# A Vision-based Deep On-Device Intelligent Bus Stop Recognition System

**Gautham Krishna Gudur**[*]
Global AI Accelerator
Ericsson
gautham.krishna.gudur@ericsson.com

**Ateendra Ramesh**[*]
Computer Science and Engineering
SUNY at Buffalo
ateendra@buffalo.edu

**Srinivasan R**
Dept. of Information Technology
SSN College of Engineering
srinivasanr@ssn.edu.in

## ABSTRACT

Intelligent public transportation systems are the cornerstone to any smart city, given the advancements made in the field of self-driving autonomous vehicles – particularly for autonomous buses, where it becomes really difficult to systematize a way to identify the arrival of a bus stop on-the-fly for the bus to appropriately halt and notify its passengers. This paper proposes an automatic and intelligent bus stop recognition system built on computer vision techniques, deployed on a low-cost single-board computing platform with minimal human supervision. The on-device recognition engine aims to extract the features of a bus stop and its surrounding environment, which eliminates the need for a conventional Global Positioning System (GPS) look-up, thereby alleviating network latency and accuracy issues. The dataset proposed in this paper consists of images of 11 different bus stops taken at different locations in Chennai, India during day and night. The core engine consists of a convolutional neural network (CNN) of size ∼260 kB that is computationally lightweight for training and inference. In order to automatically scale and adapt to the dynamic landscape of bus stops over time, incremental learning (model updation) techniques were explored on-device from real-time incoming data points. Real-time incoming streams of images are unlabeled, hence suitable ground truthing strategies (like Active Learning), should help establish labels on-the-fly. Light-weight Bayesian Active Learning strategies using Bayesian Neural Networks using dropout (capable of representing model uncertainties) enable selection of the most informative images to query from an oracle. Intelligent rendering of the inference module by iteratively looking for better images on either sides of the bus stop environment propels the system towards human-like behavior. The proposed work can be integrated seamlessly into the widespread existing vision-based self-driving autonomous vehicles.

## CCS CONCEPTS

• **Human-centered computing → Ubiquitous computing**; • **Applied computing → Transportation**; • **Computing methodologies** → *Neural networks*; *Online learning settings*; *Active learning settings.*

## KEYWORDS

Bus Stop Recognition, On-Device Deep Learning, Computer Vision, Bayesian Active Learning, Incremental Learning

**ACM Reference Format:**
Gautham Krishna Gudur, Ateendra Ramesh, and Srinivasan R. 2019. A Vision-based Deep On-Device Intelligent Bus Stop Recognition System. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3341162.3349323

## 1 INTRODUCTION

Buses are a widespread mode of transportation that are used by the majority of the public. Recently, according to APTA 2019 [1], over 47% of people use buses as their preferred public transport mode in the United States, while 76% of buses have automatic bus stop announcements. In a country like India, conventionally the conductor of the bus intimates (whistles) when a bus stop arrives and announces its location aloud, while the driver halts the bus. This scenario still exists in most buses in India, however, with increase in automatic announcing mechanisms like pre-defined queues which have been widely followed – wherein the sequence of bus stops are stocked initially, this seems to be an easier alternative to the conductor's manual announcement of bus stops.

With the recent advancements and innovations in self-driving autonomous vehicles (buses) which are mostly based on state-of-the-art computer vision techniques, it becomes

[*] A part of this work was also done when affiliated with SSN College of Engineering.

*UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3349323

an increasing necessity to not only know the sequence of bus stops, but intelligently perceive where the bus stop exactly is, in order for the bus to halt at the right place. Many current bus stops in India, particularly rural and suburban bus stops predominantly do not have bounded or localized spaces/lanes. The surroundings of the bus stops also dynamically change and evolve over time. Moreover, bus routes are periodically revised and bus stops also change based on demand and traffic patterns. This calls for an intelligent transit system to not only identify bus stops automatically, but to incrementally adapt on-device to the new dynamic surroundings of the bus stops on-the-fly with minimal human intervention, and intimate passengers about the same. To reduce the overhead of capturing images during the whole route during inference (classification of bus stop), we propose that the images are captured only when speed of the bus is below a certain ideal threshold, only after which the inference engine would capture and classify an image. The real-time speed can be acquired from the speedometer of the bus and the ideal minimum threshold (10 km/hr for instance) is subject to locality and traffic conditions.

