EasyChair Preprint

№ 14025

Drone-Based Identification of Containers and Semi-Trailers in Inland Ports

Jana Teegen, André Kelm, Ole Grasse, Maris Hillemann,
Emre Gülsoylu and Simone Frintrop

July 18, 2024

Drone-Based Identification of Containers and Semi-Trailers in Inland Ports

Jana Teegen[1], André Peter Kelm[1][0000−0003−4146−7953], Ole Grasse[2][0000−0003−1982−9436], Maris Hillemann[1], Emre Gülsoylu[1][0000−0002−3834−3645], and Simone Frintrop[1][0000−0002−9475−3593]

[1] University of Hamburg, 22527, Hamburg, Germany
{jana.teegen@studium., andre.kelm@, maris.nathanael.hillemann@studium., emre.guelsoylu@, simone.frintrop@}uni-hamburg.de
[2] Hamburg University of Technology, 21073, Hamburg, Germany
ole.grasse@tuhh.de

Abstract. This paper introduces a novel application utilizing drones and deep learning to identify containers and semi-trailers, enhancing inland port operations. With this drone-based image and text recognition system, the basic condition of the yard/storage area can be determined at any time without using (human) labor, eliminating the need for manual inspections. To our knowledge, this is the first instance of identifying containers and semi-trailers in a deep learning application through drone imagery. Automating identification through drone flights is one of the main goals of our InteGreatDrones (IGD) project. This paper lays the foundation and provides a first building block by addressing the challenges posed by the real-world data and the different drone perspectives, including the various altitudes, scenes, and viewpoints captured in this project, with the goal of cargo identification. We use a two-step recognition process, first localizing the text ID and then reading/identifying it. We take established methods such as EAST for scene text detection and TrOCR for optical character recognition and fine-tune them to enable accurate identification from drone imagery. Despite the challenging real-world images, we achieve an F1 of 0.5 for text detection and a CER of 0.16 %.

Keywords: UAV-Based Monitoring · Automated ID Recognition · Port Logistics · Loading Unit Identification · Intermodal Handling.

1 Introduction

Drone technology and artificial intelligence both have made great progress. They can be combined, as in our InteGreatDrones[3] (IGD) project, to create new applications like the one presented here to improve industrial processes towards more efficiency and effectiveness. Their applications in various fields are growing, including areas such as, but not limited to, agriculture and infrastructure

[3] https://integreatdrones.de/

surveillance. In the domain of logistics, the way is paved to optimize the operations of inland ports. Inland ports play an important role in logistical networks when connecting global sea transport to the hinterland, as well as in cases where a change of transport mode is required. Goods in intermodal transport are typically handled in different types of loading units (LUs). In this paper, we deal with the two types of containers and semi-trailers, which together make up the majority of the handled LUs in inland ports. A typical structure of an inland port with a gate, storage area and handling places is shown in Fig. 1.

