2018 15th Conference on Computer and Robot Vision

Systematic Street View Sampling

High Quality Annotation of Power Infrastructure in Rural Ontario

Kevin Dick, François Charih, Yasmina Souley Dosso, Luke Russell, James R. Green
Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada
{kevin.dick, francois.charih, yasmina.souleydosso, luke.russell, james.green}@carleton.ca

***Abstract*— Google Street View and the emergence of self-driving vehicles afford an unprecedented capacity to observe our planet. Fused with dramatic advances in artificial intelligence, the capability to extract patterns and meaning from those data streams heralds an era of insights into the physical world. In order to draw appropriate inferences about and between environments, the systematic selection of these data is necessary to create representative and unbiased samples. To this end, we introduce the Systematic Street View Sampler (S^3) framework, enabling researchers to produce their own user-defined datasets of Street View imagery. We describe the algorithm and express its asymptotic complexity in relation to a new limiting computational resource (Google API Call Count). Using the Amazon Mechanical Turk distributed annotation environment, we demonstrate the utility of S^3 in generating high quality representative datasets useful for machine vision applications. The S^3 algorithm is open-source and available at github.com/CU-BIC/S3 along with the high quality dataset representing power infrastructure in rural regions of southern Ontario, Canada.**

Keywords- Sampling Methods, Remote Sensing, Computational Complexity, Image Classification, Machine Vision, Open Source Software

I. INTRODUCTION

Remote sensing (RS) is defined as the acquisition of information about an element without the need of physical contact with the element under study. With the emergence of platforms containing a plethora of geographic data, the ability to remotely sense elements in their natural environments has become increasingly viable. While RS applications have traditionally been limited to satellite- or aircraft-based sensor technologies [1], a new instrumented vehicle is emergent: the self-driving car. A number of research groups are exploring road-based imagery as an augmentation or alternative data source for sensing the natural environment [2]–[4], broadening the definition of what constitutes an RS application.

Self-driving ground-based vehicles promise to provide an increasingly fine-grained spatio-temporal resolution in road-based imagery. While these data are primarily intended for the vehicular control systems, these images have the potential to also be used for the passive monitoring of roadside environments. Such "secondary use of self-driving vehicular imagery" is akin to the secondary use of electronic health records, where significant benefit can be derived from data collected for other purposes [5]. In anticipation of the wide-spread adoption and deployment of self-driving vehicles, for now, open-source and publically available data are considered.

Currently, Google Street View (GSV) is the most comprehensive and accessible platform to source imagery for vehicle-based RS applications. In recent years, Street View data have been used to study, compare, and contrast various facets of natural and urban environments at large scales. Naik *et al.* used a dense sampling of GSV images to quantify subjective perceptions of a neighborhood's physical appearance and how such appearance correlates with socioeconomic variables [6], [7]. Most recently, Gebru *et al.* demonstrated the feasibility of determining socioeconomic statistics and political preferences of the United States leveraging GSV imagery by detecting vehicular characteristics [8]. Beyond the deluge of machine vision applications, Google Street Maps has been broadly explored in modestly-sized applications of diverse fields such as biology for the study of animal habitats [4] and vegetation assessment [3], environmental criminology [2], environmental auditing for public health [9], and critical infrastructure protection [10].

When attempting to characterize a geographic region, the GSV imagery must be sampled in a systematic fashion to avoid the introduction of biases which may unintentionally alter inferred outcomes. A fundamental assumption of many statistical tests is the independence of observations. To obtain a sample representative of a population, all elements within that population must share an equal probability for sample selection to avoid biases. Here, we propose a framework to systematically sample the available GSV imagery in a given region to generate datasets uniformly sampled from all available data and in accordance with user-defined preferences.

The underlying GSV imagery is non-uniformly distributed due to several factors. Geographic topologies are inherently non-uniform, such as the distribution of naturally occurring bodies of water (*i.e.* hydrogeography) or man-made road infrastructure. These non-uniformities result in certain regions having variable proportions of road-ways. Furthermore, not all road-ways are guaranteed to contain GSV data leading to biases in regional representation.

978-1-5386-6481-0/18/$31.00 ©2018 IEEE
DOI 10.1109/CRV.2018.00028

While certain regions exhibit greater or lesser uniformity and density of GSV data, we sought to generate a location-independent framework enabling the collection of Street View data at any resolution for RS applications, capable of usage in conjunction with annotation frameworks.

