Automated Estimation of Burial Site Occupancy from Satellite Imagery using Crowd-Counting Models Trained with Drone Imagery Annotations

Selemani Mmbaga[1]
Selemani.Mmbaga@marquette.edu

David Kaftan[2] Kirsty Weitzel[3] Rob Chew[3]

Masabho Milali[2] Joseph Milanga[4] Kasturi Bhamidipati[2] Frey Assefa[2]

Rashaad Hasan[5] Heeva Torabi[6] Rajesh Vedanthan[2] Samson Kiware[4]

Richard Povinelli[1] Anna Bershteyn[2]

[1]Marquette University
[2]NYU Grossman School of Medicine
[3]Research Triangle Institute
[4]Ifakara Health Institute
[5]American University of Sharjah
[6]Cameron Heights Collegiate Institute

Abstract

Detecting excess mortality during epidemics is particularly challenging in areas with weak civil registration systems. Satellite images of burial sites can provide a useful estimate, but current methods often depend on slow and labor-intensive manual work. We suggest an automated system to estimate burial site occupancy from satellite images, using a computer vision model trained with drone-guided annotations. We manually annotated drone images from 14 burial sites in Dar es Salaam, Tanzania, and aligned them with satellite images using manually identified landmarks, RANSAC homography estimation, and perspective warping. We then adapted three computer vision algorithms - CSRNet, DM-Count, and P2PNet - initially developed for crowd-counting to count burial sites occupancy. We performed 10-fold cross-validation on 337 satellite images that included 8,888 annotated burial plots, P2PNet performed the best, achieving $R^2 = 0.88 \pm 0.03$*, MAE = 7.6, and RMSE = 11.8. Model performance improved greatly when trained and evaluated on drone imagery (*$R^2 = 0.97 \pm 0.02$*), indicating performance loss caused by transferring*

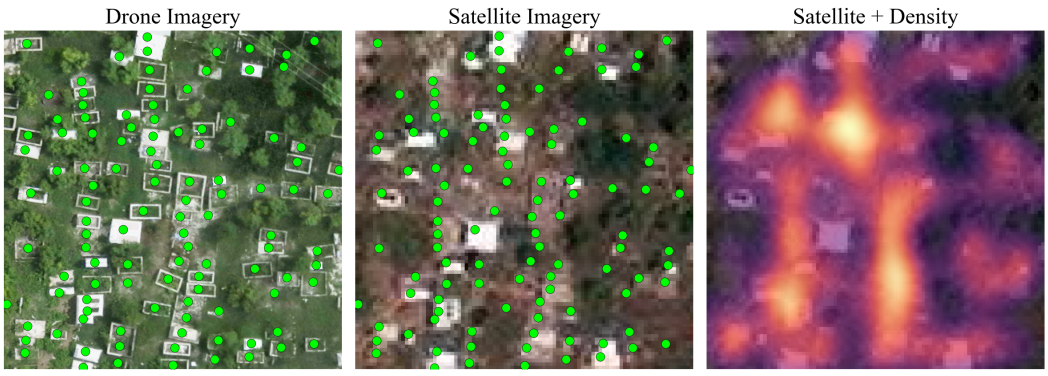

Figure 1. Example annotation of drone imagery, transfer to satellite, and generation of density maps. Graves in drone imagery were annotated using LabelBox and then transferred to satellite imagery using manual registration, RANSAC homography estimation, and perspective warp. Density maps were generated using a Gaussian filter, $\sigma = 5$ pixels.

labels to satellite imagery. Performance also varied significantly across different burial sites. Overall, these results indicate that using drone-assisted annotation along with automated counting could provide a scalable method for automated monitoring of excess mortality using burial sites from satellite imagery.

1. Introduction

Detecting and monitoring high-casualty epidemics in resource-limited settings is challenging due to inadequate disease and mortality reporting [2, 4]. Satellite imagery of burial sites has been used to monitor changes in the addition of the new burial plots as a surrogate for mortality data [2]. However, prior work relied on manual labeling of burial plots, which can be time-consuming, and the quality of these labels has not been analyzed.

To reduce the burden of manual annotation, we investigate whether high-resolution drone imagery can be used to guide the annotation of lower-resolution satellite imagery to support automated grave counting. A primary concern of this approach is that adapting labels across different imaging acquisition platforms may introduce additional sources of error. We therefore evaluate whether computer vision models adapted from crowd-counting can provide reliable estimates of grave counts despite these challenges.

