Kiran Sebastian, BSc

Object Detection for Tracking Assembly Processes in the Context of Digital Value Stream Mapping

Master's Thesis

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme:
Production Science and Management

submitted to

Graz University of Technology

Supervisor
Institute of Innovation and Industrial Management
Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Ramsauer
Ass.Prof. Dipl.-Ing. Dr.techn. BSc Matthias Wolf

Second supervisor
Fraunhofer Austria Research GmbH
Dipl.-Ing. Noël Scheder

Graz, 2024

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

…… ……

Date Signature

Acknowledgement

First and foremost, I would like to express my sincere gratitude to Univ.-Prof. Dipl.-Ing. Dr.techn. Christian Ramsauer for providing me with the opportunity to write my master's thesis at the Institute of Innovation and Industrial Management.

I am also deeply thankful to Ass.-Prof. Dipl.-Ing. Dr.techn. Matthias Wolf for his invaluable guidance and supervision throughout the course of this work. His insightful feedback was crucial in defining the structure and elevating the quality of this thesis.

I am sincerely thankful to Dipl.-Ing Mathias Nausch and Dipl.-Ing Noël Scheder from Fraunhofer Austria Research GmbH for introducing me to this research topic and for their guidance and support throughout this work.

In addition, I am grateful to the faculty members of the IIM institute and my colleagues, whose assistance in setting up the experimental environment and participation in the practical aspects of this thesis were essential to its success.

Also, I would like to thank my wife, Sandra, for her unwavering love and belief in me and for motivating me throughout my thesis journey.

I am also profoundly grateful to my family – my parents, Dr. T.J. Sebastian, and Mrs. Jc Paul, and my sister, Reshma – without their support this would not have been possible.

I am also thankful to my friends in India and Austria for their encouragement, and support throughout this journey.

Abstract

Traditional value stream mapping (VSM) relies on manual data collection and static representation of production processes, which are becoming less effective in today's dynamic and flexible manufacturing environments. As product life cycles shorten and production systems demand greater flexibility, the need for automated data collection is becoming increasingly important. While several methods have been proposed for digitizing VSM, few focus on data collection from human centric assembly processes. Although existing solutions enable data collection, often increase system complexity and disrupt workflow, negatively impacting productivity. Object detection presents a promising alternative, enabling non-intrusive data collection. While already being applied in industrial settings for various tasks, its potential to collect data specifically for supporting value stream map creation remains largely unexplored.

This thesis investigates the use of object detection for automated data collection for VSM creation. A logic was developed to use detection results to track products and workers on the shop floor and calculate KPIs such as cycle time and throughput time using the captured location and time data.

The proposed method was tested at the learning factory at TU Graz, where a custom trained YOLOv8 model successfully detected and tracked products and workers during the assembly process. The data was then processed in excel to digitally visualize the VSM. The comprehensive data enabled detailed visualizations, including Gantt charts and pie charts, providing a clearer depiction of the situation.

The non-intrusive nature of object detection technique, requiring only a camera and an algorithm, proved advantageous over other sensors-based systems. Such systems often require attaching physical sensors to the entities being tracked, adding complexity to the process, especially in environments where a high volume of items needs to be tracked. Object detection eliminates this need, reducing the workload on employees and minimizing the risk of errors associated with misplaced or malfunctioning sensors.

The test results showed that the digitally collected data closely matched the actual assembly scenario, demonstrating the viability of object detection for industrial applications. With further validation in larger and more complex environments, this approach has the potential to significantly enhance how data is collected and utilized for VSM in manual assembly processes.

Kurzfassung

Die traditionelle Wertstromanalyse (VSM) basiert auf manueller Datenerfassung und statischer Darstellung von Produktionsprozessen, was in dynamischen und flexiblen Fertigungsumgebungen zunehmend ineffektiv ist. Da kürzere Produktlebenszyklen und flexible Produktionssysteme eine automatisierte Datenerfassung erfodern, wurden zwar verschiedene Methoden zur Digitalisierung von VSM entwickelt, doch nur wenige fokussieren auf menschenzentrierte Montageprozesse. Bestehende Lösungen erhöhen oft die Systemkomplexität und beeinträchtigen den Arbeitsfluss. Die Objekterkennung bietet eine vielversprechende, interaktionslos Alternative zur Datenerfassung, deren Potenzial zur Unterstützung von Wertstromkarten jedoch noch wenig erforscht ist.

In dieser Arbeit wird der Einsatz von Objekterkennung zur automatisierten Datenerfassung für die Erstellung von VSM untersucht. Es wurde eine Logik entwickelt, um die Erkennungsergebnisse zur Verfolgung von Produkten und Mitarbeitern auf dem Shopfloor zu nutzen und anhand der erfassten Positions- und Zeitdaten KPIs wie Zykluszeit und Durchlaufzeit zu berechnen.

Die vorgeschlagene Methode wurde in der Lernfabrik der TU Graz getestet, wo ein speziell trainiertes YOLOv8-Modell erfolgreich Produkte und Arbeiter während des Montageprozesses erkannte und verfolgte. Die Daten wurden anschließend in Excel verarbeitet, um das VSM digital zu visualisieren. Die umfassenden Daten ermöglichten detaillierte Visualisierungen, einschließlich Gantt-Diagrammen und Tortendiagrammen, die eine klarere Darstellung der Situation ermöglichten.

Die nicht-invasive Natur der Objekterkennungstechnik, die nur eine Kamera und einen Algorithmus erfordert, erwies sich als vorteilhaft gegenüber anderen sensorbasierten Systemen. Solche Systeme erfordern häufig das Anbringen physischer Sensoren an den zu verfolgenden Objekten, was die Komplexität des Prozesses erhöht, insbesondere in Umgebungen, in denen eine große Menge an Gegenständen verfolgt werden muss. Die Objekterkennung beseitigt dieses Erfordernis, reduziert die Arbeitsbelastung der Mitarbeiter und minimiert das Risiko von Fehlern, die durch falsch platzierte oder defekte Sensoren entstehen können.

Die Testergebnisse zeigten, dass die digital erfassten Daten dem tatsächlichen Montageszenario sehr nahe kamen, was die Eignung der Objekterkennung für industrielle Anwendungen beweist. Mit weiterer Validierung in größeren und komplexeren Umgebungen hat dieser Ansatz das Potenzial, die Datenerfassung und -nutzung für VSM in der modernen Fertigung erheblich zu verbessern.

Table of Content

1 Introduction

The manufacturing industry is experiencing a major transformation driven by rapid advancements in digital technologies. This shift is being accelerated by growing pressure on manufacturers to find solutions that enhance flexibility while simultaneously maintaining efficiency. Traditional tools and methodologies that once ensured efficiency are increasingly becoming inadequate in addressing these challenges. Digitization is seen as a key solution to these problems. By enhancing the capabilities of traditional tools, digitization could enable manufacturers to adopt data-driven operations, resulting in greater flexibility and improved decision-making.[1]

Lean manufacturing is a widely adopted approach used by many companies to improve efficiency and productivity. It seeks to minimize waste while maximizing value throughout the production process. One of the key tools used in Lean Manufacturing is Value Stream Mapping (VSM), which is an effective methodology for identifying inefficiencies and driving continuous improvement. By creating transparency across the entire process, VSM provides valuable insights into material and information flows, ultimately aiming to reduce waste and enhance productivity.[2]

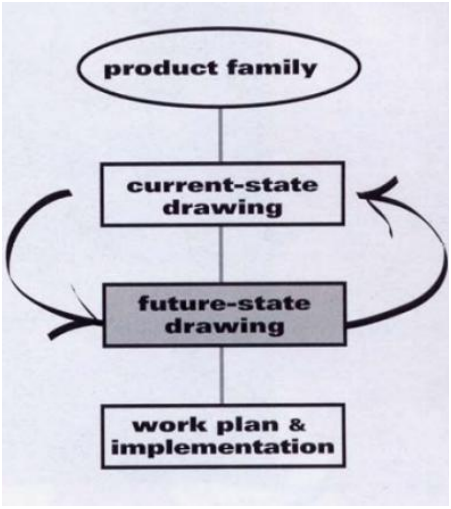


Figure 1-1: Basic steps in the value stream mapping process[3]

[1] Horsthofer-Rauch, J. et al. (2022).
[2] Tran, T.-A. et al. (2021).
[3] Rother, M., & Shook, J. (1999).

This is achieved by mapping the detailed current state of production, and then designing the future state with the goal of optimizing the process by eliminating the bottlenecks, inefficiencies and wastes identified in the current state.[4]

However, today's production environment is evolving rapidly, making it increasingly important to maintain flexible and dynamic production systems. Rising global competition and changing customer requirements have significantly affected the way products are being manufactured. To remain competitive, manufacturers are compelled to offer a high degree of customization, accommodate short-term change requests, and produce a wider variety of products. While VSM has proven effective in stable and predictable manufacturing environments, it is limited in its ability to capture the increasing complexity and dynamism that modern production environments demand.[5]

Traditional VSM relies on manual data collection, where the production data from a representative unit is captured to evaluate the performance of the production process. This static representation is often inefficient for production systems that undergo frequent changes and conducting regular data collection is impractical due to the significant effort involved in the process. The adoption of digitization, the growing trend driven by industry 4.0, has been proposed as a solution to address these challenges in value stream mapping. Digital technologies enable dynamic, real-time data collection, offering a more accurate and responsive approach to identifying inefficiencies and optimizing production flows.[6]

Real time data collection from shop floors can equip manufacturers with up-to-date insights into their production processes, allowing managers to make informed decisions and react quickly to changes in demand or production conditions.[7] Various approaches for digitizing the value stream mapping process have been suggested in the literature, including digital data collection, digital analysis, and digital visualization of the collected data. However, capturing comprehensive data digitally remains challenging, particularly in human-centered production processes, where the level of unpredictability is higher compared to automated or mechanized systems.[8] This indicates a need for further research into new methods for digital data acquisition.

This thesis aims to contribute to the field by exploring the use of object detection as a tool for automating the data collection process in value stream mapping. Object detection's capability to identify and track objects in real-time can enable the collection of diverse

[4] Rother, M., & Shook, J. (1999).
[5] Scheder, N. et al.(2023).
[6] Wang, H.-N. et al. (2021).
[7] Tran, T.-A. et al. (2021).
[8] Sullivan, B. et al. (2022).

data. Moreover, since object detection algorithms can be trained to detect a wide variety of objects of interest, it holds potential for a range of use cases in dynamic manufacturing environments.

1.1 Problem Statement

The digitization of Value Stream Mapping (VSM) has emerged as a promising solution to overcome the limitations of traditional, static VSM process. Even partial digitization of VSMs has demonstrated significant benefits, including reduced effort in the mapping process and fewer accidental errors.[9] However, the true potential of digitization lies in the ability to collect, process, and visualize value stream maps in real-time. This would enable manufacturers to implement fully flexible, dynamic production systems while maintaining the efficiency of traditional manufacturing methods.[10]

Despite several proposed approaches for real-time data collection, advanced analysis, and real time visualization, comprehensive value stream mapping of entire processes in real-time remains challenging. Many existing studies propose using digital tools for intelligent analysis and visualization of real-time data, but their usage for continuous monitoring is limited by the inability to collect real-time data continuously. While information systems such as ERP and MES can provide some of the data relevant for mapping information flow, the collection of material flow data from shop floors is more complex. [11]

Existing approaches for collecting material flow data tend to focus on machine-dependent production by monitoring machine utilization. Whereas solutions for capturing data from manual assembly processes are more limited, as these processes are typically harder to track due to the unpredictability of human actions. No single sensor maybe capable of collecting all the complexities of a production process to create complete value stream maps. However, leveraging tools that can monitor multiple KPIs of the production process could minimize the number of sensors required, thereby reducing the overall complexity of the data collection process.[12]

Fraunhofer Austria, along with Fraunhofer IPA developed and tested a sensor-based toolkit for digital data collection from shop floors. This included the necessary hardware and software to enable near-real-time data transfer from sensors for value stream analysis. The validation results showed that the manual triggering required for the sensor-

[9] Klimecka-Tatar, D., & Ingaldi, M. (2022).
[10] Frick, N., & Metternich, J. (2022).
[11] Tran, T.-A. et al. (2021).
[12] Sullivan, B. et al. (2022).

based approaches created additional workload to the production employees and affected their productivity. In addition, this creates a new potential source for errors due to misplacement or mishandling of sensors. To maximize the benefits of digitization in value stream mapping, there is a critical need for non-intrusive methods, that can gather digital data from assembly areas without affecting the productivity of the workers. Researchers have proposed object detection as a promising solution to overcome this challenge.[13]

This thesis addresses this gap by investigating the use of object detection as a tool for non-intrusive data collection from shop floors. Unlike existing sensor-based approaches, object detection enables data collection without attaching physical tags or labels on the entities being tracked. This can reduce the complexity associated with the maintenance and physical tagging of the sensors. Object detection algorithms can be trained to detect and track a variety of objects simultaneously, enabling the collection of diverse information in real-time. This could provide a more efficient and less intrusive method for data collection, without interrupting the workflow.

1.2 Aim and Objectives

The aim of this study is to evaluate the potential of object detection as a non-intrusive method for data collection to enable digital value stream mapping.

Objectives:

- To compare available object detection algorithms and select the most suitable option for application in an industrial environment.
- To train object detection models to accurately detect and identify the entities to be tracked.
- To develop a logical framework to use object detection and tracking for collecting data relevant for value stream mapping.
- To test the developed method and thereby evaluate the potential of object detection and tracking as a non-intrusive data collection approach for digital value stream mapping.

Primary Research Question: How capable is object detection to enable data collection from shop floors for digital value stream mapping?

Secondary research questions:

- How are current digital technologies being applied to enhance value stream mapping?

[13] Nausch, M. et al. (2023).

- How is object detection currently utilized in assembly processes and what considerations are essential for selecting an algorithm for an industrial application?
- Which object detection algorithms are most suitable for application in industrial environments?

1.3 Structure of the thesis

This thesis is structured into eight main chapters, each designed to guide the reader through the research journey from initial problem statement to the final conclusions and outlook. Chapter 1 introduces the thesis, outlining the problem statement, aim, objectives, as well as the research questions.

This is followed by the theoretical background in Chapter 2, which provides fundamental concepts essential for understanding the research. It covers the principle of Lean Manufacturing, the concept of Value Stream Mapping, and gives an overview of Industry 4.0 technologies

Chapter 3 explains the theory related to the work. It details the methodological approach used for the literature review and describes its result on three topics – Digitization in value stream mapping, Current applications of object detection in assembly processes, and an overview of object detection algorithms.

Chapter 4 details the practical implementation of the thesis. It begins with the conceptualization of the approach and proceeds with a thorough description of technical procedure, including model development, data acquisition and processing, KPI determination, and visualization. Finally, details of the testing conducted at the LEAD factory in TU Graz is explained in detail, covering the approach, experimental procedure, and preliminary results.

Chapter 5 presents the results of the practical experiment. It evaluates the performance of the object detection model, detailing training duration, resource efficiency, and accuracy. The results from testing in the learning factory are also analysed to assess the accuracy of the collected data and overall effectiveness of the approach.

Chapter 6 presents the discussion of the study. The implications of the results are discussed along with its limitations. The potential challenges for industrial applications are also discussed.

The thesis concludes with Chapter 7, which summarizes the key contributions, highlighting how the research question was addressed, and the objectives were met. Finally, in chapter 8, an outlook for the thesis is presented, mentioning possible future works and recommendations for improvement.

2 Theoretical Background

The theoretical background of this thesis describes the foundational concepts essential for understanding the research scope, particularly focusing on Lean manufacturing, Value Stream Mapping (VSM), and Industry 4.0 technologies.

2.1 Lean Manufacturing

Lean Manufacturing is a systematic approach aimed at maximizing the product value by minimizing waste. Originating from the Toyota production System(TPS), Lean was developed to create efficient production systems that could operate with limited resources while still ensuring high quality and flexibility. The philosophy of lean has extended beyond its automotive origins and is now widely applied across multiple industries globally.[14]

The House of TPS framework, shown in Figure 2-1, visually represents the foundational structure and guiding principles of Lean Manufacturing. The house of TPS stands on two main pillars: Just-in-Time (JIT) production and Jidoka (automation with a human touch). JIT ensures that production occurs only when there is demand, thereby minimizing inventory, storage costs, and the risk of overproduction. Jidoka, on the other hand, focuses on quality by automatically stopping production when a defect is detected, ensuring that defects are not passed down the line. In addition to JIT and Jidoka, Lean Manufacturing utilizes a range of tools and techniques that support these principles and help achieve its primary objective of maximizing value by minimizing waste.[15]

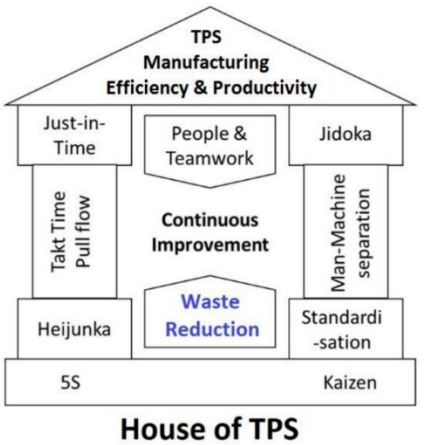


Figure 2-1: House of Toyota Production System[16]

[14] Sundar, R. et al. (2014).
[15] Lai, N. et al. (2019).
[16] Ibidem

2.1.1 Lean Principles

Womack and Jones, D. defined lean manufacturing through five core principles that guide organizations in systematically enhancing processes to maximize customer value while minimizing waste. These principles promote continuous improvement within value streams by identifying and eliminating non-value adding activities and focusing on activities that truly create value.[17]

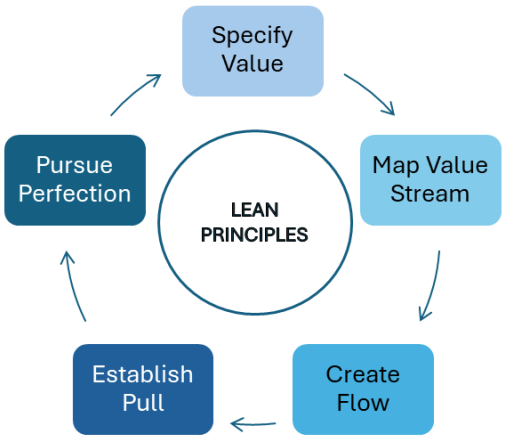


Figure 2-2: Key principles of lean manufacturing[18]

2.1.1.1 Specify Value

The first principle of lean is understanding what the customer values. Value is defined from the customer's perspective, with the goal of delivering exactly what the customer needs without adding any unnecessary complexity. Any activity that does not directly contribute to delivering this value is considered waste and should be minimized.[19]

2.1.1.2 Map Value Stream

The second principle involves mapping all the steps in the production process, with an aim to identify activities that add value and those that do not. These activities within the value stream can be classified into three categories: (i) those that add value, (ii) those that do not add value but are currently necessary, (iii) those that add no value and can be eliminated. By making this distinction, firms can immediately remove the third category

[17] Womack, J., & Jones, D. (1996).
[18] Based on Womack, J., & Jones, D. (1996)., own representation
[19] Čiarnienė, R., & Vienažindienė, M. (2012).

and eventually remove the second category of wastes in later steps to further optimize their processes.[20]

2.1.1.3 Create Flow

After identifying waste in the value stream, this principle focuses on redesigning the process to achieve a continuous, uninterrupted flow of products through each value-adding step. This is achieved by eliminating waste, which includes any activity not valued by the customer, such as excess inventory, waiting times, or unnecessary transportation of materials.[21]

2.1.1.4 Establish Pull

Lean production is driven by actual customer demand rather than predictions. In a pull system, the customer pulls the product from the manufacturer, ensuring that production aligns closely with customer needs. This approach aligns all aspects of production, from raw materials through final assembly, minimizing overproduction and inventory levels.[22]

2.1.1.5 Pursue Perfection

The final principle of lean manufacturing emphasizes the need for continuous pursuit of perfection. As the first four principles are implemented, activities within the value stream become more transparent, enabling ongoing improvements. This principle encourages organizations to regularly review and refine their processes to eliminate defects, shorten cycle times, and enhance quality.[23]

Striving for perfection creates a mindset of continuous waste elimination, where improvement is seen as an ongoing process rather than a one-time effort. By this iterative improvement approach, lean practitioners aim to reach a point where every resource and activity in the process directly contributes value to the end product.[24]

2.1.2 Lean Manufacturing Objectives

As defined in the previous section, the core objective of lean manufacturing is to identify and eliminate any unnecessary activities, referred to as muda (waste), in the production

[20] Womack, J., & Jones, D. (1996).
[21] Čiarnienė, R., & Vienažindienė, M. (2012).
[22] Thangarajoo, Y. (2015).
[23] Čiarnienė, R., & Vienažindienė, M. (2012).
[24] Ibidem

process. Seven categories of waste are defined in the traditional lean manufacturing concept as given below:[25]

Motion: Unnecessary movement of people or products, such as walking long distances between workstations or reaching for tools, that does not contribute value to the product or service is regarded as a waste of motion.[26]

Waiting: Waiting refers to the time wasted when goods or workers are idle, waiting for the next step in the production process.

Overproduction: Overproduction refers to producing more than what is needed, leading to excess inventory and increased storage costs.

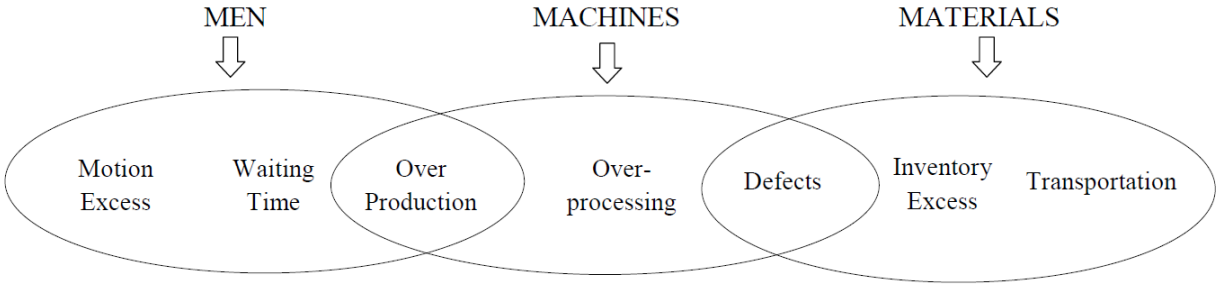


Figure 2-3: Types of wastes in manufacturing[27]

Over-processing: Overprocessing refers to doing more work than what the customer demands. Adding unnecessary features or using complex methods when simpler ones are available are wastes than does not add any value to the customer.

Defects: Producing defective products that require rework or scrapping is wasteful, as it leads to unnecessary loss of effort, money, and time.

Inventory: Inventory refers to the stock of raw materials, work-in-process items, and finished goods. Holding excess inventory ties up capital, increases storage costs, and increases the risk of becoming outdated.[28]

Transportation: It costs time and money to move materials from one place to another. Unnecessary movement of materials, which does not add value to the product is considered a waste.[29]

[25] Womack, J., & Jones, D. (1996).
[26] Kumar, N. et al. (2022).
[27] Čiarnienė, R., & Vienažindienė, M. (2012).
[28] Sundar, R. et al. (2014).
[29] Kumar, N. et al. (2022).

2.1.3 Lean Tools and Techniques

Several tools and techniques are employed to achieve lean manufacturing by reducing waste, improving efficiency, and creating value. These tools are essential for applying lean principles effectively. Some of the most commonly used lean manufacturing tools are described in this section.

2.1.3.1 5S System

5S is a method for workplace organization that aims to maintain a productive, clean, and efficient workspace. The method emphasizes that the workspace should be properly designed with a specific place for everything. This ensures that only the required items are present and those are available whenever required. The five S's stand for Sort, Set in Order, Shine, Standardize, and Sustain.[30]

2.1.3.2 Total Productive Maintenance

Total Productive Maintenance (TPM) is a lean manufacturing methodology aimed at maximizing the productivity and efficiency of equipment by involving all employees in the manufacturing process. TPM focuses on maintaining equipment in optimal working condition to prevent unplanned downtime, and reduced defects. The primary goal of TPM is to achieve zero breakdown, zero defects, and zero accidents by fostering a sense of ownership and responsibility for equipment and tools among the operators using it.[31]

2.1.3.3 Value Stream Mapping (VSM)

A Value Stream Map (VSM) is a visual tool used to map the material and information flows necessary to coordinate the activities of manufacturers, suppliers, and distributors, with an aim to deliver products to customers efficiently. It covers all actions required to bring a product through problem-solving, information management, and physical transformation stages within a business.[32]

The VSM process begins by creating a current state map, which visualizes the material and information flows within the system. This is used to identify waste sources and highlight areas for potential improvements. A future state map is then developed as an improvement plan, providing a structured approach to continuous improvement.[33]

[30] Palange, A., & Dhatrak, P. (2021).
[31] Sundar, R. et al. (2014).
[32] Ibidem
[33] Ibidem

2.1.3.1 Heijunka (Level production)

Heijunka or production levelling, is a lean manufacturing strategy aimed at reducing fluctuations in production by levelling customer demand. In volatile business environments where customer demands vary, Heijunka helps prevent issues such as man and machine idle times, quantity problems, and breakdowns caused by overburdening of capacities. This approach balances workloads by grouping products into families and scheduling their production at regular intervals. Heijunka aims to manage variability in job arrival sequences to enable higher capacity utilization while avoiding peaks and valleys in the production schedule.[34]

2.1.3.1 Just-in-Time (JIT)

Just-in-Time is a production philosophy used in lean manufacturing that minimizes inventory by closely aligning production with customer demand. Under JIT, products and parts are manufactured and delivered only as needed, following a pull-based approach where production schedules are based on the customer demand. JIT aims to provide each process with only the parts it needs at the moment they are required, reducing lot sizes, buffer inventories, and order lead times. Trained workers, efficient workspace organization, and effective equipment maintenance are some of the key components required for successfully implementing JIT manufacturing. Efficient communication and reliable relationships with suppliers are also essential, as materials must be supplied on-time according to the requirements.[35]

2.1.3.1 Kanban System

Kanban is a visual scheduling system designed to streamline production by controlling inventory levels and synchronizing the production and supply of components. It controls the flow of materials and ensures that production aligns with the demand, following the pull principle for achieving the Just-in-Time philosophy.[36]

Physical Kanban cards or digital e-kanbans are used to signal demand to ensure parts are supplied based on customer requirements. The term ‘customer’ here refers to both external customers, who are the end users of finished products, and internal customers, the production personnel at the succeeding stations in a manufacturing facility.[37]

[34] Sundar, R. et al. (2014).
[35] Kumar, V. et al. (2019).
[36] Sundar, R. et al. (2014).
[37] Palange, A., & Dhatrak, P. (2021).

2.1.3.1 Jidoka

Jidoka or ‘automation with a human touch’ is used to stop production whenever a problem occurs, enabling operators to immediately identify and correct issues. In a Jidoka system, machines are equipped to halt production automatically upon detecting abnormalities, and workers are trained to identify and address these issues. By signalling operators when a defect or abnormality occurs, Jidoka ensures that faulty products are not processed further, reducing the impact of the problem and ensuring quality standards.[38]

2.1.3.2 Poka-Yoke (Mistake Proofing)

Poka-Yoke is a lean manufacturing tools for achieving mistake proofing. It is designed to support the Jidoka system, trying to prevent defects by ensuring that processes operate under the correct conditions before the start of each process. This approach enhances control, also preventing issues like over-processing and overproduction. Poka-Yoke serves as both a preventive and detective mechanism, trying to identify potential errors early and correcting them before they impact the final product.[39]

2.1.3.1 Kaizen (Continuous Improvement)

Kaizen, a Japanese term meaning ‘continuous improvement’ is a practice of continuous, incremental improvement involving all employees from management to shop floor workers, to improve processes and eliminate muda (waste). In manufacturing, Kaizen targets inefficiencies in machinery, labour, and production methods.[40]

[38] Kumar, V. et al. (2019).
[39] Ibidem
[40] Gupta, S., & Jain, S. (2013).

2.2 Value Stream Mapping

A value stream consists of all the actions required to deliver a product to the customer. The core of value stream is to view processes from the customers perspective and evaluate them based on whether they add value to the customer. This approach focuses on the entire value creation chain rather than individual steps, allowing a high-level overview that helps identify areas for improvement across the whole process.[41]

In order to optimize the production processes effectively, it is necessary to visualize the entire value stream. Value Stream Mapping (VSM) is used to achieve this. It is designed to visualise the value stream, with an aim to analyse and optimize it by identifying and eliminating the unwanted activities.[42]

2.2.1 History of Value Stream Mapping

Similar to most of the lean manufacturing tools, the history of VSM originates from the processes adopted by Toyota. At Toyota, however, it was known as “Material and Information Flow Mapping”, and they used the tool to depict current and future states. For mapping the current and future states, Toyota defined three types of flows: the flow of material, the flow of information, and the flow of people/processes. Mike Rother and John Shook, in their book, “Learning to See”, adopted the material and information flow maps used by Toyota to present the Value Stream Mapping Method. The book provides a detailed explanation of the concept and a step-by-step procedure for implementing the method.[43]

2.2.2 Value Stream Mapping Process

Value Stream mapping (VSM) is used to visualise and analyse the flow of materials, information, and processes required to deliver a product or service to the customer. The primary objective of VSM is to identify and eliminate waste and enhance the overall efficiency of production processes. By visualising the entire processes involved in the workflow, VSM allows organizations to pinpoint inefficiencies and bottlenecks, thus supporting continuous improvement, aligning with lean principles. VSMs are created for specific product families, helping managers understand product specific operational conditions.[44] The steps involved in VSM creation is describes in Figure 2-4.

