An Image-Based Approach for Construction Site Monitoring and Documentation Using Machine Learning

Matthias W. Glunz,[1] Florian Steinbach,[1] Lina Wedekind,[1,] Martina Mellenthin Filardo[1] and Jürgen Melzner[1]

[1] Chair for Construction Engineering and Management, Bauhaus-Universität Weimar, Germany

martina.mellenthin.filardo@uni-weimar.de

Abstract

Given current shortages of skilled labour in the construction industry, this paper presents a study on the feasibility and application of an image-based, automated approach for construction site monitoring and documentation using machine learning methods. The study concentrates on object detection based on images of a specific construction site, taken multiple times a day periodically over the course of a year, that have been evaluated using the YOLOv8 technology, thus enabling progress monitoring for selected elements. Training and validation data have been created from annotated images for the object detection, which was accompanied by an evaluation of the chosen hardware and the observation viewpoint for future reference in the data acquisition. Further, a ground truth for the construction progress was generated manually to allow comparison with the results achieved by the machine learning approach.
This study demonstrated, that the expected results were achieved without the need for writing a single line of code, which is meaningful given the aforementioned labour shortages in the construction industry and highlights the fast-paced nature of the machine learning field.

Keywords: Construction Site Progress Monitoring, Machine Learning, YOLOv8, Object Detection.

1 Introduction

Image-based monitoring of construction sites using machine learning technology can be a feasible approach to assess different stages of construction. If implemented thoroughly, it can address labour shortages by making the process more transparent and therefore more robust as well as supporting decision making and facilitating documentation. To ensure this application of machine learning, two main factors must be examined: Trust in the results and the threshold level for implementation. With these goals in mind, an experiment was undertaken to confirm results from a Convolutional Neural Network (CNN). For this, a data set (described in 2.3) consisting of recurring pictures of a construction site over the period of a year was used to train a CNN (described in 2.2. as well as 3). To enable an assessment of the object detection results achieved by the CNN, a so-called ground truth of the construction process was manually generated. Through the comparison of the CNN results of the object recognition with the human-generated construction process, the construction of specific parts (columns) on the construction site illustrates the CNN results and where they deviate from the human-generated ones.

The other aspect of a broad application of machine learning in the field of construction monitoring is addressed by the tools applied to this experiment: The criteria for choosing the CNN as well as all adjacent tools was a strict ‘no coding’ policy, given that traditional construction personnel has no programming experience.

2 Background

In the Background section, the key background information for Object Detection is explained (2.1), and the technologies chosen in the experiment are outlined (2.2). Further, the used data is described in section 2.3.

2.1 Object Detection

Before the detection of the objects can begin, the object itself needs to be defined. In this experiment, the use case of target-actual comparison was chosen to demonstrate the efficiency of image-based monitoring using machine learning technology to analyse the state of construction. To exemplify the process analysis, the columns in their different

construction stages were selected. The process was classified into four states: ‘connecting reinforcement installed’, ‘reinforcement cage installed’, ‘formwork elements installed’, and ‘completion of the column’, as described in Table 1.

For image-based monitoring, a special kind of artificial neural network (ANN) was used. ANNs are computer-based systems inspired by biological nervous systems, consisting of many connected artificial neurons. The basic idea behind ANNs is that they receive information, process the given information, and forward the processed information to other neurons [1]. While in ANNs, the neurons are all connected to each other, the neurons in CNNs are partly connected to the other layers in the system. This local connection enables the CNN to capture spatial information more effectively and hierarchically learn features in images [1]. In contrast to conventional ANNs, the neurons in the layers of a CNN are organized three-dimensionally. These dimensions encompass the spatial dimensions of height and width, as well as depth. This promotes the effectiveness of pattern recognition in images. In this context, depth does not refer to the total number of layers in the network but rather to the third dimension of an activation volume. Each element in the depth of the activation volume represents a specific feature map or filter specialized in a particular feature pattern [2].

Usually, the basic structure of a CNN comprises three layers. In the input layer, the data to be processed is input. In the hidden layers, also called intermediate layers, decisions are made, influencing the final result. The output layer provides the ultimate output of the network [3].