Image data without context or ground-truth annotations are of limited value. Alas, generating ground truth annotations for image data is a repetitive, tedious, and labor-intensive task that can quickly become prohibitively expensive and lengthy. Given that systematic sampling of GSV images at high density over a large area can easily yield hundreds of thousands of images, one must consider alternative ways to collect high-quality annotations. Increased access to massive quantities of unlabeled data over the last decades has led to the emergence of crowdsourcing platforms such as CrowdFlower or Amazon's Mechanical Turk. The distributed nature of these platforms allows researchers to collect ground truth annotations at a fraction of the price and time that do-it-yourself solutions would require. These platforms are now widely used by machine learning researchers seeking to generate or annotate large datasets [11]–[14]. Amazon's Mechanical Turk (MTurk) provides a marketplace where *requesters* can post *Human Intelligence Tasks* (HITs) that *workers* will complete, typically for a low fee on the order of pennies per HIT. The MTurk platform provides requesters with a rich application programming interface (API), allowing for the implementation of complex annotation workflows and quality assurance strategies. Previous work has shown that collecting multiple annotations per image can counteract the effect of noisy or poor quality annotations. By using consensus annotations, one can achieve quality comparable to that of an expert [15]. The flexibility enabled by MTurk's API, combined with one's ability to review the reliability of the work completed, makes MTurk a powerful tool capable of adding significant value to a large set of unlabeled Street View images.

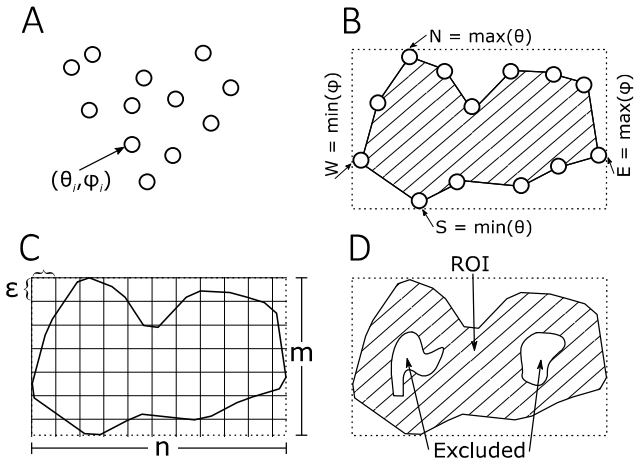

Figure 1. Pictorial Representation of the Notation and an Example Search Region. (A) illustrates an arbitrary set of coordinate points; (B) illustrates bounding set S_b and its bounds β_s; (C) depicts an arbitrary search grid parameterized by distance ϵ and resolution r; (D) shows how other bounding sets can be defined to create exclusion regions within a search area to specify the final region of interest (ROI).

GSV has afforded the world unprecedented insight into diverse environments. The S^3 framework introduced here enables researchers to sample these image data in a systematic and unbiased manner, permitting researchers to generate user-defined datasets for subsequent annotation and use in machine vision applications.

II. Methods

We developed the Systematic Street View Sampler (S^3), a software pipeline enabling individuals to systematically acquire images from the GSV service within a defined region and at a specified resolution. The S^3 serves as a free and flexible software framework to interface with various Google APIs to generate useful image datasets. While certain proprietary software tools have the capacity to study the geographical environment (*e.g.* ArcGIS) we sought to augment these functionalities by directly interfacing with Google APIs to obtain imagery representative of the queried location and increasingly democratize remote sensing applications. We first formulate our problem using set notation and planar geometries, describe the S^3 algorithm, and define it in relation to a new form of algorithmic complexity.

A. Notation

Following ISO 80000-2:2009 standards, we define a given latitude and longitude pair as (θ_i, ϕ_i) and a set of p such points as:

$$S = \{(\theta_1, \phi_1), (\theta_2, \phi_2), \dots, (\theta_p, \phi_p)\} \mid \mathrm{i} \in \{1,2, \dots, \mathrm{p}\}$$

A set S containing a given geographical region is a *bounding set*, S_b, and its bounding box is defined as $\beta_s = \{\max(\theta), \max(\phi), \min(\theta), \min(\phi)\}$ corresponding to the Northern, Eastern, Southern, and Western-most values of S_b. The bounding box delineates the *search area* of size mn where $m = \max(\theta) - \min(\theta)$ and $n = \max(\phi) - \min(\phi)$. This search area is discretized into a grid of equally spaced points (*i.e.* the grid vertices), ϵ meters apart in each cardinal direction and termed *search points*. Varying the value of ϵ results in a varying number of spaced points proportional to the square of the search area. Smaller ϵ results in a larger number of points and thereby a denser grid; larger values result in fewer and a sparser grid. The *resolution* is thus defined as $r = \frac{1}{\epsilon^2}$. Together, the total number of search points in the search area is approximated as

$$\left(\frac{m+\epsilon}{\epsilon}\right)\left(\frac{n+\epsilon}{\epsilon}\right)$$

which can be simply considered as mnr. See Figure 1 for a pictorial depiction of this notation.

B. The S³ Algorithm

Implemented in Python and leveraging four Google APIs, the S³ algorithm takes as input a bounding set of coordinates delineating the search region and a list of bounding sets of excluded regions. The latter are useful for eliminating internal regions of the search area which a user might wish to explicitly exclude from their analysis such as large bodies of water, cities, or forests. A number of additional parameters are also specifiable:

- Resolution, r, determined from ϵ
- Number of steps for the walk algorithm, w
- Image acquisition combinations, set of size c

The latter permits users to capture a set of images according to a combinatorial set of parameters. For example, specifying four different heading values, each at three different pitches would yield 12 unique images at each point ($c = 4 * 3 = 12$). These image combinations include *heading*, *pitch*, *image height*, *image width*, and *field-of-view*.