[41] Rother, M., & Shook, J. (1999).
[42] Ibidem
[43] Ibidem
[44] Jasti, N., & Sharma, A. (2014).

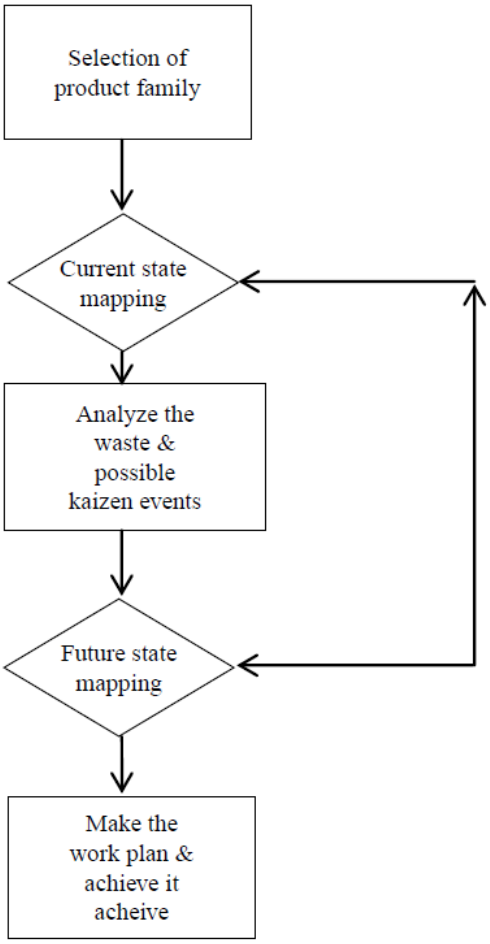


Figure 2-4: Steps involved in value stream mapping[45]

2.2.2.1 Selection of product family

When starting the VSM process, it is essential to focus on a single product family rather than attempting to map all products in the facility. VSM is a customer centric process, aiming to increase the value for the customers. They are primarily concerned with specific products they purchase and not with the entire product line. Also, it is impractical to represent every product flow on one map, especially in larger operations.[46]

[45] Jasti, N., & Sharma, A. (2014).
[46] Rother, M., & Shook, J. (1999).

A product family consists of items that share similar processing steps and utilize common equipment in the production process. Product families are to be identified from the customer end of the value stream.[47]

2.2.2.2 Current State mapping

The current state mapping begins by mapping out the entire processes in the workflow, detailing each step from start to finish. In traditional value stream mapping, the data is collected manually from the shop floor. The employee uses devices such as stop watches for time measurement and visual observation to record the process steps, material flow, and information flow as they happen in real-time. Each step that the material goes through is documented from the receipt to production to shipping.[48]

To visualize the current state of the production process, certain key performance indicators are measured and displayed within the value stream. These metrics provide insight into the efficiency, flow, and responsiveness of the production system and are outlined below:[49]

Cycle time – Cycle time represents the time taken to complete a specific task or process within a single workstation.

Throughput time – Throughput time is the total duration taken for a single unit of product to pass through all stages in the production process, from start to finish. This metric includes both value-adding and non-value adding activities included in the production process.

Lead time – Lead time is the total time taken from the point the order is received from the customer till the product is delivered to the customer. Reducing the lead time is essential to improve the efficiency of the process.

Takt time – Takt time is calculated based on customer demand and represents the ideal production rate needed to meet this demand. It serves as target time per unit, ensuring that the customer demand is satisfied without leading to over production.

Work in Progress (WIP) – Work in progress (WIP) refers to the total number of units being processed at a point of time, but not yet completed. Excessive WIP slows down production flow and increase waiting times.

Transport time – Transport time refers to the time spent moving material or products between workstations.

[47] Rother, M., & Shook, J. (1999).
[48] Tabanli, R., & Ertay, T. (2013).
[49] Rother, M., & Shook, J. (1999).

2.2.2.3 Analysis of the current state

The primary goal of analysing the current state is to understand where bottlenecks, delays, and non-value-added activities occur within the production flow. This analysis reveals which processes contribute to inefficiencies and their impact on overall performance. During this step, different activities in the value stream are classified based on their contribution in the production process, as defined below:[50]

Value Adding Activities (VAA) – These activities are fundamental to the production process and directly contribute to the delivery of a finished product to the customer. They add value from the customer's perspective.

Necessary but Non-Value Adding (NNVA) – These activities do not directly add value to the customers, but they remain essential for the current operating procedures. Removing them would require a fundamental change in the way products are currently being manufactured.

Non-Value Adding (NVA) – These activities are pure waste, meaning they add no value to the final product and should be completely eliminated. They involve activities that consume time and resources, without contributing to the production process.

This classification helps prioritize improvement efforts by clearly identifying the wastes in production. The NVAs are targeted for immediate elimination, whereas NNVAs are eventually eliminated by bringing changes in the production process. [51]

2.2.2.4 Future State Mapping

Future State Mapping is the step in the VSM process, where an ideal, streamlined version of the production process is designed. Based on insights from the current state analysis, a future state map is outlined, aiming to eliminate the identified wastes and optimizing processes to meet takt time – the ideal rate at which products must be produced to precisely meet the customer demand. This alignment ensures that production closely matches demand, reducing overproduction, excessive inventory, and other inefficiencies.[52]

Several guidelines have been defined to help achieve improved future state maps. These include setting the production rate based on Takt time, and scheduling production based on bottleneck operation. The future state is planned with an aim to maintain continuous

[50] Hines, P., & Rich, N. (1997).
[51] Ibidem
[52] Tabanli, R., & Ertay, T. (2013).

flow in production, where material and information move smoothly without stoppages or delays, minimizing lead times.[53]

2.2.2.5 Implementation

After designing the future state map, the goal is to implement the new plan. The value stream map is nearly worthless if it does not contribute to achieving improved efficiency.[54] During this stage, a detailed workplan is established, outlining the specific actions required, the personal involved, and the timeline for each step. Regular monitoring and feedback loops are essential in the implementation stage to ensure that the planned improvements are achieved.[55]

2.2.1 Benefits of Value Stream mapping

VSM offers a range of benefits to manufacturers, which is evident from the popularity of the technique. It helps manufacturers to visualize their production processes and discover inefficiencies, they never thought existed. It also eliminates potential misinterpretations, enabling everyone to see the same picture and discuss it effectively without communication barriers.[56]

Unlike monitoring individual processes separately, VSM provides a comprehensive view of the entire production flow, revealing inefficiencies across the workflow, making hidden wastes visible. It not only identifies the presence of wastes within the value stream, but also reveals the sources and highlights its root causes, offering actionable insights for improvement. By linking information flow with material flow, it gives insights into how information exchange affects production. Also, VSM equips managers with a complete understanding of the production process, enabling them to make more informed and strategic decisions. Finally, by systematically eliminating the wastes in production through its iterative approach, it helps production processes become leaner.[57]

[53] Tabanli, R., & Ertay, T. (2013).
[54] Rother, M., & Shook, J. (1999).
[55] Jasti, N., & Sharma, A. (2014).
[56] Womack, J., & Jones, D. (1996).
[57] Vaibhav, S. et al. (2013).

2.3 Industry 4.0 Technologies

The term ‘industry 4.0’ was first introduced by Siegfried Dais (Robert Bosch GmbH) and Henning Kagermann (Acatech) in 2011 at the Hannover Fair.[58] The German National Academy later published the industry 4.0 manifesto in 2013. The core element of industry 4.0 is the integration of cyber physical systems with factories, resulting in so-called Smart Factories. CPS merges the physical world with virtual cloud environments, forming a key component of Industry 4.0. Even though the concept originated in Germany, it is already taken up by the researchers and industries around the world and is now a well-researched topic.[59]

Industry 4.0 is referred to as the fourth industrial revolution, transforming manufacturing through digitization and automation. Unlike previous revolutions, which introduced mechanization, mass production, and automation, industry 4.0 aims to create interconnected, data-driven manufacturing systems. This enables ‘smart’ factories, where machines, products, and systems autonomously communicate and coordinate activities throughout the supply chain.[60]

By integrating advanced technologies such as Internet of Things (IoT), Cyber-Physical Systems (CPS), Artificial Intelligence (AI), and Big Data, industry 4.0 supports seamless data flow and decision-making, resulting in environments that are flexible, efficient and dynamic to real-time demands. This revolution redefines traditional production, focusing on agile manufacturing, predictive maintenance, and optimized resource use, that adapts to customer requirements.[61]

There are many technologies that facilitate interconnectivity and automated data transfer between physical and virtual entities. Some of the key industry 4.0 technologies are outlined in this section:

Internet of Things (IoT)

IoT is an integral part of Industry 4.0, which enables connectivity between devices, sensors, and machinery across the manufacturing floor. IoT facilitates real-time data collection and remote monitoring, which allows for better visibility and performance optimization.[62]

[58] Sanders, A. et al. (2016).
[59] Lai, N. et al.(2019).
[60] Frank, A. et al. (2019).
[61] Ibidum
[62] Lampropoulos, G. et al. (2019).

Cyber-Physical Systems (CPS)

CPS connects physical manufacturing components such as machines and tools with digital control systems, enabling real and virtual worlds to interact with each other. This enables production environments to be more agile, where systems can quickly adapt to changes in demand or process adjustments.[63]

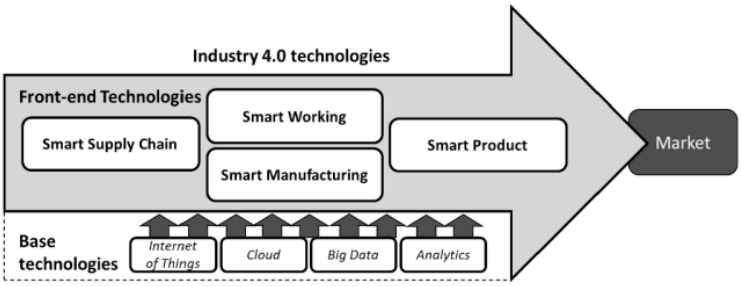


Figure 2-5: Theoretical framework of Industry 4.0 technologies[64]

Digital Twins

A digital twin is a virtual model of a physical entity that mirrors its real-time state. The literature highlights the potential of digital twins in optimizing manufacturing processes, primarily, using simulation and predictive maintenance. By simulating various scenarios, digital twins allow manufacturers to test process improvements, predict outcomes, and make data-driven adjustments without disrupting actual operations.[65]

Artificial Intelligence (AI) and Machine learning (ML)

AI and ML enable industry 4.0 technologies to make intelligent decision by interpreting large datasets and complex analysis based on patterns and predictions. One of the applications of AI in industry 4.0 include predictive maintenance where AI algorithms analyse sensor data to predict machinery failures as well as quality control, where AI detects defects in products.[66]

Machine Vision

Machine Vision, which can be considered a subset of AI, has become an important part of industry 4.0 technologies. Machine vision's ability to "see" and "understand" makes it an invaluable component of Industry 4.0. It can be used to automate processes that would otherwise rely heavily on human interpretation. This can not only substitute humans but

[63] Lampropoulos, G. et al. (2019).
[64] Frank, A. et al. (2019).
[65] Frick, N., & Metternich, J. (2022).
[66] Haffner, O. et al. (2024).

can even offer better performance. This is particularly evident in quality control, where machine vision offers high accuracy while avoiding the chances of accidental (human) errors.[67]

The industry 4.0 technologies are already transforming manufacturing, allowing companies to achieve better flexibility and efficiency than ever. When applied strategically, these technologies can further enhance production capabilities, paving the way for smarter, more adaptive manufacturing systems.

2.3.1 Compatibility of Industry 4.0 with Lean manufacturing

The interaction between industry 4.0 and lean manufacturing has been evaluated in several studies, revealing that the industry 4.0 technologies support lean principles. Lean manufacturing aims to reduce waste and improve processes to maximise value. By integrating industry 4.0 technologies, manufacturers gain enhanced visibility into operations, along with improved responsiveness, and efficiency on the shop floor. Also, it facilitates improved communication across the entire production system, as real-time data and connectivity enable seamless communication between machines, people, and systems.[68]

Industry 4.0 also enables techniques such as predictive maintenance and adaptive production scheduling, which align with lean's objectives of minimizing downtime and reducing non-value added activities. This integration allows lean initiatives to overcome traditional limitations, as digital tools continuously support waste reduction and quality improvement without requiring constant manual intervention. In this way, industry 4.0 has the potential to act as an enabler of lean, facilitating a more flexible, and efficient production environment that remains grounded in lean's foundational principles.[69]

[67] Haffner, O. et al. (2024).
[68] Sanders, A. et al. (2016).
[69] Lai, N. et al. (2019)

3 Related Theory

This chapter covers three main topics, each addressing a specific area relevant to the research. The first section, Digitization in Value Stream Mapping, examines the digital technologies currently applied in VSM. This section provides an overview of how digital tools are used for data collection in VSM and how they enhance traditional VSM method. Additionally, the current limitations and potential for improvement are analysed.

The second section, Current Applications of Object Detection in Assembly Processes explores areas within assembly processes where object detection techniques are already in use, understanding the algorithms chosen and the rationale behind their selection. This study also aimed to examine whether object detection has been specifically applied in value stream mapping.

Finally, the third section, Overview of Object Detection and Comparison of Algorithms, presents a detailed overview about object detection. In addition, this section includes a technical comparison of available object detection algorithms to identify the most suitable option for the practical implementation in this study. The findings from these topics provide essential insights for effectively conducting the practical part of this thesis.

3.1 Methodological Approach for Literature Review

This section outlines the methodological approach used to gather, filter, and analyze relevant literature for each of the three topics.

3.1.1 Digitization in value stream mapping

The purpose of this part of the study was to explore the current advancements in the digitization of value stream mapping (VSM). The review focused on identifying various digitization approaches applied to VSM and analysing their outcomes. By grouping and reviewing different digitization approaches and analysing their advantages and limitations, gaps in the existing research were identified.

To gather relevant literature on the digitization of value stream mapping (VSM), a systematic search was conducted using the query shown in Listing 3-1. The search was performed using two major academic databases: Scopus and IEEE Explore. These two databases were chosen for their comprehensive coverage of literature. Scopus, being one of the largest abstract and citation databases, indexes a wide range of publications from major sources including Elsevier, Springer, and Emerald. IEEE Explore is more focused on engineering, technology, and computer science, and returned highly relevant results for this study.

```
( "value stream map*" OR "value stream design" OR "value stream method" OR "value stream analysis" ) AND ( manufactur* OR production OR factory ) AND ( digit* OR dynamic OR automated OR "industry 4.0" )
```

Listing 3-1: Search query for digitization in value stream mapping

The search covered the title, keywords, and abstract of the papers. The review was limited to studies published in English, within the subject areas of Engineering and Computer Science, and only included publications from 2015 onwards. The procedure followed for selecting the articles is shown in Figure 3-1.

The articles were filtered based on specific exclusion criteria defined in Table 3-1. These criteria ensured that only studies directly related to the digitization of value stream mapping in manufacturing were included, while those that did not align with the research focus were excluded.

Table 3-1: Exclusion criteria for studies on digitization in value stream mapping

Exclusion Criteria	Motivation
Articles not directly related to value stream mapping (VSM) will be excluded	To exclude studies that are not relevant for the research topic
Articles that focus primarily on general lean methods, without specific focus on VSM, will be excluded	To ensure the review is focused specifically on VSM and not broader lean methodologies
Articles discussing only traditional VSM, without including digital methods for VSM, will be excluded	To exclude studies that do not explore the digitization of VSM
Articles unrelated to manufacturing or production environments will be excluded	To exclude articles that are related to the application of VSM in other fields of study
Articles that do not primarily focus on digitization or lack detailed information on the proposed digital tools and their applications will be excluded	To exclude articles that focus on topics other than digitization, such as the integration of circular economy or sustainability indicators with VSM.

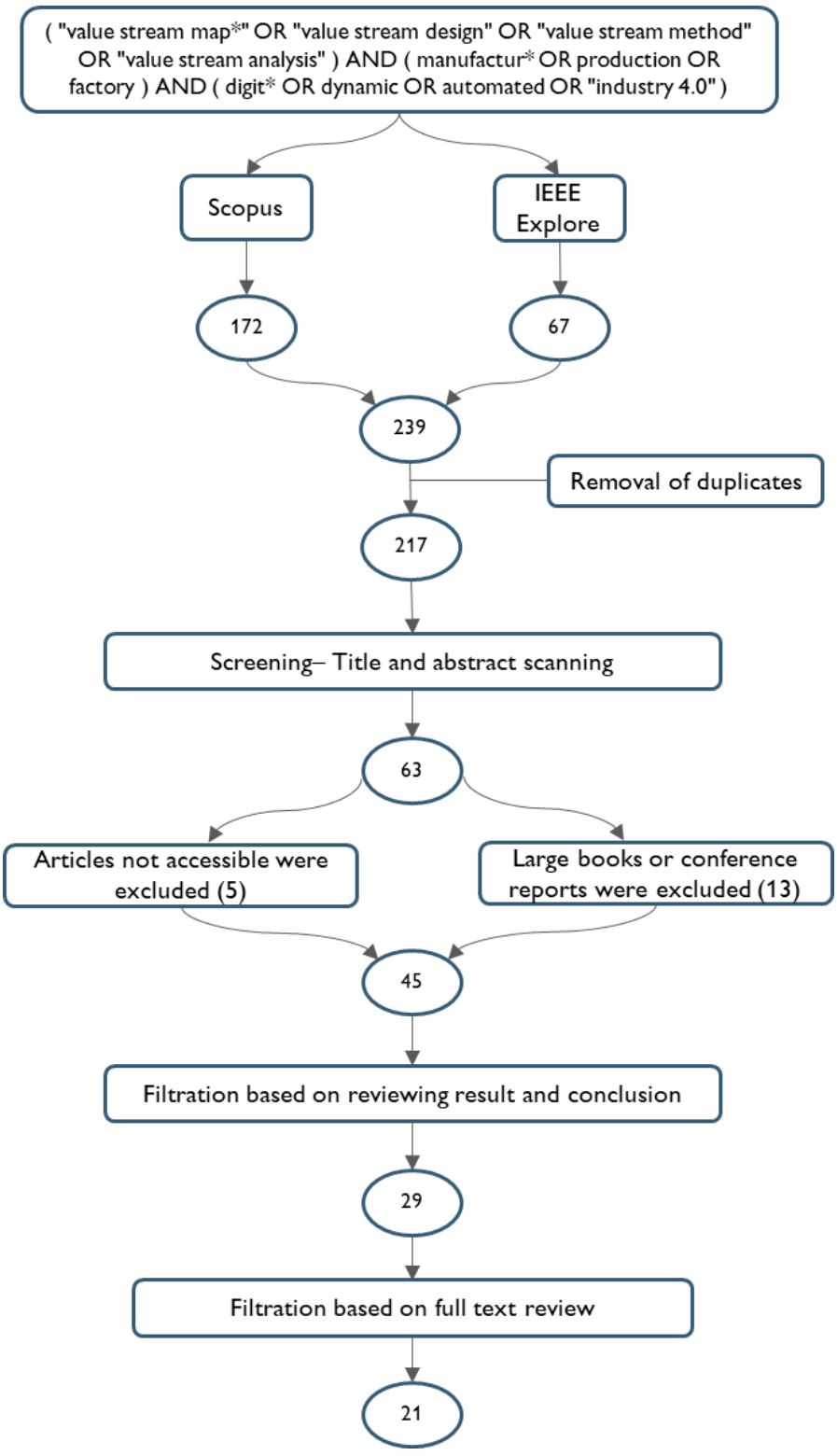


Figure 3-1: PRISMA flow diagram for literature review on digitalization in values stream mapping

3.1.2 Current applications of object detection in assembly processes

The purpose of this part of the study was to explores areas within assembly processes where object detection techniques are already in use, understanding the algorithms chosen and the rationale behind their selection. This study also aimed to examine whether object detection has been specifically applied in value stream mapping.

To gather relevant literature about object detection algorithms, a systematic search was conducted using the query shown in Listing 3-2. The search was performed using two major academic databases: Scopus and IEEE Explore.

```
("manual assembly" OR "manual operation" OR "assembly process") AND ("progress" OR "monitor*" OR "track*" OR "assist*") AND ("object detection" OR "object recognition" OR "deep learning" OR "machine vision" OR "computer vision")
```

Listing 3-2: Search query for literature review on current applications of object detection in assembly processes

The search covered the title, keywords, and abstract of the papers. The review was limited to studies published in English, within the subject areas of Engineering and Computer Science.

The articles were filtered based on specific eligibility criteria defined in Table 3-2. These criteria ensured that only studies that provide comparisons of state-of-the-art object detection algorithms were included to maintain research focus.

Table 3-2: Exclusion criteria for studies on current applications of object detection in manufacturing

Exclusion Criteria	Motivation
Articles unrelated to assembly processes will be excluded.	To remove the articles that are not related to the research topic
Articles not focused on manufacturing or production will be excluded.	To remove articles that focused on other areas such as aerial vehicles tracking or Monitoring of construction works
Articles specific to another field of operation will be excluded.	To remove articles from irrelevant fields. E.g. Retail Product recognition

Articles not focused on object detection will be excluded.	To exclude articles that focused on other deep learning technologies. E.g. Pose estimation

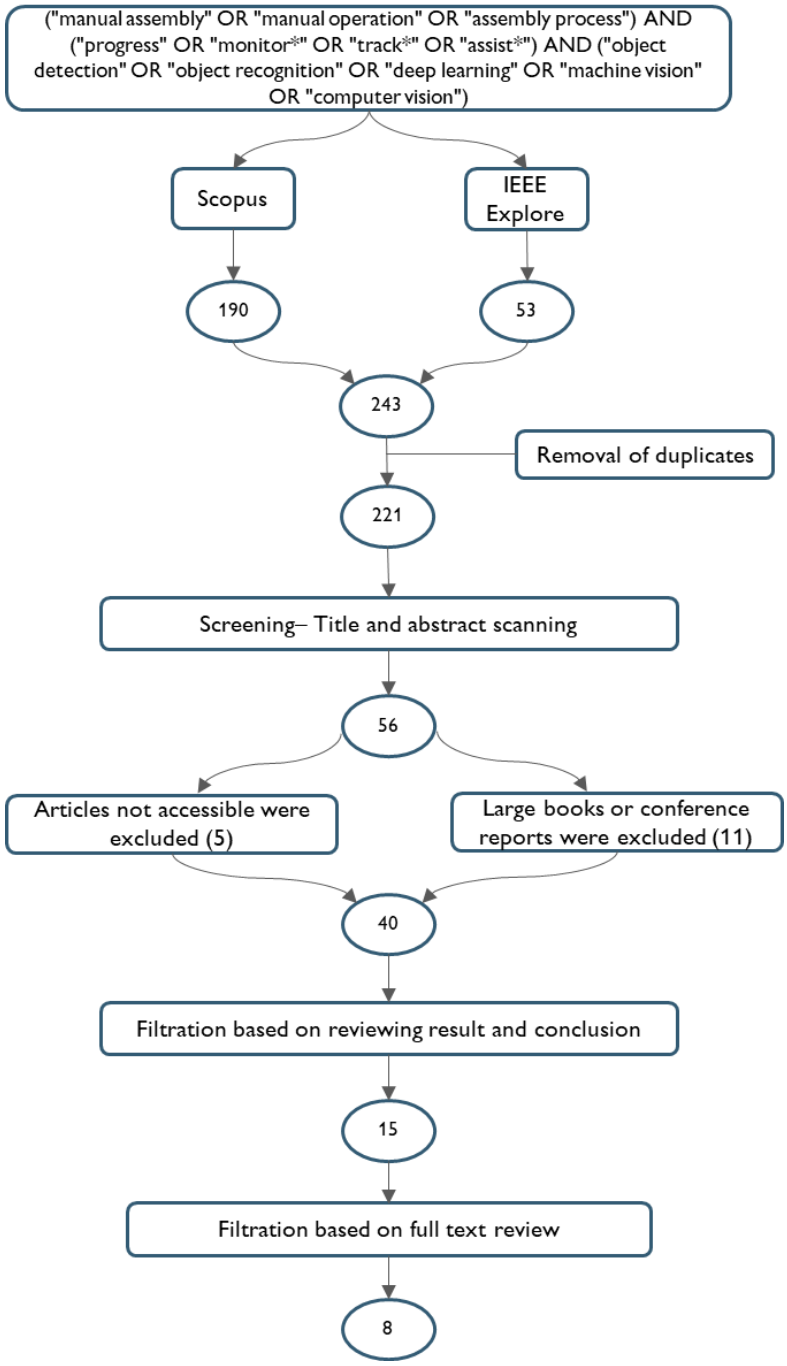


Figure 3-2: PRISMA flow diagram for literature review on current applications of object detection in assembly processes

3.1.3 An overview of Object Detection and Comparison of Algorithms

The purpose of this part of the study was to provide a comprehensive overview of the subject of object detection and to compare various object detection algorithms, with the aim of selecting the most suitable one for the practical part of this thesis. The algorithms were compared based on their performance against benchmark datasets and also their performance in custom studies found in the literature.

To gather relevant literature about object detection algorithms, a systematic search was conducted using the query shown in Listing 3-3. The search was performed using two major academic databases: Scopus and IEEE Explore.

```
("Object detection" OR "Object-detection") AND (study OR survey OR review) AND (accuracy OR speed OR performance) AND ("comparison" OR "comparative study")
```

Listing 3-3: Search query for literature review on object detection algorithms

The search covered the title, keywords, and abstract of the papers. The review was limited to studies published in English, within the subject areas of Engineering and Computer Science, and only included publications from 2020 onwards.

The articles were filtered based on specific eligibility criteria defined in Table 3-3. These criteria ensured that only studies that provide comparisons of state-of-the-art object detection algorithms were included to maintain research focus.