There are different learning paradigms for neural networks, including supervised and unsupervised learning. In unsupervised learning, no labels are used, and the network attempts to recognize patterns and structures in the input data by minimizing or maximizing a cost function. In supervised learning, the network is provided with pre-labelled input data along with corresponding goals or labels. Based on this, the network tries to learn a function that maps the input data to the correct targets. For supervised learning, a sufficiently large training dataset must be provided to the network. The size depends on how complex and variable the objects to be recognized are. In this experiment, the supervised learning paradigm was utilised [3] .

For annotation, simple bounding boxes were used to define the position and extent of the objects to be detected. In addition to the bounding boxes, there are more detailed annotation options, such as polylines, which were excluded due to their time intensity. The bounding boxes were placed as accurately as possible, with little margin, to enable the detection to be as precise as possible. All annotated images come from the same pool, which was later used as the basis for training the YOLO model.

2.2 Chosen Technology

The CNN system used in this experiment is YOLO (You Only Look Once) version 8 (YOLOv8) [4]. It was selected because of its capabilities to detect objects in images or videos in real-time. To implement and conduct object detection, the manufacturer’s platform was utilized, where various network sizes, such as ‘nano (N)’, ‘medium (M)’, and ‘small (S)’, were provided. Larger networks were not tested in this experiment. These network sizes indicate the number of hidden layers present in a CNN. The number of hidden layers can vary and depends on the complexity of the task [3]. To train the YOLOv8 model, the Google Colab platform was utilized (described in 3.2).

For the annotations, CVAT (Computer Vision Annotation Tool) was used to annotate and label data for the subsequent supervised learning step [5]. CVAT offers the advantage of being freely available and supports a variety of annotation types, including bounding boxes for object localization. Due to the absence of specific files (yaml ending) in the import from the annotation tool CVAT to the object recognition platform YOLO, information about the data's location, the number of classes, and their labels were missing (configuration data). As a result, this information had to be added manually to ensure that the model could be trained correctly.

2.3 Used Data

The object under consideration is a construction site of a multistorey building in a German city. For the following analysis, the focus was placed on the basement and the construction of reinforced concrete components (more precisely on the column construction).

The data used in this experiment comprises a storage capacity of approximately 75.7 GB, which corresponds to around 35,000 images. The storage volume includes images in the order of 3-3.1 MB per daytime image and around 1 MB for night images. The construction site was subdivided into two areas both for the ground truth generation and the YOLO results, as a division in the actual construction processes between these two areas (the back area represents area 1, the front area 2) can be recognised over time.

The choice of analysing the column production process is based on the visibility of most of the components from the camera position. Compared to the walls, for example, the step of the formwork elements can be identified more clearly. In addition, the analysed area was also limited to area 1 to avoid any distorted results. The images were captured and stored at 15-minute intervals over the course of a year. Through appropriate data management, the required storage size can be further reduced accordingly (e.g., elimination of night images). The images document the progress of the construction site and were taken from a position next to the excavation pit. Fig. 1 shows an example of the analysed construction site.

Fig. 1**.** Example image of the area under consideration, already including bounding boxes.

3 Implementation

In the implementation step, the states to be analysed are defined at the beginning. Comparative data, which is subsequently used to categorise the results, is recorded and documented using an excel file (3.1). In the next step, the existing data is integrated into the machine learning platform. The model was trained with the labelled training set accordingly in order to be deployed on the other images (3.2). Finally, the results generated by YOLOv8 are presented in 3.3.

3.1 Generating a ground truth

To analyse construction processes on a specific construction site, a shared understanding of the undergone (sub-) processes is required: ‘A process is understood to be the totality of interacting processes within a system’ [6].

Part of the control and optimisation of processes is the analysis and documentation of used resources (material, energy, and information) as well as the monitoring of progress within the process. The aim is to optimise the provision and use of available resources and to complete the process or component on time [6]. Machine learning can be a supporting tool in the control and optimisation of these processes.

For the documentation and analysis of column production, the following activities and states were identified with the help of existing images and the technological knowledge about the production of reinforced concrete work, as shown in Table 1. These were used to define the sub-processes relevant to the production of columns. Due to the time intervals between individual images (15 minutes), the activities (A1-A4) were not considered, and the focus was primarily placed on the construction stages (S1-S4). The choice of construction stages can also be explained by the fact that a hard-output-oriented description [7] can be used for the subsequent analysis with the support of machine learning.

Table 1. Process steps and construction stages used for the recognized object 'column'.