The S³ algorithm first determines the search bounds and then iterates over the search area at intervals of distance ϵ starting at the NW point, moving East- and South-wards until the final point, determined by the SE bounds. Each point in the search is evaluated in a number of ways, following a hierarchy of computations reliant upon API calls (Figure 2).

Each search point is first determined to be within the region of interest (ROI) or not, using a simple geometric query non-reliant upon external resources. If the point resides within the ROI, the first API call is made to Google Static Maps, requesting a satellite view image of size 1×1 pixels centered on (θ_i, ϕ_i). The pixel colour is compared to the corresponding pixel colour of the blue used by Google Maps to distinguish water from other geographic features (r,g,b = 163, 203, 255) to conclude whether or not the coordinate is coincident with known water features (and thereby removed from subsequent analysis). The search points are then submitted to Google Map Roads API to exploit the *Nearest Roads* feature, wherein the coordinates of the nearest road, (θ_i', ϕ_i'), are returned and used for subsequent analysis. Any (θ_i, ϕ_i) without roads in the vicinity are rejected. Having coordinates coincident with roads, we then exploit the Google Maps Javascript API to perform a "walk" of w steps. This generates a series of adjacent coordinates containing *StreetViewPanorama* objects using the *links* metadata field which points to neighbouring *StreetViewPanorama* objects:

$$\{(\theta_{i_1}', \phi_{i_1}'), (\theta_{i_2}', \phi_{i_2}'), \dots, (\theta_{i_j}', \phi_{i_j}'), \dots, (\theta_{i_w}', \phi_{i_w}')\}$$

This walk is initiated by examining the coordinates adjacent to (θ_i', ϕ_i') and selecting the point which maximizes the distance from (θ_i', ϕ_i'). Each subsequent point is similarly selected when multiple adjacent *StreetViewPanorama* objects are available and avoids cycles by registering previously visited coordinates. Finally, for each $(\theta_{i_j}', \phi_{i_j}')$ the GSV API is used to obtain images corresponding to the set of image combinations. By default, a single forward-facing (heading automatically obtained from the

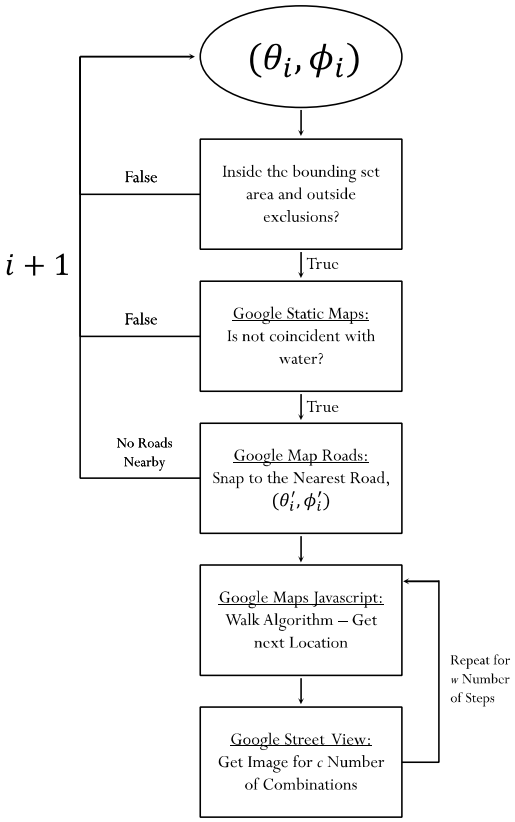

Figure 2. The S3 Algorithm Flowchart.

StreetViewPanorama object) image is captured with default parameters (pitch: 0, width: 640px, height: 360px, field-of-view: 90; $c = 1$) however various combinations of parameters can be combined to obtain a set of images for each step in the walk ($c \geq 2$).

In summary, a user with a Google API Key could utilize the S³ algorithm to define their own search area and tune parameters to obtain a systematically sampled representation of GSV images within that area.

C. Algorithmic Asymptotic Complexity

Following these definitions, we can derive the S³ algorithm complexity. Algorithmic asymptotic complexity is expressed in relation to the resources required to run the algorithm on a given input, as its size tends to infinity. Traditionally, computational time (*i.e.* time taken to run the algorithm) and computational space (*i.e.* amount of required memory) are used. From these expressions, one obtains an intrinsic understanding of the magnitude of resources necessary to compute a given input; we will refer to these as expressions of *internal complexity*. We can easily derive an expression of the internal complexity of the S³ algorithm as $O(mnrwc)$. However, given the dependency of the algorithm on external calls to the Google API services, a new limiting resource critically impacts the practicality of running the S³ on vast geographies or dense resolutions: the Google API quotas.

Here, we introduce a new expression of complexity, called *external complexity,* which is used to express S³ in relation to Google API calls as a limiting resource.