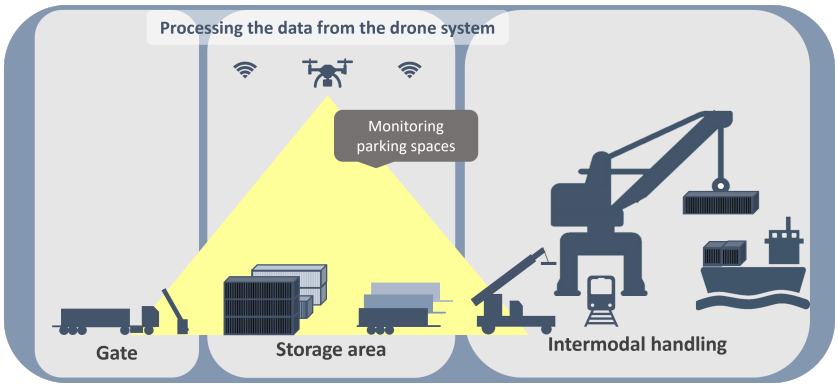


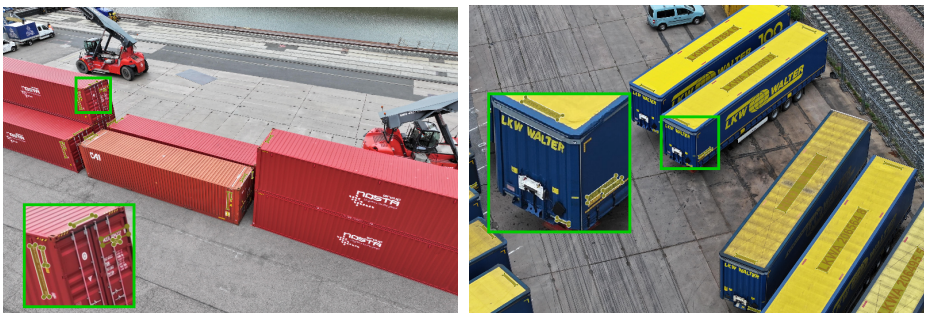
Fig. 1: Schematic representation of a typical inland port with gate, storage area and intermodal handling. In particular, the containers and semi-trailers in the storage area are captured by the drone (field of view shown in yellow).

The operations within inland ports involve the intermodal handling and the temporary storing of LUs together with necessary administration tasks. As opposed to large seaports, inland ports are usually confined in space and often do not maintain sophisticated digital infrastructure like large-scale video surveillance or optical character recognition (OCR) gates. This lack of technology generates a higher manual effort to plan, execute and ensure efficient handling. Also, inland ports operate in a competitive and more dynamic market and are faced with changing freight types and volumes resulting in regular adjustments to the layout and configuration of, e.g., the yard areas. However, technological advances in drone technology, computer vision and autonomous systems offer promising approaches to mitigate this while being reliable and affordable.

The majority of existing solutions and research papers focus on stationary scanning devices, so-called smart- or OCR-gates, which are expensive infrastructure for high throughput ports. Smart gate solutions are useful for the identification of LUs when entering and exiting enclosed port areas. However, the container code recognition task is not limited to static smart gate applications. In practice, there are use cases for both more flexible and also more challenging scenarios, such as dynamic, on-demand detection and identification of LUs

in a non-enclosed yard area or a parking lot. In some cases, the use of drones, also known as unmanned aerial vehicles (UAVs), for container code recognition can offer significant advantages over smart gates. UAVs provide a level of flexibility that can improve operational efficiency by allowing dynamic monitoring and management of containers without the constraints of fixed camera positions. Moreover, the proposed solution applying drones can provide real-time information about any LU's position in the whole yard at any given time, not limited to the terminal gate, which results in a major advantage in flexibility. From an economic perspective, the proposed UAV approach will be comparably affordable and, therefore, outperform OCR gates, especially when looking at smaller inland ports. The use of UAVs in those ports is not limited to the proposed identification task but can provide even more features based on the gathered data, such as surveillance, damage detection, etc. and, therefore, contribute significantly to the logistical efficiency and also to digitization.

The task of container code recognition falls within the field of Computer Vison (CV). Precise identification of each container and semi-trailer is integral to process optimization. This requires extracting the ISO6346 [1] compliant unique IDs, shown in Fig. 2, from the obtained drone images. As long as the container code is readable, it can be effectively addressed, as shown by the comparable work with a wide spectrum of methodologies ranging from traditional image processing to advanced deep learning techniques [4, 7]. The vertical text recognition and the recognition of codes from different LU types have also been addressed [8]. While systems such as those of Xu et al. [6] and Zhao et al. [9] provide comprehensive solutions, issues of data accessibility and generalization were not fully addressed, highlighting areas for future research and development. Although there are ready-made solutions for smart gate systems on the market, solutions for the UAV perspective are still rare. Therefore, we developed and adapted our



(a) Container with labeled identifiers (IDs). (b) Semi-trailer with labeled IDs.

Fig. 2: Drone's perspective on loading units (a) containers and (b) semi-trailers at the storage area. The IDs (annotated rotated bounding boxes - marked yellow) are used for unique identification. Green boxes indicate zoom for the areas where there are multiple IDs and give better visibility of existing details.

own system by fine-tuning and building a two-step method to recognize the IDs of LUs from a non-stationary, more challenging UAV perspective.