No.	Description
A1	Installing the connecting reinforcement
S1	Connecting reinforcement installed
A2	Production of the reinforcement cage and connection to the corresponding connecting reinforcement
S2	Reinforcement cage installed
A3	Formwork elements are installed around the reinforcement cages
S3	Formwork elements installed
A4	Concreting, formwork removal and curing of the column
S4	Completion of the column

A = activity, S = state of construction

Ground truth data in the form of a detailed schedule of the construction processes and sub-processes was used to categorise the results achieved by the machine learning approach. The ground truth data, consisting of a detailed, daily time schedule of the identified construction works (columns) within the considered area, was created manually. Fig. 2 shows the change in columns over time. In the illustration, the emphasis is on the number of columns in the corresponding construction stages.

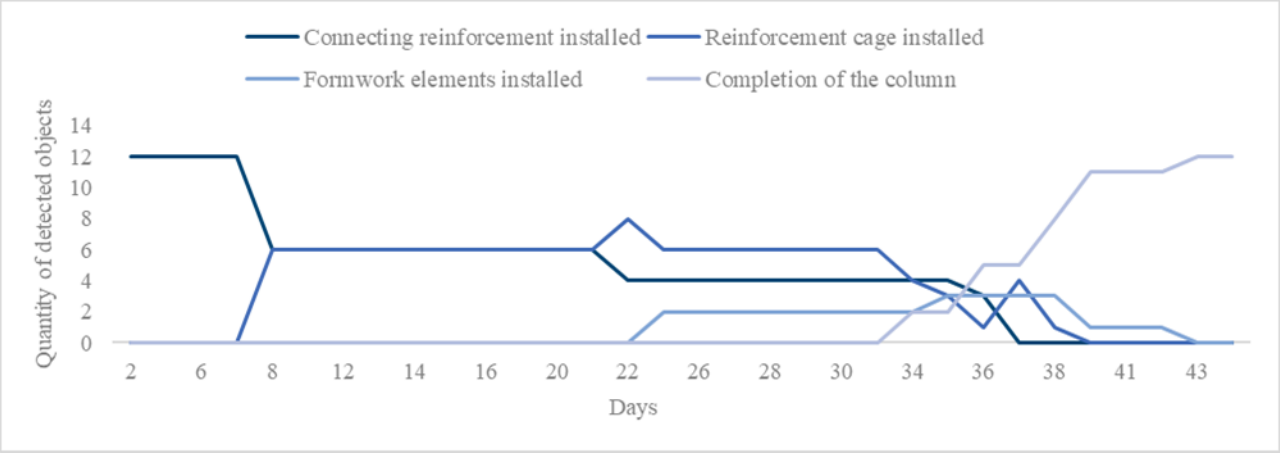

Fig. 2. States of construction for the recognized object 'column'.

3.2 Machine Learning Application

The YOLOv8 model was trained on the Google Colab platform, given that functions from a powerful graphics processing unit (GPU) were necessary for the recognition. The required GPU memory, usually around 12 GB, depends on the complexity of the training data and the desired model accuracy. By using the chosen platform, the training process was performed in a cloud infrastructure that offers scalability and resource efficiency. The model size chosen for this study was designated as 'S' (small) and a careful testing and optimisation process followed to find a balance between performance and resource consumption. Iterative training and testing procedures were carried out, with performance evaluation performed on a selected set of images.

In the context of the 'S' size model, clear trends emerged in object recognition. Excessive identification of connecting reinforcements was observed, while the detection of closed columns was found to be insufficient. The 'reinforcement cage installed' (S2) and 'formwork elements installed' (S3) classes showed comparable recognition quality, with an average maximum probability of 91.6 % for object recognition.

The evaluation of the different models showed that the model of size 'S' outperformed the others and showed a commendable balance between accuracy and resource utilisation. To further improve the performance of the model, various tuning options were explored. In particular, larger installed models were prone to overfitting, which affected their ability to generalise features and apply them to unknown data. To reduce overfitting, the complexity of the ANN had to be reduced [1].

Another key setting parameter was confidence, which indicates the reliability of the model in recognising objects. Empirical tests and comparative analyses showed that a confidence level of 12.5 % results in optimal recognition performance for number plates [8]. The Intersection over Union (IoU) setting, which indicates the correspondence between the recognised bounding box and the manually defined bounding box, was left at 25 % in this study and enabled a robust evaluation of the object recognition [9].