D. Complexity in the Web Services World

In this age of APIs, a new currency of complexity is emerging: "API Call Count". Web services follow two general usage models: free accounts with a limited daily call stack or a paid alternative limited only by one's cash stack. As a result, API calls hold an inherent value and the ability to appraise the "cost" of computing a given input size in relation to daily quotas or per-API-call pricing is an important consideration. This work utilizes four Google API services, each with a generous daily quota (TABLE *1*). Clearly, a search area and resolution resulting in 25,000+ search points is problematic, requiring more than one day's API quota to compute. This has important ramifications for the runtime of the algorithm since, irrespective of its internal time complexity, the process cannot complete until the API usage limit resets the following day without incurring expense; here, the external complexity is the bottleneck. External complexity can be optimized by maximizing the information retrieved with each API call. For example, the Google Maps Roads API, providing the *Nearest Roads* functionality, is limited to 2,500 calls per day and would be the computational bottleneck if naïvely submitting a single coordinate pair at a time. However, this API permits the submission of up to 100 individual coordinates in a single batch call. As a result, the 2,500 call limit can then process up to 250,000 independent coordinates per day, eliminating this bottleneck.

The number of calls to the API services are dependent on the size and resolution of the search area in addition to the geographical features present in the ROI which results in the conditional execution of each (θ_i, ϕ_i), what we consider as hierarchal processing in relation to each API service. The number of API calls made by S³ to each of the Google Static Maps, Google Maps Javascript, Google Maps Roads, and Google Street View services are defined as g_{sm}, g_{mj}, g_{mr}, and g_{sv} respectively. We further define p_{ROI} as the proportion of search points that fall within the ROI, S_{ROI}, in relation to the total number of search points, nmr:

$$p_{ROI} = \frac{\|S_{ROI}\|}{nmr}$$

While best- and worst-case external complexity estimates can be derived, the average-case external complexity provides the most practical expression of complexity.

1) Best-Case External Complexity

If no points fall within the ROI, we obtain a best-case complexity of $O(1)$ (*i.e.* no API calls are consumed). More practically, if the ROI fully comprises water geographical features, then each coincident search point is submitted to the Static Maps API and is rejected resulting in a best-case external complexity of $O(p_{ROI}r)$. These, however, are of little practical interest.

TABLE 1. GOOGLE API SERVICE QUOTAS

Google API Requests	Static Maps	Maps Javascript	Maps Roads	Street View
Per Day	25,000	25,000	2,500	25,000
Per 1000 Overage	$0.50	$0.50	$0.50	$0.50

2) Worst-Case External Complexity

If the ROI is equivalent to the bounding box area, no search point is coincident with water features, and each search point snaps to a nearby road containing complete Street View coverage, then we maximally utilize the Static Maps, Maps Roads, Javascript, and Street View APIs, resulting in a worst-case external complexity of $O(nmrwc)$. These conditions are not expected to be met for arbitrary areas, due to non-uniformity of geographic features and road distributions, save perhaps dense urban grid topologies such as New York City.

3) Average-Case External Complexity

The most practical expression of the S³ algorithmic external complexity is the average-case, given the non-uniformity of geographical features which varies the number of calls to each API. A scale-free term, $p \in \mathbb{R} \mid 0 \leq p \leq 1$, capturing the geographical variation within the search area, can be used to derive the average-case external complexity as a reweighting of the worst-case external complexity. From TABLE *1*, the maximum value for each of g_{sm}, g_{mj}, and g_{sv} is 25,000. It follows from the call hierarchy depicted in Figure 2 that:

$$g_{sm} \geq g_{mj} \geq g_{sv}$$

and since the value of g_{sm} is proportional to the water coverage over the ROI, the two geographical features which determine the average-case external complexity are the p_{ROI} and the p_{H_2O}: the proportion of points coincident with water over the search area. We define S_{H_2O} as the set of search points coincident with water and thus:

$$p_{H_2O} = \frac{\|S_{H_2O} \cap S_{ROI}\|}{\|S_{ROI}\|}$$

$$p = p_{ROI} - p_{H_2O}$$

We then obtain an average-case external complexity of $O(prwc)$.

Furthermore, term p can be empirically estimated using a Monte Carlo approach for varying values of r. The proportion of points in the ROI and not in water over all search points will converge on a single value of p with increasingly fine-grain resolutions.

III. DATASET GENERATION & ANNOTATION

To illustrate the use of the S³ algorithm for generating datasets of Street View images useful for machine vision applications, we generated a high-quality dataset of images representative of rural regions of southern Ontario containing power infrastructure and their annotated consensus ground truths. Such a dataset can be used to train convolutional neural networks to segment power-related infrastructure for passive monitoring applications.