Global Positioning System (GPS) look-up can be used bus stops identification, however latency issues in the network, along with accuracy and privacy concerns make GPS look-up disadvantageous. Also, it might be hard to localize, identify and halt the vehicle right in front of the bus stop using GPS. Thus, in order to address and alleviate the aforementioned concerns, we propose a novel on-device vision-based solution. Given the significant developments in the field of vision-based deep learning in self-driving modes of transport [2], our system could seamlessly integrate with such systems with minimal processing power. This enables on-device feasibility and remote recognition in real-time.

Furthermore, given that the working of this paper focuses on real-time recognition of bus stops, real world unlabeled data necessitates the requirement of ground truth for various bus stop images. Hence, the authors utilize Bayesian Active Learning strategies by leveraging recent advancements of *Bayesian Neural Networks* to conveniently model uncertainties, making it easier to combine various acquisition functions to learn unlabeled data efficiently, thereby reducing the load of querying the oracle. Incremental Active Learning mechanisms employed on-device help overcome the dynamic nature of real-time bus stop data, and also enable scalability of bus stops. The proposed mechanism would in turn learn the most informative acquired images from uncertainty estimations, with minimal human intervention.

The motivation for building the inference system is to simulate the thinking of the human brain, i.e., a way that imitates or mimics a person's reaction when inquired about a particular bus stop. Typically, the individual would look to their left and right and then make a decision, and when the person is unsure of the bus stop, he/she would look for further frames in order to become more confident. This idea has been encompassed here, wherein the system would not classify a bus stop if not fully confident with only two images, which further augments its performance and intelligent decision making capabilities, in turn making it robust and efficient. The key contributions of this paper include:

- Proposing a dataset of bus stop images acquired from cameras placed atop a bus in few select locations in the city of Chennai, India, and a light-weight model for the same to perform on-device inference.
- Incorporating Incremental Learning capabilities to enable scalability, and dynamically adapt to evolving surroundings across various bus stops on-device. Data augmentation techniques to handle class imbalance for new classes are also studied.
- A study of *Bayesian Active Learning* using Bayesian Neural Networks to model uncertainties, by examining several acquisition functions to acquire labels on-the-fly and substantially reducing labeling load on oracle.
- An intelligent inference engine which mimics human-like thinking.

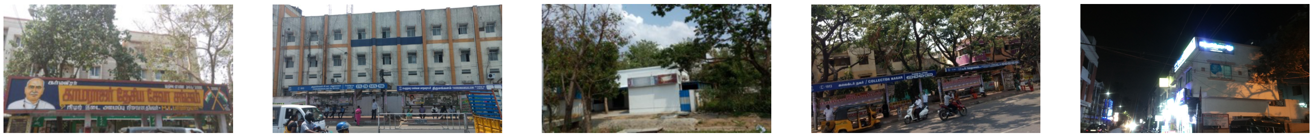
## 2 RELATED WORK

Rapid developments have been made in field of image recognition in tandem with intelligent and autonomous vehicles. However, in the scope of bus stop recognition, very few systems employ such learning models that are extremely capable of learning the dynamically evolving surroundings of bus stop over time.

The system proposed by Pan et al. [13] focuses on an image based (HOG algorithm with SVM), however this work is used for identification of a bus and its number, rather than the bus stop. The authors from [16] and [9] proposed intelligent bus stop identification systems using trajectories from GPS traces of bus routes in smartphones with impressive efficiencies. However, the use of GPS in systems increase the network latency and communication overheads.