To our knowledge, this work is the first approach that uses UAV imagery to identify containers and semi-trailers with a deep learning application. Unlike existing methods focused on ground-based perspectives, our system is intended to handle images taken from arbitrary angles, ensuring that the model is robust in real-world scenarios, because the point of view of UAVs can be highly variable.

2 Methodology

To identify LUs and overcome the different challenges posed by different altitudes, scenes, viewing angles, and ID print qualities, it seems advantageous to apply deep learning in several stages. Our two-step recognition process is described by the workflow diagram in Fig. 3 to identify both types of LUs.

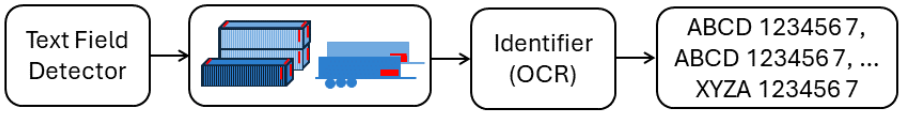
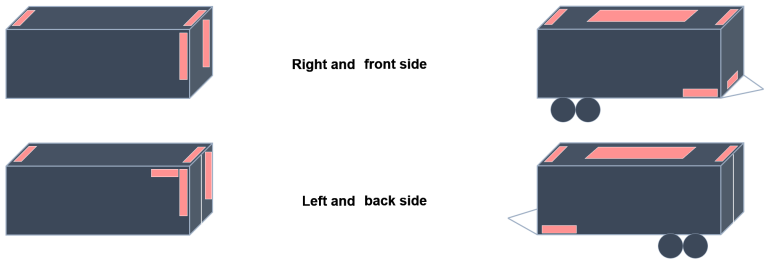


Fig. 3: Workflow of the two-step approach. First, a text box detector provides bounding boxes for the ID (marked in red). Second, for each bounding box, an OCR method reads the text content within the bounding box area. There is redundancy as multiple IDs are printed on the same LU and are therefore read more often. In future work, this will be used to further support robustness.

End-to-end systems are suitable for smart gate solutions where the task is only to read the container code. On the other hand, UAV applications require more flexibility which can be obtained by two-step approaches. There may be scenarios where the recognition step is not possible due to the camera's position, but text detection is still important to locate the text on an LU and adjust the pose of the UAV for the best text recognition view. For example, if the text is too far to be successfully recognized, the UAV can go closer to provide a better view for the recognition step. One-stage approaches cannot provide this flexibility as they take an image as an input and gives the recognized text as an output.

2.1 ID Detection - Step 1

The approach introduced by this paper is based on the Efficient and Accurate Scene Text Detection (EAST) [10] system for text detection in natural scenes. EAST utilizes a fully convolutional network (FCN) architecture specifically tailored for scene text detection tasks. Fundamentally, EAST employs the concept of a multi-channel FCN in which each layer consists solely of convolutional operations. Through a set of convolutional layers, the FCN extracts both early and



(a) Container IDs. (b) Semi-trailer IDs.

Fig. 4: Placements of IDs (light-red) on (a) containers and (b) semi-trailers.

late features, allowing the model to effectively detect text of different sizes, orientations and styles. A further key aspect of the EAST architecture is the feature merging strategy. To reduce computational costs while maintaining performance, feature maps generated by convolutional layers are progressively merged. This merging process involves concatenating feature maps from different layers, followed by upsampling and convolution operations to refine the feature representations. By integrating information from multiple layers, EAST achieves a balance between efficiency and accuracy in text detection, keeping the computational expenses relatively low.

The output layer of EAST produces confidence values and geometries for predicted bounding boxes per pixel. Using thresholding and non-maximum suppression (NMS), overlapping predictions are merged in a subsequent process. The results are rotated bounding boxes represented by five feature maps containing the distance to and orientation of an edge of the rectangle per pixel.

EAST is fine-tuned with dice loss function that evaluates the similarity between the predicted bounding boxes and the ground truth. The training was based on 325 annotated images, of which 70 percent, thus 227 images, were used for training and the remaining 98 images for testing. The different locations in which the IDs can be found also need to be considered, as displayed in Fig. 4. The training included over 900 epochs, ensuring comprehensive exposure to training images.