A. *Standardized Sampling of Street View Images in Ontario*

The province of Ontario was selected given the concentration of infrastructure in its southern region and for being of sufficient size and diversity to capture typical road infrastructure representative of the rest of Canada. The set of coordinates comprising the Ontario bounding set were obtained from [16]. For the purposes of generating a dataset suitable for machine vision applications, we sought to minimize the number of confounding features in the environment. We focused on rural areas as a balance between overly-dense scenes in urban environments, and the sparsity of remote regions. To this end, we truncated the province of Ontario at the 45.7[th] latitude, given that half of the Canadian population lives below this line [17]; we consider it the Northern bound of our search area to limit remote regions. To avoid sampling urban regions, we excluded all cities having a population greater than 100,000. The bounding set of coordinates for each city was obtained using OpenStreetMaps [18] where available; 23 cities were part of the exclusion set. Various resolutions were used, $\epsilon \in \{1000, 2000, 5000, 10000\}$ in meters, and the finest-grained resolution $\epsilon = 1000$ was selected to produce our final dataset. To limit the similarity between scenes (prevention of classifier overfitting) we do not leverage the walk algorithm feature and instead select a path length of $w = 1$. For each $\left(\theta_{i_j}', \phi_{i_j}'\right)$ we obtain images in two configurations ($c = 2$): the forward-facing image and the rear-facing image, both with the default parameters (Figure 4). This maximizes the visual information available for a given coordinate while minimizing the duplication of such elements. In this way, images can be considered as independent, thereby facilitating segregation into a training and testing set of images. Figure 4Figure 3 illustrates this sampling strategy; Figure 3 the search area.

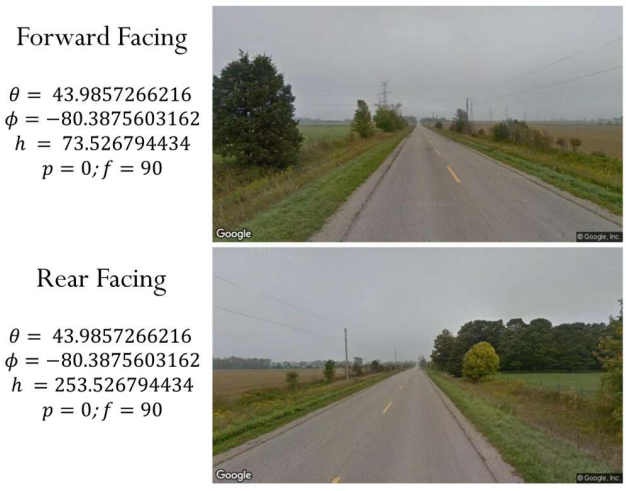

Figure 4 Sample Street View Images. Parameters h, p, and f are the *heading*, *pitch*, and *field-of-view* respectively.

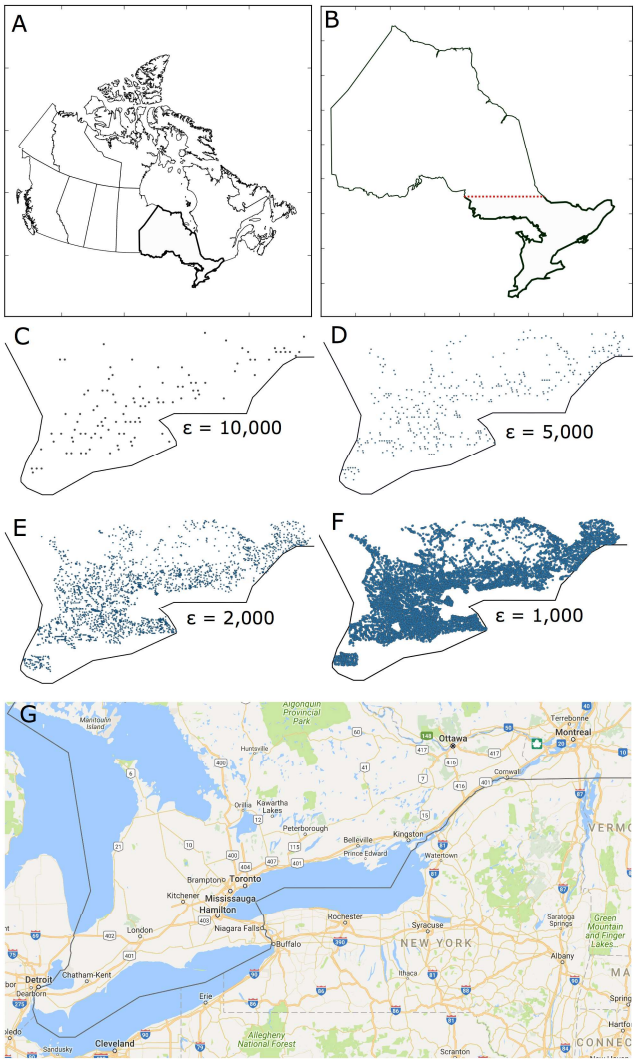

Figure 3. Visualization of the Standardized Sampling of Street View Images in Southern Ontario. (A) depicts Ontario in relation to the rest of Canada; (B) illustrates the Northern-bound of the search region (shaded) at the 45.7[th] latitude; (CDEF) plot the coordinates of the respectively fine-grained resolutions; (G) is the Google reference map of the region under study.

