Next Best View for Text Detection and Recognition in Port Monitoring Unmanned Aerial Vehicles

Emre Gülsoylu*,[1], Niklas Fiedler[2], and Simone Frintrop[1]

I. Introduction

Next-Best-View (NBV) optimisation is a critical capability for autonomous unmanned aerial vehicles (UAVs) tasked with perceiving complex, occluded, or partially observable environments. By iteratively selecting viewpoints that maximise informativeness, NBV enables efficient data acquisition for downstream tasks, such as 3D reconstruction [1], object detection [2], structure inspection [3], exploration [4], search and rescue [5], and crop harvesting [6]. However, its application to scene-text detection and recognition remains underexplored, particularly in industrial settings where textual identifiers (e.g., identifier (ID) codes, labels) are prone to occlusion or degradation. Addressing this, we formalise NBV optimisation for text detection and recognition, a problem with direct implications for logistics and port monitoring.

Efficient identification of intermodal loading units (ILUs), such as containers and trailers, in ports is critical for streamlining logistics operations, yet existing methods often focus on fixed camera perspectives. On the other hand, UAVs offer a scalable solution by capturing high-resolution imagery from adaptive viewpoints, enabling robust ILU identification [7]. For this task, we propose two types of missions. First, *survey mission* to capture nadir-view images. Survey missions are limited by their reliance on top-down perspective, which can fail to recognise ID codes in degraded or occluded conditions. To overcome these problems, we introduce a second type of mission, *targeted mission*, where a UAV captures sides of ILUs, guided by an NBV optimisation strategy.

By integrating 3D pose estimation of ILUs with a novel Legibility Score (LS), which balances viewing angle, distance, and line of sight constraints, we optimise UAV mission paths to maximise ID code legibility while minimising flight time. By tackling the ILU identification task as a case study, we demonstrate how domain-specific utility function design can extend NBV to other applications where textual information is critical.

II. UAV Missions for ILU Identification

Images captured by survey missions are processed using Open Drone Map (ODM) [8] to generate a georeferenced 3D point cloud and an orthophoto (a georeferenced overview image), which is then fed into a state-of-the-art ILU identification pipeline, Three-stage Identification of Transportation UnitS (TITUS), for ILU segmentation, text detection, and recognition. Although they provide a comprehensive overview about the predefined area, survey missions have certain limitations. Since only the top surfaces of ILUs are visible, damage to upper surfaces may obscure or distort ID codes, leading to missed text detections or mistakes in text recognition. Such cases can create information gaps for the terminal operations and result in operational inefficiencies. To mitigate this, a second mission type, *targeted mission* was introduced. Unlike the survey mission, this approach focuses on specific areas flagged as unidentified ILU (i.e., an ILU segment without an associated ID code) or non-compliant ID (i.e., an ILU segment associated with an ID code non-compliant with the ISO6436 standard).

A. Survey Mission

The survey mission relies on the Divide Areas Algorithm for Optimal Multi-Robot Coverage Path Planning (DARP) algorithm [9], [10] to generate optimised flight paths for one or more UAVs, ensuring full coverage of the predefined area. The input for the DARP algorithm is determined by the operational constraints. For example, grid representation of the area is defined by the three key factors. (1) UAV's flight altitude suggested by the terminal operator, (2) camera's field of view, and (3) expected image overlap for creating a point cloud and an orthophoto. Similarly, the home position of UAV and *no-fly* zones like cranes areas are defined by the terminal operator.

Among these parameters, image overlap is critical for point cloud and orthophoto generation. Higher overlap between the images improves feature alignment during point cloud and orthophoto generation but at the cost of extended flight duration and increased processing time for point cloud and orthophoto generation [11]. Since processing time does not scale linearly with image volume, excessive overlap can disproportionately slow down the pipeline. To balance the quality and efficiency, we adopted empirically validated parameters [12], refined through both simulation testing and real-world experiments. These settings ensure sufficient overlap for robust point cloud and orthophoto generation while minimising redundant data collection.