2.2 ID Recognition - Step 2

With the successful application of EAST, the next step is text extraction using OCR for the detected IDs. This step is performed using Microsoft's Transformer-based optical character recognition (TrOCR) [3] model. The image data and the rotated bounding boxes provided by EAST are used as input. Since the IDs may not be axis-aligned, an alignment process is performed beforehand by rotating and cropping the image so that the image solely consists of the ID itself. This pre-processing step is essential as TrOCR specializes in text recognition rather than text detection in natural scenes. By leveraging the strengths of TrOCR

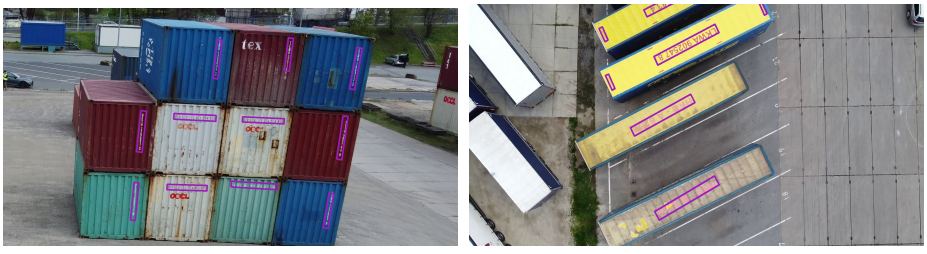
in text recognition, this approach complements the focus of EAST on scene text detection, facilitating accurate and efficient ID extraction from complex environments. For the fine-tuning of TrOCR, 2580 text fields were cropped out based on the ground truth annotations, each of which contains a single ID. The dataset was split into training data with 70 percent, validation data, holding 15 percent and test data, also consisting of 15 percent of the total images.

3 Evaluation

During the calculation of the following metrics intersection over union (IoU) threshold was defined as 0.5. The EAST model which was pre-trained on ICDAR2015 [2] achieved a low precision (0.034) and a low recall value (0.029), indicating that fine-tuning the model for ID detection from UAV images is required. Fine-tuning EAST with the batch size of 10 led to a noticeable performance gain, as precision and recall has increased to 0.71 and 0.38, respectively, at 900th epoch as shown in Table 1. The F1-score of the model (0.50) reflects adaptability, increasing from 0.031 before training to over 0.50, implying a balanced detection ability, although minimizing false negative detections remains difficult. The UAV perspective, with its considerable distance from the objects, prevents the detection of smaller IDs, contributing to the observed low recall.

As shown in Figure 5a and 5b, the EAST model can detect text fields of the IDs of both containers and trailers regardless of their orientation when the UAV flies at the lower altitudes. However, the EAST model has difficulties in making line-level predictions, due to its pre-training on the ICDAR2015 dataset, which contains word-level annotations. The number of images fall short to make the model adapt line-level annotations by fine-tuning.

The evaluation of the recognition step, conducted with the ground truth annotations of the text fields to eliminate the effects of the detection model. The recognition precision and recall were both high, at 0.81 and 0.82 respectively, resulting in an F1-score of 0.81. The character error rate (CER) was relatively low at 0.16, indicating that the model accurately recognized the majority of characters. Although TrOCR was originally not able to recognize vertical text,



(a) Containers with detected IDs. (b) Semi-trailers with detected IDs.

Fig. 5: ID detection performance of EAST on (a) containers and (b) semi-trailers.

Table 1: Summarized results for detection, and recognition steps for containers and semi-trailers. For the recognition step, all metrics were calculated at the character level after the removal of whitespace.

Step	**Precision**	**Recall**	**F1-Score**	**CER**
Detection (EAST)	0.71	0.38	0.50	-
Recognition (TrOCR)	0.81	0.82	0.81	0.16

fine-tuning the model on our dataset enabled recognition of vertical as well as horizontal IDs text.