B. *Data Annotation Process*

Applying the S[3] algorithm to the region of interest previously described resulted in 18,883 unlabelled Street View images. We sought to investigate whether crowdsourcing can be leveraged to generate accurate annotations for large sets of images collected with the S[3] algorithm. To this end, we designed, implemented, and posted a power infrastructure classification task on Amazon's MTurk platform. Workers were asked to review training images illustrating what qualifies as an electric pole or power line, and to assign 18,883 Street View images to one of four classes: *contains both, contains electric poles*, *contains power lines*, *contains*

neither. To restrict access to reliable workers, we permitted only workers with at least 1,000 approved HITs and a minimum HIT approval rate of 98%. Workers were offered 0.0025 USD in exchange for every annotation. The images were displayed in an ergonomic and responsive interface that allows workers to zoom in on ambiguous regions and to rapidly assign classes by means of keystrokes, in an attempt to motivate workers to provide quality annotations. We collected additional metadata including a timestamp corresponding to the annotation time for every image, in addition to information indicating whether or not the worker had consulted the labelled training images prior to completing the annotation task.

Inter-rater agreement can provide insight into the degree of difficulty of a classification task and reveal whether one can expect to obtain consistent results through crowdsourcing-based annotation. To determine the inter-rater agreement among the workers for this power infrastructure classification task, we computed Fleiss' kappa statistic [19], suitable for classification tasks involving more than two raters, with the "raters" package for R [20]. We interpreted the resulting kappa statistic using Landis and Koch's scale [21], which classifies kappa values between 0.41-0.60 as indicative of "moderate" agreement, 0.61-0.80 of "substantial" agreement, values greater than 0.81 of "almost perfect" agreement.

Subsequently, to determine the degree of reliability of the annotations provided by the workers, we compared the consensus annotation to a set of ground truth annotations that we generated for a random subset of 200 images. More specifically, we compared the accuracy of the annotations for images where three, four, and five MTurk workers agreed against our high-quality ground truth annotations. The objective of this analysis was to gauge the benefit of outsourcing the work to four or five workers instead of three to generate a consensus annotation.

IV. Results & Discussion

The emergence of self-driving vehicles will generate tremendous potential for the passive monitoring of environments. In order to develop machine vision models capable of utilizing this new data stream, road-based imagery representative of the roadways driven by these newly instrumented vehicles is needed to train, validate, and test those models in anticipation of their deployment. To collect a representative sampling of imagery to capture the inherent variation within an ROI, we developed the S^3 to remotely sense these regions in a systematic fashion enabling the generation of datasets suitable to statistical studies. Prior work looking to achieve the same have necessitated the manual curation of coordinates or the use of browser emulators. The resulting images are warped from the spherical panorama thereby requiring the application of equirectangular projections [8]. The S^3 algorithm resolves both the need to develop sampling strategies and to post-process the resulting images and can consistently be applied within and between geographical regions.

A. External Complexity in the Age of APIs

The S^3 framework highlights the implications of a new limiting resource in the world of web services: API call count. This budget of calls (free or otherwise) has considerable bearing over the acquisition of images at various scales. As such, a new expression of algorithmic asymptotic complexity was required and which we introduce as *external complexity*. To the extent of our knowledge, we are the first to propose this form of complexity and believe it will have important ramifications given the recent growth in API service offerings. Notably, any software pipeline analogous to the S^3 will be dependent on pay-per-call rates or the quotas limiting their API usage and a similar expression of external complexity will provide an estimate of the extent of API usage for a given input. Internal complexity is optimized with the use of efficient representations of data and intelligent use of computational resources; the same applies to external complexity wherein maximizing the value of each API call and intelligently reducing the number of required calls results

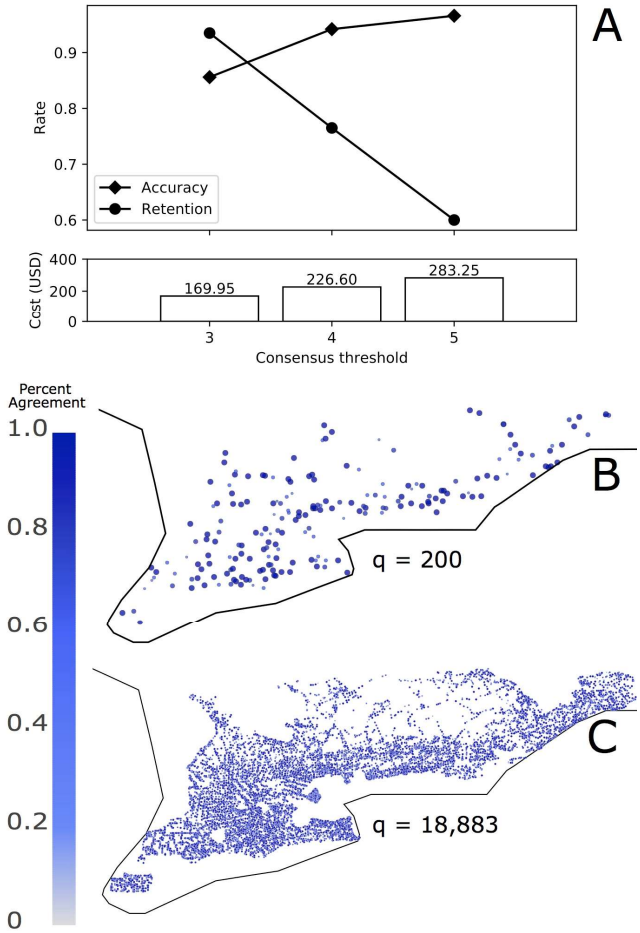

Figure 5 Summary of MTurk Worker Agreement. (A) depicts the data quality-cost trade-off over 200 images in relation to the ground truth annotations; (BC) illustrate the distribution of coordinates and percent agreement in the 200 and 18,883 image datasets (intra-panel bubble size proportional to the agreement). The difficulty of the annotation task appears uniformly distributed.

