Preprint *Krylov V.A., Dahyot R. (2019) Object Geolocation from Crowdsourced Street Level Imagery. In: Alzate C. et al. (eds) ECML PKDD 2018 Workshops. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11329. Springer, Cham. https://doi.org/10.1007/978-3-030-13453-2_7*

Object Geolocation from Crowdsourced Street Level Imagery *

Vladimir A. Krylov and Rozenn Dahyot

ADAPT Centre, School of Computer Science and Statistics,
Trinity College Dublin, Dublin, Ireland
{vladimir.krylov,rozenn.dahyot}@tcd.ie

Abstract. We explore the applicability and limitations of a state-of-the-art object detection and geotagging system [4] applied to crowdsourced image data. Our experiments with imagery from Mapillary crowdsourcing platform demonstrate that with increasing amount of images, the detection accuracy is getting close to that obtained with high-end street level data. Nevertheless, due to excessive camera position noise, the estimated geolocation (position) of the detected object is less accurate on crowdsourced Mapillary imagery than with high-end street level imagery obtained by Google Street View.

Keywords: Crowdsourced street level imagery · object geolocation · traffic lights.

1 Introduction

In the last years massive availability of street level imagery has triggered a growing interest for the development of machine learning-based methods addressing a large variety of urban management, monitoring and detection problems that can be solved using this imaging modality [1, 2, 4, 5]. Of particular interest is the use of crowdsourced imagery due to free access and unrestricted terms of use. Furthermore, Mapillary platform has recently run very successful campaigns for collecting hundreds of thousands of new images crowdsourced by users as part of challenges in specific areas all over the world. On the other hand the quality of crowdsourced data varies dramatically. This includes both imaging quality (camera properties, image resolution, blurring, restricted field of view, reduced visibility) and camera position noise. The latter is particularly disruptive for the quality of object geolocation estimation which relies on the camera positions for accurate triangulation. Importantly, crowdsourced street imagery typically comes with no information about spatial bearing of the camera nor the

* This research was supported by the ADAPT Centre for Digital Content Technology, funded by the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and the European Regional Development Fund. This work was also supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No.713567.

Preprint *Krylov V.A., Dahyot R. (2019) Object Geolocation from Crowdsourced Street Level Imagery. In: Alzate C. et al. (eds) ECML PKDD 2018 Workshops. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11329. Springer, Cham. https://doi.org/10.1007/978-3-030-13453-2_7*

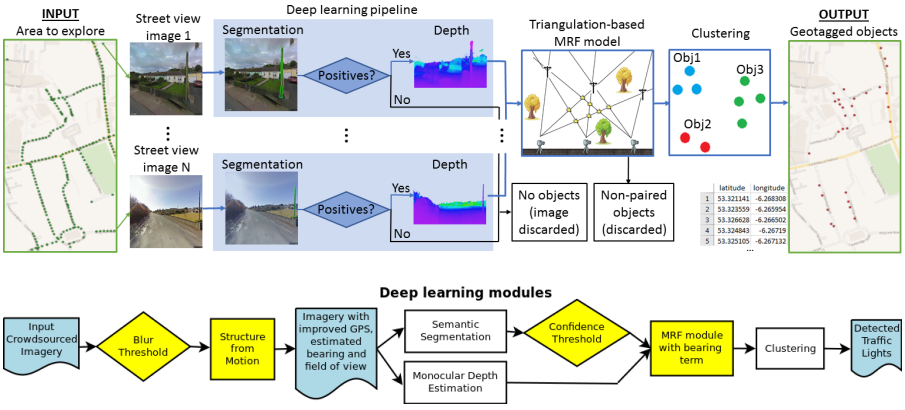

Fig. 1. Top: The original street level image processing pipeline proposed in [4] for object geolocation. Bottom: The modified pipeline with yellow components inserted to process crowdsourced street level imagery.

information about the effective field of view (i.e. camera focal distance), which requires estimation of these quantities from the image data.

The expert street level imaging systems, like Google Street View (GSV), ensure comparable data quality by using calibrated high-end imaging systems and supplementing GPS-trackers with inertial measurement units to ensure reliable camera position information, which is of critical importance in urban areas characterized by limited GPS signal due to buildings and interference. Here, we modify and validate the object detection and geotagging pipeline previously proposed in [4] to process crowdsourced street level imagery. The experiments are performed on Mapillary crowdsourced images in a study case of traffic lights detection in central Dublin, Ireland.