After creating a point cloud and an orthophoto of the port, the orthophoto is processed by the TITUS pipeline with image tiling. TITUS pipeline involves three stages: (1) segmenting ILU instances, (2) detecting their ID text area, (3) extracting the ID code from detected text areas

[1]University of Hamburg, Faculty of MIN, Department of Informatics, Computer Vision Group Hamburg, Germany
*emre.guelsoylu@uni-hamburg.de
[2]University of Hamburg, Faculty of MIN, Department of Informatics, Technical Aspects of Multimodal Systems, Hamburg, Germany

and associating the extracted ID with the corresponding ILU segmentation mask.

While survey missions provide an overview through orthophotos, they often fail to capture legible ID codes on stacked or damaged ILUs due to occlusions or suboptimal lighting conditions and viewing angles. Such ILU segments are flagged as unidentified (i.e. an ILU segment without an associated ID code) or non-compliant (i.e. an ILU segment associated with non-ISO6436 ID code). These flagged segments or target coordinates are the focus of targeted missions.

B. Targeted Mission

Targeted missions aim to resolve identification gaps by navigating UAV to optimal viewing positions for ID text detection and recognition. Unlike survey missions, which assume a flat 2D plane, targeted missions use 3D point clouds generated after the survey mission.

To define the optimum waypoints, we extend the TITUS pipeline to estimate the pose of each ILU using point clouds generated by ODM. This process involves four stages: (1) using the 2D segmentation masks from the TITUS pipeline, we crop the global point cloud to isolate the points corresponding to each ILU prediction, (2) for each isolated point cloud, a rectangular prism representing an ILU is fitted by using the perspective n-point algorithm [13] to determine ILU's 3D bounding box and therefore its pose relative to the cropped point cloud coordinate system, (3) by reallocating these cropped point clouds into the global point cloud, we identify the four sides of each ILU in the port's coordinate system, and (4) finally, for each face, f, we define a truncated half cone, c_f, projecting outwards from the face normal:

$$\begin{aligned} c_f = \big\{(x,y,z) \in R^3 \,|\sqrt{x^2+y^2} \leq kz, \\ y \geq 0, \\ d_{\text{min}} \leq z \leq d_{\text{max}}\big\}, \end{aligned} \tag{1}$$

where z represents the distance from the face along the normal $\hat{n}_f$, k is the aperture coefficient defining the maximum allowable off-axis viewing angle, $y \geq 0$ restricts the volume to the upper half of the cone as the lower half of the cone is not safe to fly, d_{min} and d_{max} are the distance thresholds that crop the cone to prevent Field of View (FOV) clipping and avoid low resolution of text areas due to distance.

For each of the four faces, five candidate waypoints are randomly sampled within the cone and the one with the highest legibility score is selected as the optimum waypoint.

C. Legibility Score

We propose a Legibility Score (LS) for NBV optimisation. This function differs from prior NBV work as it focusses on scene-text detection and recognition by evaluating each candidate viewpoint.

Let $\mathcal{T} = \{T_1, T_2, ..., T_n\}$ be the set of n target ILU coordinates in 3D space. For each T_i, four faces (sides) $F_i = f_1, f_2, f_3, f_4$ are predicted and for each, f_i, define a truncated half cone, c_{f_i}. Within these cones, sample a set of waypoints $\mathcal{W}_i = \{w_1, w_2, ..., w_n\}$, where $w_i \in R^3$. For a waypoint w_j relative to face f_i, we define the LS as follows:

$$LS(w, f_i) = \underbrace{\left(\frac{1 - \hat{n}_i \cdot \hat{v}_w}{2}\right)}_{\text{Angle Term}} \cdot \underbrace{\left(\frac{1}{d(w, f_i) + 1}\right)}_{\text{Distance Term}}, \tag{2}$$

where $\hat{n}$ is the surface normal, $\hat{v}$ is the camera's viewing direction, $d(w, f_i)$ is the Euclidean distance between w and f_i.

Using LS, we select the optimum waypoint w^* for a given face:

$$w^* = \arg\max_{w \in \mathcal{V}_f} \left(LS(w, T_i) \cdot V(w, f)\right), \tag{3}$$

where $V(w, f)$ is the binary visibility function that determines whether a Line of Sight (LOS) exists between the UAV's camera at waypoint w and the target ILU face f.