In [5], the authors propose a system that recognizes painted patterns/messages on the road using a Kalman filter and pattern matching, however this system is computationally expensive. Moreover, road signs might not be a dependable means of pinpointing a bus stop in most places. Most previous works assume that the incoming bus stop data are labeled previously, which puts forth the requirement to label real-time data on-the-fly. Active Learning (AL) techniques effectively identify the most informative data points and query the oracle (user) for ground truth.

Traditional algorithms in AL [14], [4] are statistically proven for low-dimensional data, but do not generalize



**(a) 7H - Mogappair** **(b) Thirumangalam** **(c) Gurunath** **(d) Collector Nagar** **(e) Pari Salai (Night)**

**Figure 1: Examples of Bus Stops**

across deep neural networks, which are inherently high-dimensional. Approximations using uncertainty sampling has been an active area of research for many years now, with efficient Bayesian AL strategies [8] and [12] recently proposed. These proposed works primarily deal with image data, and help in identifying the most uncertain images to be queried by the oracle. Moreover, the authors from [10] propose Incremental Active Learning for wearable on-device scenarios, and our proposed work also takes motivation from the same.

The rest of the paper is organized as follows. We propose our dataset in Section 3 and our Bayesian model architecture in Section 4. Section 5 discusses about the various acquisition functions used for querying the oracle during Bayesian Active Learning. The baseline efficiencies for the model with existing and new classes with data augmentation are elucidated in Section 6. This is followed by systematic evaluation in resource-constrained Incremental Active Learning scenarios in Section 7. A novel intelligent inference mechanism is presented in Section 8, and Section 9 concludes the paper.

## 3 DATASET

The dataset primarily consists of images of 8 bus stops from Chennai, India, of which five are public bus stops and three are taken inside SSN College of Engineering (SSNCE) over different days. These images were acquired using two 5 MP cameras placed in opposite directions on buses during the day. In addition to these eight bus stops, three bus stops were captured during night. We propose using two different classifiers for day and night separately, for which a light sensor can be used to switch between daylight and night-time (including dark and overcast times).

For each bus stop during the day, 60 images were acquired for each side of the bus (left & right), thus, resulting in a total of 960 images. The classification model was trained with 720 images stratified across each bus stop, and the rest 240 were used for testing the same. The same amount of images per bus stop on either sides were acquired for the night. Figure 1 illustrates examples of different bus stops in the dataset. We focus on the results acquired during the day, and the same methodology and model can be scaled during night.

Furthermore, the images were resized to 32 × 32 × 3 and normalized (divided RGB pixel values by 255 for easier model convergence), in order to facilitate hardware-friendly computation with lesser memory overhead.

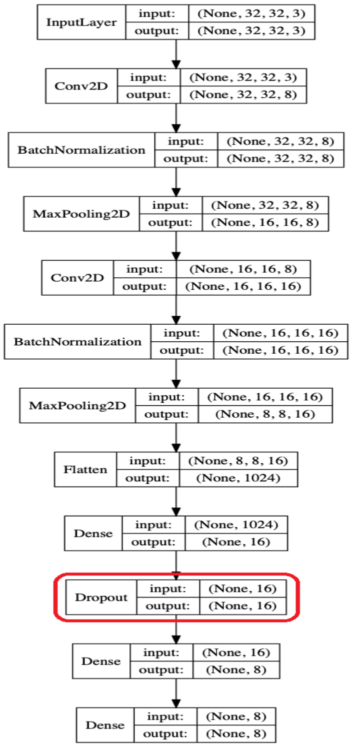

InputLayer input: (None, 32, 32, 3) output: (None, 32, 32, 3)
Conv2D input: (None, 32, 32, 3) output: (None, 32, 32, 8)
BatchNormalization input: (None, 32, 32, 8) output: (None, 32, 32, 8)
MaxPooling2D input: (None, 32, 32, 8) output: (None, 16, 16, 8)
Conv2D input: (None, 16, 16, 8) output: (None, 16, 16, 16)
BatchNormalization input: (None, 16, 16, 16) output: (None, 16, 16, 16)
MaxPooling2D input: (None, 16, 16, 16) output: (None, 8, 8, 16)
Flatten input: (None, 8, 8, 16) output: (None, 1024)
Dense input: (None, 1024) output: (None, 16)
Dropout input: (None, 16) output: (None, 16)
Dense input: (None, 16) output: (None, 8)
Dense input: (None, 8) output: (None, 8)