2. Methods

Our methods address two practical challenges. First, because graves are difficult to annotate directly in low-resolution satellite imagery, we use high-resolution drone imagery to generate annotations and transfer them to aligned satellite images. Second, because this transfer process may introduce error, we evaluate three crowd-counting models and design experiments to assess their performance, robustness, and sensitivity to cross-platform label transfer.

2.1. Datasets

Drone imagery was acquired for fourteen burial sites in Dar es Salaam, Tanzania, from the OpenAerialMap Project [11]. Drone imagery was captured in April 2015 as 5 cm/pixel RGB images. Corresponding satellite imagery was acquired from Maxar as 50 cm/pixel RGB images. Burial plots in drone imagery of burial sites were manually labeled by local staff in duplicate using LabelBox to establish ground-truth counts. Drone and satellite imagery were aligned by manually annotating pixel pairs. The homography matrix was estimated using Random Sample Consensus [5], and annotations were realigned to the satellite imagery (Figure 1) using a perspective warp in OpenCV[3]. Satellite images were tiled into 96x96 pixel patches, and image portions without valid drone labels were masked.

2.2. Counting Frameworks

We evaluated counting frameworks from the related field of crowd counting. Similar to grave counting, crowd counting deals with spatial counting, typically in congested scenes. We applied three crowd counting frameworks to grave counting.

2.2.1. Congested Scene Recognition Network

The Congested Scene Recognition Network (CSRNet) is a density-based counting model [9], formulating counting as a density map estimation problem, a common technique in object counting [8]. CSRNet achieved state-of-the-art performance in crowd-counting, and captures large contextual regions while preserving spatial detail. While other models have since outperformed CSRNet in crowd counting, it remains a well established and understood benchmark [1, 10, 17]; therefore we adapt it to this novel application. To apply CSRNet, density maps of burial plots were created by passing a Gaussian filter across each burial plot location, using a sigma (standard deviation of the Gaussian kernel) of 5 pixels (Figure 1). The model was trained by minimizing the Euclidean loss between the predicted and the Gaussian-generated ground-truth density maps. Predicted density map spatial integration yielded the object count.

2.2.2. Distribution Matching for Crowd Counting

Distribution Matching for Crowd Counting (DM-Count) is another density-based crowd counting method [16]. DM-Count was applied using the same density map preparation as CSRNet. However, rather than a density-based loss function, DM-Count regresses on a loss function based on Optimal Transport [15], matching the distribution of the proposed density map to the actual point annotations. Count predictions are obtained through spatial integration of the predicted density map. At the time of its publication, DM-Count outperformed CSRNet on the ShanghaiTech and UCF-CC-50 benchmark datasets [6, 18].

2.2.3. Point to Point Network

Point to Point Network (P2PNet) deviates from the popular density map approach and performs object counting and localization using a purely point-based framework [13]. P2PNet was applied to the same tiled satellite images as CSRNet and DM-Count; unlike the density-based models, it uses point annotations directly without requiring density map generation. P2PNet outputs a set of proposals, which are matched with point annotations using the Hungarian algorithm [7, 14]. The network is trained using a combination of a Euclidean localization loss and a Cross-Entropy classification loss. Count predictions are calculated as the sum of proposals with a classification score > 0.5. At the time of its publication, P2PNet achieved state-of-the-art performance on the ShanghaiTech benchmark dataset, outperforming DM-Count and CSRNet. Additionally, the local-

ization mechanism present in P2PNet can lead to more interpretable results than density map estimation techniques.

2.3. Experiment

In our main analysis, model performance was evaluated using 10-fold cross-validation. Data from all cemeteries were pooled prior to splitting, as the intended application of this framework is the longitudinal monitoring of known sites rather than zero-shot generalization to unseen locations. Primary outcomes were mean absolute error, root mean squared error, and coefficient of determination (R^2).

In secondary analyses, we examined generalization, calibration, and the impact of transferring labels from high-resolution drone imagery to low-resolution satellite imagery. All secondary analyses were performed using CSRNet as a counting framework. To examine generalization, we stratified results from our main analysis across burial site. We additionally reran our cross-validation, holding out entire burial sites rather than random samples of images. To assess calibration, we estimated the slope and bias of predicted vs. actual counts, and adjusted error metrics using 10-fold cross-validation. To evaluate the impact of transferring labels from high-resolution drone imagery to low-resolution satellite imagery, we repeated the experiment using the original drone imagery, tiled to match the exact coordinates of corresponding satellite imagery tiles.