Table 3-3: Exclusion criteria for studies on object detection algorithms

Exclusion Criteria	Motivation
Articles not directly related to object detection will be excluded	To exclude studies that are not relevant to the research focus
Articles that do not discuss or compare two or more object detection algorithms will be excluded	To ensure that the review covers comparative studies of object detection algorithms
Articles that do not compare performance characteristics (e.g., speed, accuracy,) of algorithms will be excluded	To focus on studies that provide meaningful evaluations of algorithm performance

Articles that discuss modifications to existing algorithms to create variations will be excluded	To exclude articles that are totally unrelated to the goal of the review
Articles discussing only outdated algorithms (<= YOLOv3 or Fast R-CNN) will be excluded	To focus on reviewing state-of-the-art algorithms and avoid studies based on outdated methods

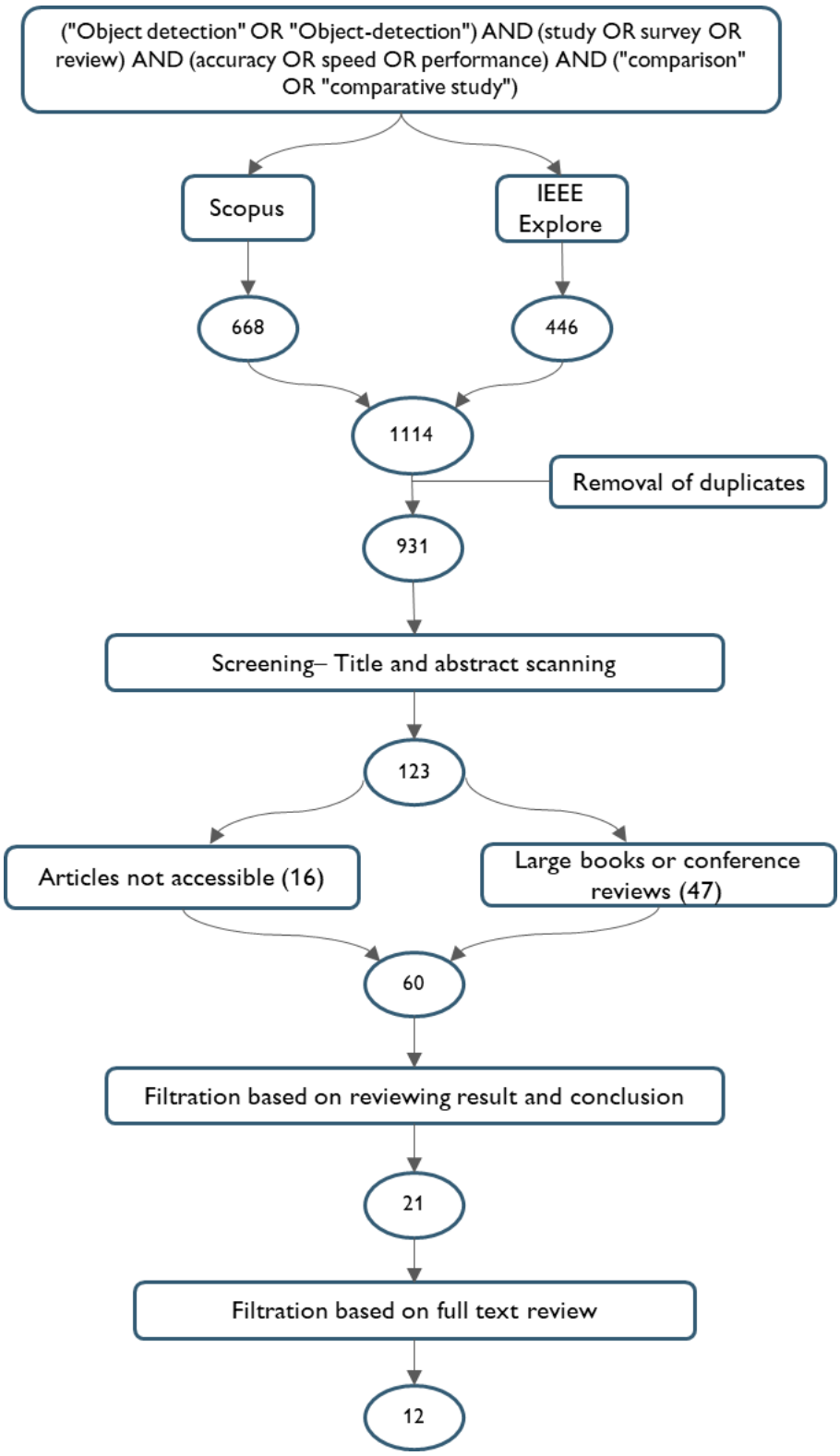


Figure 3-3: PRISMA flow diagram for literature review on object detection algorithms

3.2 Digitization in value stream mapping

The digitization of manufacturing, driven by the rise of Industry 4.0 has transformed the way products are being manufactured. The shift toward automation, data analytics, and interconnected systems is enabling manufacturers to optimize processes, thereby become more flexible, efficient, and meet the growing demand for customized products. Industry 4.0 technologies, such as IoT, cloud computing, and artificial intelligence have helped manufacturers in optimizing processes and enhancing productivity.[70]

This trend toward digital transformation has impacted many aspects of manufacturing including the widely used tool of value stream mapping (VSM). Value stream mapping is a lean management tool that facilitates the analysis of value streams and the identification of opportunities for optimization. Value stream maps visualize essential process steps and key figures, which improves the understanding of the actual as-is process and provides insights about potential areas for improvement.[71] Traditionally, VSM has been a manual process, relying on direct observation and static data collection to map material and information flows across production systems. While this approach has been effective in stable and predictable manufacturing environment, it is limited in its ability to capture the increasing complexity and dynamism of modern production environments.[72]

Digitization has the potential to keep value stream mapping relevant in today's manufacturing by incorporating digital data acquisition, automation, and advanced analytics into the process. Employing technologies such as the Internet of Things (IoT), different sensors, and simulation tools could enable manufacturers to continuously monitor production activities and gain deeper insights into their operations. This shift from static to dynamic VSM aligns with the broader goals of Industry 4.0, enabling companies to achieve higher levels of flexibility, responsiveness, and operational efficiency.[73]

3.2.1 Need for digitization

Value stream mapping was developed as a paper and pen tool to see and understand the material and information flow as the product moves through the values stream.[74] This concept has been used for decades and have been very effective in helping the companies in improving their efficiency. Traditionally, production systems are designed for a predefined set of products and production volumes. Adhering to this workflow will

[70] Mariappan, R. et al. (2023).
[71] Frick, N., & Metternich, J. (2022).
[72] Horsthofer-Rauch, J. et al. (2022).
[73] Ferreira, W. et al. (2022).
[74] Rother, M., & Shook, J. (1999).

ensure optimum efficiency, as any deviation from this can lead to losses. Therefore, periodic checks are to be carried out to determine the status of the production and to look for areas for improvement. Value stream mapping has served as an efficient tool for this purpose in relatively stable production scenarios. However, the requirements are changing, and the production systems are evolving accordingly. The traditional data collection methods reach its limits when it is required to adjust the value stream quickly and objectively to new challenges.[75]

According to the literature, lean philosophy in production is currently being successfully implemented in only two-thirds of the organizations and only a few are able to sustain it in the long run. This shows that the multi-varieties and small batch-oriented manufacturing has deteriorated the effectiveness of lean implementation.[76] Manufacturers have to ensure high degree of individualization, flexible assembly assets, and short-term change requests to remain competitive in today's market.[77] The static nature of the conventional value stream mapping lacks the capability to efficiently manage this challenge. The singular data recording on site, used in traditional value stream mapping process, may not necessarily reflect the actual situation in the shop floor. This approach often struggles in capturing the dynamic nature of modern manufacturing environments and provides only a limited view of the current state. As a result, these maps often fail to provide accurate insights for decision-making in dynamic manufacturing settings. [78]

Another major challenge is the considerable effort required in manual data collection, which is a tedious and time-consuming process. This issue is further compounded in highly flexible and reconfigurable environments, where products, processes, and workflows change frequently. In such situations, regularly updating the VSMs become essential to ensure they accurately reflect the current state of the production. However, due to the labour-intensive nature of manual data collection, it is often impractical to repeat these processes as frequently as necessary. Digitization offers a way to meet this need by enabling the automated collection of data from machines, sensors, and IoT devices, eliminating the delays and inaccuracies associated with manual methods.[79]

The reliance on the worker's expertise is another limitation of the traditional approach. Although this method does not require expensive equipment and relies on basic tools like stopwatches and visual analysis, it requires workers to be highly experienced, particularly

[75] Frick, N. et al.(2024).
[76] Iyer, S. et al. (2023).
[77] Scheder, N. et al. (2023).
[78] Frick, N., & Metternich, J. (2022).
[79] Horsthofer-Rauch, J. et al. (2022).

in dynamic and complex environments.[80] There is also a risk of human errors and subjectivity during manual data collection. Any resulting inaccuracies can lead to misinterpretation of the problem and suboptimal decision-making, undermining the effectiveness of lean initiatives. Digitization shifts the focus from manual observation to data-driven systems. Digital tools capture objective, consistent data that can be analysed in real-time, minimizing the risk of human error and ensuring that decision-making is based on comprehensive, up-to-date information.[81]

Digitization can also enable manufacturers to fully leverage the data already available within many modern production systems. In today's manufacturing environments, much of the data beneficial for value stream mapping is already captured in various digitized systems such as Enterprise Resource Planning (ERP), Manufacturing Execution Systems (MES) and Supply Chain management (SCM) systems. By connecting VSM to these existing information sources, key data such as supplier data, customer orders, and production schedule can be updated in real-time.[82]

The scope of digitization extends beyond automated data acquisition. It also enables manufacturers to employ advanced tools that can transform the collected data into actionable insights. Integration of technologies such as process mining, digital twins, and simulations, allows manufacturers to analyse data in real-time, identify inefficiencies, and aid in decision making. Digitization can help in better visualization and analysis of the mapped value streams.[83] Making variations to the production processes can be expensive and may also affect other processes in different ways. Therefore, it would be smart to leverage the existing digital technologies to analyse the impact of any variations to the value stream using simulation technologies. This can also help in analysing and comparing various possible scenarios thereby aiding the managers in decision making.[84] A study cited by (Frick and Metternich) states that 66% of lean experts believe that the advancement of value stream mapping by the incorporation of industry 4.0 technologies is beneficial.[85]

3.2.2 Areas of Digitization

There are many approaches in the literature aiming to improve the value stream mapping method by the incorporation of digitization. Different approaches focus on different areas within value stream mapping. While some focus on material flow, others focus on the

[80] Wang, H.-N. et al. (2021).
[81] Klimecka-Tatar, D., & Ingaldi, M. (2022).
[82] Wang, H.-N. et al. (2021).
[83] Trebuna, P. et al. (2019).
[84] Liu, Q., & Yang, H. (2020).
[85] Frick, N., & Metternich, J. (2022).

information flow. Some focus on the development of current state maps digitally, while other focus on the design, comparison and optimization of future state maps. Different areas of digitization present in the literature and the existing approaches are outlined in this section.

3.2.2.1 Digital Data Acquisition

One of the primary motivations for digitizing value stream mapping (VSM) is to eliminate the inefficiencies associated with manual data collection methods. Not only is manual data collection time-consuming and prone to human errors, but also provides only a snapshot of the production process, failing to capture the dynamic nature of activities on the shop floor.[86] To overcome the challenges with traditional data acquisition, many approaches have been presented in the literature to digitally collect data from the shop floor.

RFID tags

Radio Frequency Identification (RFID) systems have been proposed by many researchers for enhancing VSM by enabling real-time data collection and dynamic monitoring of production processes in the shop floor. RFID technology uses radio waves to automatically identify and track items. The system consists of tags, readers, and a software and all the three components work together to facilitate the tracking of an item. RFID tags when attached to objects, transmit data to RFID readers, which is then processed by software to filter any errors and provide clean data for analysis.[87]

A Dynamic Value Stream Mapping (DVSM) system was proposed by M. Ramadan, Z. Wang and B. Noche, integrating RFID technology into traditional value stream mapping. Unlike static VSM, which provides only a snapshot of the production process, the RFID-enabled DVSM offers continuous real-time data collection. RFID tags are attached to objects such as products, equipment, or tools, allowing them to be tracked throughout the production flow. RFID gives this information from the production floor to the already built VSM, facilitating comparison with the planned flow. This allows managers to respond to situations such as inventory depletions, production delays, or bottlenecks on time.[88]

Wang et al. also suggested integrating RFID technology with value stream mapping. They suggested a multi sensor approach where RFID tags, or barcode systems attached to workers map the workflow information, whereas PLCs or other sensors attached to

[86] Wang, H.-N. et al. (2021).
[87] M. Ramadan et al. (2012).
[88] Ibidem

machines collect the machine data , and other sensors to collect data related to machines, personnel, and materials.[89]

RFID has the advantages of its relatively low cost, wide coverage, and ease of implementation, which makes it a good choice for monitoring production processes. However, RFID does have its limitations. One significant drawback is its limited precision, as it does not offer high level of accuracy required for tasks that demand exact positioning of objects. Furthermore, there is a need to physically attach tags to the objects being tracked, which introduces additional step in the workflow and increases the workload for workers. This can affect the value adding time, especially in environments where large volumes of items need to be tagged. The manual handling of tags also opens up the possibility for error, such as misplacement or forgetting to attach the tags, which could disrupt the data collection process.[90]

Real-Time Location Systems (RTLS)

RTLS have emerged as a promising solution for enhancing data acquisition for VSM. These systems enable the continuous tracking of assets such as products, materials, and operators, within a facility, providing real-time location data along with corresponding timestamps. By utilizing RTLS, manufacturers can track movement and material flow more accurately, offering a more dynamic and detailed perspective of the value stream. RTLS operates by using active tags , which continuously communicate their positions to a network of anchors installed within the facility.[91]

Tran, Ruppert and Abonyi proposed the use of Indoor Positioning Systems (IPS) as a valuable addition to the data sources for VSM. They demonstrated that positional data acquired from IPS can be transformed to determine KPIs used in Lean Manufacturing and developed a framework for extracting and analysing manufacturing data through process mining. Ultrawide band (UWB) – based RTLS systems which uses active tags and anchors for localization were used for the case study where the method was validated. IPS tags were attached to carts carrying semi-finished products and to operators, with positional data calculated using the Time Difference of Arrival (TDoA) method. The integration of this system with existing IT systems like MES allows for comprehensive data collection relevant to VSM, providing better visualization of the production flow. The study also highlighted how integrating other sensors, such as RFID, optical or vibration sensors could further enrich data collection.[92]

[89] Wang, H.-N. et al. (2021).
[90] M. Ramadan et al. (2012).
[91] Sullivan, B. et al. (2022).
[92] Tran, T.-A. et al. (2021).

On the other hand, Sullivan et al. focused on a stand-alone implementation of UWB - based RTLS by developing an excel-based semi-automated VSM creation tool. This method allowed for the generation of current state maps using positional data and timestamps collected from RTLS sensors. The semi-automated tool was validated in a case study. While it still required some human intervention, the approach proved to be less time-consuming and less prone to errors compared to traditional methods. Although not all the inputs required for a complete value stream map could be derived using the sensors, the accuracy of the collected data closely reflected actual production conditions, demonstrating the potential of RTLS to reduce manual efforts and to generate a more dynamic VSM.[93]

RTLS systems offer high precision and accuracy, particularly the UWB RTLS systems, which provides centimetre level accuracy. This makes them exceptionally suitable for manufacturing environments where exact positioning is critical. UWB signals also penetrate most obstacles and cover large distances without significant losses.[94]

However, these systems are relatively expensive and the cost increases with the number of items being tracked, limiting their use in high-volume applications. They also consume considerable power, increasing operational costs, particularly in large-scale implementations where many tags need to be tracked. Additionally, the presence of metallic objects in industrial environments may block or distort the signals from these sensors, leading to reduced accuracy.[95]

Also, similar to the case with RFID, the need to physically attach RTLS tags to the objects being tracked could increase the workload for workers. This can affect the production efficiency, especially in environments where large volumes of items need to be tagged. The manual handling of tags also opens up the possibility for error, such as misplacement or forgetting to attach the tags, which could disrupt the data collection process.[96]

Other IoT devices and sensors

Many studies proposed the integration of other industry 4.0 technologies with value stream mapping to enable it tackle today's challenges in manufacturing. IoT devices, PLCs, sensors, and Human-Machine Interfaces(HMIs) are among the tools that enable continuous data collection. Mariappan et al. proposed an intelligent VSM (IVSM) model, which incorporates IoT devices, sensors, HMIs and machine interlocks into an Integrated Efficiency Monitoring System (IEMS). This system allows for seamless data collection

[93] Sullivan, B. et al. (2022).
[94] Ibidem
[95] Ibidem
[96] Tran, T.-A. et al. (2021).

and analysis, using cloud computing to store and process data. The real-time monitoring capabilities of the IVSM framework enable faster and more informed decision-making. The framework also incorporates alerts and insights to managers, allowing them to react to bottlenecks and inefficiencies on time. The framework was validated in an automotive electromechanical component manufacturing company in India.[97]

Huang et al. proposed a multi-agent system for dynamic value stream. This system composed of several embedded Arduino units as agents and a Raspberry Pi as the core agent, which when combined with switch sensors can map the material flow. Node-RED, an open-source flow-based software, was introduced as a visualization layer.[98] In a later study by the same author, more focus was given on upgrading legacy machines with cyber-physical systems under the framework of industry 4.0. Companies, particularly with limited resources are often reluctant to make huge investments in latest machineries. These companies continue to use a high number of non-networked legacy machines due to their reliable performance and relatively high replacement costs. The study suggested equipping these machines with low-cost sensors to collect and share data, allowing them to be integrated into modern production networks and to contributing to the development of dynamic value stream maps. The framework was validated in a case study conducted at an Australian manufacturer of customized alloy ute canopies, where switch sensors were used to gather machine data and thereby determine the processing time.[99]

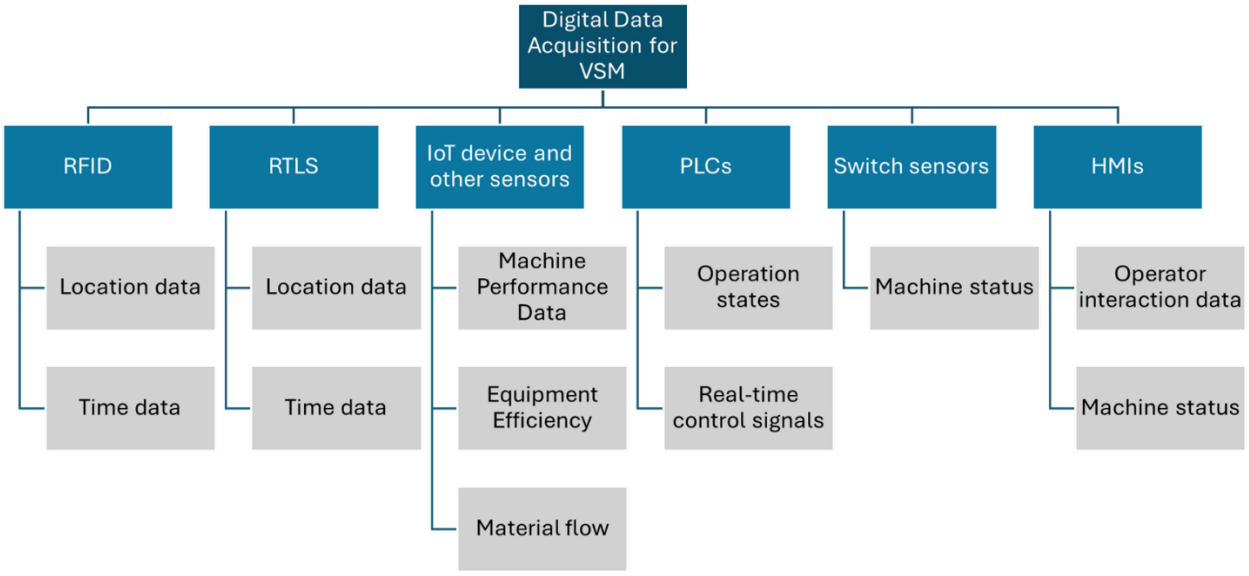


Figure 3-4: Overview of Digital Data Acquisition Tools for Value Stream Mapping

[97] Mariappan, R. et al. (2023).
[98] Huang, Z. et al. (2019).
[99] Ibidem

Leitão et al. proposed a real-time KPI monitoring tool specifically designed to enhance strategic decision-making in manufacturing environment. The tool's Data Service Module gathers and standardizes data from multiple sources, including PLCs, IoT devices, and sensors connected to legacy machines. Since data is collected from a mix of old and new technologies, technological adapters are used to convert the native data structure into a unified data model. The system then analyses this data to detect trends, deviations, and potential bottlenecks. The tool was validated in a real-world factory setting producing microwave ovens. However, the exact sensors or IoT device used for data collection were not mentioned.[100] Klimecka-Tatar and Ingaldi evaluated the potential benefits of digitization in small and medium sized enterprises (SMEs) within the manufacturing sector. As SMEs face barriers such as limited resources and resistance to change, making huge investments for digitization is often impractical. The study highlights that even partial digitization can offer significant improvements. They enriched the VSM with data regarding Overall Equipment Effectiveness (OEE) that was collected using a software configured to the machine. The operation data of the machine was downloaded automatically, while information about defects was manually entered into the software by the operator. The study claimed that this enhanced process visualization by providing accurate data on machine performance and reduced the errors caused by manual data collection.[101]

The central focus of these studies is the integration of Industry 4.0 technologies to automate data collection directly from machines and equipment. Modern interconnected machinery, equipped with technologies such as PLCs and IoT devices, generate a wealth of data that is crucial for determining key performance indicators (KPIs). These studies provide a framework for leveraging that data for digitized value stream mapping. At the same time, the integration of low-cost sensors allows even legacy machines, which may not be equipped with digital capabilities, to be incorporated into the value stream mapping process.

Summary of Digital Data Acquisition Methods

The application of several sensors is suggested in the literature to facilitate data acquisition from shop floors. These sensors collect a variety of data, from the location of products, tools, and workers to monitoring operational parameters like equipment performance, environmental conditions, and machine states. These systems are capable of collecting data in real-time, enhancing productivity and enabling faster responses to issues in production environments. Technologies like RFID, RTLS, and barcodes are

[100] Leitão, P. et al. (2019).
[101] Klimecka-Tatar, D., & Ingaldi, M. (2022).

commonly used to track the location and movement of entities on the shop floor, offering automatic tracking of material flow.

On the other hand, PLCs, IoT sensors, HMIs, and other machine-focused sensors primarily gather machine data, such as uptime, cycle times, and operational states, which help derive key KPIs like processing times and equipment efficiency. These sensors enable the digital data collection from both modern machinery and legacy machines, ensuring comprehensive data collection across the entire shop floor.

3.2.2.2 Advanced digital tools for Intelligent Analysis and Decision Support

Digitization has transformed value stream mapping by not only facilitating efficient and dynamic data acquisition, but also by providing tools for deeper analysis of the current state, offering decision support to managers for optimizing future state maps. Many studies in literature propose the use of various digital tools to enhance value stream mapping. These tools range from digital twins and simulations to process mining and machine learning, all aimed at improving process visibility, supporting decision-making, and designing efficient future state maps.

Digital twins and Digital Shadows

Digital twins and digital shadows are digital representations of physical systems. Digital twins include a bidirectional flow of data between the physical entity and its digital counterpart, meaning that not only does the digital twin update in real-time based on changes in the physical system, but it can also send data back to influence or optimize the system. On the other hand, digital shadows have a unidirectional data flow, reflecting the current state of a system, without influencing the physical entity directly. Different researchers proposed integrating digital twins or digital shadows with VSM to realize real-time visualization and predictive capability.[102]

Frick and Metternich proposed a framework for the development of a digital value stream twin (DVST), which acts as a comprehensive and dynamic digital representation of a value stream. The aim of the DVST is to continuously capture, store, and transfer real-time or near-real-time data from the shop floor to the virtual layer. This enables comparison of the real-time shop floor data with target states in VSM, and to generate proposals for improvement using optimization algorithms. However, the proposed framework was not validated.[103]

[102] Frick, N., & Metternich, J. (2022).
[103] Ibidem

Building on this framework, the same author later introduced the Design Model for the Digital Shadow of a Value Stream, narrowing the focus to the digital representation of the current state and historical analysis. In this project, the data flow from the physical to the digital object was only considered. The study excluded the automated optimization concept and focused more on visualization, real-time monitoring, and historical analysis. The model was validated at the learning factories at TU Darmstadt and the users were able to design a digital shadow of the value stream using the model.[104]

The use of digital twins was also proposed by Iyer, Sangwan and Dhiraj to enhance value stream mapping. Although referred to as a digital twin, the framework lacked an automated data flow from the digital to the physical object. However, the study puts forward a comprehensive architecture that covers everything from the data acquisition to the visualization of the value streams to the analysis and optimization of the future state maps. This architecture was designed to enhance decision-making at both operational and strategic levels by providing real-time insights into production processes. The study suggested a solution that incorporates low-cost sensors for the collection of data necessary for the development of the digital twin. The information from the digital twin is visualized on a multi window virtual dashboard that displays the status of each station, overall equipment effectiveness (OEE), and different KPIs. The framework was validated in an industry 4.0 complaint automated assembly line with five processing stations. The digital twin, developed using the AnyLogic software, enabled the simulation of multiple 'what-if' scenarios to analyse and compare future state maps before any changes were implemented.[105]

For the effective implementation of these methods, researchers have identified three critical layers – the physical layer for data collection, the virtual layer for data storage and analysis, and the connection layer to manage the data flow between the other two layers.[106] While existing IT systems such as Enterprise Resource Planning systems (ERP) or Manufacturing Execution Systems (MES) could provide critical data to the physical layer, the use of other data acquisition methods are often necessary, particularly in human-centred production processes. This can be achieved by the integration of different sensors. Therefore, a multi-model data acquisition system is necessary to ensure comprehensive data collection from various sources to obtain enough information about the production process.[107]

[104] Frick, N. et al. (2024).
[105] Iyer, S. et al. (2023).
[106] Frick, N., & Metternich, J. (2022).
[107] Frick, N. et al. (2024).

Application of simulation

For companies not yet ready to adopt full digital twin implementations, simulations provide a more accessible method to model, analyse, and optimise value stream maps. Several studies have explored it's use in VSM, without all the complexities of a digital twin. Simulation can be used to model and simulate the current-state and future-state maps of the value stream. Data required for these simulations must be collected either using the traditional methods or using modern digital data acquisition techniques. The primary role of simulation in VSM is to model the processes, enabling the analysis of different scenarios and optimization strategies. This enhances process visualization, improving clarity and understanding. This is particularly useful in complex production scenarios as it can handle the dynamic nature of the modern manufacturing environment better. Simulation can therefore help managers understand the potential impact of proposed improvements before they are implemented on the shop floor. [108]

In order to emphasize the potential of using software tools for enhancing classical methods such as value stream mapping, Trebuna, Pekarcikova and Edl used Tecnomatix plant simulation software to digitize VSM process in a case study that involved the production of steel cords. The data collected using traditional methods is entered into the software, which creates a digital model of the production process. This can visualize the dynamic fluctuations in production due to variation in batch size, procedure, product type, or other faults. The model can analyse current maps and can propose and compare future-state maps.[109]

Other researchers tried to combine various approaches to enhance the effectiveness of the solution. Ferreira et al. introduced a hybrid simulation - value stream mapping (HS-VSM) framework, integrating discrete-event simulation and agent-based modelling with value stream mapping. DES has been widely adopted to enhance VSM by modelling the material and information flow through a system. Combining it with agent-based modelling improves its potential to model and analyse complex, decentralized and dynamic production systems typical of industry 4.0. The approach, therefore, compliments the integration of industry 4.0 technologies in production and facilitates the transition to Industry 4.0 by modelling complex interactions within a distributed production system. This method was validated using AnyLogic software at an SME in the furniture manufacturing sector, demonstrating the potential to optimize future-state maps by leveraging simulation tools.[110]

[108] Liu, Q., & Yang, H. (2020).
[109] Trebuna, P. et al. (2019).
[110] Ferreira, W. et al. (2022).

Meanwhile, Liu and Yang suggested the idea of integrating Grey Taguchi method with simulation to improve multiple attribute decision making. In this method, in addition to visualizing the future state maps using the simulation model, it also prioritizes them by ranking multiple future state maps based on the predefined parameters. Therefore, the integration of the Grey Taguchi method enables multiple-attribute decision-making that involves the evaluation of multiple decision metrics simultaneously. This results in a structured approach to decision-making that allows to evaluate and prioritize multiple lean initiatives based on a range of performance criteria. In this study, FlexSim 2019 simulation software was used to model scenarios at a footwear manufacturing company, simulating different flows of material and information to evaluate performance under various conditions. This combination of simulation and decision-making methods offered more structured insights into optimizing value streams.[111]

Process mining

Process mining is another tool recommended in the literature for enhancing value stream mapping (VSM). By analysing collected data, process models can be automatically generated to discover, monitor, and improve real-world processes. This method offers an effective way to efficiently identify inefficiencies, bottlenecks, and deviations from the expected workflow. Visualizing these deviations allows managers to quickly identify inefficiencies and take corrective actions.[112]

Several researchers have proposed frameworks that integrate process mining with VSM to improve process visibility and efficiency. Horsthofer-Rauch et al. introduced a framework that integrates process mining with VSM to automatically generate real-time process models and KPIs by utilizing event logs recorded in systems such as Manufacturing Execution Systems (MES) and Enterprise Resource Planning (ERP). This framework enables the continuous monitoring of production flows and helps align actual production processes with strategic goals by automating the generation of KPIs. Visualizing the KPIs such as cycle time, throughput time, and resource utilization in real-time, allows for more effective and timely decision-making.[113]

While process mining offers significant benefits, its effectiveness relies heavily on the availability and quality of event data. Incomplete or inconsistent logs can result in inaccuracies in the generated process models. Tran, Ruppert and Abonyi proposed a framework to integrate process mining with IPS data to optimize flexible manufacturing processes. The study suggested that incorporating location data captured by IPS

[111] Liu, Q., & Yang, H. (2020).
[112] Horsthofer-Rauch, J. et al. (2022).
[113] Ibidem

provides an additional layer of information for process mining, enhancing its effectiveness beyond relying solely on data from information systems. This integration is particularly useful in environments with high variability, such as those involving manual assembly tasks.[114]

Decision tree algorithms

In addition to the tools discussed above, use of algorithms has also been proposed as an effective method for automated root cause analysis in value stream mapping (VSM). Wang et al. suggested the integration of decision tree algorithms to extract actionable knowledge from the data collected in a production workshop. Decision trees are a type of machine learning algorithm that uses a hierarchical, tree-like structure to guide decision-making processes based on data. The algorithm analyses real-time production data to identify the causes of inefficiencies such as machine downtime, delays in material delivery, or operator-related issues. It can analyse heterogeneous production data from various sources such as machine sensors, RFID systems and can identify patterns helping to understand the underlying issues.[115]

The proposed framework was validated in a furniture manufacturing workshop, where the decision tree algorithm was used to automatically diagnose the root cause of a low machine utilization rate.[116]

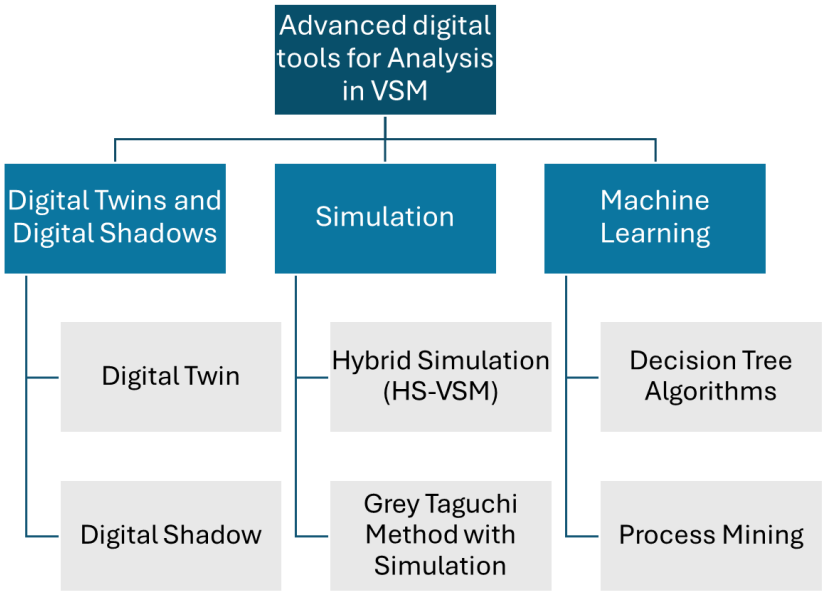


Figure 3-5: Advanced Digital Tools for Intelligent Analysis in VSM

[114] Tran, T.-A. et al. (2021).
[115] Wang, H.-N. et al. (2021).
[116] Ibidem

Summary of Tools for Intelligent Analysis

The tools described in this section offer manufacturers powerful methods to analyse, monitor, and optimise production processes in real time. These tools enhance the visibility and understanding of value stream mapping (VSM) by transforming collected data into actionable insights. By leveraging real-time data, these tools help manufacturers respond quickly to issues in production, optimise resource utilization, and design future state maps more efficiently.

The effectiveness of these models depends heavily on the data available for analysis. Traditional manual data collection methods are insufficient for capturing the complexity and rapid changes of modern production processes. Automated data collection from sensors, IoT devices, and existing information systems is essential for feeding these advanced tools with accurate, real-time data.