**Figure 2: *Bayesian Convolutional Neural Network (CNN)* Architecture which can model uncertainties**

## 4 MODEL ARCHITECTURE

Bus-Stop Recognition in real-time requires identification of discriminative features of the acquired bus stop scenes for efficient classification. In this paper, we utilize Convolutional Neural Networks (CNNs) which are powerful mechanisms for distinctive spatial representations and offer automatic, effective feature learning capabilities using a series of Convolutional and Pooling layers. The convolutional layers take in the images in three-dimensions and perform convolutional operations to the input sequence with various kernels (receptive fields) and desired amount of filters (feature maps).

The model consists of two stacked two-layered convolutional network comprising of 8 and 16 filters each, and receptive field size of 2x2. Each convolutional layer is followed by a Batch Normalization and a Max Pooling layer of size 2x2 each. This is followed by two Fully-Connected (FC) layers with 16 and 8 neurons with weight regularization (L2-regularizer with a weight decay constant), and ReLU activation functions. The *Dropout* regularization technique [17] is used between the FC layers with a probability of 0.3, and finally a Softmax layer is used for calculating the negative log-likelihood probability estimates. The categorical-cross entropy loss of the model is minimized using Adam optimizer, with a learning rate of $10^{-3}$, and implemented using the TensorFlow framework.

Dropout is used at both train and test times between FC layers for multiple stochastic forward passes (*T=10* in this paper), to sample the approximate posterior as stated in [7]. Since the weights in *Bayesian (Convolutional) Neural Networks* are probability distributions (Gaussian priors) instead of point estimates – equivalent to performing dropout for $T$ iterations, they can efficiently model uncertainty estimates, which can be used with existing deep acquisition functions for Active Learning. The model architecture can be observed from Figure 2.

## 5 ACQUISITION FUNCTIONS FOR ACTIVE LEARNING

Given a model $M$, real-time pool data $D_{pool}$, and inputs $x \in D_{pool}$, [8] states that an acquisition function $a(x, M)$ is a function of $x$ that the Active Learning system uses for inference of the next query point:

$$x^* = argmax_{x \in D_{pool}} a(x, M).$$

Acquisition functions are used in AL to quantify uncertainties and arrive at the most efficient set of data points to query from $D_{pool}$. We examine the following acquisition functions, whose detailed results are observed in Section 7:

*Bayesian Active Learning by Disagreement (BALD).* Information about model parameters are maximized under the posterior that disagree the most about the outcome [11].

$$\mathbb{I}[y, \omega | x, D_{train}] = \mathbb{H}[y|x, D_{train}] - E_{p(\omega|D_{train})}\left[\mathbb{H}[y|x, \omega]\right]$$

where $\mathbb{H}[y|x, \omega]$ is the entropy of $y$, given model weights $\omega$.

*Variation Ratios.* Utilizes the Least Confident method for uncertainty based pool sampling [6].

$$variation - ratio[x] := 1 - \max_y p(y|x, D_{train})$$

*Max Entropy.* Predictive entropy is maximized by appropriately chosen pool points. [15].

$$\mathbb{H}[y|x, D_{train}] := -\sum_c p(y = c|x, D_{train}) \log p(y = c|x, D_{train})$$

*Random Sampling.* Equivalent to choosing a random point from a uniform pool distribution.

## 6 BASELINE EXPERIMENTS AND RESULTS

Initially, only 7 bus stops are taken into consideration and called existing classes, while the eighth class (SOMCA bus stop) is treated as the new unseen bus stop for illustrating scalability.