2 Methodology

We rely on the general processing pipeline proposed in [4], with semantic segmentation and monocular depth estimation modules operating based on custom-trained fully convolutional neural networks on street level images (Fig. 1). A modified Markov Random Field (MRF) model is used for fusion of information for object geolocation. The MRF is optimised on the space $\mathcal{X}$ of intersections of all the view-rays (from camera location to object position estimation via image segmentation). For each intersection location x_i with state z_i ('0' discarded, '1' included in the final object detection map), the MRF energy is comprised of several terms. The full energy of configuration $\mathbf{z}$ in $\mathcal{Z}$ is defined as sum of all energy contributions over all sites in $\mathcal{Z}$:

$$\mathcal{U}(\mathbf{z}) = \sum_{\forall x_i \in \mathcal{X}} \Big[c_d u_d(z_i) + c_c u_c(z_i) + c_b u_b(z_i) \Big] + c_m \sum_{\forall x_i, x_j \text{on the same ray}} u_m(z_i, z_j),$$

Preprint *Krylov V.A., Dahyot R. (2019) Object Geolocation from Crowdsourced Street Level Imagery. In: Alzate C. et al. (eds) ECML PKDD 2018 Workshops. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11329. Springer, Cham. https://doi.org/10.1007/978-3-030-13453-2_7*

with parameter vector $C = (c_d, c_c, c_b, c_m)$ with non-negative components subject to $c_d + c_c + c_b + c_m = 1$. The unary term $u_d(z_i)$ promotes consistency with monocular depth estimates, and the pairwise term $u_m(z_i, z_j)$ penalizes occlusions. These are defined as in [4]. To address the specific challenges of the crowdsourced imagery the other two terms are modified compared to [3, 4]:

• A second unary term is introduced to penalize more the intersections in the close proximity of other intersections (inside clusters):

$$u_c(z_i|\mathcal{X}, \mathcal{Z}) = z_i \Big[\sum_{\forall j \neq i} I(||z_i - z_j|| < C) - C \Big],$$

where I is the indicator function. Practically, the fewer intersections are found in C meters vicinity of the current location x_i, the more it is encouraged in the final configuration, whereas in intersection clusters the inclusion of a site is penalized stronger to discourage overestimation from multiple viewings. This term is a modification of high-order energy term proposed in [4], and has the advantage of allowing the use of more stable minimization procedures for the total energy.

• The crowdsourced imagery is collected predominantly from dashboard cameras with a fixed orientation and limited field of view (60-90 degrees). Hence, a unary bearing-based term is added to penalize intersections defined by rays with a small intersection angle because these are particularly sensitive to camera position noise. This typically occurs when an object is recognized several times from the same camera's images with a fixed angle of view (in case of dashboard camera, as the vehicle is approaching the object the corresponding viewing bearing changes little). In case of several image sequences covering the same area this term stimulates mixed intersections from object instances detected in images from different sequences. The term is defined as:

$$u_b(z_i|\mathcal{X}, \mathcal{Z}) = z_i(1 - \alpha(R_{i1}, R_{i2})/90), \quad x_i = R_{i1} \cap R_{i2},$$

with $\alpha(R_{i1}, R_{i2})$ — the smaller angle between rays R_{i1} and R_{i2} intersecting at x_i.

Optimal configuration is reached at the global minimum of $\mathcal{U}(\mathbf{z})$. Energy minimization is achieved with Iterative Conditional Modes starting from an empty configuration: $z_i^0 = 0, \forall i$, see in [4].

3 Experimental study and conclusions

We demonstrate experiments on Mapillary crowdsourced image data. We study the central Dublin, Ireland, area of about 0.75 km^2 and employ the 2017 traffic lights dataset [3] (as ground truth). All together, 2659 crowdsourced images are available collected between June 2014 and May 2018. We first remove the strongly blurred images identified by weak edges (low variance of the response to Laplacian filter), which results in 2521 images. We then resort to Structure from Motion (SfM) approach, OpenSfm (available at `https://github.com/mapillary/OpenSfM`) developed by Mapillary, to adjust camera positions and recover estimates of image bearing, field-of-view for cameras. This results in 2047

Preprint *Krylov V.A., Dahyot R. (2019) Object Geolocation from Crowdsourced Street Level Imagery. In: Alzate C. et al. (eds) ECML PKDD 2018 Workshops. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11329. Springer, Cham. https://doi.org/10.1007/978-3-030-13453-2_7*

Fig. 2. Examples of successful and failed traffic lights segmentation on Mapillary data.