3.2.2.3 *Digitization in VSM Visualization*

With the integration of industry 4.0 technologies into value stream mapping, the traditional paper-based VSM visualization is no longer sufficient. It is necessary to have digital VSMs capable of visualising the situation in the shop floor in real-time. Digital VSMs must also incorporate additional elements that represent the additional data and metrics that becomes available. Digitizing the VSM also opens the opportunity to enhance the visualization by integrating various visual aids, such as graphs, charts, and dashboards, which provide deeper insights to the decision makers.[117]

Several studies have proposed different ideas to improve value stream visualization. Scheder et al. proposed an approach to structure the information gathered digitally into three perspectives in a way to provide maximum value to the user.[118] Lewin, Voigtländer and Fay proposed an extended approach to VSM including additional elements and symbols to better represent the large amount of data that becomes available with the digitization of VSM. With the introduction of Swimlanes and new symbols, it visualizes the data flow and its source, which improves clarity and transparency.[119] The capability for real-time monitoring and the subsequent reactivity within the shop floor is the primary benefit that comes with the implementation of industry 4.0 solutions. Therefore, Arey, Le and Gao also included the time taken for detecting deviations from the defined Key Performance Characteristics as a metric in the extended VSM. In addition, the extend of

[117] Wang, H.-N. et al. (2021).
[118] Scheder, N. et al. (2023).
[119] Lewin, M. et al. (2017).

digitization in the transfer of information is visualized by the introduction of an information flow scoring system.[120]

Finally, different tools for representing VSMs digitally have been suggested in the literature. A Visio-based software, VASCO was developed by Fraunhofer Austria to visualize value stream maps digitally. Although the data required for the visualization is to be collected using the conventional VSM approach, the tool enables easy creation of VSMs.[121] On the other hand, Fernandes et al. developed a dynamic web application to visualize VSMs. The application featured a multi-window approach to toggle between current state maps, future state maps, and a page for graphical comparison between the two. It integrated the features and symbols of traditional VSM and offered a user-friendly interface to interact with the results. A bar graph was also included to visualize KPIs for each operation, complementing the conventional VSM.[122]

3.3 Current application of object detection in assembly processes

As outlined in the previous section, manufacturing sector has undergone significant digitization with new use cases being integrated frequently due to the potential improvement in efficiency and flexibility. This trend extends across industries, where digital technologies are being adopted rapidly to enhance productivity and efficiency. One such technology is object detection, which has already found widespread application in various industries such as logistics, automotive, healthcare, as well as in manufacturing. This is due to the general push towards more automation as well as due to the rapid advancements in the computer vision field. New innovations, such as self-driving cars, autonomous robots, and automated video surveillance demand greater accuracy and performance, which has significantly accelerated the research in this field.[123]

The manufacturing industry has also begun to adopt object detection as a tool to improve productivity and reduce human errors. In this section, the current applications of object detection within manufacturing, particularly in manual assembly processes is examined. This is to understand the extent of incorporation of object detection in assembly processes and to examine if the concept of using it for digital data acquisition from the shop floor in the context of value stream mapping has already been studied.

A survey by Ahmad and Rahimi offers a comprehensive overview of object detection algorithms and techniques, along with their applications in smart manufacturing. The

[120] Arey, D. et al. (2021).
[121] Scheder, N. et al. (2023).
[122] Fernandes, E. et al. (2023).
[123] Chen, W. et al. (2024).

three main applications highlighted in the study are defect detection, personal protective equipment detection and surveillance. The study emphasize that object detection is being used in defect detection in a variety of manufacturing sectors including steel, aluminium, and fabric manufacturing. Object detection enables the detection of cracks and other defects on the products more quickly and cost-effectively than the conventional quality inspection methods. Another critical application is in ensuring the safety of the workers. Ensuring that workers use the required safety equipment and adhere to safety standards, such as avoiding restricted areas and unauthorized tasks, is critical in manufacturing. Slight errors or carelessness from workers can result in danger to the health and safety of workers and in financial losses.[124]

Object detection can be applied for automated surveillance of workers to quickly identify the absence of personal protective equipment or deviations from safety standards, offering a faster and more convenient solution than manual inspection. The survey also highlights many challenges faced and future directions. A particular challenge in using object detection in manufacturing is the lack of large datasets in the industrial environment and it is important to increase the availability of industrial datasets. Also, the object detection algorithms should be improved to achieve more accurate results particularly for real-time applications.[125]

3.3.1 Object detection in assisting assembly processes

One use case where object detection is being applied in manufacturing is in assisting assembly processes. Researchers highlight mass customization as a major challenge in modern manufacturing, resulting in smaller batch sizes and more process variations. This complexity severely reduces worker efficiency, resulting in increased errors and lead times.[126]

To minimize this impact, Raj et al. proposed combining object detection with augmented reality to offer a guidance system for workers to assist them in the assembly task. They claimed that object detection could be used to identify and locate components in real-time, eliminating the need for manual placement of markers on each part, which was a limitation of earlier AR systems. The system was applied to the assembly of pneumatic cylinders, where the AR system, enhanced by object detection, was used to overlay holographic instructions directly onto the workspace to guide workers through the assembly. YOLO v5 algorithm was chosen for the object detection task due to its

[124] Ahmad, H., & Rahimi, A. (2022).
[125] Ibidem
[126] Raj, S. et al. (2024).

relatively higher speed and accuracy.[127] On the other hand, Zamora-Hernández et al. proposed a system combining object detection with action recognition to monitor and guide operators during assembly. The system captures the motion and interaction between the operator and the tools and generates action commands to verify if the assembly process is being followed correctly. YOLO algorithm was used for the object detection task and an action recognition module based on Deep Activity Description Vector (D-ADV) was employed for the action recognition task.[128]

3.3.2 Object detection for monitoring assembly processes

Some studies have evaluated the use of object detection for monitoring manual operations. Lou et al. suggested using object detection along with a counting algorithm to monitor and count repetitive tasks in manual assembly processes. In this two-stage approach, the object detection algorithm identifies and classifies manual operations, while a sliding window counter algorithm counts the repetitive tasks based on boundary points. The method was validated by using YOLOv4 for detection and a counting algorithm for counting repetitive tasks. The authors highlighted that object detection enables contactless, real-time monitoring, offering a significant advantage in smart manufacturing environments.[129]

Kitsukawa et al. also proposed the use of object detection for monitoring assembly process, but with the combination of deep metric learning for progress estimation. The primary goal was to achieve real-time progress estimation of assembly processes without the need for attaching sensors to products or adding additional steps to the assembly process. They also followed a two-step approach, where object detection is used in the first step to locate the product and crop in to focus on the relevant parts of the assembly from images captured by fixed cameras. Faster-RCNN was used for the object detection task. In the next step, a deep metric learning model is employed to estimate the current progress of the assembly by comparing the captured images to the predefined steps of the assembly. The method was validated in a desktop PC assembly experiment, achieving an accuracy of 91.8%. However, further studies are needed to assess its effectiveness in more complex assembly processes.[130]

In addition to estimating the progress of assembly, automated monitoring also serves to ensure the safety of workers. This is particularly useful in human-robot collaborative (HRC) environments. Kozamernik et al. proposed a system that incorporates machine

[127] Raj, S. et al. (2024).
[128] Zamora-Hernández, M.-A. et al. (2021).
[129] Lou, P. et al. (2022).
[130] Kitsukawa, T. et al. (2023).

vision, deep learning, and stereo vision to enhance safety and quality monitoring in HRC assembly processes. The system aims to achieve safety of workers through a combination of real-time posture tracking, hand detection, and intelligent robot motion control. A kinetic depth camera tracks the body posture of the workers to monitor their position relative to the robot to ensure a safe distance. An object detection algorithm detects the hands of the workers in real-time to prevent the robot from operating when the hands of the workers are detected on or near the workspace. This is achieved by the robot motion control that takes input from the posture tracking and hand detecting systems.[131]

3.3.3 Object detection for defect detection

Chen et al. and Kozamernik et al. suggested using deep learning techniques for detecting and identifying components as they are assembled, ensuring correct sequence and alignment of the parts. In Chen et al.'s study, a 3D Convolutional Neural Network (3D CNN) was employed to recognize actions such as twisting screws, hammering, and other manual tasks, identifying missing or incorrect actions in real-time. Additionally, a Fully Convolutional Network (FCN) that used depth images for parts recognition and segmentation was used to identify missing or misaligned components in the assembly process. Kozamernik et al. took a more straight forward approach, utilizing the YOLOv3 object detection algorithm to detect different classes of errors, such as missing components or incorrect part placement during assembly, and displaying it on the computer monitor, warning the operator about potential errors in assembly processes. Both these methods can be used to achieve quality control during the assembly to minimize errors. In addition, the deep learning techniques can again be employed in the final inspection to check for any errors in the final assembled product.[132]

Tao et al. also proposed using object detection for automating the final quality inspection processes in manufacturing. The system captures images of the product surfaces using high resolution cameras and processes them using deep learning techniques to identify and locate potential defects. The use of robotic arms is suggested to position cameras and sensors accurately to ensure thorough coverage of the surface area of the product being inspected. The method was tested on industrial products to detect surface defects on steel plates. YOLOv3 was employed for initial detection of surface defects in industrial products, followed by a level set algorithm that refines the location and nature of the detected defects. [133]

[131] Kozamernik, N. et al. (2023).
[132] Chen, C. et al. (2020).
[133] Tao, J. et al. (2022).

3.4 Object detection

Object detection is an important task in computer vision which enables identification and localization of objects from an image or video streams. It is already being used in numerous applications, ranging from autonomous driving and surveillance systems to medical imaging and industrial automation. The core objective of object detection is to determine both the location of the objects of interest using bounding boxes and to predict the class of the detected object.[134]

3.4.1 Evolution of object detection techniques

Over the years, object detection has advanced from traditional methods to modern approaches that utilize deep learning techniques. In traditional approaches, object detection was carried out in three phases – selection of region, extraction of features, and classification. These methods resulted in slow and inaccurate detection. In addition, the sliding window approach used in these methods for generating bounding boxes were computationally expensive. They had limited ability to generalize across varying conditions and scales of objects. Their limitations in handling complex scenarios with multiple objects led to the need for more sophisticated methods.[135]

The introduction of Deep Convolutional Neural Networks (DCNNs) significantly improved the capabilities of object detection. Deep learning-based object detection frameworks are designed to detect objects through an end-to-end learning process, which means that the model learns feature extraction, object localization, and classification simultaneously. These methods learn feature representation from data automatically, which results in improved detection accuracy as well as computational efficiency. [136]

One of the notable breakthroughs in deep learning-based object detection came with the introduction of the Regions with Convolutional Neural Networks (R-CNN) family. These two-stage detectors first generate region proposals and then classify them, resulting in good accuracy but often at the cost of speed. On the other hand, one-stage detectors, like YOLO and SSD, perform detection and classification in a single pass, prioritizing speed without sacrificing accuracy. The performance of object detection models has been improved significantly in recent years, making them applicable to a variety of domains

[134] Kaur, R., & Singh, S. (2023).
[135] Chen, W. et al. (2024).
[136] Kaur, R., & Singh, S. (2023).

and real-time applications.[137] A detailed comparison of different state-of-the-art algorithms are discussed in a later section.

3.4.2 Two-Stage vs Single-stage detectors

Modern object detectors can be broadly classified into two-stage and one-stage detectors. Two-stage detectors solve object detection as a classification problem where the module classifies candidates as either an object or a background. On the contrary, once-stage detectors consider object detection as a regression problem and directly predict the image pixels as objects along with its bounding box attributes. Both approaches have their own advantages and disadvantages, and the choice of method depends on the use case. [138]

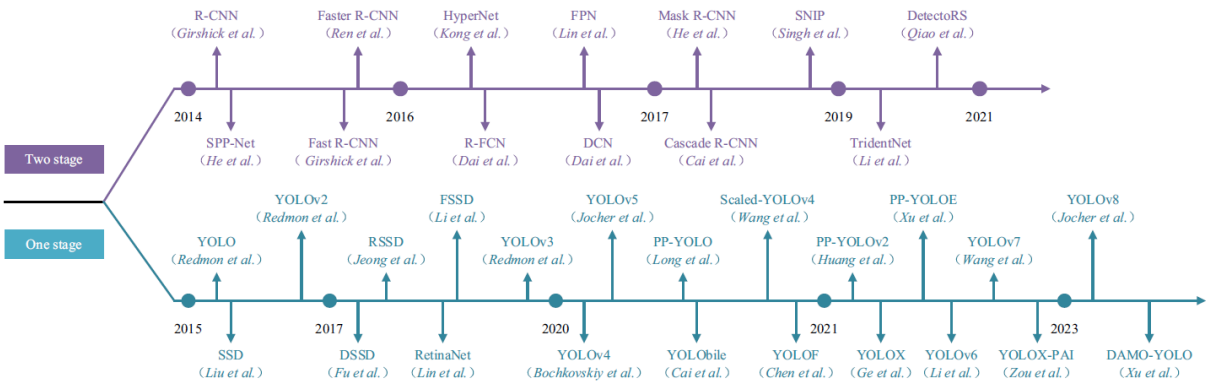


Figure 3-6: Timeline of two-stage and one-stage object detection algorithms[139]

3.4.2.1 Two-Stage Detectors

In two-stage detectors, the detection process is split into two phases. In the first phase, potential object locations in the image are identified using region-based proposals. These proposals are refined in the second phase, and objects are classified into different classes. Two-stage detectors are known for their high accuracy, particularly in detecting small or overlapping objects. However, this approach is computationally intensive due to the additional steps required for region-based analysis of the images and are relatively slower making it less suitable for real-time applications.[140]

[137] Kaur, R., & Singh, S. (2023).
[138] Zaidi, S. et al. (2022).
[139] Chen, W. et al. (2024).
[140] Kaur, R., & Singh, S. (2023).

This concept was first introduced in the Region-based Convolutional Neural Network (R-CNN) model. R-CNN achieved higher accuracy than traditional methods but was computationally expensive and slow. Fast R-CNN and Faster R-CNN, which came as successors to this model improved speed of detection. This also resulted in higher accuracy, particularly in detecting small objects. However, due to the approach of carrying out the detection in two stages, the process was still slow and computationally intensive, affecting its application in real-time applications.[141]

R-CNN (Region-based Convolutional Neural Networks)

R-CNN was one of the first models to successfully apply CNNs for object detection. In this approach, the image is first passed through a region proposal module, which uses a selective search method to identify approximately 2000 candidate regions from an image where objects might be present. Each candidate region is then resized and fed into a CNN to extract features and to classify objects. A regression model is then used to define the bounding box coordinates. While R-CNN significantly improved detection accuracy over traditional methods, it was slow and computationally expensive, with each image requiring multiple forward passes through CNN.[142]

SPP-net (Spatial Pyramid Pooling Networks)

As an attempt to address the inefficiency of R-CNN, SPP-net introduced a spatial pyramid pooling layer to allow the network to accept inputs of varying sizes, thus eliminating the need to resize each region proposal to a fixed size. This enabled faster processing by computing the CNN features only once for the entire image, and then applying the region proposals directly on these features. Although it improved speed compared to R-CNN, the model still required separate training for the classifier, bounding box regressor, and region proposal network and did not allow for end-to-end training, making it less efficient.[143]

Fast R-CNN

Fast R-CNN further streamlined the object detection process by introducing an end-to-end training approach. It eliminates the need for separate feature extraction for each region proposals by applying region proposals directly to a shared convolutional feature map, reducing computational overhead. Additionally, it uses a single stage softmax classifier and a bounding box regressor. This resulted in a model that is faster and more

[141] Kaur, R., & Singh, S. (2023).
[142] Ibidem
[143] Ibidem

accurate than its predecessors, but the region proposal step still relied on external algorithms like selective search, which limited real-time performance.[144]

Faster R-CNN

Eliminating the reliance on external algorithms for region proposal, Faster R-CNN incorporated a Region Proposal Network (RPN) into the architecture, making the region proposal process a part of the CNN itself. Faster R-CNN achieved state-of-the-art performance by reducing the time needed for region proposal generation, making it nearly real-time while maintaining high accuracy. However, its reliance on a two-stage process still makes I slower compared to one-stage detectors.[145]

R-FCN (Region-based Fully Convolutional Network)

R-FCN improves the speed of two-stage detectors by sharing almost all computations within the network, unlike previous two-stage detectors which relied on resource intensive techniques on each proposal. Instead of fully connected layers, it employs position-sensitive score maps for better localization. R-FCN enhances the R-CNN framework by making the entire network convolutional, which significantly speeds up detection while maintaining similar accuracy.[146]

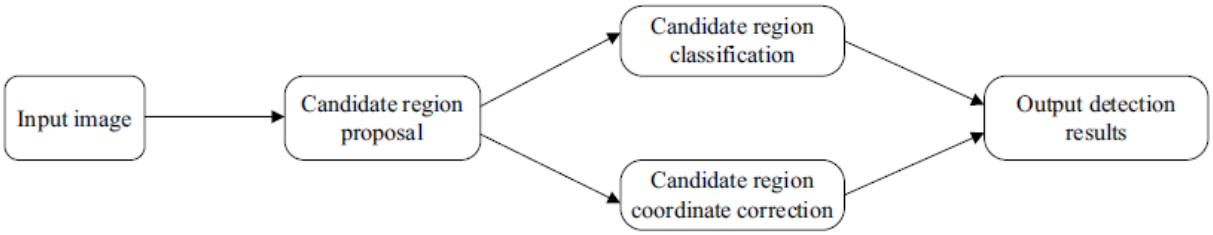


Figure 3-7: Process of object detection using two-stage detectors[147]

3.4.2.2 One-Stage Detectors

One-stage detectors simplify the object detection process by eliminating the region proposal step. Instead of first proposing regions and then classifying them, these detectors predict bounding boxes and object classes directly from the input image in a single step. This approach offers much faster detections, making one-stage detectors ideal for real-time applications.[148]

[144] Kaur, R., & Singh, S. (2023).
[145] Ibidem
[146] Ibidem
[147] Chen, W. et al. (2024).
[148] Ibidem

YOLO (You Only Look Once)

YOLO revolutionized object detection by reframing it as a regression problem, predicting bounding boxes directly from the input image rather than relying on region proposals like two-stage detectors. It performs object detection by dividing the input image into a grid, where each grid is responsible for predicting bounding boxes and class probabilities for objects whose center falls within the cell.[149] A grid cell predicts multiple bounding boxes, with each prediction consisting of coordinates for the center of bounding box, the width and height of the box, as well as the confidence score. From these overlapping bounding boxes, the one with the highest IOU is selected, while the others are removed. This approach of performing detection in a single step makes YOLO much faster than the two-stage detectors and enables it to process images in real-time. The initial version of YOLO lagged behind other prominent models in terms of accuracy, particularly for detecting small objects and it also had limitations on the number of objects per cell. However, these issues were addressed in the later versions through techniques like anchor boxes and multi-scale predictions.[150]

YOLOv2 and YOLOv3 improved upon their predecessor, with better detection accuracy. However, a significant improvement was observed with the release of YOLOv4 which incorporated several enhancements from modern deep learning practices, like CSPNet, mish activation and mosaic augmentation. YOLOv5 became popular due to its ease of use, compatibility with PyTorch, and improved accuracy. It also offered various model sizes to balance speed and accuracy providing more flexibility to the users. Newer models were released in short intervals with improved performance. At the time of this thesis, the latest model was YOLOv8 which combined features of previous YOLO versions with newer innovations, offering improved performance. Also, the compatibility with PIP and the inclusion of command line interfaces has made it even easier to use.[151]

SSD (Single Shot MultiBox Detector)

Single Shot MultiBox Detector (SSD) was the first single stage detector to match the accuracy of prominent two stage detectors like Faster R-CNN, while still achieving real-time speed. It introduced a multi-scale detection approach by using a series of convolutional layers at different depths to detect objects of varying sizes.[152] The design of SSD combines YOLO's regression approach with the anchor mechanism from Faster R-CNN. By incorporating YOLO's regression, SSD reduces the computational complexity of the neural network, enabling real-time performance. At the same time, the use of

[149] Zaidi, S. et al. (2022).
[150] Kaur, R., & Singh, S. (2023).
[151] Hussain, M. (2023).
[152] Chen, W. et al. (2024).

anchors allows SSD to capture features at different sizes and aspect ratios, ensuring detection accuracy. However, SSD was relatively weaker in detecting small objects.[153]

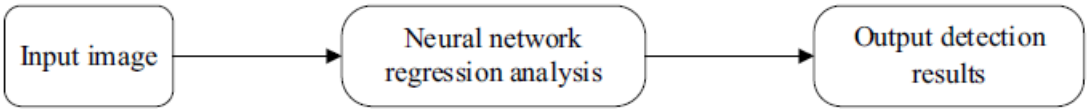


Figure 3-8: Process of object detection using one-stage detectors[154]

3.4.3 Evaluation Metrics

Different metrics are available to evaluate object detection models based on their performance in accurately detecting objects in images. These metrics help assess the ability of the model to localize and classify the objects in images and serve to compare various object detection algorithms with each other. Various metrics used are defined in this section to understand what they represent

3.4.3.1 Intersection over Union (IoU)

Intersection over Union (IoU) is the standard metric used to evaluate the localization accuracy of an object detection model. It measures the overall between the predicted bounding box and the ground truth bounding box.[155]

$$IoU = \frac{Area\ of\ intersection\ of\ predicted\ and\ ground\ truth\ bounding\ boxes}{Area\ of\ union\ of\ predicted\ and\ ground\ truth\ bounding\ boxes}$$

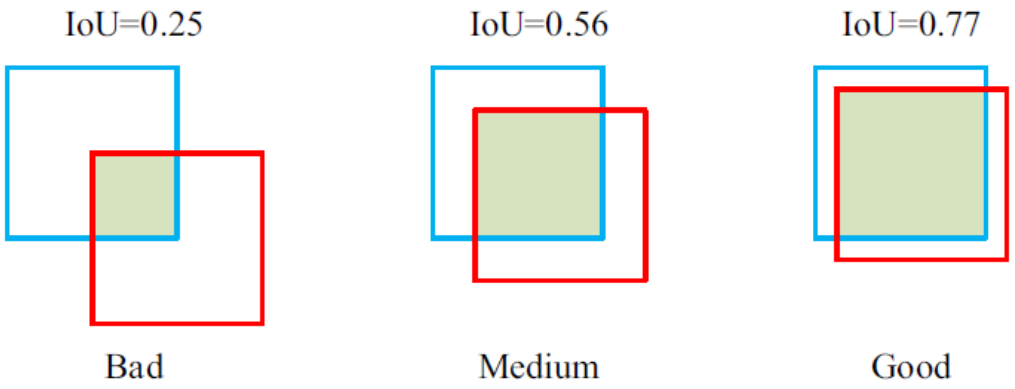


Figure 3-9: Illustration of IoU calculation[156]

[153] Kaur, R., & Singh, S. (2023).
[154] Chen, W. et al. (2024).
[155] Kaur, J., & Singh, W. (2022).
[156] Chen, W. et al. (2024).

The IoU value ranges from 0 to 1, with 1 indicating a perfect match between the predicted and actual bounding boxes, whereas 0 indicating no intersection between the two. A threshold value of IoU is used to determine whether the detection can be considered successful or not. A common threshold is IoU > 0.5, meaning that at least 50% of the predicted bounding box must overlap with the actual bounding box. This metric directly influences the calculation of other metrics.[157]

3.4.3.2 Accuracy, Precision, Recall, and F1 Score

Accuracy, Precision, Recall, and F1 score are some basic metrics for evaluating the detection performance. The results obtained from calculating Intersection over Union (IoU) is used to calculate these metrics. For this, every predicted bounding box must be classified as True Positive, True Negative, False Positive, or False Negative. True Positives (TP) are correctly predicted detections that match the ground truth objects. True Negatives (TN) occur when the model correctly identifies the absence of objects in areas where no objects are present. False Positives (FP) are incorrectly predicted detections, where the model detects objects that either do not exist or do not correspond to the actual ground truth. False negatives (FN) occur when the model fails to detect objects that actually exists in the image or video.[158]

Predict \ Fact	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Figure 3-10: Confusion matrix illustrating classification outcomes[159]

[157] Kaur, J., & Singh, W. (2022).
[158] Chen, W. et al. (2024).
[159] Ibidem

Accuracy

Accuracy measures the overall proportion of correct predictions made by the model, considering both True Positives and True Negatives.[160]

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision (P)

Precision measures the fraction of correctly predicted positive samples out of all samples predicted as positive.[161]

$$Precision = \frac{TP}{TP + FP}$$

Recall (R)

Recall measures the fraction of correctly predicted positive samples out of all actual positive samples.[162]

$$Recall = \frac{TP}{TP + FN}$$

F1 Score

Precision and Recall alone provide only a partial view of a model's performance. High precision means that when the model makes a prediction, it is likely correct, resulting in fewer false positives. High recall means that the model successfully detects most of the objects that are present, resulting in fewer false negatives. However, improving one of these metrics often leads to a decline in the other. The F1 score is often used as a single metric to summarize both precision and recall. It is the harmonic mean of precision and recall, giving a balanced measure of the model's performance. [163]

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.4.3.3 Average Precision (AP)

Average Precision (AP) is one of the most widely used metrics in object detection for evaluating the precision-recall tradeoff at different IoU thresholds. It is derived from the precision-recall curve which plots precision against recall at various thresholds. To

[160] Kaur, R., & Singh, S. (2023).
[161] Ibidem
[162] Ibidem
[163] Ibidem

compute the precision-recall curve, the model's detections are ranked by confidence score, and precision and recall are calculated for different confidence levels. The precision-recall curve provides insight into how the model's precision varies with recall. As recall increases, precision typically decreases because of chances for more false positives.[164]

AP is calculated as the area under the precision-recall curve. In some cases, the AP is computed at a single IoU threshold, such as 0.5, and is referred to as AP@0.5. In other cases, it may be averaged over multiple IoU thresholds, such as [0.5-0.95], to provide a more comprehensive evaluation of the model's performance across different levels of localization accuracy.

$$AP = \int_0^1 P\,(r)\,dr$$

where P (r) denotes the precision value when the recall is r.[165]

3.4.3.4 Mean Average Precision (AP)

Mean Average Performance (mAP) is a metric used in object detection to evaluate the overall performance of a model across all object classes. It is the mean of average precision values computed for each class individually, providing a single value that summarizes the overall detection accuracy of the model.

$$mAP = \frac{\sum_{c=1}^{c} AP\,(c)}{C}$$

Where c denotes the total number of classes detected and AP (c) denotes the AP value of the model on the c th class.[166]

3.4.3.5 Inference time

Inference time refers to the amount of time required for the object detection model to process an image or a single frame from a video to make predictions. It is usually expressed in milliseconds. It is an important metric for applications that require real-time processing as it directly affects the model's responsiveness and speed. The lower the inference time, the faster is the model.[167]

[164] Kaur, R., & Singh, S. (2023).
[165] Ibidem
[166] Ibidem
[167] Chen, W. et al. (2024).

3.4.3.6 Frames Per Second (FPS)

The performance of object detection models in terms of speed is commonly evaluated using the Frames Per Second (FPS) metric. It measures how many frames an object detection model can process in one second, providing a clear indication of the model's inference speed. High FPS values indicate faster processing, which is essential for applications requiring real-time detections.[168]

3.4.4 Current challenges in object detection

Even though object detection advanced significantly in recent years, several challenges continue to impact the effectiveness of current algorithms in real-world applications. These challenges, along with the possible strategies proposed by researchers to minimize them are discussed below.

- Data Annotation

Currently, object detection algorithms primarily rely on supervised learning, which requires human-annotated data for training. This is a very time consuming and tedious task. In reality, large amounts of unannotated data are available, and it is very inefficient to annotate these manually. Even though data annotation programs currently offer tools for semi-automated annotation, it is still very labor intensive. Weakly supervised algorithms aim to address this issue by training object detectors using only image-level annotations, eliminating the need for precise border annotations. Although these algorithms currently face challenges with accuracy and positioning precision, their advancement could significantly reduce the effort required for object detection.[169]

- Small Object Detection

Detecting small objects remains one of the most difficult tasks in object detection. Object detection algorithms often struggle with small objects due to the lack of adequate feature representation in deep convolutional networks, as small objects occupy only a few pixels in the overall image. Low-resolution images further worsen the performance as they can only carry finite contextual details. To address this issue, solutions such as data augmentation or increasing the model's input resolution have been suggested. For applications where detecting small objects is critical, capturing data from a closer distance, if feasible, can help ensure the object occupies a sufficient portion of the image frame.[170]

[168] Chen, W. et al. (2024).
[169] Ibidem
[170] Kaur, R., & Singh, S. (2023).

