. Mahajan and J. Chitode, "Waste bin monitoring system using integrated technologies," *IJIRSET*, vol. 3, no. 7, 2014.

[16] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, "Streetscore-predicting the perceived safety of one million streetscapes," in *CVPR*, 2014, pp. 779–785.

[17] S. Ning *et al.*, "Soil erosion and non-point source pollution impacts assessment with the aid of multi-temporal remote sensing images," *J. of Environmental Management*, vol. 79, no. 1, pp. 88–101, 2006.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014, pp. 1717–1724.

[19] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.

[20] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[21] L. J. Riccio, J. Miller, and G. Bose, "Polishing the big appple: Models of how manpower utilization affects street cleanliness in new york city," *Waste Manag. Res*, vol. 6, no. 1, pp. 163–174, 1988.

[22] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *PAMI*, vol. 28, pp. 1619–1630, 2006.

[23] H. Samet, *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[24] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*. IEEE, 2003, pp. 1470–1477.