If the ILU face is not visible within the defined truncated half cone, that face is skipped, resulting in less than four waypoints for that target coordinate.

After defining the waypoints for each target coordinate, these waypoints are represented as a node in a graph to be used in the Ant Colony Optimisation algorithm [14]. The edges between nodes are weighted by the cost of travelling from one waypoint to another. Using the LS as part of the heuristic information to guide the ants, the pheromone update rule is designed to favour paths that maximize LS and minimise travel distance. The cost function is defined as follows:

$$\text{Cost}(w_i, w_j) = \alpha \cdot d(w_i, w_j) + \beta \cdot \frac{1}{LS(w_j, f)}, \tag{4}$$

where α and β are weights to balance the importance of distance and legibility.

An ordered sequence of waypoints that optimises both path length and legibility is the output of the Ant Colony Optimisation algorithm.

III. Future Directions

The proposed legibility score balances recognition accuracy and mission path length but its effectiveness in port environments requires validation in both simulation and real-world environments. One approach could involve integrating adaptive weighting for the angle and distance terms (α and β parameters). The proposed framework can be extended to accommodate multi-agent systems by dynamically assigning ILUs to UAVs based on the proximity, and battery levels. Finally, the NBV for text should be validated in different domains and applications.

Acknowledgements

The project is supported by the Federal Ministry for Digital and Transport (BMDV) in the funding program Innovative Hafentechnologien II (IHATEC II).

References

[1] H. Dhami, V. D. Sharma, and P. Tokekar, "Pred-NBV: Prediction-guided next-best-view planning for 3d object reconstruction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7149–7154.

[2] M. Grotz, D. Sippel, and T. Asfour, "Active vision for extraction of physically plausible support relations," in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2019, pp. 439–445.

[3] R. Border and J. D. Gammell, "The surface edge explorer (SEE): A measurement-direct approach to next best view planning," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1506–1532, 2024.

[4] M. Naazare, F. G. Rosas, and D. Schulz, "Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3779–3786, 2022.

[5] S. H. Strand, T. Wiedemann, B. Burczek, and D. Shutin, "Enhancing UAV Search Under Occlusion Using Next Best View Planning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 1085–1096, 2025.

[6] R. Menon, T. Zaenker, N. Dengler, and M. Bennewitz, "NBV-SC: Next best view planning based on shape completion for fruit mapping and reconstruction," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4197–4203.

[7] E. Gülsoylu, A. Abdelhalim, D. K. Boztas, O. Grasse, C. Jahn, S. Frintrop, and J. Edinger, "Automatic Intermodal Loading Unit Identification using Computer Vision: A Scoping Review," *arXiv preprint arXiv:2509.17707*, 2025.

[8] OpenDroneMap Authors, "ODM - a command line toolkit to generate maps, point clouds, 3D models and DEMs from drone, balloon or kite images."

[9] A. C. Kapoutsis, S. A. Chatzichristofis, and E. B. Kosmatopoulos, "DARP: Divide areas algorithm for optimal multi-robot coverage path planning," *Journal of Intelligent & Robotic Systems*, vol. 86, no. 3, pp. 663–680, 2017.

[10] S. D. Apostolidis, P. C. Kapoutsis, A. C. Kapoutsis, and E. B. Kosmatopoulos, "Cooperative multi-UAV coverage mission planning platform for remote sensing applications," *Autonomous Robots*, vol. 46, no. 2, pp. 373–400, 2022.

[11] N. T. Pham, S. Park, and C.-S. Park, "Fast and efficient method for large-scale aerial image stitching," *IEEE Access*, vol. 9, pp. 127 852–127 865, 2021.

[12] S. Wallat, "Implementation and Evaluation of an End-to-End Image Processing Pipeline for UAV-Based Surveying of Inland Ports," Master's Thesis, University of Hamburg, MIN-Faculty, Department of Informatics, Hamburg, Germany, 2025.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] M. Dorigo and T. Stützle, "Ant colony optimization: overview and recent advances," *Handbook of metaheuristics*, pp. 311–351, 2018.