3.3 YOLOv8 Results

The analysed domain parallels human analysis, with images retained in their unaltered state. Consequently, no image processing, such as adjustments to brightness or contrast, was executed. The relevance of natural illumination guided image selection, precluding the inclusion of images with pronounced contrast or shadowed regions encroaching upon the analysis area, given that no construction work takes place in the dark. Recognition quality was deemed susceptible to shadows or nocturnal images, although such factors did not result in complete failure or non-detection of objects. To maintain standardisation, images captured within a specific time window (04:30 a.m. to 07:00 p.m., as per metadata) were chosen. Furthermore, image subsections were not provided, eliminating predefined boundaries for object location in subsequent analyses, thereby mitigating potential distortions. Notably, explicit definitions are warranted for elements like formwork at the construction site periphery, detailing the extent to which these should factor into the evaluation.

Aligning with human evaluation standards, identical classes were employed. Emphasis was placed on achieving a hard-output-oriented description and measurability to ensure clarity and unambiguous results. Fig. 3 shows the results of the object recognition within the employed platform illustratively.

While objects were theoretically recognized, variations in recognition reliability were discerned among different classes. The expectation for classes such as ‘connection reinforcement installed’ (S2) and ‘completion of the column’ (S4) to exhibit maximum column count was not uniformly met. Conversely, classes like ‘formwork elements installed’ (S3) and ‘completion of the column’ (S4) demonstrated heightened certainty in correct recognition.

Fig. 3. Output from the Ultralytics-platform with results.

In summary, the image evaluation process was semi-automated, requiring methodological decisions. Despite encountering limitations and challenges, a foundational object recognition capability was attained. Recognition reliability, however, exhibited variability across distinct classes. The outcomes were notably influenced by perspective and material utilisation. The insights derived from this analysis serve as a groundwork for subsequent investigations and the refinement of future evaluation methodologies.

4 Comparison of YOLOv8 Results with Ground Truth

The results from the comparison between the ground truth data and the ‘generated’ results using machine learning show that all four classes/construction stages (S1-S4) were not fully recognised (see Fig. 4). As part of the study, 299 objects were correctly recognised with machine learning. A total of 91 objects were recognised incorrectly or not at all.

A further 87 objects were under-recognised in the experiment in relation to the comparison data. This indicates that machine learning has potential for improvement.
These results suggest that the support provided by machine learning is not yet optimal and that corresponding potential for improvement needs to be tested to find the precise necessary adjustments.

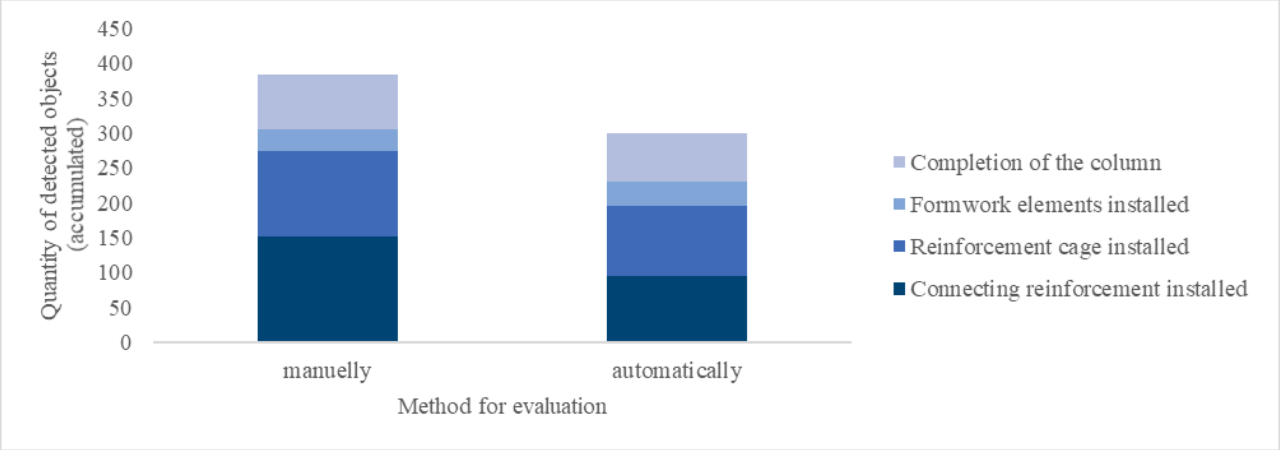

Fig. 4. Difference in the number of recognized classes.