- Occlusions

Occlusion refers to situations in object detection where an object is partially or fully blocked, due to different reasons. It is one of the most challenging problems in computer vision because the object detector is expected to recognize the object even when it is not fully visible. This can lead to errors in detecting, classifying and localizing objects, making it a significant issue for many applications. Training object detection models using data that includes occlusion scenarios can help the model learn to recognize partially occluded objects.[171]

- Intraclass Variation

Intraclass variation refers to differences in different instances of the objects within the same object class. These variations, influenced by both inherent factors such as difference in size, shape, color, and material and environmental factors such as lighting, perspective, and camera quality, pose significant challenges for object detection algorithms. Ensuring that the training data includes a wide range of these factors can help mitigate this issue by providing the model with more varied examples during training.[172]

- Efficiency

Computational complexity handled by object detection models increases with the number of object classes to be detected. This rise in complexity requires high computational resources to process numerous locations within a single image. Therefore, high performance GPUs often become necessary both during training and inference to ensure sufficient performance, particularly for complex models.[173]

- Generalization Issues

Generalization problems are another challenge in object detection, which is caused by either underfitting or overfitting of the model. Underfitting happens when the model fails to learn the patterns in the data and performs poorly on both the training data and new, unseen data. This typically occurs during the early stages of training and can be addressed by increasing the number of epochs or the complexity of the model. Overfitting occurs when a machine learning model learns the training data too well, resulting in exceptional performance on training data, but poor performance on new, unseen data. This can be mitigated through techniques such as data augmentation, early stopping, or using regularization methods. [174]

[171] Kaur, J., & Singh, W. (2022).
[172] Ibidem
[173] Kaur, R., & Singh, S. (2023).
[174] Ibidem

- Video Object Detection

In real-world applications, object detection is often performed on video sources. In this case, object detection algorithms analyze each frame of the video to detect and localize objects. However, frames from video sources may often lack the quality of the images captured from a camera due to lack of focus or presence of blurred parts due to movement of objects. The difference in quality in the images used for training and the frames on which object detection is performed can cause performance issues. Including frames from video sources in the training dataset can help reduce this problem.[175]

- Inference Speed

Real-time object detection demands not only high accuracy but also fast processing speeds. Speed is a critical factor, especially for applications requiring real-time detection from video feeds. While several modern object detectors are able to achieve real-time performance, they still fall short of achieving speeds comparable to human perception. Therefore, it is necessary to improve the speed further.[176]

- Class Imbalance

Irregular data distribution in the dataset is referred to as class imbalance. This can either be a foreground-background imbalance or foreground-foreground imbalance. Foreground-background imbalance refers to situations where there is a large disparity between the number of pixels or regions representing objects and backgrounds in the image, making it harder for models to differentiate between actual objects and the background. [177]

Foreground-foreground imbalance, on the other hand, arises when the number of instances across different object classes are uneven. This occurs when certain object classes dominate the dataset, while others are significantly underrepresented. This can cause the model to become biased toward the majority class, often resulting in poor detection of the minority class. Both of these issues can affect the overall accuracy of the model and can be mitigated through techniques like data augmentation, under sampling, or oversampling.[178]

[175] Kaur, J., & Singh, W. (2022).
[176] Ibidem
[177] Kaur, R., & Singh, S. (2023).
[178] Ibidem

3.4.5 Comparison of Object detection algorithms

In this section, major object detection algorithms are compared to each other. First, their strengths and weaknesses are outlined based on findings from review papers. Next, their performance is assessed on benchmark datasets, focusing on accuracy and speed. Finally, the algorithms are evaluated through comparison results from studies using custom datasets, providing insights into their effectiveness in specific, real-world applications.

3.4.5.1 Strengths and weaknesses of major algorithms

Various object detection algorithms have been developed over the last decade, each aiming to improve upon the performance of its predecessors. While accuracy remains a crucial metric in object detection, the speed of detection is often just as important, especially in applications where real-time processing is required. In order to achieve optimal performance, researchers aim to find the right balance between accuracy, speed, and resource efficiency. The following tables summarize the strengths and weaknesses of different popular two-stage algorithms.

Table 3-4: Strengths and weaknesses of major two-stage object detection algorithms

Algorithm	Strengths	Weaknesses	Sources
R-CNN	First Neural Network based on region proposal Significant performance improvement over the traditional methods	Complexity training High time and space expenditures	Kaur and Singh (2023) Chen et al. (2024)
SPP-Net	Extracts the features of entire image at once Faster than RCNN	High computational costs No end-to-end training	Kaur and Singh (2023) Chen et al. (2024)
Fast R-CNN	Faster and accurate than previous models Reduce training time and feature storage space	Not fast enough for real-time application No end-to-end training	Kaur and Singh (2023) Chen et al. (2024)
Faster R-CNN	Reduce training time and improved detection efficiency Provides end-to-end training	Inefficient for real-time applications	Kaur and Singh (2023) Chen et al. (2024)

		Poor detection of small and multi-scale objects	
R-FCN	Faster than other two-stage detectors Accurate positioning	Poor detection of multi-scale objects	Zaidi et al. (2022) Chen et al. (2024)

Table 3-5: Strengths and weaknesses of major one-stage object detection algorithms

Algorithm	Advantages	Disadvantages	Sources
YOLO	Simple network structure Removes the concept of region proposal Fast detection speed	Low detection accuracy for dense and small objects Limited number of objects per cell	Kaur and Singh (2023) Chen et al. (2024)
SSD	Accuracy on par with Faster R-CNN Faster than YOLO	Poor detection of small objects Slow model convergence	Kaur and Singh (2023) Chen et al. (2024)
YOLOv2	Faster than YOLO High classification accuracy	Complex training Poor detection of small objects	Kaur and Singh (2023) Chen et al. (2024)
YOLOv3	Better multi-scale detection accuracy Improved small object detection accuracy	Large dataset recommended for training High false negative rate (missed detections)	Kaur and Singh (2023) Chen et al. (2024)
YOLOv4	Single GPU training Enhanced accuracy Automatic hyper-parameter optimization	High model complexity	Zaidi et al. (2022) Kaur and Singh (2023)
YOLOv5	Real-time detection Better accuracy	Performance can be improved	Kaur and Singh (2023) Chen et al. (2024)

YOLOv6	Trade-off between speed and accuracy for different industrial scenario applications	Needs further improvement to adapt to more demanding scenarios	Chen et al. (2024)
YOLOv7	Improved accuracy Real-time detection	Higher calculation volume Increased training costs	Chen et al. (2024) Terven, Córdova-Esparza and Romero-González (2023)
YOLOv8	Improved small-object detection Ease of use (pip install, command line interface) Real-time detection	Higher training time for larger models	Chen et al. (2024) Terven, Córdova-Esparza and Romero-González (2023)

3.4.5.2 Performance comparison of major algorithms on benchmark datasets

A comparison of the performance of major two-stage and one-stage object detection algorithms is presented in this section. The evaluation focuses on the accuracy and speed of the algorithms. The performance on pre-trained models of each algorithm is evaluated on the PASCAL VOC and MS COCO datasets, two widely used benchmarks. These datasets serve as benchmarks due to their comprehensive nature and the diversity of objects they contain. However, some algorithms were primarily developed and benchmarked on specific datasets, and therefore, their benchmarks are not available across both datasets. For example, older algorithms tend to lack benchmarks on the MS COCO dataset, which was released later and presents a higher level of complexity due to a larger number of object classes and complex environments. Similarly, recent algorithms were primarily benchmarked on COCO and lack results on PASCAL VOC, as it is no longer considered a leading benchmark for modern models.[179]

Table 3-6 provides a comparison of different algorithms based on available benchmark results, providing insights into their performance in terms of detection accuracy and inference speed. The metric used for measuring accuracy is mean Average Precision (mAP) at an IoU threshold of 0.5 for PASCAL VOC and mAP averaged across IoU

[179] Chen, W. et al. (2024).

thresholds from 0.5 to 0.95 for COCO. Speed is measured in terms of frames per second (fps).

Table 3-6: Performance comparison of algorithms on benchmark datasets[180]

Algorithm	PASCAL - mAP (0.5)	COCO - mAP (0.5-0.95)	FPS
R-CNN	58.5	-	0.03
SPP-Net	59.2	-	2
Fast R-CNN	70.0	19.7	3
Faster R-CNN	73.2	21.9	5
R-FCN	83.6	27.6	5.9
YOLO	63.4	-	45
SSD	79.8	28.8	19.3
YOLOv2	76.8	21.6	67
YOLOv3	-	33.0	19.6
YOLOv4	-	43.5	23
YOLOv5	-	50.7	82.6
YOLOv6	-	52.5	98
YOLOv7	-	51.2	161
YOLOv8	-	53.9	283

From the above table, it is clear that there is a clear progression in terms of performance for the newer models in both accuracy and speed. Among the two-stage detectors, Faster R-CNN and R-FCN showcased much better performance compared to earlier models, both in terms of speed and accuracy. However, they are not ideal for real-time applications.

[180] Kaur, R., & Singh, S; Chen, W. et al. (2023); (2024).

Among the one-stage detectors, the YOLO models show a clear progression in both accuracy and speed. From the initial version, YOLO has tried to achieve a balance between speed and accuracy. The latest versions, YOLOv5 to YOLOv8, show significant improvements over earlier models, with YOLOv8 achieving the highest accuracy of 53.9% on the COCO dataset. Additionally, with a maximum of 283 fps, YOLOv8 achieved significantly higher inference speed than all the previous models, making it a great choice for applications requiring real-time detections.

3.4.5.3 Experimental comparisons of major algorithms on custom datasets

The evaluation of the object detection algorithms on benchmark datasets like PASCAL VOC and MS COCO provide a good overview of the potential of the algorithms in detecting a variety of objects in different backgrounds and conditions. However, since the pre-trained models of these algorithms are trained to detect a limited category of objects, it maybe often required to train custom object detection models to achieve specific use cases. In order to evaluate the potential of algorithms in learning to detect domain-specific objects, it will be helpful to evaluate the performance of the algorithms on custom datasets. During the literature review, 7 studies were identified, that have experimentally compared the latest object detection algorithms on custom datasets. The key findings from the studies are provided in the Table 3-7 below.

Table 3-7: Comparison of major algorithms on custom dataset

Test objective and Environment	Algorithms compared	Results	Key takeaways	Study
semiconductor defect detection[181]	Faster R-CNN, DINO, RetinaNet, YOLOv7	DINO had the highest accuracy, whereas YOLOv7 came close second. However, inference time of DINO (108.7ms per image) is very high compared to that of YOLOv7 (20.2ms per image)	With the 2nd best mAP and lowest inference time, YOLOv7 offers a balanced solution	Dehaerne, E. et al. (2022).

[181] Dehaerne, E. et al. (2022).

Hand gesture recognition[182]	YOLOv5,YOLOv6, YOLOv8	All the tested algorithms showcased very high accuracy with YOLov8 being the most accurate	Custom YOLOv8 model returned the most accurate results	Herbaz, N. et al. (2023).
Threat detection (firearms, knives, fire)[183]	Faster R-CNN, SSD, YOLOv4, YOLOv5, YOLOv7,YOLOv8	All models showcased good performance. YOLO v8 achieved the highest accuracy and speed, followed by YOLOv7.	YOLOv8 is the most accurate and fastest among the compared models.	Hasan, M. et al. (2023)
People detection using fish-eye cameras[184]	YOLOv8, Mask R-CNN	Model pre-trained on COCO dataset did not provide sufficient accuracy for the use-case. The custom-trained models gave good accuracy	YOLOv8 is more accurate and more resource efficient than Mask R-CNN	Telicko, J & Jakovics, A. (2023)
object detection from remote sensing satellite images[185]	Faster R-CNN, YOLOv6, YOLOv7, YOLOv8	YOLOv8 exhibited superior performance in Precision, Recall, and mAP, while also achieving the shortest inference time.	YOLOv8 outperformed other models trained with the same dataset	Adegun, A. et al. (2023).
Pothole detection under diverse weather condition[186]	Mask R-CNN, CASCADE R-CNN, SPARSE R-CNN, YOLOv5, YOLOv6, YOLOv7	YOLOv7 offered best performance, the smallest size and the fastest inference time. At night, Cascade R-CNN showcased better mAP.	YOLOv7 performed better in all conditions except at night	Jakubec, M. et al. (2023).

[182] Herbaz, N. et al. (2023).
[183] Hasan, M. et al. (2023).
[184] Telicko, J & Jakovics, A. (2023).
[185] Adegun, A. et al. (2023).
[186] Jakubec, M. et al. (2023).

Automatic reading in smart metering system[187]	YOLOv5, YOLOv6, YOLOv7, YOLOv8	YOLOv8 outperformed the other models in accuracy. YOLOv5 – shortest training time, YOLOv7 – longest training time	YOLOv8 delivered highest accuracy with a reasonable training time	Hattak, A. et al. (2023).

3.5 Key Takeaways from Literature Review

The first section of the literature review reveals a variety of tools for digitizing value stream mapping. Although majority of the articles focused on the use of advanced technologies for analysis, there are a lot of approaches proposed for digital data acquisition.

While many of these are machine centric, two sensors, RFID and RTLS, has been mainly proposed for data acquisition in human-centric assembly processes. They enable tracking material flow in real time, but they come with limitations, particularly the need for manual tagging and interference with the workflow. This makes it difficult to use these sensors for continuous tracking over long durations.

Second section of the literature review reveal that there are already many applications of object detection in manufacturing. Object detection can also be used to collect time and location data from the shop floor, without interfering with the workflow. It has the potential to collect a variety of data and therefore its application in VSM is promising. However, studies linking object detection with value stream mapping was not found in the literature. This research aims to fill this gap by exploring the potential of object detection for non-intrusive digital data collection for value stream mapping.

Selection of Algorithm

In the benchmark evaluations, two-stage detectors like Faster R-CNN and R-FCN demonstrated significant improvements over their predecessors in terms of accuracy and speed but continued to struggle with offering real-time detection capabilities. On the other hand, the one-stage YOLO family consistently evolved to balance speed and accuracy, with YOLOv8 achieving the highest mAP on the COCO dataset and showcasing exceptional inference speed, making it ideal for real-time applications.

The evaluation of algorithms on custom datasets provides results similar to those on benchmark datasets. The newer versions of the YOLO detector outperform other

[187] Hattak, A. et al. (2023)

detectors in both accuracy and speed. In certain applications, such as low light conditions, algorithms like Faster R-CNN and Cascade R-CNN offer competitive performance, but they often fall short in terms of speed and resource efficiency when compared to the YOLO family of models.

The state-of-the-art one-stage detection models, YOLOv5, YOLOv6, YOLOv7, and YOLOv8 exhibit accuracy levels that are fairly close, though YOLOv7 and YOLOv8 show slightly better results. In terms of inference speed, YOLOv8 offers a substantial improvement over the others. YOLOv5 has the shortest training time, making it a good option for resource-constrained environments, while YOLOv7 requires the longest training time among the three. YOLOv8, with a training time that falls between the two, strikes a good balance.[188] Additionally, YOLOv8 offers flexible installation and usage options, such as pip installation and a command-line interface, thereby making it easier to use.

In conclusion, results indicate that most of the state-of-the-art algorithms are capable of delivering good accuracy and real-time performance. However, YOLOv8 proves to be the best choice, offering a balance between accuracy, inference speed, resource efficiency and ease of use. Particularly in an industrial environment, where lighting conditions and backgrounds doesn't undergo significant changes, YOLOv8 should offer excellent performance. Therefore, YOLov8 is chosen for the object detection task in the practical part of this thesis.

[188] A. Hattak et al. (2023).

4 Practical Implementation

This chapter outlines the proposed solution for digital data acquisition from shop floors, utilising object detection in the context of digitization of value stream mapping (VSM). The conventional VSM approach uses a representative unit, visualizing the value stream based on production data from a single unit within a product class. This may not reflect the full dynamics of the process, especially in complex production systems. In contrast, the proposed method captures production data over a defined period of time, and thereby visualize the actual situation in the shop floor. The approach uses object detection to collect the location and corresponding timestamp of detection of products, and operators in the shop floor during the assembly process.

Unlike other sensor-based approaches discussed in the literature that require attaching sensors or labels to products and operators to gather location data, object detection enables data collection without physical attachments or additional workload for operators. Therefore, this method enables non-intrusive data collection from the assembly area without affecting the productivity of the operators.

4.1 Conceptualization

The proposed approach aims to enable digital data collection from the shop floor through the use of object detection. The core idea is to use object detection algorithms to track products as they move across various stages of the assembly process. Object detection models can identify different objects and return their class names, along with their positions in the form of bounding boxes. By systematically saving this data, the information required to determine Key Performance Indicators (KPIs) relevant for value stream mapping is obtained. The methodology involves collecting video data of the assembly process, detecting objects of interest, tracking their movement, and processing the results to extract meaningful insights.

This is achieved by recording the assembly process using a camera and feeding the video input into an object detection model. State-of-the-art object detection algorithms are capable of real-time detection, allowing the model to infer live video from the camera. Alternatively, the assembly process can be recorded and fed to the object detection model after the completion of the assembly process. To achieve optimal results, the camera must be positioned to provide a comprehensive view of the assembly process, ensuring maximum coverage of workers and products thereby minimizing any possible occlusions.

The output of object detection algorithms indicates the presence of objects in the frame by displaying the predicted object's class and marking its location with a bounding box. However, this data directly does not provide information about the progress of the product through the assembly process. To address this, the concept of region of interest (ROI) is introduced. ROIs represent distinct zones on the shop floor, such as workstations and waiting areas. By defining these zones, the presence of products in these areas are easily identified, and this information is used for further processing. Each time a product is detected inside an ROI, the detection information such as the product class, track ID and the corresponding timestamp are saved to a csv file. This data is then processed to determine the KPIs relevant to VSM.

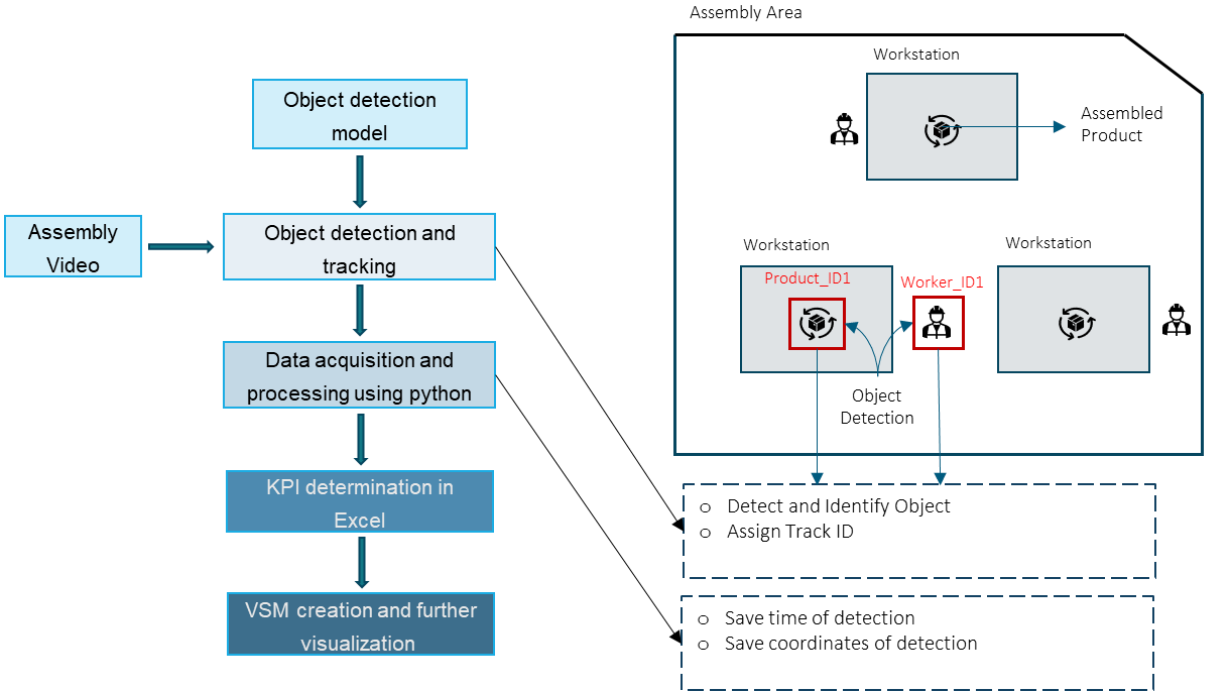


Figure 4-1: Conceptualization

A significant challenge in using this approach is ensuring that products are continuously detected within and across workstations. During the assembly process, products undergo continuous modifications, causing changes in their shape and appearance as they progress along the assembly line. To ensure continuous detection, a base component is identified for each product type. This base part remains unchanged across different stages of the assembly, allowing for consistent detection. If the assembly process includes certain workplaces that does not share any common parts with others, a different base part is to be identified for that workplace and is classified as a sub-class of the product class. The object detection model is trained to detect these base component(s), ensuring that the model recognizes and tracks the product even as its overall shape changes during the assembly process.

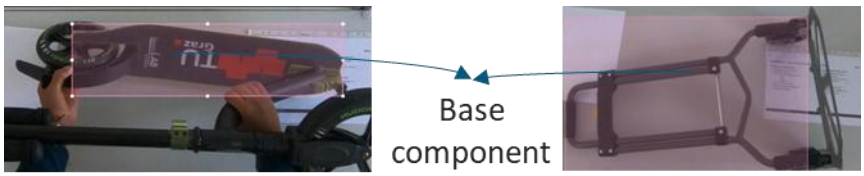


Figure 4-2: Approach for tracking products

Another important requirement is the ability to differentiate between individual units of the same product class. When multiple units of the same product type are being assembled simultaneously across different workplaces, it is essential to distinguish between them. This is achieved through the track functionality offered by object detection algorithms. When a unit of a product is first detected, it is automatically assigned a unique track ID by the tracker, which serves as an identifier for that unit. The tracker ensures that the same unit is consistently recognized and assigned the same ID across multiple frames. This is achieved by comparing each detection to the previous detections, and assigning new or existing IDs based on a matching threshold.

4.2 Technical procedure

Therefore, the proposed method consists of the following key steps – training an object detection model to detect the required objects, developing a python script to run inference on the assembly video and to save the detection results, and KPI determination and VSM creation using excel based on the collected data. The process is visualised in Figure 4-3

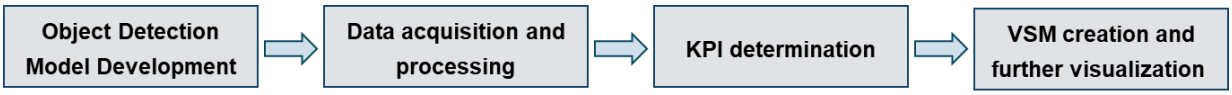


Figure 4-3: Technical Procedure

The following section provides a detailed explanation of each step in the proposed approach.

4.2.1 Model Development

Object detection algorithms are pre-trained on large, publicly available datasets and can detect many of the common objects such as humans, cats, dogs, cars and others. These models can be used out-of-the-box to detect objects that belong to classes that are included in these public datasets. However, despite the rapid growth of high-quality public datasets, the number of classes detectable by pre-trained models are still limited. Object

detection is currently used in a variety of applications, and it is very often necessary to detect objects that are not included in these classes. Therefore, object detection algorithms offer users the ability to train custom object detection models.

The results from the literature review underscore the necessity of training custom models for real-world applications. Even for object classes that are included in pre-trained models, performance may not meet expectations due to differences in camera types, angles, and object-to-camera distances, all of which can affect detection accuracy. Therefore, to achieve reliable performance, it is often beneficial to train models using samples captured with similar cameras and angles as those in the intended application.

For this project, object detection models were developed to detect the workers and the two products assembled in the learning factory of TU Graz. Yolov8 was chosen as the object detection algorithm for the study based on the result of the literature review described in section (describe section). Yolov8 offers different model sizes namely Yolov8n (nano), Yolov8s (small), Yolov8m (medium), Yolov8l (large), and Yolov8x (extra-large) offering various trade-offs between speed, accuracy and computational requirements. For this work, the dataset was trained on Yolov8n, Yolov8s, and Yolov8m to compare performance and select the best model. The two largest models were excluded due to computational limitations.

The following section explains the stages involved in developing a custom object detection model.

4.2.1.1 ***Data collection process***

The first step in training an object detection model is to prepare a high-quality dataset. This dataset can be composed of images captured specifically for the project, images available in public datasets, or open-source images available online. The images for the dataset are to be carefully chosen as the quality and diversity of the dataset directly influence the model's ability to learn and generalize across various real-world scenarios.

The best practice is to gather a representative set of images that include the objects of interest in different environments, backgrounds, angles, lighting conditions, and sizes. However, the extent of variety also depends on the use case of the model. For very complex tasks, a large and diverse dataset is essential, covering a broad range of conditions such as indoor and outdoor settings, varying lighting and weather conditions, and different environmental complexities.[189]

[189] Kaur, J., & Singh, W. (2022).

On the contrary, in more controlled environments, such as industrial use cases, the need for variety may be limited to the specific situations in which the model will be applied. For example, in an industrial setting where objects are consistently viewed from fixed angles under controlled lighting conditions, the dataset can be more focused, reducing the need for diverse images. This approach ensures that the model is optimised for the particular conditions of the intended application.

4.2.1.2 *Dataset Description*

The dataset used in this work is a custom dataset collected specifically for this project. The images were extracted from multiple videos recorded in the learning factory at TU Graz. The video data was collected using a GoPro Hero7 camera, recorded at 30 frames per second. The first set of data was collected using the GoPro mounted on the ceiling of the learning factory, providing a comprehensive view of the entire assembly area. This placement ensures clear visibility of both products and workers, with minimal chances of occlusions. A video of the assembly of scooters and hand trucks, the two products assembled in the learning factory, were recorded in a similar setup that is used for testing. Frames captured from this video were included in the dataset used for training the models for detecting both workers and products. A total of 1067 images for training, and 238 images for validation were included in the dataset used for training the worker detection model.

In manual assembly processes, even with standardized procedures in place, the way workers handle products can vary slightly from one assembly to another. This variations in product handling result in the products appearing in different orientations within the camera's field of view, necessitating the model's ability to detect products under these varying conditions.

To improve the model's generalization capability, additional images were generated by combining various angles of the products with a frame captured from the ceiling-mounted camera. For this purpose, additional videos were captured in which scooters and hand trucks were handled at various angles within the workplace. Close-up images of scooters and hand trucks captured from these videos were then layered onto an image of the entire assembly area. This method simulated the actual appearance of the products in the workplaces, as seen from the overhead camera. This approach allowed the model to learn how the products would look on the camera while accounting for potential variations in product orientations as handled by workers.

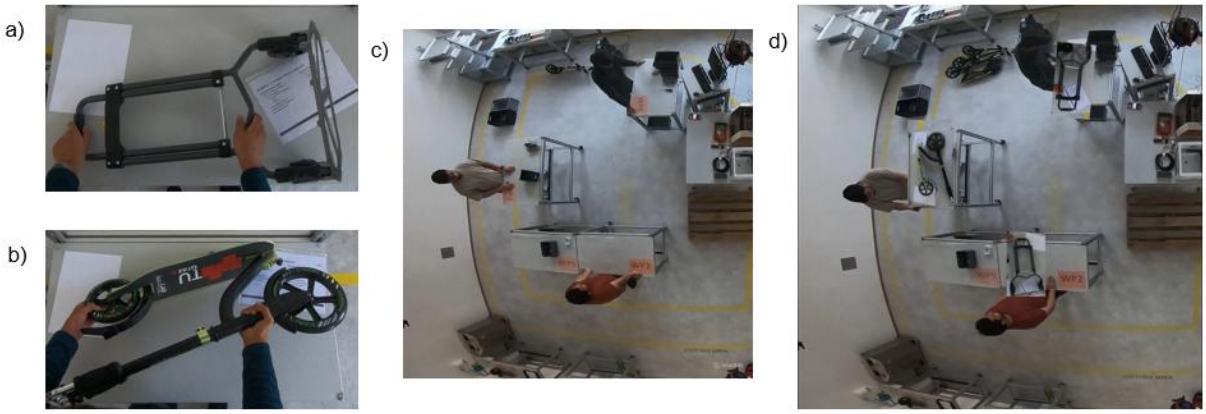


Figure 4-4: Creation of combined images to address scale differences in the training dataset.
a) Close-up view of a hand truck, b) Close-up view of a scooter, c) Overhead view of the assembly area, d) Combined image mimicking the actual appearance of the products from the overhead camera

To enhance training variability, 1670 images generated using this method and 3321 images captured from the assembly video mentioned earlier were included in the dataset used for training the product detection model. The total dataset was divided into training, validation, and testing sets with a ratio of 70:20:10, respectively.

4.2.1.3 ***Annotation of images***

For training an object detection model, the image dataset must be accompanied by annotation files that provide information about the location of objects in each of the images. Yolov8 uses the Yolo annotation format, where each image has a corresponding text file containing details about the objects within it. Each line in these text files represent each instance of the objects in the image. This information includes the class ID, which denotes the object class of the instance and the bounding box coordinates for that instance. The bounding box coordinates are specified by the centre point (x,y) of the bounding box, and its dimensions (width, height), all relative to the image size. An example of the .txt file is shown in Listing 4-1 below.

```
1 0.897556 0.559486 0.129333 0.069898
0 0.714065 0.603662 0.030000 0.098676
0 0.285361 0.418005 0.032556 0.116639
```

Listing 4-1: Example of the contents of a YOLO annotation file

Rules for Annotation

The dataset annotation process involves drawing bounding boxes around the objects of interest in each image and assigning the correct class label. Annotation accuracy directly impacts the performance of the model. Proper care must be taken to ensure that the bounding boxes accurately enclose the objects, and the annotation patterns are consistent throughout the dataset. For the annotation of products and workers in this work, a set of annotation rules was defined to ensure that the annotations remained consistent throughout the dataset. The rules followed are outlined in Table 4-1 below.

Table 4-1: Defined rules for annotation

Rules for Annotation
Objects must be fully enclosed by the bounding box.
Bounding boxes must be as tight as possible, leaving no unnecessary empty spaces.
Ensure that the label IDs are consistent throughout the dataset.
If objects overlap, each object should have its own bounding box. Do not group multiple objects into a single bounding box.
Partially occluded objects (up to 50%) must be annotated, while anything more than that should be left unannotated.
If objects are blurred, they must be annotated only if they are still recognizable
If parts of an object are overlapped by other objects (e.g. A worker’s hands overlapping a product), the entire object, including the overlapping part, should be annotated.

Tools for Annotation

Several tools are available to annotate images and to create the annotation files. Multiple tools were used and compared during the study and the result of comparison is described in the Table 4-2 below.

Table 4-2: Comparison of common annotation tools

Annotation tools	Advantages	Disadvantages
CVAT	Dashboard for project and task management, open source, feature rich, semi-automated annotation, image & video annotation	May not be as intuitive to use in the beginning compared to others
Roboflow	Easy to use, labelling assistance, version control, built-in augmentation features	In the free version, all datasets are made publicly available to all users
Makesense.ai	Easy to use, no sign in required, very convenient for quick and short tasks	No dataset organization capabilities, no assistance for labelling, no video annotation

Considering the ease of managing different datasets for the project, video annotation capability, and other advanced functionalities for assisting in annotation process, the open-source annotation tool, CVAT was used for the annotation task.