in improved asymptotic external complexities. For example, by obtaining the bounding set for large bodies of water in the ROI we preemptively avoid unnecessary calls to Google Static Maps, thereby reducing g_{sm}. Future work could investigate the modelling of this process using a Bayesian framework which could leverage priors to optimize the selection of query points. Sampling search points from a coarser resolution and utilizing information about whether the point falls within water, snaps to a road, or contains Street View imagery would optimize the selection of subsequent and adjacent points at finer resolutions.

B. Exploring the Data Quality-Cost Trade-Off

Redundancy, or collecting multiple annotations per instance, is a widely used strategy to ensure annotation quality on crowdsourcing platforms [22]. However, the optimal number of replicates to collect per instance depends on the difficulty of task. We posted a preliminary HIT on MTurk to gauge the number of replicate annotations required per image and estimate both the confidence in the resulting annotations and total cost for our power infrastructure classification task. We collected five sets of annotations for a subset of 200 randomly sampled images from the complete southern Ontario dataset (Figure 5). While setting the vote threshold to full consensus among the five raters only resulted in a consensus label for 60% of the images (the retention rate), the accuracy achieved jointly by the workers was exceptionally high (0.958). Compared to the accuracy achieved from setting the vote threshold to three (0.860), or four (0.923) workers, we justified the relatively small added cost of collecting five annotations per image, exemplifying the data quality-cost trade-off inherent to annotation tasks.

C. From Images to Insights: The Labelling Process

Using Amazon's MTurk platform, we were able to rapidly convert a large set of unlabelled 18,883 images into a fully labeled dataset. Every image had been classified by five distinct workers within nine hours of posting the HITs for a total cost of 283.25 USD. The workers spent an average of 16 minutes on a HIT comprising 200 images; this is equivalent to an hourly compensation of 1.88 USD/hour. The distributed nature of the MTurk platform enabled the inexpensive and rapid acquisition of annotations from 88 independent contributing workers. We evaluated the annotation performance in relation to the accuracy, retention rate, and total cost across three, four, and five worker class consensuses and determined that our estimates scaled as expected (TABLE 2). We note that the retention rate dropped considerably indicative of the volatility arising from the variability in the worker pool that we address using vote thresholding. The scalability and flexibility of the MTurk platform makes it ideal for use in conjunction with the S^3 algorithm which can collect GSV images at arbitrary resolution over regions of interest of arbitrary size. Altogether, these observations demonstrate the completion of quick and economic classification tasks, even with the imposition of stringent worker qualification criteria.

TABLE 2 MTURK WORKER METRICS OVER 200 IMAGE DATASET FROM THE FINAL ANNOTATION PROCESS OF 18,883 IMAGES

Consensus Threshold	Class	Precision	Recall	Accuracy	Retention Rate
3 of 5	Both	0.867	0.982	0.880	0.955
	Poles	0.800	0.462		
	Lines	0.833	0.625		
	Neither	0.952	0.909		
4 of 5	Both	0.928	0.990	0.942	0.775
	Poles	0.857	0.500		
	Lines	1.00	0.500		
	Neither	1.00	0.972		
5 of 5	Both	0.983	1.00	0.988	0.420
	Poles	1.00	0.500		
	Lines	N/A	N/A		
	Neither	1.00	1.00		

N/A: No consensus label for this class

D. Post-Annotation Quality Assessment

While MTurk's scalability makes it an ideal option for use along the S^3 algorithm, the issue of annotation quality remains. Fleiss' kappa statistic for our task indicates that the degree of agreement between the workers was "moderate" across all images in the dataset ($\kappa = 0.526 \pm 0.02$; 95% confidence intervals). This may suggest that the task is difficult or contains subjectivity. This figure is typical of previous estimates of inter-rater reliability for similar image classification tasks posted on the MTurk marketplace [15]. Given that a certain degree of variability in reliability is to be expected for any non-trivial or partially subjective classification task, we investigated the use of vote thresholding as a means to mitigate this effect (TABLE 2). Unsurprisingly, increasing the vote threshold required to produce a consensus annotation resulted in improved accuracy and precision, consistent with our preliminary analysis over a random subset of 200 images. At a 5-votes threshold, the recall over the two most prevalent classes (*contains both*, *contains neither*) is perfect. We note, however, that requiring perfect agreement among the five annotators on the same subset of images reduced the image retention rate to 42% over our subset of pre-annotated images (84 out of 200), and to 47% over the entire southern Ontario dataset (8,948 out of 18,883). This is lower than we anticipated based on our preliminary estimates, but still acceptable, given that these consensus annotations are of very high-quality. Furthermore, the ability to rapidly and systematically acquire large image datasets facilitates the decision to accept lower retention rates in favour of higher confidence annotations; the region can always be expanded or sampled at higher resolutions to extract additional images. All things considered, one can expect to convert large sets of unlabelled images collected through S^3 into high-quality and machine learning-ready datasets very cheaply and reliably by combining MTurk and vote thresholding. Both the S^3 framework and the MTurk templates are freely available at github.com/CU-BIC/S3 to facilitate the generation of user-specified datasets and problem-specific HITs.

