NORDDESIGN CONFERENCE
REYKJAVIK, ICELAND, 12-14 AUGUST 2024

Intelligent Part Comparison in Computer-Aided Design

Fabian Hanke[1], Karan Moallim[2], Ruslan Bernijazov[3], Riza Demir[2], Jörg Brünnhäußer[2], Roman Dumitrescu[1], Kai Lindow[2]

[1]Fraunhofer IEM, Germany
[2]Fraunhofer IPK, Germany
[3]Heinz-Nixdorf-Institut, Germany

Abstract: Increasing product complexity and customisation lead to a growing number of parts in computer-aided design, which poses a challenge for part management. This paper presents an approach for automated part comparison using part geometry representation. The solution involves converting CAD models, generating shape distribution histograms and using bounding box dimensions for part comparisons. The method, tested on an open-source dataset and a large industrial dataset, shows significant improvements in the efficiency of identifying similar parts, making it highly applicable for industrial use.

Keywords: Product Development, Computer Aided Design (CAD), Artificial Intelligence (AI), Systems Engineering (SE)

1. Introduction

Systems engineering faces increasing challenges due to the growing complexity of modern systems, which are often characterized by high levels of size, coupling, diversity, variability, and uncertainty (Hennig et al., 2022). Complexity can affect system performance, reliability, robustness, adaptability, and evolvability, as well as the system development process, cost, schedule, and risk (Parrend and Collet, 2020). The increasing complexity in systems engineering is a significant challenge, especially in the field of mechanical engineering. The work of Summers and Shah (2010) provides valuable insight into this issue. The authors identify three fundamental aspects of complexity: size, coupling and solvability. These aspects are expanded in terms of the three elements of design: problem, process, and product (Summers and Shah, 2010).

- Size: This refers to the number of components or elements in a system. The size of a system can significantly affect its complexity, with larger systems generally being more complex (Summers and Shah, 2010).
- Coupling: This refers to the degree of interdependence between the components or elements of a system. Systems with high levels of coupling tend to be more complex due to the increased potential for interactions and dependencies (Summers and Shah, 2010).
- Solvability: This refers to the difficulty of solving a design problem. Problems that are more difficult to solve are considered more complex (Summers and Shah, 2010).

This paper focuses on reducing the number of system components by identifying common parts, reducing duplication and ensuring the minimum number of parts. Typically, mechanical designs of parts are created in computer-aided design (CAD) tools, stored in databases and managed in product lifecycle management (PLM) tools. CAD refers to the concept of using computers to design the physical form of a system and its components (Narayan et al., 2013). Modern CAD systems use 3D solids to represent individual components that can be combined into more complex assemblies (Stroud and Nagy, 2011). The solids that make up the atomic components are called parts.

Finding the same or similar parts becomes increasingly difficult as product complexity and variants increase (McKinsey & Company, 2022). However, finding similar existing parts is crucial to business success because it helps avoid redundant internal resource intense work processes as well as optimising supply chains by design (Lindquist, 2023). In addition, similar parts are often not found because the master data is not maintained correctly (Stewart, 2018). Parts are described using inconsistent naming conventions, and some are expressed in different languages as design teams sometimes work at different company locations around the world, making it difficult to find identical components due to language barriers and the need to adhere to a standardised naming convention. Identifying similar part can also optimise the supply chain by challenging prices for duplicates to identify inappropriate peaks and evaluate offers based on known reference prices and by reducing the number of suppliers for similar parts (Lindquist, 2023; McKinsey & Company, 2001). Therefore, effective parts management is essential for developing complex technical systems. Typical approaches to identifying similar or identical parts include naming conventions and part databases.

However, current trends, such as the increasing complexity of technical systems, the growing demand for individualised products (Ponn et al., 2004), and shorter development cycles, lead to severe challenges for traditional part management approaches. Similarly, individualised products require engineers to design custom parts, increasing the total number of existing parts. At the same time, shorter design cycles reduce the time available to search for similar parts in part databases. As a result, companies must improve the efficiency of their parts management processes to remain competitive. Yet existing methods do not fully exploit the potential of digitisation for parts management. In particular, there is a lack of data analysis approaches for automated searches for similar and identical parts.

In this paper, our overall goal is to improve the search for similar parts and reduce the number of parts in the inventory database in the medium term. To address this issue, we adapt an existing solution to ensure a scalable solution for real-world application of geometric similarity assessment. We propose a general computational pipeline for the computation and comparison of part histograms, discuss various design options for the individual components of this pipeline, and perform several experiments on real and test data to analyse the validity of this approach. We aim to provide a reproducible and comprehensive foundation by using open-source data for evaluation and applying the proposed solution on an industrial dataset provided by the agricultural machinery manufacturer CLAAS KGaA mbH (AI Marketplace, 2020). This concept was developed and tested under real operation conditions as an integrated extension of the cad development environment with CLAAS engineers. It showed great potential and is currently being put into production by the company (AI Marketplace, 2020).