The ‘formwork elements installed’ (S3) and ‘completion of the column’ (S4) classes are nonetheless used to analyse and document construction progress. It was found that more columns were recognised (13 in total) than were planned according to the original design (ground truth data: 12 columns). One possible cause can be attributed to e.g., inconsistent planning data. In this case, more columns were produced on the construction site than the original planning had anticipated. This particular column was therefore not included in the manual analysis. Nevertheless, a clear tendency can be recognised from the available data, as the graphs of the detected columns are close together or overlap. It would therefore be conceivable to report on this indicator. Despite the described differences, clear trends could be recognised through this experiment. This can be argued by the overlapping or slight deviations (or something similar).

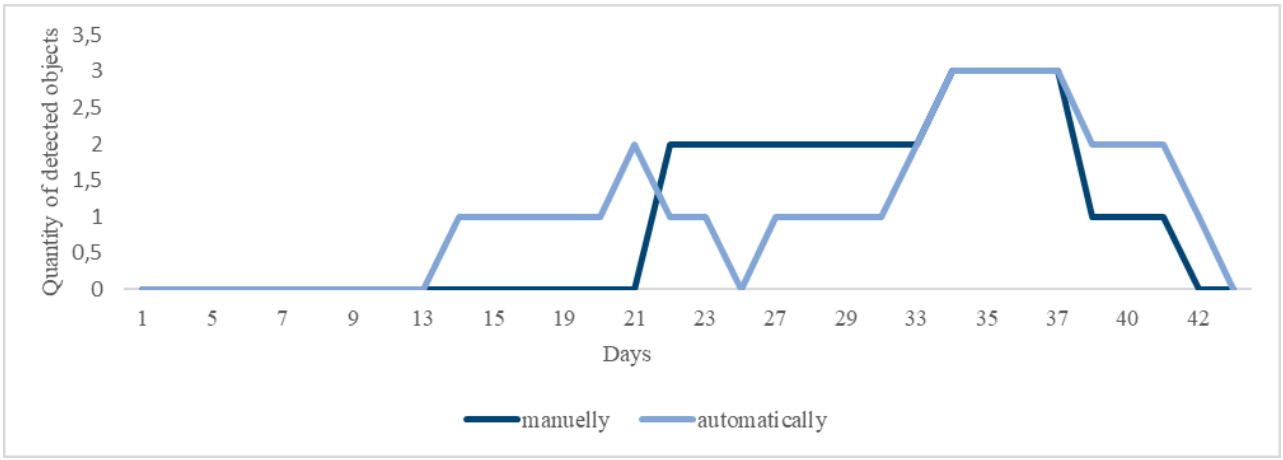

Fig. 5. Recognized formworks per day.

The highest agreement between human evaluation and machine learning was found in the ‘formwork elements installed’ (S3) class (see Fig. 5). Only three misclassifications were recognised in this class over the entire observation period. In addition, all formwork elements used (also in different types) were recognised, which can be seen as a positive result. Fig. 5, which shows both the manually (ground truth) and automatically (YOLOv8 results) recognised formwork objects, shows that no formwork could be identified at day 25, even though it is clear from the ground truth, that formworks had already been built on that date. One possible cause could be, i.e., that formwork elements in the rear

area could be concealed by occlusion from existing reinforcement cages. As a result, the columns and formwork in the background would not be recognised. Fig. 6 shows the same graph for columns.