### Existing Classes

The experiments are performed with 630 train images and 210 test images stratified across all seven classes. This is generally a one-time training from scratch on server for establishing the baseline accuracies, and it can be observed from Figure 3 that the training and testing modules give high accuracies of ~97% and ~96% respectively.

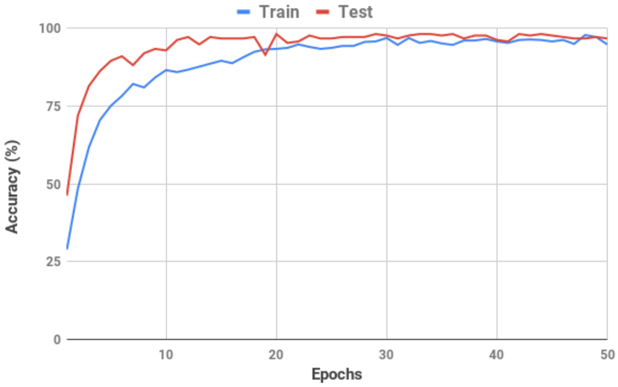


Figure 3: Baseline Train, Test Accuracies vs Epochs

### Data Augmentation for New Classes

When a new bus stop is being added to the route, it becomes a necessity for the model to scale and handle the incoming real-time data of the new class. The acquired number of images from the new bus stop is predominantly lesser than that of existing classes which might result in a class imbalance. Hence, image augmentation techniques are applied in order to ensure stratified training across all classes. Techniques like zoom, shear and rotation by small fractions are utilized for generating new images [3], which almost resemble the images acquired from a real-time camera.

Initially, we assume only 4 data points were collected from either side of the new bus stop. From Figure 4, it can be seen that the accuracy of the model with just the 8 data points, retrained from scratch gives an accuracy of 86.25%. However, the model with new bus stop data augmented to sufficiently higher images such that no class imbalance exists, achieves an accuracy of 96.7% which is a ~10.5% increase in accuracy.

This shows the effectiveness of data augmentation strategies for new classes, thereby ensuring scalability of bus stops.

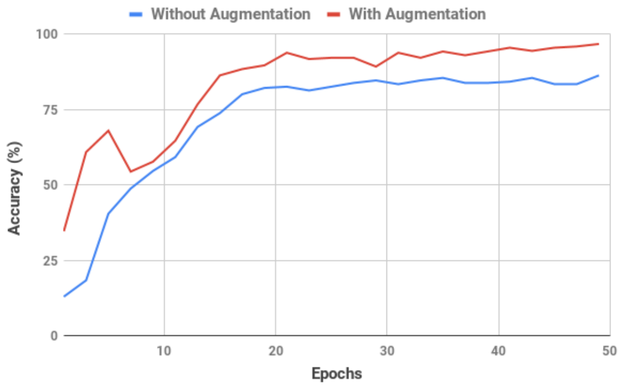


**Figure 4: With and Without Data Augmentation for New Class; Test Accuracies vs Epochs**

## 7 INCREMENTAL ACTIVE LEARNING

In order to reduce oracle's load in labeling the incoming real-time bus stop images, we perform Bayesian AL using various acquisition functions as mentioned in Section 5. The system is deployed on a Raspberry Pi 2 in real-time, and the stocked model weights are updated on-device in an incremental manner. In the event of there being a change in scene in an existing bus stop environment, incremental training will emphasize on learning the most recent and salient features of that bus stop. Moreover, when a new bus stop is introduced, the existing model with the newly learned weights are updated.