Process Description of Annotation

The open-source annotation tool, CVAT was used for annotating images for this work. The general process for annotation is very similar across annotation tools expect for slight variations in the way the tool is designed. The following steps outline the detailed process followed to create good quality annotations for training the object detection models.

Dataset Upload: The selected images and videos were uploaded to CVAT for annotation. The tool provides efficient project management features which allows to store the dataset in a structured format. The tool automatically separates each frame in a video into subsequent images to facilitate annotation.

Bounding Box Placement: Each object of interest, including both products and workers, was annotated by placing bounding boxes around each of them in every selected image. It was ensured that the bounding boxes tightly enclosed the objects, ensuring minimum background was included in the bounding box while ensuring the comprehensive coverage of the object.

Class Label Assignment: For each bounding box, the relevant class label was assigned (Scooter, Hand Truck, or Worker).

Review: To maintain annotation quality, the annotated frames were reviewed periodically to confirm that the bounding boxes were placed based on the predefined rules.

Export: After completing annotations and reviewing the accuracy of the bounding boxes, the annotations were exported. CVAT offers multiple formats for exporting the labels. Since YOLOv8 was chosen as the object detection algorithm for the experiment, labels were exported in YOLO format.

CVAT webapp interface and examples of annotations using the tool is shown in Figure 4-5 and Figure 4-6.

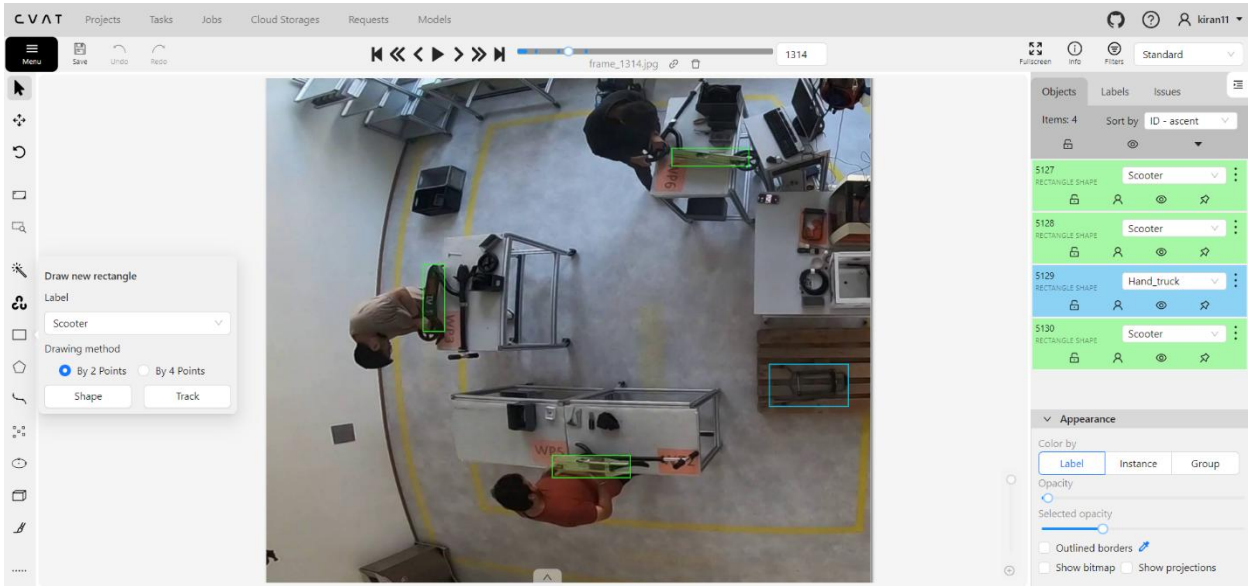


Figure 4-5: Example annotation of products using CVAT

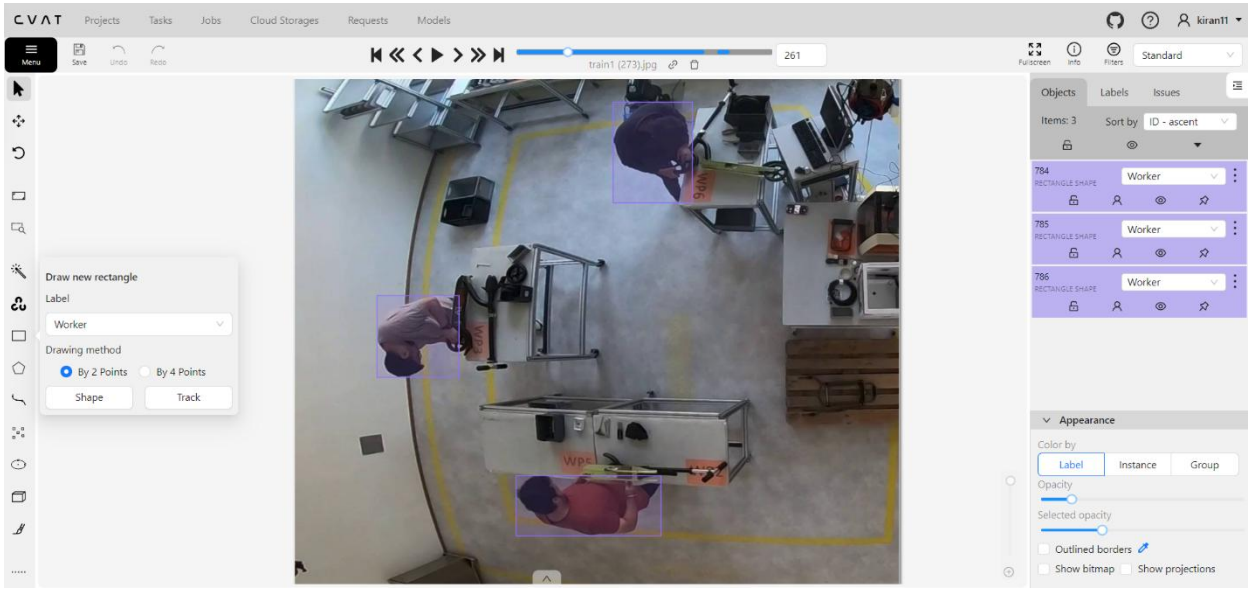


Figure 4-6: Example annotation of workers using CVAT

The choice of the annotation tool does not directly influence the accuracy of the models, as all these models offer the annotation results in Yolo format which is used for training. However, the choice of annotation tool can greatly influence the time taken for the annotation task, which is one of the main labour-intensive parts of training an object detection model.

CVAT offers different tools for semi-automatic annotation in video data. Two of these tools proved to be handy and helped in making the annotation process easier. The first is a basic track functionality which allows users to create tracks by manually annotating an object in one frame, with the rectangle being automatically interpolated on the next frames. This functionality is useful only if the object stays at almost same position in the subsequent frames. When the object changes its position or orientation, the rectangle is to be modified manually to ensure the box accurately covers the object. CVAT offers a concept called keyframes, to reduce the effort required in these manual corrections. They claim that it is only necessary to manually annotate frames in specific intervals (keyframes), and the tool will automatically interpolate the frames in between.[190]

This gives best results, if the movement of the object is in a single direction, and therefore, predictable. For annotating assembly data in a shop floor, this didn't prove to be very helpful. However, this feature was still useful as the object moves only slightly between consecutive frames in a video and adjusting the existing bounding boxes were much easier that drawing new bounding boxes each time. The second tool was an OpenCV

[190] Sekachev, B. et al. (2020).

assisted tracker offered by CVAT, which attempts to predict and adjust the object's movement across frames, thus automatically moving the bounding box to follow the object. While this prediction is also not accurate and still requires manual corrections, it was more efficient than the basic track feature in predicting the object position when it moves significantly between consecutive frames.

4.2.1.4 Data augmentation techniques

Data augmentation techniques are employed to make training samples more diverse thereby improving the model performance and generalization. A problem often encountered in object detection tasks is the overfitting of the model to the training data, meaning it learns the details of the training data too well which causes the model to perform poorly on unseen data. Augmentation introduces variations that prevent the model from memorizing specific patterns in the training images.[191]

This is achieved by applying various transformations to the initial dataset to artificially increase its diversity. Data transformation and data synthesis techniques are two key approaches used to accomplish this. Data transformation techniques modify the existing training data by applying transformations, whereas data synthesis is applied when there is a lack of training data, creating new training samples from scratch using synthetic methods. The most common data transformation techniques include geometric transformations such as flipping, rotation, translation, scaling, as well as photometric transformations that vary the brightness, contrast, and saturation of images. These transformations help ensure that the model is more robust to variations in orientation, scale and lighting.[192]

Additionally, multi-image combined augmentation techniques are often employed to further enhance the dataset by blending or combining multiple images in the original dataset to generate new, varied samples. Some of the popular methods that combine content from different images are CutMix, Mosaic, and Mixup. These methods significantly improve model performance by preventing the model from memorizing specific positions, backgrounds or scale of the objects in the images. Instead, they encourage the model to generalize better by learning to recognize objects in diverse contexts and conditions.[193]

Many of these augmentation techniques are integrated into the Yolov8 algorithm and are applied during training. Some techniques are enabled by default, while others can be

[191] Zeng, W. (2024).
[192] Mumuni, A., & Mumuni, F. (2022).
[193] Zeng, W. (2024).

customized by users based on their requirements. The extent of augmentation can be controlled by adjusting the relevant hyperparameters.[194]

In this work, since object detection was intended to be used in a controlled environment, additional augmentation techniques were not considered necessary. Therefore, no further augmentation was performed before training and, the default settings were used during training.

*4.2.1.5 **Training Yolov8 models***

After collecting the required data, and annotating the objects of interest in the images, the final step involves training an object detection model, allowing it to learn to detect the workers and products assembled in the learning factory. Using yolov8 for training a custom model is relatively straightforward. However, certain prerequisites need to be met for using yolov8 in a local environment. This is outlined in Table 4-3.

Table 4-3: Prerequisites for using YOLOv8[195]

Prerequisites for YOLOv8 installation	Version/Requirement
Python	3.8 or higher
PyTorch	1.7 or higher
CUDA Toolkit	Recommended for GPU acceleration
Python package manager, PIP	Installed
Ultralytics package and dependencies	Installed

Although Yolov8 supports training using CPUs, it is highly recommended to use a GPU for better model performance and reduced training times. The hardware and software specification used for this work are outlined in Table 4-4 and Table 4-5 , respectively.

[194] Jocher, G. et al. (2023).
[195] Ibidem

Table 4-4: Hardware Specifications Used

Hardware Component	Specification
Processor (CPU)	Intel Core i5-12450H
Graphics Processing Unit (GPU)	NVIDIA GeForce RTX 4050
RAM	16 GB
Storage	512 GB SSD

Table 4-5: Software Specifications Used

Software / Library	Version
Python	3.11.5
PIP	23.2.1
PyTorch	2.2.0
CUDA	12.1
cuDNN	8.8.1
Ultralytics (YOLOv8)	8.2.48

The model is trained using a set of annotated images that contain the objects of interest, along with their corresponding labels. During the training process, the model learns to identify patterns and features from the labelled data, allowing it to generalize and detect objects in new, unseen images. The model goes through multiple epochs, during which it processes the entire dataset in each epoch. After each epoch, its performance is evaluated on a validation dataset. The training continues until either the predefined number of epochs is completed, or an early stopping criterion is triggered.[196]

A YAML file is used to configure the dataset for training and define the object classes to be identified. It specifies the paths to the training and validation datasets, as well as labels

[196] Terven, J. et al. (2023).

for the annotated object classes. An example command to train a model and the YAML file configuration are shown in Figure 4-7 and Figure 4-8, respectively.

```
model = YOLO("yolov8m.yaml").load("yolov8m.pt")

results = model.train(data="data_product.yaml", epochs=100, device=0, patience=50)
```

Figure 4-7: Example of command used for training YOLOv8 model

```
path: D:/Workspace/Thesis/Data/product_dataset
train: images/train
val: images/val
test: images/test

# classes
names:
  0: Scooter
  1: Hand_truck
```

Figure 4-8: Example of YAML file configuration

Two different models were trained for detecting workers and products. To compare the performance across YOLOv8 model sizes, three separate models were trained for product detection using YOLOv8n, YOLOv8s, and YOLOv8m. These models were trained using pretrained weights, as recommended in the official YOLOv8 documentation. This approach leverages transfer learning, where the model benefits from prior training on large, diverse datasets. Although the specific target objects were not part of the pretrained dataset, the pretrained models had already learned general features that enables efficient fine-tuning and faster learning on the new dataset. The time taken for training model sizes of YOLOv8n, YOLOv8s, and YOLOv8m for 100 epochs are compared in Table 4-6.

Table 4-6: Comparison of training time and model sizes for YOLOv8 variants

Metric	YOLOv8n	YOLOv8s	YOLOv8m
Training time (hours) 100 epochs	1.331	2.119	23.383
Model size (mb)	6.3	22.5	52

YOLOv8n, the smallest model, completed training in 1.33 hours and has a model size of 6.3 MB, making it suitable for using with edge devices having limited hardware capabilities. YOLOv8s took 2.12 hours to train and has a size of 22.5 MB. YOLOv8m is

much larger than other two models and took 23.38 hours to train, making it very resource intense. It would require good hardware performance to train and use YOLOv8m models.

4.2.1.6 YOLOv8 model performance

The models were evaluated on a test dataset containing 307 images extracted from another assembly video with the same layout and background as the images used for training. The results show that all three models performed were able to detect both the products most of the time.

YOLOv8n is clearly the fastest, capable of processing 200 frames per second (FPS), making it ideal for real-time applications. However, it has slightly lower accuracy, especially in mAP@50:95, which measures performance across various IoU thresholds. The performance comparison of the models across various metrics is presented in Table 4-7.

YOLOv8s offers a balanced trade-off between accuracy and speed, achieving a higher mAP@50:95 (72%) while maintaining a reasonable inference speed of 94.3 FPS.

YOLOv8m delivers a slightly better accuracy (73.3% mAP@50:95), but it is much slower, processing images at 46.5 FPS. This makes it more suitable for applications where accuracy is prioritized over real-time performance.

Table 4-7: Comparison of YOLOv8n, YOLOv8s, and YOLOv8m model performance

Metric	YOLOv8n	YOLOv8s	YOLOv8m
Precision (P)	0.95	0.986	0.951
Recall (R)	0.942	0.924	0.939
mAP@50	0.978	0.976	0.984
mAP@50:95	0.683	0.72	0.733
Inference time (MS)	5.0	10.6	21.5
Frames per second (fps)	200	94.3	46.5

4.2.2 Data acquisition and processing

In this section, object detection is used to analyse the video footage of the assembly process to collect key information from the shop floor. The custom trained yolov8 model, discussed in the previous section, is trained to detect both the workers and the two products assembled in the learning factory at TU Graz.

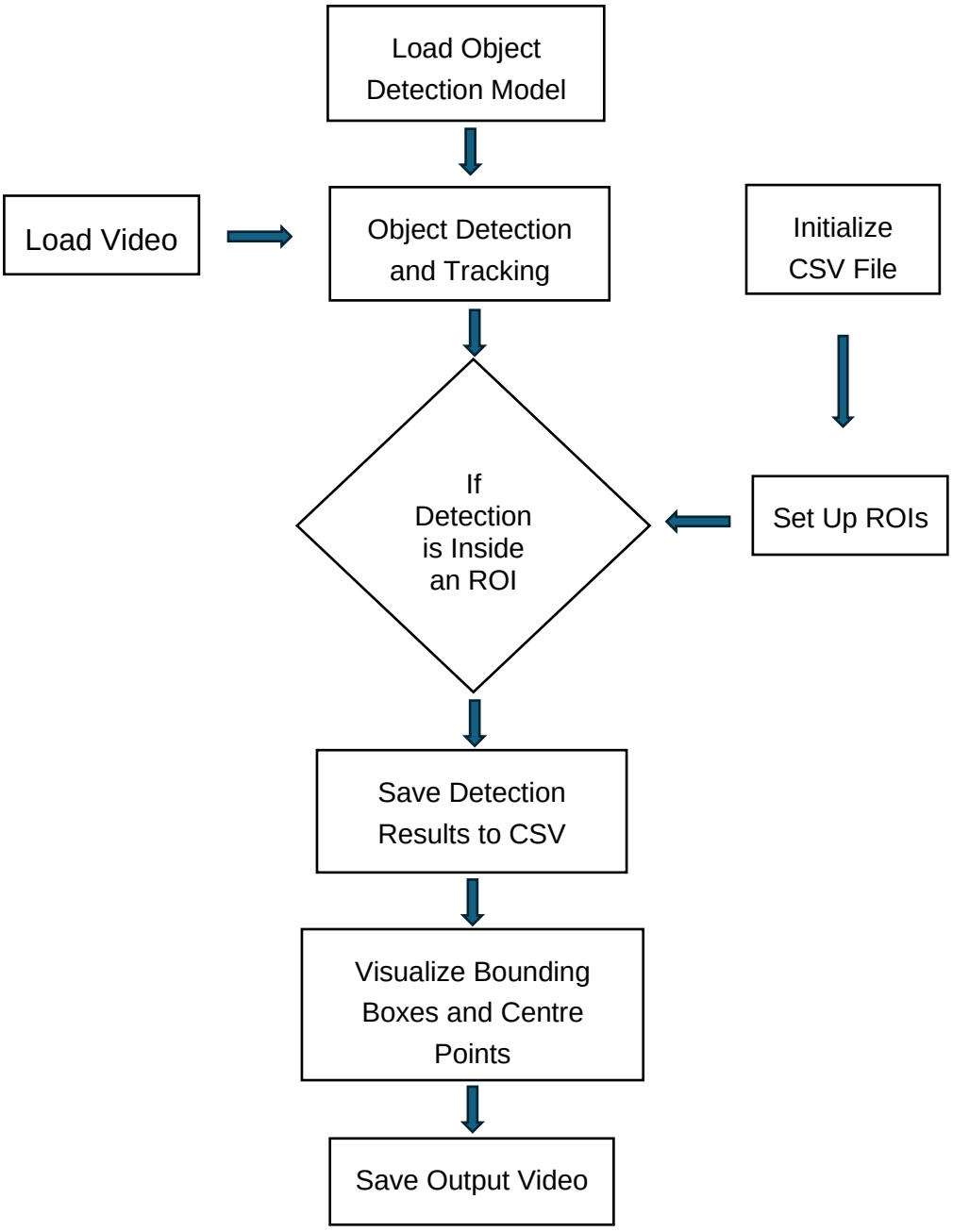


Figure 4-9:Workflow for data collection and processing using object detection

These models analyse each frame of the input video, and it returns a bounding box for each detected instance, along with the predicted class name. A python script is developed, utilizing OpenCV tools, to capture and transform this detection data into structured information, suitable for determining KPIs of the assembly process. The workflow for achieving this is defined in Figure 4-9.

4.2.2.1 Tracking Object Movement

Object detection algorithms detect and locates objects of interest from the input source and returns the output with bounding boxes, class names and confidence scores for each of the detected instances. The bounding box denotes the location of the detected objects in each frame, the class name denotes the predicted object class, and the confidence score denotes how confident the model is that the object belongs to the predicted class.[197]

But this does not differentiate between different units of product that belong to the same product class. Tracking functionality addresses this limitation by allowing the system to identify and differentiate between multiple instances of the same object class across consecutive frames. While object detection provides information about the presence and location of objects in each individual frame, the tracking functionality assigns a unique ID to each detected object, enabling the system to track them across multiple frames. This ensures that each object is consistently tracked as it moves through the assembly area.[198]

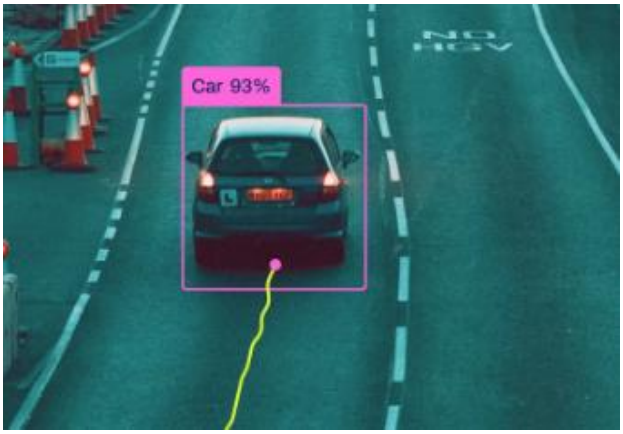


Figure 4-10: A general example of object tracking[199]

In this work, object tracking is implemented using the tracking functionality of YOLOv8, which supports BoT-SORT and ByteTrack tracking algorithms. These trackers can be

[197] Kaur, R., & Singh, S. (2023).
[198] Chen, W. et al. (2024).
[199] Jocher, G. et al. (2023).

enabled by passing the relevant YAML configuration file during inference.[200] In this work, the Bot-SORT tracker is used, which is the default tracker in YOLOv8. To apply the tracking functionality, the track() method is called, allowing various arguments to be passed based on the preferences, such as the object classes to detect, confidence threshold, and the tracker to be used. Listing 4-2 shows how the track () method is applied to detect and track objects.

```
results = model.track(
    frame, persist=True, device=0, tracker="botsort.yaml")
```

Listing 4-2: An example use of the track functionality

When the detection model processes the first frame, it detects objects, and the tracker assigns each detected object a unique track ID. As the video progresses to the next frame, the tracker attempts to match the objects detected in the new frame with those from the previous frames. The tracker does this by comparing the position of objects in the current frame with the position of the objects in the previous frame. If the tracker finds that an object in the current frame closely matches an object detected in the previous frame (based on its IoU, size, confidence threshold), it assigns the same track ID to that object, ensuring consistency across frames.[201]

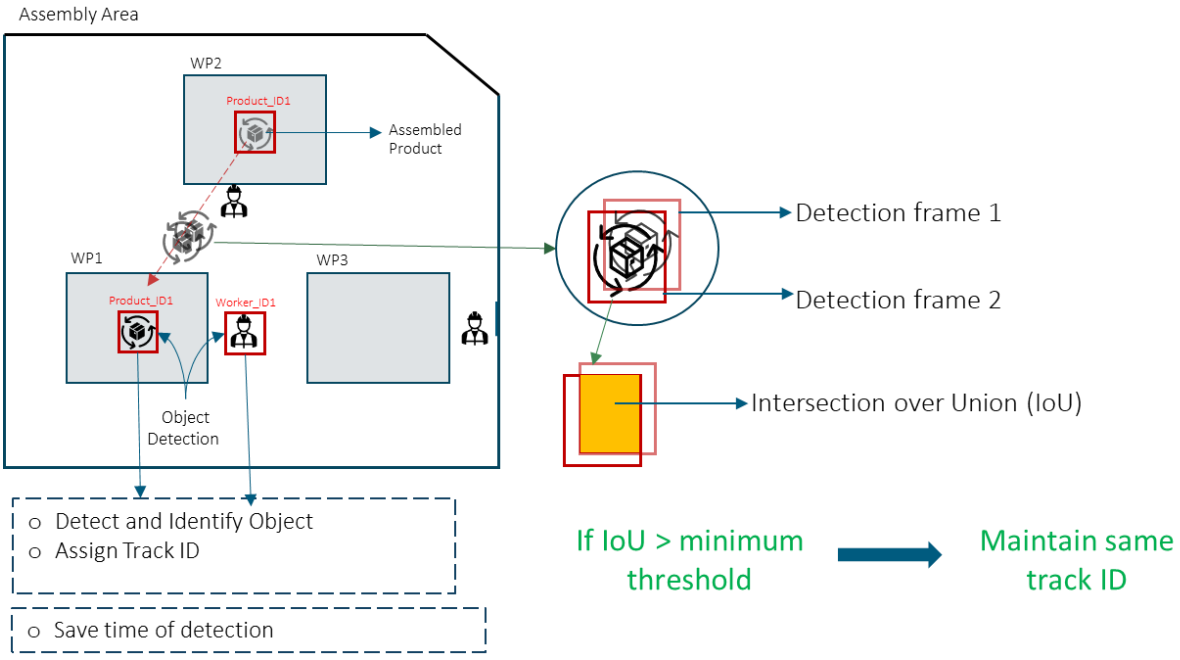


Figure 4-11: Basic representation of how the tracker maintains the track ID for a detected object

[200] Jocher, G. et al. (2023).
[201] Ibidem

The tracker continues to maintain the same ID until the object either leaves the frame or is no longer detected. If a new object appears in the frame, or the tracker can't match the current object to any of the objects in the previous frames, it assigns it a new track ID.

4.2.2.2 Region of Interest (ROI) and Data Logging

These outputs form the basis for collecting the required data from the shop floor. However, just knowing the coordinates at which the product is detected in the frame does not provide information about the assembly progress. The concept of region of interest (ROI) is used in this approach to understand the progress of the product through the assembly line. This is used to define the layout of different workplaces in the assembly area and to structure the output data accordingly. To identify the coordinates of the vertices of the required ROIs, a function was included in the script that leverages OpenCV's mouse event handling functionality (Bradski, 2000). By moving the mouse over the video frame, the coordinates of the vertices of the desired ROI can be captured and printed to the console. These coordinates are subsequently used to define the boundaries of each ROI. This method ensures that the ROIs can be altered easily whenever a change occurs in the layout of the assembly process. The function used is shown in the Listing 4-3 below.

```
def find_coordinates(event, x, y, flags, param):
    if event == cv2.EVENT_MOUSEMOVE:
        mouse_position = [x, y]
        print(mouse_position)
```

Listing 4-3: Function used to find the coordinates of the ROIs

Different ROIs are defined as polygons within the video frame by passing the coordinates corresponding to their vertices as tuples. An example snippet is shown in Listing 4-4.

```
WP1 = [(459, 378), (286, 376), (285, 540), (458, 542)]
WP2 = [(503, 40), (347, 36), (345, 160), (493, 159)]
WP3 = [(258, 197), (82, 196), (79, 320), (248, 321)]
```

Listing 4-4: Defining polygon coordinates for workplaces

A detection logic is implemented to determine whether a detected object's centre falls within any of the predefined polygons. When an object is detected by the Yolov8 model, an if-statement checks whether the detection is inside any of the defined ROIs. This is achieved using cv2.pointPolygonTest function (Listing 4-5), which checks if the detected object's centre falls inside, on or outside the boundary of any of the predefined ROIs.

```
if (cv2.pointPolygonTest(np.array(WP1, np.int32),
    ((center_x, center_y)), False)>= 0):
    save_data_to_WP1(product_class, obj_id, video_timestamp)
```

Listing 4-5: Logic to check if an object's center is within an ROI

If the detection is found to be within any of the defined ROIs, the detection details are logged to a CSV file in real-time along with the detected object's class (e.g. 'Scooter, 'Hand Truck'), its unique track ID assigned by the Yolov8 tracker, and the timestamp of detection. The object class and track ID are recorded in the column corresponding to the ROI where the detection occurred, while the timestamp is saved in a separate column for each detected instance. For real-time applications, the system timestamp can be used to log the exact time of detection. When detection occurs in a recorded video of the assembly process, the video timestamp is derived by dividing the frame number at which the detection occurs by the frame rate of the video. A sample snippet of the applied logic is shown in Listing 4-6 and Listing 4-7 below. Figure 4-12 shows the structure in which the data is saved to the csv file.

```
def save_data_to_WP1(product_class, obj_id, video_timestamp):
    with open(csv_file, mode="a", newline="") as file:
        writer = csv.writer(file, delimiter=";")
        writer.writerow(
            [video_timestamp,
                f"{product_class}-id{obj_id}",
                "",
                "",])
```

Listing 4-6: Logic applied for logging detections within the ROIs to a CSV file

```
def get_video_time(cap, frame_number):
    fps = cap.get(cv2.CAP_PROP_FPS)                # gets the fps of the video
    elapsed_time_sec = frame_number / fps
    video_timestamp = str(
        timedelta(seconds=int(elapsed_time_sec))) # converts to hh:mm:ss format
    return video_timestamp
```

Listing 4-7: Logic for determining the video timestamp based on the frame number

A	B	C	D
Timestamp	WP1	WP2	WP3

Figure 4-12: Structure of saving detection results

In addition to the time data, the coordinates at which each object is detected in each frame is also saved to a csv file. This data is used for drawing spaghetti diagrams to visualize the movement of products and workers through the assembly area.

By tracking worker movement, inefficiencies such as excessive walking and unnecessary waiting can be identified. Unnecessary movement is a waste according to lean principle and optimizing it helps to increase efficiency. Ensuring worker safety is another key area in which this data is crucial. Presence of workers in hazardous or prohibited areas can be identified and corrective actions can be initialized.

Similarly, visualizing product movement helps to map the path taken for the transfer of products. While the time data is sufficient to determine KPIs, monitoring the path helps to uncover inefficiencies that goes unnoticed otherwise.

In addition to saving detection information, visual indicators are added to the output video by drawing bounding box around each detected object, and a circle at the centre of the object. These visual indicators improve the clarity and ease of understanding of the detection results during the review of the output video.

4.2.3 KPI determination and Value Stream Map Visualization

The objective of this work is to evaluate the potential of using object detection methods to enable digital data collection from shop floors for value stream mapping. The previous sections explained the procedure for training object detection models and for extracting data from the shop floor using the detection results. The final step in the proposed approach is to determine various KPIs of the assembly process and visualize them in a VSM. This section outlines the process of calculating KPIs and describes how they are visualized to provide actionable insights into the assembly process.

The output from the python script contains location and time data for all products assembled on the shop floor. This data can be further processed using any of the data processing tools and visualized according to the requirement. For this work, Microsoft excel was chosen due to its simplicity and ease of use. The location and time data from both worker and product detections are exported into an excel file for further calculation and visualization in a value stream map to gain actionable insights.