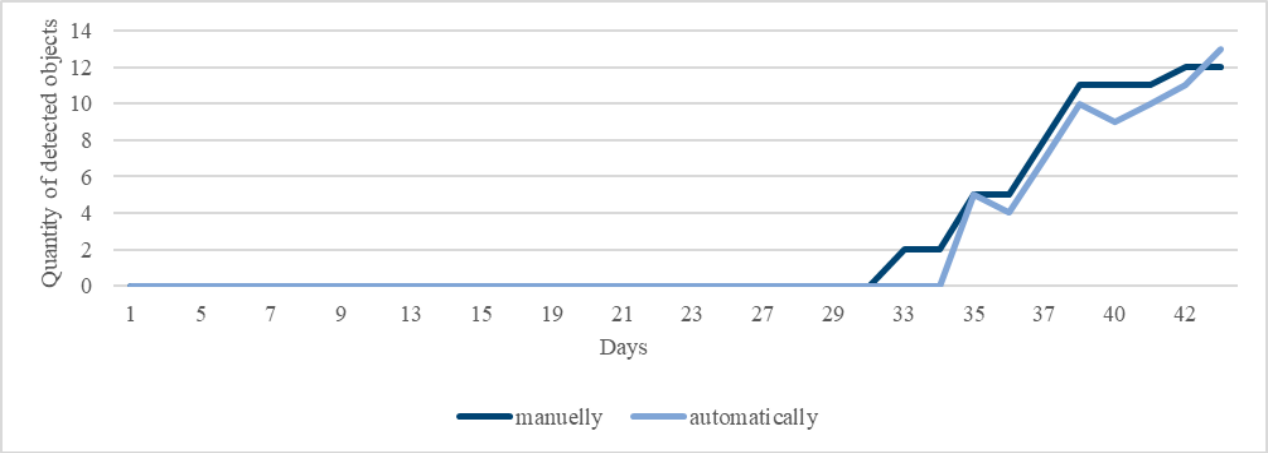

Fig. 6. Completed columns per day.

The ‘connecting reinforcement installed’ (S1) class showed the least agreement between human evaluation and machine learning. Particularly at the beginning of the evaluation, significant discrepancies to the actual value were found, with a total of nine unrecognised connection reinforcements. The achieved results can be deemed unsuited for the aimed use case of construction progress monitoring. As shown in Fig. 7, the only common result between ground truth (manually graph) and the YOLOv8 results (automatically graph) is the recognition on day 37, from which on there are no more connecting reinforcements on the construction site. Recognising the connecting reinforcement understandably poses a particular challenge, as they only protrude a few centimetres from the floor slab and can easily be concealed by other objects. In addition, their colouring is very similar to the underlying floor slab. The increase in recognised connecting reinforcements towards the end of the evaluation could not be logically explained and could be due to incorrect recognition by other objects, such as reinforcements for walls.

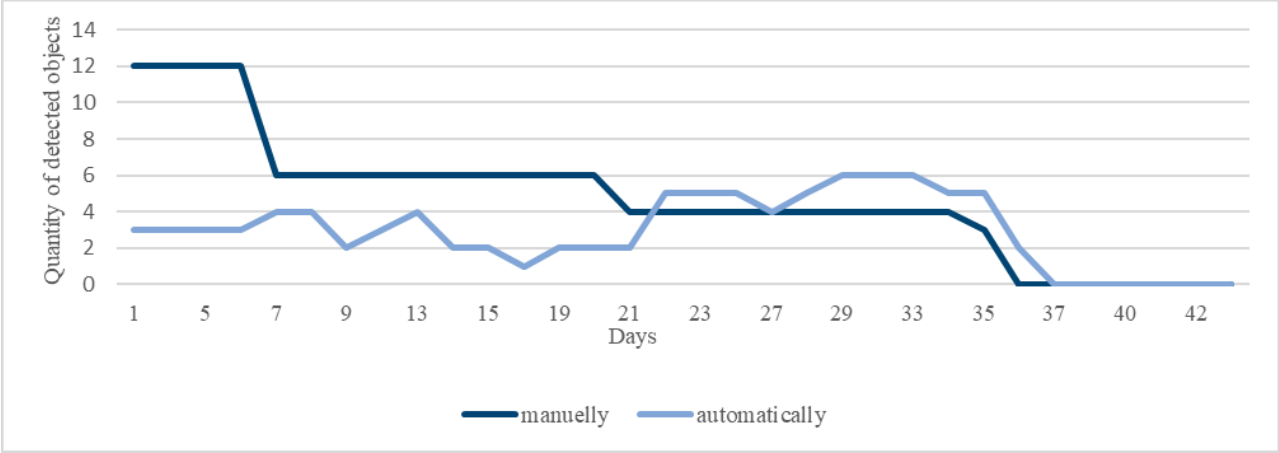

Fig. 7. Number of recognized connection reinforcements per day.