3. Results

Satellite tiles ($N = 337$) had a mean (standard deviation) of 26.4 (33.4) burial plots per tile. In total, 8,888 burial plots were identified across all tiles. Of these, 75% ($N = 254$) contained at least one burial plot. Among tiles with at least one burial plot, the mean number of burial plots was 35.0 per tile.

Table 1. Counting Performance Across Models

Experiment	MAE	RMSE	$\mathbf{R^2}$
CSRNet	8.4 ± 1.4	12.7 ± 2.3	0.86 ± 0.03
DM-Count	8.9 ± 1.7	12.7 ± 2.3	0.86 ± 0.05
P2PNet	$\mathbf{7.6 \pm 1.6}$	$\mathbf{11.8 \pm 2.3}$	$\mathbf{0.88 \pm 0.04}$

CSRNet exhibited skill in estimating grave counts from satellite imagery (R^2 = 0.86 ± 0.03). Mean absolute error was roughly one-third of the average true count (8.45 burial plots), and the root mean squared error was slightly higher (12.4 burial plots), indicating a long tail in the error distribution. DM-Count performed similarly, with identical R^2 and root mean squared error. Mean absolute error was slightly elevated (8.9 ± 1.7). P2PNet performed the best of all counting frameworks, with an R^2 of 0.88 ± 0.04, a mean absolute error of 7.6 ± 1.6, and a root mean squared error of 11.8 ± 2.3.

Table 2. Comparative Performance Study

Experiment	MAE	RMSE	$\mathbf{R^2}$
Baseline	8.4 ± 1.4	12.7 ± 2.3	0.86 ± 0.03
Hold-out Cemetery	12.5 ± 5.4	19.8 ± 7.3	0.65 ± 0.65
Drone	4.0 ± 0.7	6.7 ± 1.1	0.97 ± 0.02
Calibration (CV)	8.7 ± 1.4	12.6 ± 2.3	0.86 ± 0.03

Performance was highly sensitive to burial site characteristics. In a stratified main analysis (randomized split, 10-fold cross-validation) of CSRNet, the best-performing cemetery had an R^2 of 0.97, while four of the 14 cemeteries had an $R^2 < 0.5$. The worst-performing burial-site predictions had a mean absolute error of 11.2 burial plots per tile, and the best-performing had a mean absolute error of 2.7 burial plots per tile. The relative absolute error calculated at the cemetery level varied from 0.16 to 2.33. The worst performing cemeteries, as measured by R^2 and relative absolute error, had the lowest burial plot density averaged across all tiles. Poor generalization across burial sites persisted when the entire burial site was held out from training (Table 2). R^2 dropped to 0.65, mean absolute error increased to 12.5, and root mean squared error increased to 19.8.

CSRNet was able to count burial plots much more accurately using the original drone imagery. R^2 increased to 0.97, mean absolute error decreased sharply to 4.0, and root mean squared error similarly decreased to 6.7.

CSRNet predictions were well calibrated (Figure 2). The calibration regression yielded a bias term of 1.39 (95% CI -0.342 – 3.129) and a slope of 1.014 (95% CI 0.969 – 1.058). Consequently, calibration had a limited impact on 10-fold cross-validation results (Table 2).

4. Discussion

Our findings suggest that it is possible to count burial plots in satellite images using crowd-counting models trained on drone-guided annotations. This could provide a practical, automated alternative to slow manual labeling. It could be especially useful for quick mortality monitoring in areas where traditional data systems are weak or delayed. These discoveries could support a fully automated epidemic early warning system.

Our model performance assessment had both strengths and important limitations. Overall errors were moderate, and predictions followed the true counts reasonably well. Additionally, predictions were well calibrated, despite some prior work in object counting identifying calibration as a potential issue. [12]. However, performance was substantially lower on satellite imagery when compared to drone