4.2.3.1 KPI Determination

Since all the ROIs in the assembly area are defined, and the presence of products in these areas are already known, this data can be used to calculate various KPIs of the assembly process. To calculate the KPIs, the entry and exit times of each individual product at different workplaces or waiting areas are determined. This is done by fetching the first and last detection times for each individual product, identified by its track ID, in each of the ROIs.

A combination of conditional and array functions in excel are used to achieve this. First, all the unique track IDs detected across the workplaces are identified using the UNIQUE and TOCOL functions in excel (Listing 4-8). This formula collects all the track IDs found within the detection data, ensuring that each ID appears only once.

```
=UNIQUE(TOCOL(B3:D1000; 1))
```

Listing 4-8: Formula to collect all unique track IDs across workplaces

To find the first occurrence of each product in a specific ROI, the MATCH function is used in combination with the INDEX function. The formula searches for the first appearance of the product ID in a specific workplace (ROI) and retrieves the corresponding timestamp. An example of the use of the formula is shown in Listing 4-9. The value returned by this formula denotes the time of entry of that specific product into the specified ROI.

```
=IFERROR(INDEX(A:A; MATCH(H3; B:B; 0)); "")
```

Listing 4-9: Formula to retrieve the timestamp (from column A) of the first appearance of a product (mentioned in cell H3) in a specific workplace (column B)

To find the last occurrence of each product in a specific ROI, a combination of MAX and IF functions is used to locate the last row where the product ID appears. The

corresponding timestamp is then retrieved using the INDEX function. The value returned by this formula denotes the time of exit of that specific product from the specified ROI.

```
=IFERROR(IF(MAX(IF(B:B=H3; ROW(B:B)))>0; INDEX(A:A;
MAX(IF(B:B=H3; ROW(B:B)))); ""); "")
```

Listing 4-10: Formula to retrieve the timestamp (from column A) of the last appearance of a product (mentioned in cell H3) in a specific workplace (column B)

By comparing the entry and exit times of each product through the workplaces, the time spend by the product within workplaces and in transit can be easily identified. This is the foundation for calculating the KPIs relevant for VSM process.

Identifying unexpected deviations during assembly process

To ensure that the calculated KPIs are valid, it is essential to identify any unexpected behaviour and to validate that the material flow for each product follows the defined workflow. If a product deviates from the planned workflow due to mistakes from the workers or any other unplanned event, it can lead to inaccurate KPI measurement and misinterpretation of the actual scenario. Therefore, a checking criterion is implemented during processing, to detect any unexpected behaviour.

This is done by analysing the sequence in which products enter different workplaces. For this, the time of entry of each product at each workplace is listed. The entry times are then arranged in ascending order using the SMALL function in excel. Once the entry times are arranged, the sorted times are matched with the corresponding workstations using INDEX and MATCH functions. The MATCH function locates the position of the sorted timestamp within the original detection data, and INDEX retrieves the corresponding workplace name from the header row. This visualizes the actual flow of the product through the assembly area.

```
=INDEX($F$8:$H$8;MATCH(F14;$F$9:$H$9;0))
```

Listing 4-11: Example formula to match the workplace name to product's sorted timestamp using the original detection data

Detection data					
	WP1	WP2	WP3		
S1	00:00:04	00:02:59	00:01:43		
H1	00:01:56	00:04:13	00:07:10		
Observed Workflow					
S1	WP1	WP3	WP2		
	00:00:04	00:01:43	00:02:59		
H1	WP1	WP2	WP3		
	00:01:56	00:04:13	00:07:10		
Defined Workflow					
Scooter	WP1 Handle assembly	→	WP3 R - wheel assembly	→	WP2 F - wheel assembly
Handtruck	WP1 Handle-bar assembly	→	WP2 Plate assembly	→	WP3 Wheel assembly

Figure 4-13: Example visualization comparing observed workflow with the defined workflow

After obtaining the sequence of workflow based on entry times, the observed workflow is compared with the predefined workflow. If any discrepancies are found, it is flagged as a deviation from the standard process. This ensures that any unexpected events on the shop floor, which could impact the accuracy of the calculated KPIs are identified, preventing any misinterpretation of the data.

KPI Calculation

Once the entry and exit times are determined, the KPIs are calculated as defined in Table 4-8

Table 4-8: Calculation of Key Performance Indicators

KPI	Formula
Cycle time for process (i)	Exit time from WP (i) - Entry time in WP (i)
Throughput time	Exit time from WP (n) - Entry time in WP (1)
Waiting time for process (i)	Exit time from WA (i) – Entry time in WA (i)
Transport time from WP (i) to WP (i+1)	[Entry time in WP (i+1) - Exit time from WP (i)] – Waiting time
Number of workers	Count of Worker IDs in WP (i)

Worker idle time	Worker detection time in WP(i) – Total processing time in WP (i)
Worker utilization	Total processing time in WP (i) / Worker detection time in WP(i)
Workplace utilization	Total processing time in WP (i) / Total available time
Value adding time for product (i)	SUM of cycle times for product (i)
Non-value adding time for product (i)	SUM of waiting times for product (i) + SUM of transport times for product (i)

Explanation of terms used in the calculation formulas:

WP : Workplace

WA : Waiting Area

Entry time in WP (i) : The time when the product or worker enters the workplace (i).

Exit time from WP (i) : The time when the product or worker leaves the workplace (i).

Entry time in WP (1) : The time when the product or worker first enters the initial workplace.

Exit time from WP (n) : The time when the product or worker leaves the final workplace (n).

Entry time in WA (i) : The time when the product or worker enters the waiting area (i).

Exit time from WA (i) : The time when the product or worker leaves waiting area (i).

Worker detection time in WP (i) : It is the total time a worker is detected in workplace (i). It is calculated by summing the total number of times the worker's ID is recorded in that workplace. Since exactly one entry is made per second, this directly corresponds to the total time the worker spends in the workplace.

Total processing time in WP (i): This represents the total time during which value adding activities are performed in workplace (i). It is the sum of the cycle times of all products assembled in that specific workplace over a defined period.

Total available time: This represents the total time during which the workplace is available for production activities. In the context of testing, the time period considered for evaluation is regarded as the total available time.

4.2.3.2 Visualization

This section presents the visualization techniques used to analyse and interpret data collected from the shop floor using object detection. The thesis aims to evaluate the potential of object detection for collecting data relevant for digital value stream mapping. To achieve this, the collected data is first visualized in a digital VSM. This gives a holistic view of the production process. Subsequently, the collected data is used for further visualizations to provide a more detailed view on individual KPIs. Different tools such as Gantt charts and pie charts are used to get in-depth insights into specific processes.

Digital Value Stream Map

Once the KPIs are calculated, the final step is to visualize the production process through a digital value stream map (VSM). The approach developed in this thesis focuses on creating a current state map, which is critical for capturing fact-based data directly from the shop floor.

The VSM is created and visualized in excel, mirroring the structure of a traditional VSM. However, unlike static paper-based VSMs, the digital VSM, created using data collected over a period of time, reflects the actual situation on the shop floor. This dynamic representation allows for more insightful analysis of the production flow and process inefficiencies.

Since both data processing and visualization are performed digitally, the map can be easily updated with the most recent data. Furthermore, object detection's real-time detection capability means that if real-time processing is realized, it would be possible to collect and visualize real-time data directly from shop floors. This offers the potential for continuous monitoring and instant decision-making, further enhancing process optimization.

Further Process Visualization

Beyond the creation of a digital VSM, additional visualizations can provide deeper insights into the assembly process, highlighting key areas for improvement. The dynamic data collected from the shop floor can be transformed into detailed charts and graphs, making information more accessible and visible. Visualizing factors such as total processing time, worker and workstation utilization, and the movement of both workers and products, allows for easier identification of inefficiencies, and bottlenecks. These enhanced visualizations make it possible to extract more value from the VSM, as they

provide a clearer and more comprehensive understanding of the assembly process. Therefore, digital data availability also enables the creation of a more digital data makes it possible to create more visualizations that enhance the use of value stream maps.

4.2.4 Challenges and Solutions

This study explored the potential of object detection and tracking for collecting data from shop floors. Object tracking was employed to ensure that individual unit of an object is recognized consistently, even when multiple instances of the same object are present. This is achieved by assigning a unique track id to each detected object by the tracker. During each frame, the tracker tries to match the detected objects to detections in the previous frame. This is done by comparing the object's position and movement between frames using metrics like Intersection over Union (IoU), which measure the overall between bounding boxes in consecutive frames.

However, challenges arise when the object is missed by the detection algorithm for multiple consecutive frames. In such cases, the object may move significantly between detections causing its new bounding box to have a lower IoU score with the last detected bounding box. As a result, the tracker may not be able to confidently match the object to its previous detection and may assign a new track ID considering it as a newly introduced object. This reassignment of track IDs due to continuous false negatives pose a challenge for collecting accurate data.

When being processed within a workplace, products remain generally visible, so missing detections are less frequent, assuming the object detection model has sufficient accuracy. However, during transit between workplaces, occlusions or missed detections can occur due to sudden handling of the product by workers or because the product may be briefly covered by a worker. These situations can cause the tracker to assign a new track ID when the product reappears, which complicated the accurate tracking of products throughout the assembly process.

To address this issue, a VBA script was implemented in excel to match and correct track IDs across different workplaces. The script leverages the predefined workflow of the assembly process, where the expected sequence of transitions between workplaces is already known. The script uses a simple logic for matching IDs. When a track ID disappears from a workplace (represented in a specific excel column), the code checks for a new detection in the next expected workplace. If a new detection occurs within a defined time period, it is assumed to be the same unit moving between workplaces according to the defined workflow. This approach ensures that, even if a track ID is reassigned during transit, it can be corrected during post processing, maintaining consistent identification of each product.

4.3 Testing

A method was developed in the previous section to collect data from shop floors using object detection and to generate digital value stream maps (VSMs) by systematically processing the data using Python and Excel. This chapter presents the testing of this method, emphasizing the potential of object detection to collect dynamic data form the shop floor to support digital value stream mapping. The method was tested at the learning factory at TU Graz.

4.3.1 Learning Factory at TU Graz

The learning factory of TU Graz, referred to as the LEAD Factory, is being operated by the Institute of Innovation and Industrial Management (IIM) since 2014. The name of the factory, LEAD, reflects its focus on Lean, Energy efficiency, Agility, and Digitization, and provides academic education, company training, and hands-on research opportunities.[202]

As a miniature industrial manufacturing site, it is equipped with industry standard tools and continuously integrates new technologies to stay aligned with the latest advancements. Its digital infrastructure includes RFID-based process control, digital work instructions, smart meters for energy monitoring, augmented reality glasses, human and process simulation, and RTLS based workflow tracking. A digital shop floor management board (SFMB) is also used to visualize the process data.[203]

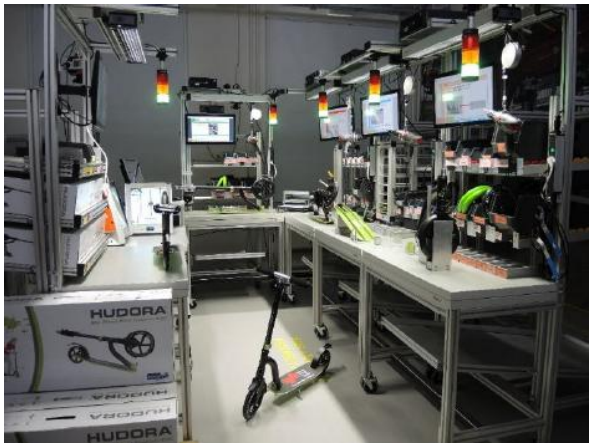
a)  b) 

Figure 4-14: LEAD Factory - (a) Optimized digital state[204], (b) Sub-optimal state

The factory operates in three different configurations, representing a sub-optimal initial state, an advanced lean state, and an optimized digital state. This setup enables

[202] Rantschl, M. et al. (2023).
[203] Ibidem
[204] Ibidem

participants to gain hands-on experience in transforming inefficient production processes into lean, digitized systems, with efficiency improving as more lean and digital tools are implemented. The LEAD factory originally produced a self-branded TU Graz scooter and recently introduced a hand truck to its product line. This addition increased operational complexity, mimicking the real-world challenge of managing multi-product lines. Such complexity enhances the learning experience by providing trainees with hands-on practice in managing the varied demands of modern manufacturing environments.[205]

In addition to offering practice-oriented training, the facility functions as a research platform for modern manufacturing technologies and processes. It simulates the key complexities of real-world manufacturing environments, making it an ideal testing environment for state-of-the-art innovations. This makes the LEAD Factory particularly suited for testing the object detection-based method for digital value stream mapping developed in this study.

4.3.2 Testing approach

For testing the method, assembly of products in the LEAD Factory was monitored using object detection to collect the location and time data of the products assembled in the shop floor. A new assembly configuration was introduced for testing, in which products were fully assembled across three workplaces, supported by three workers.

a)  b) 

Figure 4-15: Products Assembled - a) Scooter, b) Hand Truck

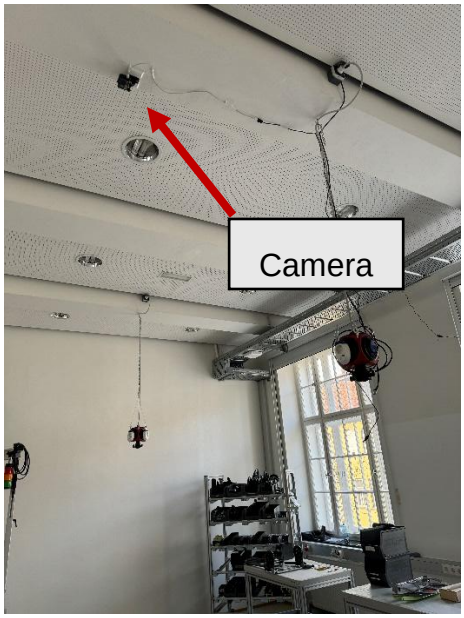
This setup involved the simultaneous assembly of both scooters and hand trucks, aiming to assess the capability of the object detection model to accurately detect and classify multiple products simultaneously throughout the assembly process.

[205] Rantschl, M. et al. (2023)

Both products moved through the same workplaces and were assembled by the same three workers. Each worker was assigned a specific role during the assembly process, with Worker 1 responsible for workplace 1 (WP1), worker 2 for workplace 2 (WP2), and worker 3 for workplace 3 (WP3). After completing each task, the workers moved either to the next workplace or to the storage area to transfer or store the product. To introduce additional complexity, two different workflows were defined for the products. The hand truck followed a sequential workflow from WP1 to WP2 and then to WP3, whereas the scooter followed a different path, moving from WP1 to WP3, and then to WP2. This arrangement aimed to test the method's ability to manage varying product flows.

4.3.3 Experimental Procedure

The assembly process was recorded using a ceiling-mounted camera, which provided a comprehensive view of the entire assembly area. This placement was chosen to ensure a clear view of the movement of both products and workers, with minimal chances for occlusions. This also helps to minimize the privacy concerns associated with recording the assembly processes as the face of the operators are not directly visible. A GoPro Hero7 was used for recording, capturing the footage at a resolution of 1080p and a frame rate of 30 frames per second. The recording aimed to capture the entire assembly process of both the products, ensuring that the complete movements of products and workers through the assembly area are documented. The recording duration of approximately 9 minutes provided enough data for the analysis. The camera installation and the field of view of the camera is shown in Figure 4-16.

a) 

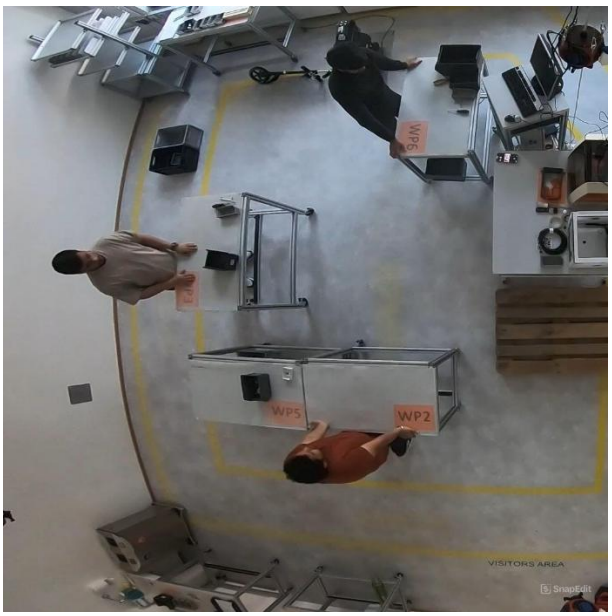
b) 

Figure 4-16: Camera setup - a) installation, b) field of view of the camera

4.3.3.1 Data Extraction from Video

The recorded video was then inferred using the custom trained Yolov8 model to detect and track both products and workers. Since the appearance of workers remain identical throughout the assembly process, the normal approach for tracking objects was followed. However, the appearance of the products gets altered throughout the assembly as parts gets assembled together. To ensure continuous detection of products within and across workplaces, the base plate of the scooter and the handle of the hand truck were chosen as the base parts to be tracked, as described in the methodology. The yolov8 model detected these parts continuously to track the assembly progress.

a)  b) 

Figure 4-17: Base part to be tracked for a) Scooter and b) Hand Truck

The model performed well in detecting and classifying the products across the assembly area. The tracker accurately tracked the products by maintaining a unique id for each of the detected products. This worked particularly well inside each of the workplaces, where movement or chances of occlusions were limited. However, during transit between workstations, the product was not detected in some frames due to occlusion or fast handling by the workers, resulting in the product getting reassigned with new ids. This issue was resolved during post-processing by matching the IDs based on the planned workflow using VBA code, as outlined in the methodology section.

The python script, described in the methodology section was used to systematically save the time and location information of the assembled products and workers. The csv output generated by the script included the timestamp at which each unit of product was detected in each workplace. A sample of this output is show in the Figure 4-18. Further data processing was done in excel.

Timestamp	WP1	WP2	WP3	WP1	WP2	WP3
		Product data			Worker data	
0:00:00			Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:01			Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:02			Scooter-id1		Worker-id3	Worker-id1
0:00:03			Scooter-id1		Worker-id3	Worker-id1
0:00:04			Scooter-id1		Worker-id3	Worker-id1
0:00:05		Scooter-id3	Scooter-id1		Worker-id3	Worker-id1
0:00:06		Scooter-id3	Scooter-id1		Worker-id3	Worker-id1
0:00:07		Scooter-id3	Scooter-id1		Worker-id3	Worker-id1
0:00:08		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:09		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:10		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:11		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:12		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:13		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:14		Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:15	Scooter-id4	Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:16	Scooter-id4	Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:17	Scooter-id4	Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1
0:00:18	Scooter-id4	Scooter-id3	Scooter-id1	Worker-id4	Worker-id3	Worker-id1

Figure 4-18: Output displaying detection times of products and workers across the three workplaces

The entry and exit times of the products at different workplaces were determined by fetching the first and last occurrence of each product ID in the respective workplaces.

Ids	Product	WP1		WP2		WP3	
		Entry time	Exit time	Entry time	Exit time	Entry time	Exit time
Scooter-id1	Scooter1			0:01:18	0:02:14	0:00:00	0:01:15
Scooter-id3	Scooter2			0:00:05	0:00:57		
Scooter-id4	Scooter3	0:00:15	0:01:39	0:03:00	0:04:08	0:01:43	0:02:56
Hand_Truck-id5	Hand_truck1	0:01:58	0:04:09	0:04:14	0:07:08	0:07:11	0:08:33

Figure 4-19: Entry and exit times of products in different workplaces

The time spent by each product in each workplace, as well as the time spent in transit between workplaces, were then calculated by comparing the entry and exit times. This information enabled the straightforward calculation of various KPIs, such as cycle times, waiting times, transport times, throughput times, total value adding time, and non-value adding time. Additionally, by analysing the worker detection data alongside the product detection data, metrics such as number of workers in each workplace, worker idle time, and worker utilization were calculated.

4.3.4 Test Results

This section presents the test results from using object detection to collect value stream relevant data from the shop floor. The KPIs calculated using this data are visualised in a value stream map created in excel. Additional visualizations such as, Gantt charts and pie charts, provide a comprehensive overview of the assembly process, showing the capability of object detection to effectively capture dynamic data from the shop floor.

4.3.4.1 Object Detection Enabled Digital Value Stream Map

This data was then visualized in a digital value stream map using Excel, providing a clear representation of the assembly process. Since both products shared the same workplaces and workers, their data was combined into a single current stream map. This approach provides a holistic view of the entire production process, enabling a clear understanding of how workstations and workers are utilized across the production of both products. It also facilitates the comparison of the performance of the products and helps identify bottlenecks, allowing for the optimization of the value stream.

The map visualizes the flow of the products from the start to the end of the assembly. The average processing time, number of assembled units, and worker idle time at each workplace were displayed separately for each product type over the analysed period. Additionally, the map shows the total number of assembled units and the utilization rate for each individual workplace, accounting for both products. It also highlights the average value adding and non-value adding times for each product type, as well as the percentage of time spent on value-adding activities. The created digital value stream map is presented in Figure 4-20.

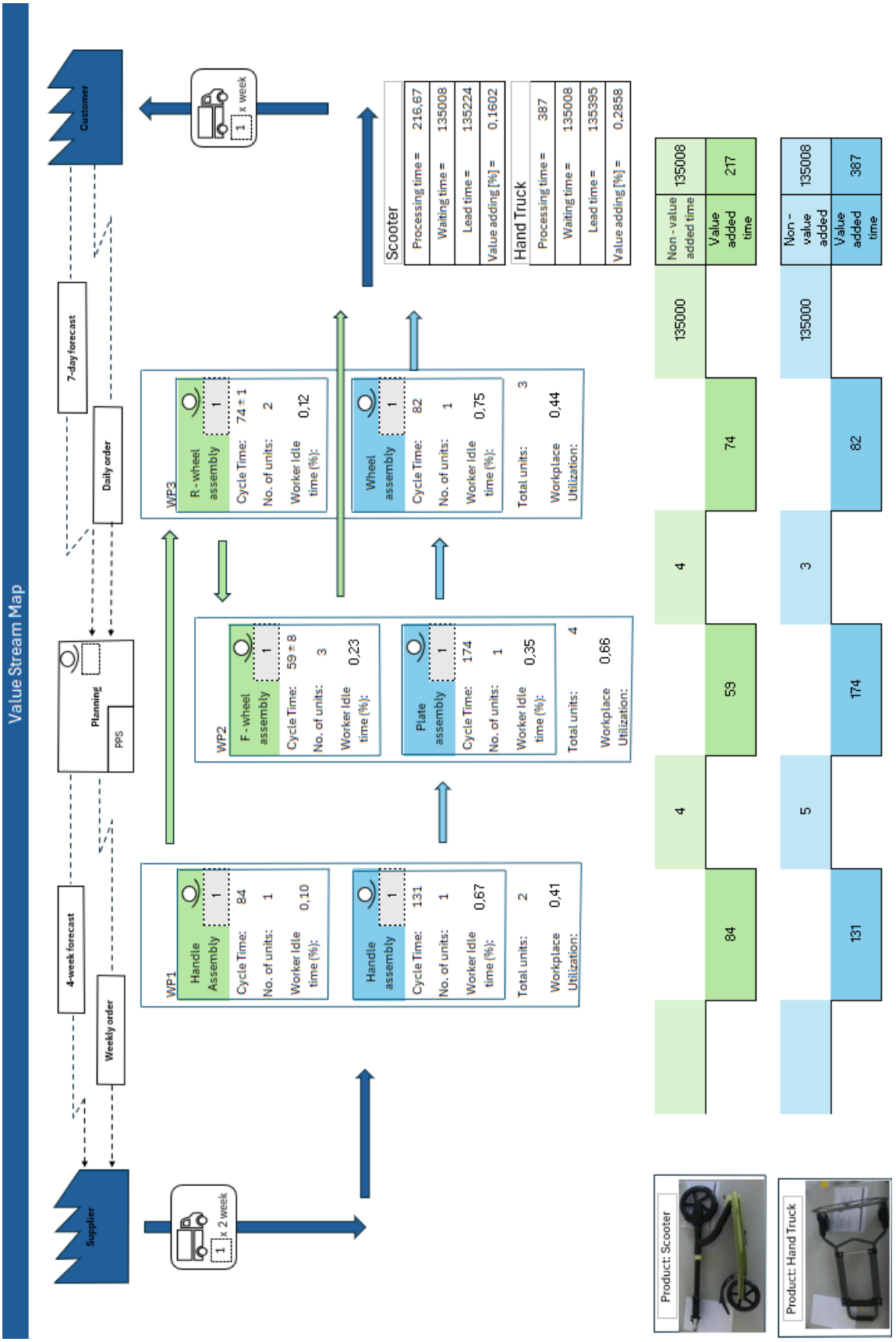


Figure 4-20:Digital value stream map based on the data collected using object detection

4.3.4.2 Detailed Process Visualization Based on Collected Data

In addition to the digital VSM, several Gantt charts and pie charts were created using the data collected using object detection. These visualizations provide more insights into the assembly process.

The entire process flow for both products was visualized using two Gantt charts. These charts, based on the data collected through object detection, displays the flow of each product from the start to the end of the assembly process. Each block in the Gantt charts represents the time a detected product spent in a particular region. Periods during which a product was not detected in any of the workplaces were also visualized, representing the transit time of the product between workplaces.

To ease the comparison of the time spent by each product at each stage of the assembly, the average cycle times of both the products were also visualized using a Gantt chart. Such visualizations make inefficiencies in the current state more visible, encouraging efforts to optimize the assembly process and improve line balancing. The Gantt charts are presented in Figure 4-21 and Figure 4-22.

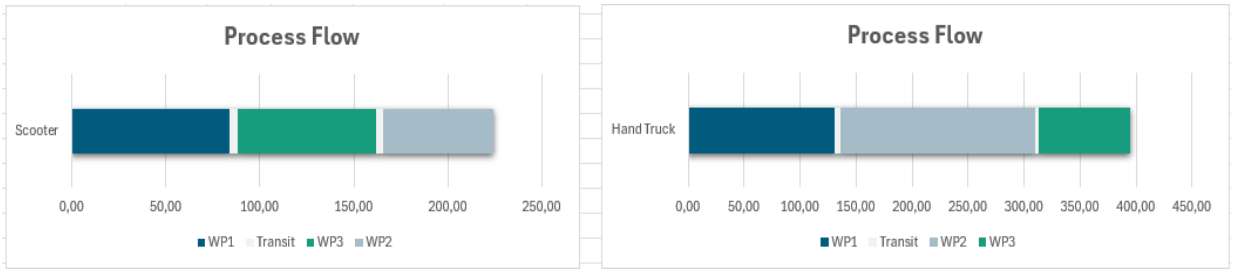


Figure 4-21: Process flow of the products

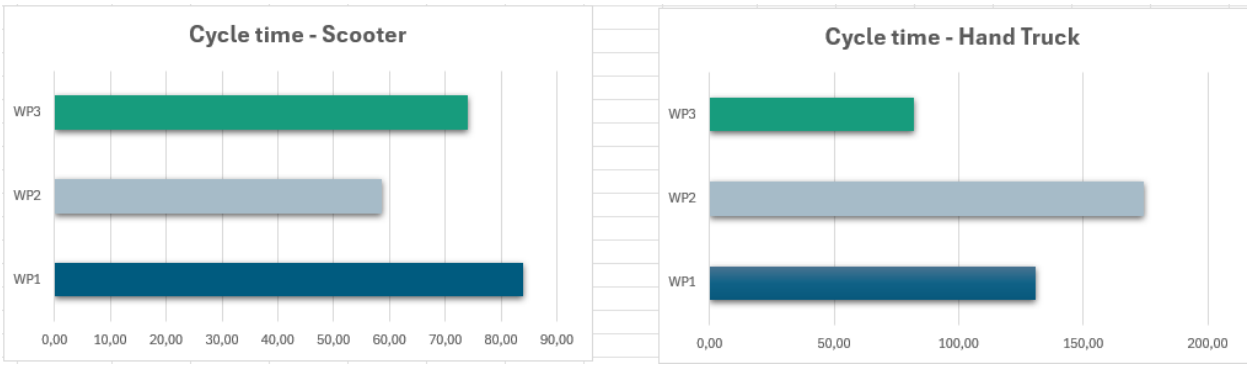


Figure 4-22: Cycle times comparison

Similarly, the average throughput times for the products were visualized (Figure 4-23), enabling a comparison of the overall production durations. The pie charts in Figure 4-24 illustrate the distribution of throughput times across different assembly stages for both the scooters and hand trucks.