5 Conclusion and Outlook

The experiment investigated the extent to which machine learning, in this particular case object detection results from the YOLOv8 technology, can be employed for construction site progress monitoring, mainly through detection of individual construction stages of defined components.

In general, it was established that object detection on the chosen data set can be achieved and the respective construction stages could be identified using the example of the column construction process. Differences were identified between the detection of individual construction stages (in YOLOv8: classes). As considered in more detail in Section 4, the formwork class showed the greatest match with the comparison data (ground truth data). On the other hand, there were differences between the ground truth data and the results from YOLOv8 in the construction stages 'connecting reinforcement' and 'reinforcement'.

These findings demonstrate the potential of object detection with the support of machine learning for construction site progress monitoring. Moreover, it illustrates not only that, but also how it can be achieved without any coding, thus dismantling a presumed threshold implementation barrier.

For the future practical application and implementation in daily tools of construction project management or construction supervision, it is necessary to automatically analyse a minimum number of images. The increased use will result in corresponding time savings as the number of analysed images increases. With a small number of images, manual evaluation is more efficient, as the initial training and definition of the classes is correspondingly more time-consuming ('initial investment'). On a side note, the conducted experiment showed that recognition with machine learning was possible after around 30 images, validating the general assumption, that the actual manual recognition is more time-consuming than the automated, machine learning approach. An additional programming-based implementation, e.g., Python-based, can be used in a subsequent experiment to further investigate the extent to which this addition influences time saving.

This study rests upon the utilization of image material generated for the specific objectives set out for this study, but without previous knowledge of machine learning specifics. The application of machine learning within this study is confined to a discrete subset of images, employed both for the training of the model and subsequent evaluation processes. To augment result accuracy in future analyses, the used dataset should be extended. This extension should encompass diverse construction site viewpoints to prevent some of the encountered limitations (e.g., occlusions). The incorporation of different construction sites could address the generalizability of the developed models, thereby enhancing their efficacy in practical applications as well as their robustness.

Future studies should prioritise an examination of the transferability of the developed models to projects beyond the scope of the training dataset. Scrutiny of the extent to which the trained network can be seamlessly applied to distinct construction sites without requiring specialized retraining is essential. The objective of this evaluation should be authenticating the applicability of the developed models in varied contexts and ascertaining their general validity for construction site progress monitoring. Another focus of future investigations should be the systematic analysis of viewpoint perspectives and its consequential effects on the accuracy of the object recognition.

6 Acknowledgement

This research as developed in collaboration with the Chair of Construction Chemistry and Polymer Materials at our University, whom we would like to thank in particular for the data acquisition. We would also like to acknowledge the collaboration of Anton Streit and Dominik Reimann.

References

1. Frick D., Gadatsch A., Kaufmann J. et al. (eds) (2021): Data Science: Konzepte, Erfahrungen, Fallstudien und Praxis. Springer Vieweg, Wiesbaden, Heidelberg.
2. LIU W., SHAO Y., ZHAI S. et al. (2023): Computer Vision-Based Tracking of Workers in Construction Sites Based on MDNet. IEICE Trans Inf & Syst E106.D:653–661. https://doi.org/10.1587/transinf.2022DLP0045.
3. Mockenhaupt A. (2021): Maschinelles Lernen. In: Mockenhaupt A (ed) Digitalisierung und Künstliche Intelligenz in der Produktion. Springer Fachmedien Wiesbaden, Wiesbaden, pp 133–163.
4. Ultralytics Inc. (2023): YOLOv8 Repository. https://github.com/ultralytics/ultralytics. Accessed 15 Dec 2023
5. CVAT.ai Corporation (2023) CVAT Repository. https://github.com/opencv/cvat. Accessed 15 Dec 2023.
6. (2019) Gabler Wirtschaftslexikon, 19. Auflage. Springer Gabler, Wiesbaden, Heidelberg.
7. Erlei M., Leschke M., Sauerland D. (2007): Neue Institutionenökonomik, 2. Aufl. Schäffer-Poeschel Verlag, s.l.
8. Laroca R., Severo E., Zanlorensi L. A. et al. (2018): A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector. In: 2018 International Joint Conference on Neural Networks (IJCNN): 2018 proceedings. IEEE, Piscataway, NJ, USA, pp 1–10.
9. Terven J., Cordova-Esparza D. (2023): A Comprehensive Review of YOLO: From YOLOv1 and Beyond.