### Incremental Active Learning on Existing Classes

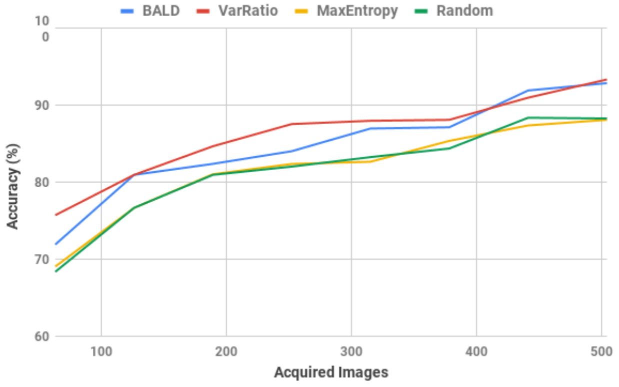


**Figure 5: Acquired Images vs Accuracy on existing classes**

The training data with existing 7 classes are split into pool ($D_{pool}$) and train ($D_{train}$) (80-20 ratio) for simulating the Bayesian Incremental Active Learning framework, as an approximation for real-world data. $D_{test}$ is used for evaluation purposes only. The initial accuracy with just 20% of train data is observed to be 64.28%. After Incremental Active Learning, various acquisition functions are utilized for evaluation of the most informative queries. From Figure 5, we can infer that Variation Ratios performs the best, achieving ~88% with just less than 250 data points (less than 50% of total $D_{pool}$), which is a good trade-off between accuracy and images actively acquired. Random Sampling has the least incremental accuracy as expected.

### Incremental Active Learning on Augmented Classes

A similar training mechanism (80-20% – $D_{pool}$-$D_{train}$ split) as that of Section 7 is followed for existing and augmented classes together as well, with $D_{test}$ used for evaluation. We can infer from Figure 6 with an effective trade-off between accuracy and acquired images that Variation Ratios again performs the best again, with a classification accuracy of ~90% with just ~180 images (~37% of total $D_{pool}$).

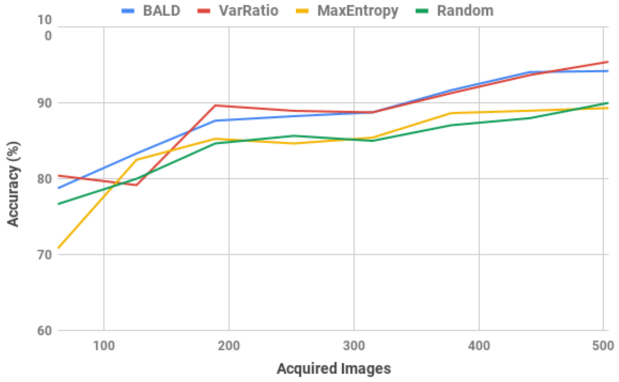


**Figure 6: Acquired Images vs Accuracy on augmented classes**

After the first acquisition iteration/experiment, the further incremental updates would typically require substantially fewer number of actively queried data points from AL to achieve on-par base classification accuracies of ~96%.

## 8 INTELLIGENT INFERENCE

*Inference Methodology.* The authors propose an intelligent inference approach that is built on human intuition. In this approach, the model classifies multiple iterative bus stop images acquired on demand as opposed to conventional classification. Let the number of images captured and classified during inference be $n$, which is initially set to 2 and capped at 10 ($2 < n < 10$). In order to facilitate intelligent decision making, we propose a *confidence factor* $\alpha$ – which is the ratio of mode of predicted classes to $n$. Just like a human brain, the system acquires images and classifies them iteratively until it can assure a confidence of at least $\alpha$. Typically, the threshold for $\alpha$ is set to a majority among the classified ($\alpha > 0.5$ for 2 images, $\alpha \geq 0.67$ for 3 images, and so on). In order to simulate real-time testing, a classifier is trained on the aforementioned 720 images and tested iteratively on examples at random from the test data.

The results showcased in Section 7 is a conventional way of testing with stratified splits during Incremental Active Learning. On the contrary, the accuracy of the proposed intelligent inference mechanism would steer the model towards near-100% accuracy, while the bus stop is deemed misclassified only when $n > 10$. However, when simulated across numerous trials, the performance of the model did not falter with the maximum value of $n$ reaching 5.