E. Future Work

This work is open-source and easily extensible. Future studies will compare the generalizability of models trained on various resolutions of sampled data, explore the ramifications of external complexity, and evaluate alternate geographical sampling strategies such as the use of a Monte Carlo approach (II.D.3), a Bayesian framework (IV.A), or using hexagonal packing to obtain a denser coverage of an area.

V. Conclusion

The availability of Google Street View and the emergence of self-driving vehicles offer unprecedented access to street-based imagery with which we can make inferences about and between remotely sensed environments. To facilitate the systematic selection of this data to create representative and unbiased samples, we introduce the Systematic Street View Sampler (S^3) framework enabling researchers to produce their own user-defined datasets of Street View imagery, open-sourced and available at github.com/CU-BIC/S3. This work also introduces a novel expression of algorithmic asymptotic complexity, termed *external complexity*, in relation to a new limiting computational resource: API Call Count. We exemplify the usage of the S^3 in conjunction with the Amazon MTurk annotation environment for machine vision-related applications by generating a high-quality dataset capturing power-related infrastructure in southern Ontario, found at [23]. We anticipate the use of the S^3 framework in a broad range of fields and diverse applications.

Acknowledgment

This work was supported by Natural Resources Canada.

References

[1] F. F. Sabins, *Remote Sensing: Principles and Applications*. Waveland Press, 2007.

[2] C. Vandeviver, "Applying Google Maps and Google Street View in Criminological Research," *Crime Sci.*, vol. 3, no. 1, p. 13, Dec. 2014.

[3] E. Deus, J. S. Silva, F. X. Catry, M. Rocha, and F. Moreira, "Google Street View as an Alternative Method to Car Surveys in Large-Scale Vegetation Assessments," *Environ. Monit. Assess.*, vol. 188, no. 10, p. 560, Oct. 2016.

[4] P. P. Olea and P. Mateo-Tomás, "Assessing Species Habitat Using Google Street View: A Case Study of Cliff-Nesting Vultures," *PLoS One*, vol. 8, no. 1, p. e54582, Jan. 2013.

[5] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities.," *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2010, pp. 1–5, Mar. 2010.

[6] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, "Streetscore - Predicting the Perceived Safety of One Million Streetscapes." pp. 779–785, 2014.

[7] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo, "Computer Vision Uncovers Predictors of Physical Urban Change," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 29, pp. 7571–7576, Jul. 2017.

[8] T. Gebru *et al.*, "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 50, pp. 13108–13113, Dec. 2017.

[9] C. M. Kelly, J. S. Wilson, E. A. Baker, D. K. Miller, and M. Schootman, "Using Google Street View to Audit the Built Environment: Inter-Rater Reliability Results," *Ann. Behav. Med.*, vol. 45, no. S1, pp. 108–112, Feb. 2013.

[10] K. Dick, L. Russell, Y. Souley Dosso, F. Kwamena, and J. R. Green, "Deep Learning for Critical Infrastructure Resilience: A Case Study," *Submitt. to ASCE J. Infrastruct. Syst.*, 2018.

[11] P. Nakov *et al.*, "Developing a Successful SemEval Task in Sentiment Analysis of Twitter and Other Social Media Texts," *Lang. Resour. Eval.*, vol. 50, no. 1, pp. 35–65, 2016.

[12] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," pp. 1–9, 2016.

[13] A. Karpathy and F. F. Li, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015.

[14] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8828, no. c, pp. 1–14, 2017.

[15] S. Nowak and S. Rüger, "How Reliable are Annotations via Crowdsourcing," *Proc. Int. Conf. Multimed. Inf. Retr. - MIR '10*, p. 557, 2010.

[16] Statistics Canada, "Census Subdivision Boundary File, Reference Guide," 2017.

[17] Statistics Canada, "Census Profile," 2016.

[18] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ." 2017.

[19] J. L. Fleiss, "Measuring Nominal Scale Agreement Among Many Raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[20] P. Quatto and E. Ripamonti, "raters: A Modification of Fleiss' Kappa in Case of Nominal and Ordinal Variables." 2014.

[21] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data on JSTOR," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[22] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality Management on Amazon Mechanical Turk," *Proc. ACM SIGKDD Work. Hum. Comput. - HCOMP '10*, p. 64, 2010.

[23] K. Dick, F. Charih, and J. Green, "High Quality Annotations of Power Infrastructure in Rural Ontario," doi:10.5683/SP/0YOVH1, Scholars Portal Dataverse, 2018.