Because of the language model that TrOCR use, the pre-trained model predicts some words from natural languages and not ISO6346 compliant ID code. After fine-tuning the model on our dataset the model was able to learn about the general structure of the ID. Therefore, even if an ID cannot be recognized completely, the structure of the predicted text will be correct so that certain mistakes can be corrected in post-processing.

Table 1 summarizes the results for both the detection and recognition steps. The detection step using the EAST model is affected by its training limitations and the UAV's distant perspective, leading to lower recall and moderate precision. On the other hand, the recognition step using the TrOCR model demonstrated robust performance, with high precision and recall on character level, and a low CER, effectively recognizing characters once they were detected. These results highlight the necessity for improved detection methods tailored to the specific challenges posed by UAV imagery to enhance overall system performance.

4 Conclusion

From a logistical perspective, the presented approach promises huge value for operators and their customers, especially when focusing on smaller inland ports (or combined transport terminals likewise). Established OCR solutions as stationary infrastructure are often not available for financial reasons and do also often not fit the flexibility demands of smaller ports. Providing them with an affordable, aerial identification solution providing the same information quality in combination with an even higher information availability will mitigate their use of manual labor, boost their digitization endeavor and will also increase the efficiency of logistical networks and hinterland logistics.

From the CV perspective, our initial two-step approach shows significant potential for improving logistics in inland ports. The first step can guide an automatic UAV control based on detected text, reaching an F1 score of 0.5 for text detection, while the second step recognizes text in arbitrary orientations with a CER of 0.16 %. This serves as a solid foundation for future work.

One of the biggest challenges is still the significant size difference between the UAV's HD images and the mostly very small IDs in them, known in the literature as the small object detection problem [5]. We plan to further mitigate this

problem by extending our two-stage approach and adding an object detector for containers and semi-trailers as the first step. Additionally we see great potential in exploring subtasks such as optimal navigation and automatic zooming to capture target objects/IDs from the best view, which humans perform intuitively. Achieving the best view for IDs simplifies text recognition, creating a dynamic interplay between optimal view finding and improved scene text detection and recognition methods.

Even this early stage of the application shows, there is an enormous potential for small to medium inland ports. Flexible "on the fly" identification of LUs through the combined application of CV and UAV cameras to port logistics enables low-effort digital support and a reasonably cheap alternative to existing stationary solutions at the same time. The continued exploration and enhancement of these techniques promise to bring robustness and enhanced applicability for inland port operations in the future.

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. Freight containers — Coding, identification and marking. Standard, International Organization for Standardization, Geneva, CH (2022)
2. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V., Lu, S., et al.: Icdar 2015 competition on robust reading. In: ICDAR. pp. 1156–1160. IEEE (2015)
3. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: TrOCR: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI. vol. 37, pp. 13094–13102 (2023)
4. Nguyen, H.S., Huynh, C.D., Bui, N.Q.: Digital transformation for shipping container terminals using automated container code recognition. TELKOMNIKA **21**(3), 535–544 (2023)
5. Wilms, C., Frintrop, S.: Attentionmask: Attentive, efficient object proposal generation focusing on small objects. In: Proceedings of the ACCV. pp. 678–694 (2019)
6. Xu, Y., Liang, Z., Liang, Y., Li, X., Pan, W., You, J., Long, Z., Zhai, Y., Genovese, A., Piuri, V., et al.: Data-driven container marking detection and recognition system with an open large-scale scene text dataset. IEEE TETCI (2024)
7. Yang, D., Wang, G., Liu, M., Yue, S., Zhang, H., Chen, X., Zhang, M.: Lightweight container code recognition based on multi-reuse feature fusion and multi-branch structure merger. Journal of Real-Time Image Processing **20**(6), 108 (2023)
8. Zhang, R., Bahrami, Z., Liu, Z.: A vertical text spotting model for trailer and container codes. IEEE Trans. Instrum. Meas. **70**, 1–13 (2021)
9. Zhao, J., Jia, N., Liu, X., Wang, G., Zhao, W.: A practical unified network for localization and recognition of arbitrary-oriented container code and type. IEEE Trans. Instrum. Meas. (2024)
10. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: EAST: An efficient and accurate scene text detector. In: Proceedings of the CVPR (2017)