*Inference Engine.* The entire deep convolutional network model is sized 266 kB, thus making it ideal for effective on-device training and inference. Dropout at inference time is also used owing to the Bayesian nature of the model for active querying using uncertainty-based acquisition functions. The Incremental Active Learning module can be customized depending on the locality and bus usage characteristics, like periodic per trip update, per day/night update, etc. The various metrics and Inference times on Raspberry Pi 2 are observed in Table 1. Also, many such incremental updates would happen in real-time, and we can observe that for a total of *T=10* dropout iterations, ~12 seconds were used for querying most uncertain data points during one such acquisition iteration.

**Table 1: Time taken for Execution**

| Process | Computational Time |
|---|---|
| Inference time | 11 ms |
| Incremental Learning per epoch | ~1.7ms |
| Dropout iteration | ~1.2ms |

## 9 CONCLUSION AND FUTURE WORK

This paper proposes an intelligent on-device Bus Stop recognition engine which can accurately classify bus stop images in real-time, thereby eliminating the need for GPS network and latency overheads. By systematically optimizing upon different Bayesian Incremental Active Learning methodologies involving multiple acquisition functions, the proposed system adapts to the dynamically evolving nature of bus stops using periodic updates. Variation Ratios acquisition function is observed to perform the best during Active Learning, furthermore, data augmentation strategies are introduced for new bus stops to ensure scalability. Hence, a resource-friendly unified framework for Bus Stop Recognition which facilitates seamless integration with existing self-driving vehicles with image/video recognition capabilities is proposed in this paper. In future, we aim to create a unified (master) model across all buses traveling in the same route with periodic incremental updates from every bus, which enables information sharing across buses, thereby making bus stop recognition truly ubiquitous.

## REFERENCES

[1] American public transportation association (apta) 2019 fact book. https://www.apta.com/wp-content/uploads/APTA_Fact-Book-2019_FINAL.pdf, 2019. Online.

[2] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

[3] CireÅ§an, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation 22*, 12 (2010), 3207âĂŞ3220.

[4] Dasgupta, S., Kalai, A. T., and Monteleoni, C. Aanalysis of perceptron-based active learning. In *International Conference on Computational Learning Theory* (2005), Springer, pp. 249–263.

[5] Franke, U., and Ismail, A. Recognition of bus stops through computer vision. In *IV2003 Intelligent Vehicles Symposium. Proceedings* (2003), IEEE, pp. 650–655.

[6] Freeman, L. C. *Elementary applied statistics: for students in behavioral science.* John Wiley & Sons, 1965.

[7] Gal, Y., and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (2016), ICML'16, pp. 1050–1059.

[8] Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (2017), ICML'17, pp. 1183–1192.

[9] Garg, N., Ramadurai, G., and Ranu, S. Mining bus stops from raw gps data of bus trajectories. In *10th International Conference on Communication Systems Networks (COMSNETS)* (2018), IEEE, pp. 583–588.

[10] Gudur, G. K., Sundaramoorthy, P., and Umaashankar, V. Activeharnet: Towards on-device deep bayesian active learning for human activity recognition. In *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications* (2019), EMDL '19, ACM, pp. 7–12.

[11] Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).

[12] Kendall, A., and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems* (2017), pp. 5574–5584.

[13] Pan, H., Yi, C., and Tian, Y. A primary travelling assistant system of bus detection and recognition for visually impaired people. In *International Conference on Multimedia and Expo Workshops (ICMEW)* (2013), IEEE, pp. 1–6.

[14] Settles, B. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning 6*, 1 (2012), 1–114.

[15] Shannon, C. E. A mathematical theory of communication. *Bell system technical journal 27*, 3 (1948), 379–423.

[16] Srinivasan, K., and Kalpakis, K. Intelligent bus stop identification using smartphone sensors. In *14th International Conference on Machine Learning and Applications (ICMLA)* (2015), IEEE, pp. 954–959.

[17] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15* (2014), 1929–1958.