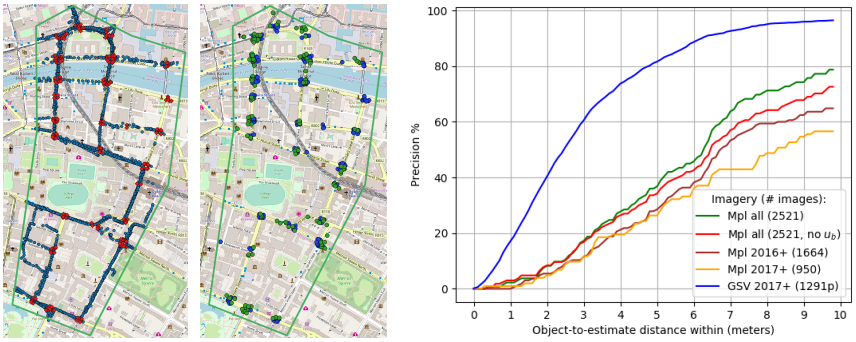

Fig. 3. Left: Dublin TL dataset (♦) in 0.75 km^2 area inside green polygon, and Mapillary image locations (●). Center: detection on Mapillary (●) and on GSV (●) imagery. Right: Precision plots as function of distance between estimates and ground truth.

images post-SfM, with the rest being discarded due to failure to establish image matches using ORB/SIFT image features. The image resolutions are 960x720 (12%), 2048x1152 (34%), and 2048x1536 (54%), these are collected from cameras with estimated fields of view ranging from 58 to 65 degrees. Object detection is performed at the native resolution via cropping square subimages. Pixel level segmentations are aggregated into 1180 individual detections, of which 780 with mean CNN confidence score of above .55 after Softmax filter, see examples in Fig. 2. In this study contrary to [4] we adopt a threshold based on the CNN confidence due to variation in detection quality from different camera settings and imaging conditions. In the reported experiments, the energy term weights are set to $c_d = c_m = 0.15, c_b = 0.3, c_c = 0.4$, $C = 5$ meters in the u_c energy term.

To compare the performance of the proposed method we also report the results of traffic lights detection on GSV 2017 imagery (totaling 1291 panoramas) in the same area. The object recall reported on Mapillary (GSV) dataset reaches 9.8% (51%) at 2m threshold (ground truth object is located within such distance form an estimate), 27% (75%) at 5m and 65% (91%) at 10m. As can be seen in Fig. 3 the coverage of the considered area is not complete and several traffic light clusters are not covered or by very few Mapillary images. This caps the possible recall to about 94% on the given dataset. The precision is plotted for increasing object detection radii in Fig. 3 (right) for the complete Mapillary dataset (inclusive of 2521 images) and smaller subsets to highlight the improvement associated

Preprint *Krylov V.A., Dahyot R. (2019) Object Geolocation from Crowdsourced Street Level Imagery. In: Alzate C. et al. (eds) ECML PKDD 2018 Workshops. ECML PKDD 2018. Lecture Notes in Computer Science, vol 11329. Springer, Cham. https://doi.org/10.1007/978-3-030-13453-2_7*

with increased image volume. The latter is done by restricting the years during which the Mapillary imagery has been collecting: 950 on or after 2017, 1664 on or after 2016, out of 2521 total images inside the area. It can be seen that the introduction of the bearing penalty u_b improves the detection and the precision grows with larger image volumes. Our preliminary conclusion after using crowdsourced imagery is that in high volume, these data can potentially allow similar detection performance but with a potential loss on geolocation estimation accuracy.

Future plan focuses on the analysis of multiple sources of data (e.g. the mixed GSV + Mapillary, Twitter, as well as fusion with different imaging modalities, like satellite and LiDAR imagery) and scenarios to establish the benefits of using mixed imagery for object detection and position adjustment with weighted SfM methods.

References

1. Bulbul, A., Dahyot, R.: Social media based 3d visual popularity. Computers & Graphics **63**, 28 – 36 (2017)
2. Hara, K., Le, V., Froehlich, J.: Combining crowdsourcing and google street view to identify street-level accessibility problems. In: Proc. SIGCHI Conf. Human Factors Computing Syst. pp. 631–640. ACM (2013)
3. Krylov, V.A., Dahyot, R.: Object Geolocation using MRF-based Multi-sensor Fusion. In: Proc. IEEE Int Conf. Image Process.air (2018)
4. Krylov, V.A., Kenny, E., Dahyot, R.: Automatic discovery and geotagging of objects from street view imagery. Remote Sens. **10**(5) (2018)
5. Wegner, J.D., Branson, S., Hall, D., Schindler, K., Perona, P.: Cataloging public objects using aerial and street-level images — urban trees. In: Proc IEEE Conf on CVPR. pp. 6014–6023 (2016)