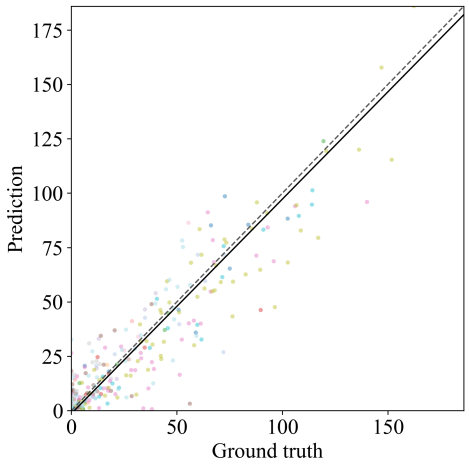

Figure 2. CSRNet Prediction vs ground truth plot. Each tile is represented as a point, and points are colored by cemetery. The dashed black line represents the regression, while the diagonal is represented in light grey for reference. Predictions were strongly correlated with ground truth (R^2 = 0.86). Predictions were calibrated, with a bias of 1.39 (95% CI -0.342 – 3.129), and slope of 1.014 (95% CI 0.969 – 1.058).

imagery, indicating the value of high resolution imagery in counting graves. Importantly, other factors may have played a role in this performance drop-off. There may have been cases where burial plots were obscured by vegetation in the drone imagery, and therefore gone unlabeled. Future validation of drone imagery labels by physical inspection of burial sites could help to quantify this effect.

Additionally, we found that model's performance varied across burial sites, limiting the generalizability of our findings. We observed substantial performance degradation when an evaluated burial site was fully held out from the training data. This finding suggests that it is crucial to validate individual burial sites before "enrolling" them into an epidemic early-warning system, and that it may be necessary to fine-tune the system for specific sites.

Finally, due to the limited availability of high-resolution drone data at multiple time-points, our study was cross-sectional. To validate this framework for detecting changes in grave counts, longitudinal drone data must be collected.

Despite these limitations, our findings show that using satellite-based burial plot counting trained on drone imagery-assisted annotations has real potential to fully automate mortality monitoring systems in areas with high needs but limited on-the-ground data. We recommend additional data collection of drone imagery alongside on-site validation of labels in future work.

References

[1] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive Dilated Network With Self-Correction Supervision for Counting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4593–4602, Seattle, WA, USA, 2020. IEEE. 2

[2] Emilie S. Koum Besson, Andy Norris, Abdulla S. Bin Ghouth, Terri Freemantle, Mervat Alhaffar, Yolanda Vazquez, Chris Reeve, Patrick J. Curran, and Francesco Checchi. Excess mortality during the COVID-19 pandemic: a geospatial and statistical analysis in Aden governorate, Yemen. *BMJ Global Health*, 6(3), 2021. 2

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2

[4] Michael Chasukwa, Augustine T. Choko, Funny Muthema, Mathero M. Nkhalamba, Jacob Saikolo, Malebogo Tlhajoane, Georges Reniers, Boniface Dulani, and Stéphane Helleringer. Collecting mortality data via mobile phone surveys: A non-inferiority randomized trial in Malawi. *PLOS global public health*, 2(8):e0000852, 2022. 2

[5] Martin A. Fischler, View Profile, Robert C. Bolles, and View Profile. Random sample consensus. *Communications of the ACM*, 24(6):381–395, 1981. 2

[6] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 2

[7] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[8] Victor Lempitsky and Andrew Zisserman. Learning To Count Objects in Images. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2010. 2

[9] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, Salt Lake City, UT, 2018. IEEE. 2

[10] Xinyan Liu, Guorong Li, Zhenjun Han, Weigang Zhang, Yifan Yang, Qingming Huang, and Nicu Sebe. Exploiting sample correlation for crowd counting with multi-expert network. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3195–3204, 2021. ISSN: 2380-7504. 2

[11] OpenAerialMap. OpenAerialMap: The Open Collection of Aerial Imagery. https://openaerialmap.org, 2026. Accessed: 2026-03-13. 2

[12] Matthew D. Ritch, Bailey G. Hannon, A. Thomas Read, Andrew J. Feola, Grant A. Cull, Juan Reynaud, John C. Morrison, Claude F. Burgoyne, Machelle T. Pardue, and C. Ross Ethier. AxoNet: A deep learning-based tool to count retinal ganglion cell axons. *Scientific Reports*, 10(1):8034, 2020. 3

[13] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 3365–3374, 2021. 2

[14] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016. 2

[15] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009. 2

[16] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, pages 1595–1607. Curran Associates, Inc., 2020. 2

[17] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to Scale: Generating Multipolar Normalized Density Maps for Crowd Counting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8381–8389, 2019. ISSN: 2380-7504. 2

[18] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 2