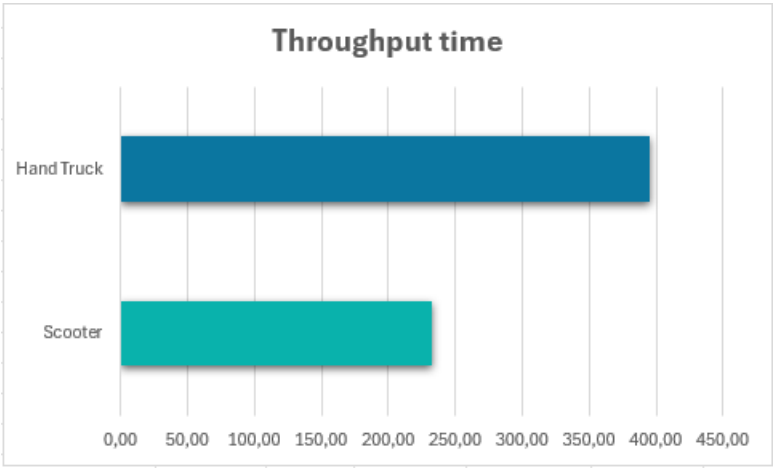


Figure 4-23: Throughput time for each product

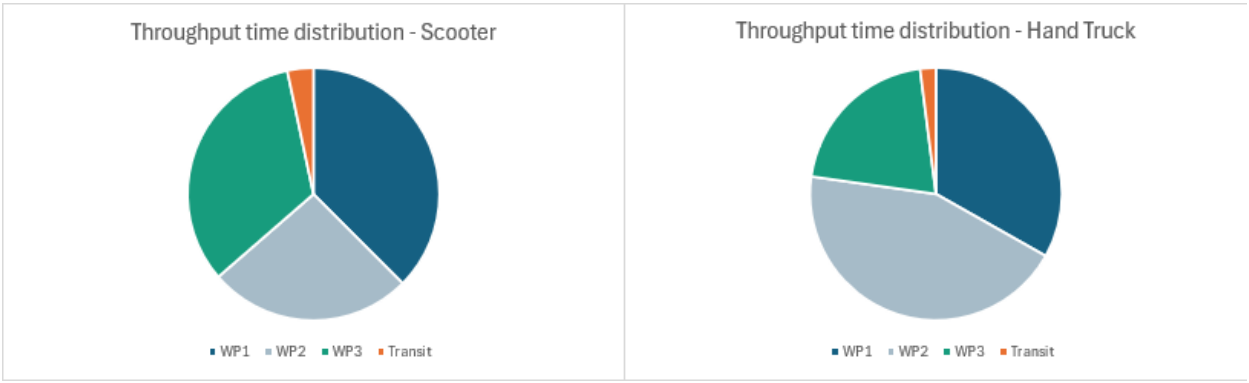


Figure 4-24: Distribution of throughput time across assembly stages

Two sets of pie charts were generated to visualize the utilization of both the workplaces and the workers involved in the assembly process. The first set, displayed in Figure 4-25, illustrates the workplace utilization by showing the proportion of value-adding time versus idle time for each workplace. The second set, shown in Figure 4-26, visualizes worker utilization by breaking down the worker's total time into working time, transport time, and idle time. These visualizations are particularly useful for assessing how effectively labour resources are deployed, helping to identify periods of underutilization or overburdening of workers.

Figure 4-25: Workplace Utilization

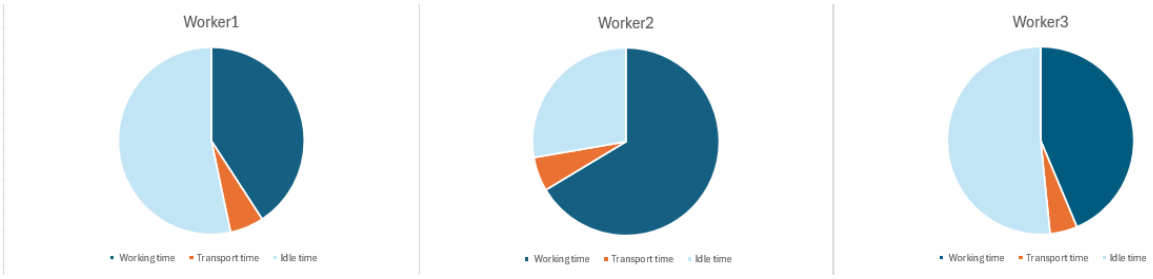


Figure 4-26: Worker Utilization

4.3.5 Summary

The testing outcomes confirmed the effectiveness of the object detection-based method in supporting digital value stream mapping. By tracking the assembly of scooters and hand trucks in a multi-product environment over a defined period, the method effectively captured dynamic data and managed varying workflows. The entire data was collected without the need for attaching any physical sensors to products or workers. This non-intrusive approach ensured that worker's tasks were not disrupted, and no additional workload was imposed on them.

5 Result

This chapter presents the results of the object detection and tracking experiments conducted to evaluate the feasibility of using object detection for digital data collection in value stream mapping. The results outline both the performance of the custom-trained object detection models in detecting and tracking objects in a dynamic assembly environment as well as the results from the testing, depicting the potential of object detection as a tool to enable digital data collection for value stream mapping.

5.1 Object Detection Model Performance

The object detection models were trained using a custom dataset consisting of images from the shop floor, with annotations for workers, scooters, and hand trucks. The training process was aimed at achieving maximum performance in the standardized environment where the method was tested.

5.1.1 Training duration and Resource Consumption

The training of the YOLOv8 models were conducted using an intel i5 processor with an NVIDEA GeForce RTX 4050 6GB GPU. The training data of YOLOv8s models, both for worker detection and product detection, are shown in Table 5-1.

Table 5-1: Training results

Metric	Worker	Products
Training time (hours)	0.895	2.119
Dataset size (number of images)	1305	4494
Epochs	100	100
Model size (mb)	21.98	22.5

5.1.2 Model Performance

The performance of the trained YOLOv8 models were evaluated using key metric, including precision, recall, mAP, and speed (fps). The models demonstrated strong performance on the test dataset, with maximum precision values of 0.975 for scooters and 0.996 for hand trucks. The results are shown in Table 5-2. The model was able to

accurately identify and classify these objects during a new assembly process, under the same environment. The choice of YOLOv8 allowed for a balance between detection speed and accuracy, making it suitable for real-time data collection on the shop floor.

Table 5-2: Performance results of YOLOv8s models in detecting specific classes

Metric	Worker	Scooter	Hand truck
Precision (P)	0.989	0.975	0.996
Recall (R)	0.995	0.998	0.85
mAP@50	0.994	0.995	0.957
mAP@50:95	0.847	0.656	0.783
Inference time per image (MS)	10.2	10.6	10.6
Frames per second (FPS)	98.04	94.3	94.3

5.2 Results from testing at the learning factory

The object detection method was tested at the learning factory of TU Graz, which is designed to reflect the real-world complexities of manufacturing environments. During the experiment, the YOLOv8 model was able to consistently detect and track objects throughout the assembly area.

The detection accuracy remained high even when different product classes were assembled simultaneously. However, occasional issues with object tracking were observed during periods of brief occlusions or false negatives, resulting in object IDs being incorrectly reassigned. However, such reassignment happened during transmit between workplaces and were corrected during post processing using a VBA code, leveraging the knowledge of the predefined workflow.

Object detection enabled the collection of a comprehensive set of data relevant for value stream mapping. This data was processed and visualized in a VSM using excel. The diverse range of data collected allowed to plot multiple Gantt charts and pie charts, providing greater visibility into the current state of the assembly process.

5.2.1 Accuracy Evaluation of the Collected Data

To evaluate the accuracy of the data collected using object detection, the determined KPIs were compared with the values derived from manual video analysis. The tables Table 5-3 and Table 5-4 below provide a comparison between the object detection data and manually analysed data for scooter and hand truck assembly.

Table 5-3: Comparison of automatically collected and manually collected data for scooter assembly

KPI	Automatic collection (seconds)	Manual collection (seconds)	% difference
Processing time (P1)	84	88	4,54
Transit + Waiting time (P1 – P2)	4	5	20,00
Processing time (P2)	73	73	0,00
Transit + Waiting time (P2 – P3)	4	7	42,85
Processing time (P3)	68	64	6,25
Throughput time	233	237	1,69

Table 5-4: Comparison of automatically collected and manually collected data for hand truck assembly

KPI	Automatic collection (seconds)	Manual collection (seconds)	% difference
Processing time (P1)	131	129	1,55
Transit + Waiting time (P1 – P2)	5	9	44,44
Processing time(P2)	174	170	2,35
Transit + Waiting time (P2 – P3)	3	7	57,14
Processing time(P3)	82	78	5,12
Throughput time	395	393	0,50

For most KPIs, the difference between automatic and manual data collection is minimal, demonstrating a high level of accuracy for the automated data collected using object detection.

For processes with discrepancies, these differences were primarily not due to inaccuracies in object detection itself. The automated method calculates processing times as the total time the product spends in the workplace. However, this may not always reflect the actual processing time, as it may take a few seconds before workers begin their tasks. In the automated process, this time is also included in the processing time, causing slight discrepancies. This is a common problem with all location-based tracking methods. To achieve more precise results, more advance technologies such as pose estimation or action recognition would have to be used.

Nevertheless, such differences did not significantly impact the throughput times, as any discrepancies in the calculated production times were typically balanced out by corresponding differences in the transit and waiting times.

5.3 Summary of Results

The results presented in this chapter demonstrate the feasibility and effectiveness of using object detection for dynamic data collection for value stream mapping. The performance of an object detection algorithm, YOLOv8 was evaluated in an industrial environment. The algorithm showed good detection accuracy and proved to be adaptable to the specific needs of industrial applications, making it relatively easy to train models for detecting custom objects. From the comparison of the first three size variants of YOLOv8, the small variant YOLOv8s) was found to strike a good balance between speed, accuracy, and resource requirements.

The testing results showed that the automatic data collection method closely matched the data collected through manual analysis, confirming the applicability of the method.

By collecting data continuously over a period of time, rather than relying on static snapshots used in traditional VSM, more comprehensive data was captured, enabling enhanced visualization of the current state. This improved visibility of processes helps in identifying inefficiencies more effectively. Overall, the results validate the potential of object detection as a valuable tool for enabling digital collection of dynamic data for value stream mapping.

6 Discussion

This chapter presents the key findings from the object detection-based data collection experiment for value stream mapping, discussing the potential and limitations of the applied methodology. The discussion focusses primarily on the capability of the object detection method to collect data from shop floors and realize digital value stream maps. In addressing this, the requirement for implementing object detection, the type of data that can be captured, the benefits of using this method, and the accuracy of the collected data are evaluated. Finally, the challenges of implementing the method in industrial environments are discussed, along with recommendations for further improvement.

The research successfully achieved its aim of evaluating the potential of object detection for non-intrusive data collection for digital value stream mapping. Time and location data required for determining KPIs were collected using object detection without inferring with the workflow. Through the systematic approach adopted for conducting the practical part of the thesis, each objective was met as outline below:

- Selection of suitable algorithm for industrial applications – An extensive comparison of object detection algorithms was conducted. Based on the comparison YOLOv8 was chosen as the algorithm for object detection in the experiment.
- Model training – Object detect models were trained for detecting workers and the two products, scooters and hand trucks assembled in the LEAD factory.
- Development of a Logical Framework – A logic for collecting time and location data to track the assembly process using object detection was developed.
- Testing and evaluation of the method – The method was tested by using object detection to track an assembly process. The model successfully detected and tracked workers and products assembled in the factory with acceptable accuracy. The data required for the determination of the most important KPIs for value stream mapping were successfully collected and were used to visualize a digital VSM.

➢ **Requirements for Object Detection Implementation**

The main requirements for using object detection for digital data collection are the video footage of the assembly process and an object detection model. While public datasets for industrial components are still limited, they are expanding, which will eventually reduce the effort needed for future implementations. Currently, most industrial applications require training custom models from scratch or fine tune existing ones to meet specific needs.

In this experiment, detecting workers and products required custom-trained models, using large, annotated datasets, which is a time-consuming process. The data collection and annotation are the most labour-intensive aspects, while model training takes only a few hours depending on the dataset size and GPU capability. For training larger models of YOLOv8 or any other computationally demanding algorithms, GPU acceleration is essential.

The time needed for annotating images depends heavily on the type of data being annotated, number of classes in each image, total instances on each image, as well as the size and shape of the objects on the image. The annotation of worker dataset took on average 1 hour to annotate 150 images, with three workers per image, whereas product dataset, was more time consuming due to the presence of multiple classes and the smaller size of the objects. Using video data for annotation reduced this time by half as it was able to semi-automate the process since the objects appear in similar position across consecutive frames.

The smaller variants of YOLOv8 require minimal computational resources, reducing hardware demands while offering fast inference rates. Therefore, the effort required with using object detection is mainly associated with the collection and annotation of datasets and training models. Once the model is trained, no additional costs are required and scalability becomes free, unlike physical sensors where costs increase with the number of items being tracked.

- **Type of Data Collected Using Object Detection**

Object detection was used to collect the location and time data of workers and products in the assembly area, enabling the calculation of KPIs relevant to VSM. The most important metrics, such as cycle time, throughput times, transport times, waiting times, total value adding and non-value adding times were captured using the approach. In addition, indicators such as worker and workplace utilization were also determined.

However, some KPIs commonly included in VSMs, such as changeover time, and setup time could not be calculated. Since the method relies on the location of the product on the assembly area, the progress of the product through the assembly area is primarily mapped. It cannot distinguish whether a product is actively being processed or simply resting in a workplace while the worker is working on other tasks, such as tool setup or reading instructions.

- **Benefits of using Object Detection for Data Collection**

The benefits of digital data collection over traditional methods are well documented in the literature. This study emphasized the advantages of object detection over other location-based methods, such as RFID tags and RTLS systems.

Similar to the other location-based sensors, object detection allows for continuous data collection over a period of time, providing insights about actual situation in the shop floor. In the experiment, YOLOv8n achieved a speed of 200 FPS, demonstrating that real-time detection is not an issue. With real-time processing, object detection can realize real-time visualization of value streams.

A key benefit of this method is its ability to collect data without interfering with the workflow. An overhead camera records video, which is then processed using object detection. This eliminates the need for installing and handling of sensors which increases the workload for operators and reduces the available time for value adding activities. In addition, continuously attaching physical sensors to products and workers to track their movement is labour intensive and can be prone to errors over extended periods. Even though the cost of individual sensors is not high, the total cost increases with the number of items being tracked, making it expensive for high volume productions. Whereas object detection models can continuously track any number of products, without additional costs. Therefore, object detection enables interaction-free, low cost, continuous data collection from shop floors.

- **Accuracy of the Automatically Collected Data**

The KPIs determined using the object detection data was carefully analysed by evaluating the video manually. The comparison results showed that the automatic data collection closely matched the actual scenario, with differences for processing times typically under 6%. This indicates that object detection can serve as an alternative to manual methods, offering reliable representation of the actual situation on shop floor.

- **Challenges in Industrial Environments**

While object detection proved to be effective in this experiment, its performance on a larger industrial environment still needs to be studied. Some of the possible challenges in implementing this method in an industrial environment is discussed below.

- Occlusions and Tracking Errors

A major challenge encountered during the experiment was the issue of object occlusion, where products were temporarily blocked from the camera's view, leading to incorrect reassignments of object IDs. Although this problem can be resolved during post-processing using predefined workflows, they may still impact real-time monitoring. Since the chances of such occlusions are higher on a more complex industrial environment, very efficient detection models, capable of detecting objects even during partial occlusions may be necessary to achieve the best results.

- Varying Environments

During the testing, models where trained using images captured from a specific assembly area. For industrial applications, the model may be used in different assembly locations, and therefore, the background, lighting, and other environmental parameters may change. In order to ensure consistent performance, the models should be trained using data from a variety of situations, that reflect the diverse environments in which they will be used.

- Performance During Continued Operation

The testing in this experiment was conducted over a nine-minute assembly process, which provided enough data for the analysis. However, the performance of the method in continuous industrial operations remains to be studied. Particularly for real-time data collection, object detection systems need to run for extended periods and the effect of this sustained computational load on its performance is not known. Further research is needed to assess how the method performs during prolonged use and to identify any optimisations necessary for continuous monitoring in industrial environments.

- **Recommendations for Improvement**

The following recommendations are proposed for improving the application of object detection in value stream mapping

- Integration of Advanced Object Tracking Algorithms

In the experiment, the ne of the state-of-the-art tracking algorithms, Bot-SORT was used with YOLOv8 to achieve tracking functionality. Even though it performed well, the potential of using more advanced trackers to address the occlusion issue needs to be studied. Improving the tracking capability can significantly improve the results from the proposed method.

- Integrating Pose Estimation

A limitation of the proposed approach was its inability to distinguish whether products were being actively processed or simply resting on the workplaces. This prevented the determination of KPIs such as set-up time and changeover time. Relying solely on location data cannot solve this issue. Pose estimation is an important computer vision task, which in addition to detecting an object, also tracks the key points on the object to infer its orientation or movement. Pose estimation of humans is an actively researched topic and integrating it with object detection can help in identifying KPIs that current location-based systems fail to do.

7 Conclusion

This thesis evaluated the potential of using object detection technology to facilitate non-intrusive data collection for digital value stream mapping (VSM). The study demonstrated that object detection could serve as an effective tool for tracking objects such as workers and products on a shop floor without interfering with the workflow.

A key advantage of this approach compared to other location-based tracking systems is its non-intrusive nature. Unlike systems that require physical sensors to be attached to the entities to be tracked, this method enables the collection of a variety of data using just a camera and an algorithm. This interaction-free approach ensures that the employees are not burdened with additional work and can focus on their primary tasks. It also eliminates the risk of errors caused by mishandling or misplaced sensors.

To test the method, an object detection algorithm, YOLOv8, was selected based on the results of the literature review. The approach was tested at the learning factory at TU Graz, where a custom-trained YOLOv8 model successfully detected and tracked the workers and products throughout the assembly process. The location and time data collected using object detection was processed in excel to dynamically visualize the VSM, reflecting the situation on the shop floor. This comprehensive data enabled more detailed visualizations, using Gantt charts and pie charts, providing a clearer depiction of the situation. Visualizing key KPIs individually made the data more accessible, offering deeper insights into the process. The holistic perspective provided by value stream map was enhanced by the focused view provided by these charts.

An analysis of the accuracy of the collected data through video review showed that the digitally collected data closely matched the actual scenario during the assembly process, reinforcing the viability of object detection for industrial applications. By offering dynamic insights into production processes, this approach has the potential to offer manufacturers with the ability to optimize their operations more efficiently. With further validation in larger and more complex environments, object detection has the potential to largely improve how material data is collected and utilized in manual assembly processes.

8 Outlook

This study explored the potential of object detection and tracking for non-intrusive data collection in the context of digital value stream mapping (VSM). While the developed method has shown promising results in controlled environments, further validation in complex industrial settings is crucial. It would assess the capability of object detection to adapt to more complex assembly processes and diverse environmental conditions providing insights into the robustness of the approach.

In addition, since object detection algorithms are capable of real-time detections, it is possible to achieve real-time visualizations, if data transfer and processing can also occur in real-time. Therefore, implementing real-time data transfer and analysis for live visualization of value stream maps is proposed as a future step.

Occasional tracking errors and occlusions were the challenges observed during the experiment. To mitigate these issues, more sophisticated solutions for achieving better tracking results could be explored.

Finally, integrating pose estimation with object detection is proposed for improving the accuracy of the collected data. Relying solely on the location of the product may cause misinterpretations, as the presence of a product in a workplace cannot guarantee that the worker is actively working on the product. Pose estimation could address this limitation by providing additional information, such as worker orientation and movement, to predict whether active processing is occurring.

9 References

Hattak, A., et al. (2023). Benchmarking YOLO Models for Automatic Reading in Smart Metering Systems: A Performance Comparison Analysis. *In: 2023 International Conference on Machine Learning and Applications (ICMLA).*

Adegun, A., et al. (2023). State-of-the-Art Deep Learning Methods for Objects Detection in Remote Sensing Satellite Images. *Sensors 23* (13).

Ahmad, H. & Rahimi, A. (2022). Deep learning methods for object detection in smart manufacturing: A survey. *Journal of Manufacturing Systems 64,* 181–196.

Arey, D., et al. (2021). Lean industry 4.0: a digital value stream approach to process improvement. *Procedia Manufacturing 54, 19–24.*

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools.*

Chen, C., et al. (2020). Monitoring of Assembly Process Using Deep Learning Technology. *Sensors 20* (15).

Chen, W., et al. (2024). A review of object detection: Datasets, performance evaluation, architecture, applications and current trends. *Multimedia Tools and Applications* 83 (24), 65603–65661.

Čiarnienė, R. & Vienažindienė, M. (2012). Lean Manufacturing: Theory and Practice. *Economics and Management* 17 (2).

Dehaerne, E., Dey, B., Halder, S. (Eds.) (2022). A Comparative Study of Deep-Learning Object Detectors for Semiconductor Defect Detection, *ICECS 2022 - 29th IEEE International Conference on Electronics, Circuits and Systems, Proceedings.*

Fernandes, E., et al. (2023). A Flexible and Intelligent Production System for Process Planning and Enterprise Performance Optimization. In: *Advances in Transdisciplinary Engineering*, 482–491.

Ferreira, W., et al. (2022). Extending the lean value stream mapping to the context of Industry 4.0: An agent-based technology approach. *Journal of Manufacturing Systems 63*, 1–14.

Frank, A., et al. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics* 210, 15–26.

Frick, N. & Metternich, J. (2022). The Digital Value Stream Twin. *Systems 10* (4).

Frick, N., et al. (2024). Design Model for the Digital Shadow of a Value Stream. *Systems 12* (1).

Gupta, S. & Jain, S. (2013). A literature review of lean manufacturing. *International Journal of Management Science and Engineering Management 8 (4)*, 241–249.

Haffner, O., et al. (2024). Applications of Machine Learning and Computer Vision in Industry 4.0. *Applied Sciences 14* (6).

Herbaz, N., El Idrissi, H., Badri, A. (Eds.) (2023). Deep Learning Empowered Hand Gesture Recognition: using YOLO Techniques, *Proceedings - SITA:14th International Conference on Intelligent Systems: Theories and Applications.*

Hines, P. & Rich, N. (1997). The seven value stream mapping tools. *International Journal of Operations & Production Management 17 (1)*, 46–64.

Horsthofer-Rauch, J., et al. (2022). Digitalized value stream mapping: review and outlook. *Procedia CIRP 112*, 244–249.

Huang, Z., et al. (2019). Industry 4.0: Development of a multi-agent system for dynamic value stream mapping in SMEs. *Journal of Manufacturing Systems 52*, 1–12.

Hussain, M. (2023). YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines 11* (7).

Iyer, S., et al. (2023). Digitalization: a tool for the successful long-term adoption of lean manufacturing. *Procedia CIRP 116*, 245–250.

Jakubec, M., et al. (2023). Comparison of CNN-Based Models for Pothole Detection in Real-World Adverse Conditions: Overview and Evaluation. *Applied Sciences (Switzerland)* 13 (9).

Jasti, N. & Sharma, A. (2014). Lean manufacturing implementation using value stream mapping as a tool. *International Journal of Lean Six Sigma 5* (1), 89–116.

Jocher, G., et al. (2023). Ultralytics YOLOv8 2023. Available online at https://github.com/ultralytics/ultralytics.

Kaur, J. & Singh, W. (2022). Tools, techniques, datasets and application areas for object detection in an image: a review. *Multimedia Tools and Applications 81* (27), 38297–38351.

Kaur, R. & Singh, S. (2023). A comprehensive review of object detection with deep learning. *Digital Signal Processing 132*, 103812.

Kitsukawa, T., et al. (2023). Camera-based Progress Estimation of Assembly Work Using Deep Metric Learning. In: *2023 IEEE/SICE International Symposium on System Integration (SII)*, 1–6.

Klimecka-Tatar, D. & Ingaldi, M. (2022). Digitization of processes in manufacturing SMEs - value stream mapping and OEE analysis. *Procedia Computer Science* 200, 660–668.

Kozamernik, N., et al. (2023). Visual quality and safety monitoring system for human-robot cooperation. *The International Journal of Advanced Manufacturing Technology 128 (1)*, 685–701.

Kumar, N., et al. (2022). Lean manufacturing techniques and its implementation: A review. *Materials Today: Proceedings 64*, 1188–1192.

Kumar, V., et al. (2019). Is the Lean Approach Beneficial for the Manufacturing Sector: Review on Literature. In: *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 379–383.

Lai, N., et al. (2019). Industry 4.0 Enhanced Lean Manufacturing. In: *8th International Conference on Industrial Technology and Management (ICITM)*, 206–211.

Lampropoulos, G., et al. (2019). Internet of Things in the Context of Industry 4.0: An Overview. *International Journal of Entrepreneurial Knowledge 7*, 19 - 4.

Leitão, P., et al. (2019). A Lightweight Dynamic Monitoring of Operational Indicators for a Rapid Strategical Awareness. In: I*EEE International Conference on Industrial Cyber Physical Systems (ICPS)*, 121–126.

Lou, P., et al. (2022). Real-time monitoring for manual operations with machine vision in smart manufacturing. *Journal of Manufacturing Systems 65*, 709–719.

Hasan, M. et al. (2023). Comparative Study of Object Detection Models for Safety in Autonomous Vehicles, Homes, and Roads Using IoT Devices. In: *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–6.

Lewin, M. et al. (2017). Method for process modelling and analysis with regard to the requirements of Industry 4.0: An extension of the value stream method. In: 43rd *Annual Conference of the IEEE Industrial Electronics Society,* 3957–3962.

Ramadan, M. et al. (2012). RFID- Enabled dynamic Value Stream Mapping. In: *Proceedings of 2012 IEEE International Conference on Service Operations and Logistics, and Informatics*, 117–122.

Mariappan, R., et al. (2023). Intelligent VSM Model: a way to adopt Industry 4.0 technologies in manufacturing industry. *The International Journal of Advanced Manufacturing Technology 129 (5)*, 2195–2214.

Mumuni, A. & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. Array 16, 100258.

Nausch, M., et al. (2023). Nachhaltiges und echtzeitnahes Wertstrommanagement NeW3. Fraunhofer Austria Research GmbH; Fraunhofer IPA; Fraunhofer IZM. (accessed December 2023).

Palange, A. & Dhatrak, P. (2021). Lean manufacturing a vital tool to enhance productivity in manufacturing. *Materials Today: Proceedings 46*, 729–736.

Liu, Q. & Yang, H. (2020). An Improved Value Stream Mapping to Prioritize Lean Optimization Scenarios Using Simulation and Multiple-Attribute Decision-Making Method. *IEEE Access 8*, 204914–204930.

Raj, S., et al. (2024). Augmented reality and deep learning based system for assisting assembly process. *Journal on Multimodal User Interfaces 18* (1), 119–133.

Rantschl, M., et al. (2023). Extension of the LEAD Factory to address Industry 5.0. In: *13th Conference on Learning Factories* (CLF 2023).

Rother, M. & Shook, J. (1999). Learning to See: value stream mapping to add value and eliminate muda.

Sanders, A., et al. (2016). Industry 4.0 implies lean manufacturing: Research activities in industry 4.0 function as enablers for lean manufacturing. *Journal of Industrial Engineering and Management 9 (3)*, 811–833.

Scheder, N., et al. (2023). Concept for Value Stream-Oriented Analyses of Event-based Data in Three Perspectives. In: 2023 *International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 228–233.

Sekachev, B., et al. (2020). *opencv/cvat:* v1.1.0 2020.

Sullivan, B., et al. (2022). Digital Value Stream Mapping: Application of UWB Real Time Location Systems. *Procedia CIRP 107*, 1186–1191.

Sundar, R., et al. (2014). A Review on Lean Manufacturing Implementation Techniques. *Procedia Engineering 97*, 1875–1885.

Tabanli, R. & Ertay, T. (2013). Value stream mapping and benefit–cost analysis application for value visibility of a pilot project on RFID investment integrated to a manual production control system—a case study. *The International Journal of Advanced Manufacturing Technology 66 (5)*, 987–1002.

Tao, J., et al. (2022). Utilization of Both Machine Vision and Robotics Technologies in Assisting Quality Inspection and Testing. *Mathematical Problems in Engineering 2022*.

Telicko, J. & Jakovics, A. (2023). Comparative Analysis of YOLOv8 and Mack-RCNN for People Counting on Fish-Eye Images, *International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2023*.

Terven, J., et al. (2023). A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction 5 (4)*, 1680–1716.

Thangarajoo, Y. (2015). Lean Thinking: An Overview. *Industrial Engineering and Management 04*.

Tran, T.-A., et al. (2021). Indoor Positioning Systems Can Revolutionise Digital Lean. *Applied Sciences 11 (11)*.

Trebuna, P., et al. (2019). Digital value stream mapping using the tecnomatix plant simulation software. *International Journal of Simulation Modelling 18 (1)*, 19–32.

Vaibhav, S., et al. (2013). Manufacturing System Performance Improvement by Value Stream Mapping a Literature Review.

Wang, H.-N., et al. (2021). Framework of automated value stream mapping for lean production under the Industry 4.0 paradigm. *Journal of Zhejiang University: Science A 22 (5)*, 382–395.

Womack, J. & Jones, D. (1996). Lean Thinking : Banish Waste and Create Wealth in Your Corporation. *Journal of the Operational Research Society 48*.

Zaidi, S., et al. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing 126*, 103514.

Zamora-Hernández, M.-A., et al. (2021). Deep learning-based visual control assistant for assembly in Industry 4.0. *Computers in Industry 131*, 103485.

Zeng, W. (2024). Image data augmentation techniques based on deep learning: A survey.

Mathematical Biosciences and Engineering 21 (6), 6190–6224.

10 List of Figures

11 List of Tables

12 List of Listings

13 List of Abbreviations

TU	University of Technology
AI	Artificial Intelligence
ML	Machine Learning
IT	Information Technology
IIM	Institute of Innovation and Industrial Management
KPI	Key Performance Indicator
ERP	Enterprise Resource Planning
MES	Manufacturing Execution System
SCM	Supply Chain Management
JIT	Just-in-Time
TPS	Toyota Production System
TPM	Total Productive Maintenance
VVA	Value Adding Activities
WIP	Work in Progress
IoT	Internet of Things
CPS	Cyber Physical Systems
IPS	Indoor Positioning System
PLC	Programmable Logic Controller
UWB	Ultra-Wide Band
ROI	Region of Interest
CSV	Comma-Separated Values
VBA	Visual Basic for Applications
HMI	Human Machine Interface
VSM	Value Stream Mapping
DVSM	Digital Value Stream Mapping
IVSM	Intelligent Value Stream Mapping
IEMS	Integrated Efficiency Monitoring System

NVVA	Non - Value Adding Activities
NNVA	Necessary Non - Value Adding Activities
RFID	Radio Frequency identification
RTLS	Real-Time Location System
TDoA	Time Difference of Arrival
YOLO	You Only Look Once
CVAT	Computer Vision Annotation Tool