This paper is structured according to the Action Design Research (ADR) (Sein et al., 2011). The ADR is a methodology that combines the practice of action research with the theoretical insights of design research. It unfolds through four distinct stages:

1) Problem Formulation, where the problem at hand is identified and contextualised;
2) Building, Intervention, and Evaluation (BIE), an iterative cycle of developing, applying, and evaluating solutions;
3) Reflection and Learning, a stage dedicated to analysing outcomes and extracting insights; and
4) Formalising Learning, which aims to encapsulate the knowledge gained into theories, models, or best practices.

This approach is particularly valuable in applying theoretical knowledge to the real problems of increasingly complex parts management.

Chapter 1 lays the foundation for the proposed approach and defines the underlying problem. This is followed by a discussion of related work in Chapter 2 and the general solution concept in Chapter 3. This chapter focuses on the building and intervention of the solution. Chapter 4 presents the experiments carried out to evaluate the impact of individual components of the solution and the overall performance of the system. Therefore, the chapter addresses the evaluation reflection and learning and aims to provide best practices. Finally, Chapter 5 concludes the thesis and provides an outlook for future research.

2. Related Work

The related work has been identified by conducting a literature review. We aimed to identify, evaluate, and interpret existing research and research questions in a topic area. In this work, the literature review is used to identify relevant publications and the current state of the art regarding approaches to geometry comparison of design models. The literature review carried out does not claim to be complete, even if the procedure described is similar to a structured literature review, it must be clearly distinguished from it. The state of the art presented is intended to provide the reader with an initial overview of the subject area. The IEEE and Scopus literature collections were searched, representing the essential collections for the given research question. The search string consists of three thematic keywords linked by "AND" and "OR" operators. The search string is as follows:

(geometry) AND (similarity OR comparison OR histogram) AND (intelligence OR machine OR learning OR smart)

This search string returns 2348 papers, split between IEEE 1264 and Scopus 1084 (as of 04/05/2022). Four steps were then taken to reduce the number of papers. In the first step, the titles are checked for duplicates, and identified duplicates are removed. At the same time, posts that do not contain at least one of the keywords in the title are filtered out. This reduces the number of posts to 230. The following three steps to reduce the number are manual checks of the posts. This check is done first based on the title (32), then based on the abstract (13), and finally based on the entire post (3). The remaining three papers represent the relevant state of the art approaches for intelligent geometry comparison of structural components. One additional paper was manually identified, "Shape distribution-based approach to comparing 3D CAD assembly models" by Kim et al. (2017). These relevant papers are listed below.

2.1. The Use of Geometric Histograms for Model-Based Object Recognition

Evans et al. (1993) present a novel method for representing shapes by capturing pairwise geometric relationships between local features using geometric histograms derived from image data. This method supports recognition under difficult conditions such as fragmentation noise and occlusion. Unlike previous approaches, it matches features based on distributions of geometric relationships within shapes, enabling principled shape similarity metrics. The paper describes the geometric histogramming scheme and a simple parallel matching strategy. The matching algorithm correlates histograms to identify correspondences, enabling robust recognition in different scenarios. The study introduces the D2 shape function for measuring distances between random points on the surface of a model, generating random sample points, calculating shape distribution histograms and comparing them using curve matching techniques. (Evans et al., 1993)

2.2. 3D Shape Histograms for Similarity Search and Classification in Spatial Databases

Ankerst et al. (1999) propose the use of shape distributions to simplify the comparison of 3D shapes and demonstrate their effectiveness in discriminating between different groups of models. They advocate histograms as a more intuitive method and compare them with existing techniques. The study uses a dataset of proteins from the Brookhaven Protein Data Bank (PDB). The authors present three decomposition techniques for space partitioning: the shell model, the sector model and the combined model. The paper illustrates successful basic similarity search and classification based on shape similarity, achieving high levels of accuracy. It also examines the runtime performance, noting improvements over current systems. In addition, the authors outline future directions, such as refining the space decomposition, developing cost models, and improving visualisation to increase user understanding and confidence. (Ankerst et al., 1999)

2.3. Using Shape Distributions to Compare Solid Models

Yiu Ip et al. (2002) describe a method for adapting shape distributions used for 2D objects to 3D solid models from CAD systems. The method involves creating a polyhedral approximation of the solid model and computing shape distribution histograms using the D2 shape function, which measures the distance between two random points on the model surface. This function provides robust metric distance measures that are invariant under shape preserving transformations. Random sample points are generated uniformly across the surface of the polyhedral approximation, and shape distribution histograms are computed from these points. These histograms, which represent the distances between pairs of random points, are compared using curve matching techniques such as Minkowski LN or Earthmover's distance. The technique works well for distinguishing broad categories of models (e.g. planes, boats, people, animals), but struggles with shapes that have similar gross properties but different detailed characteristics. As models become more complex, the shape histograms tend to become more normal, making it difficult to distinguish models with different topological properties. As a result, the technique often produces false positives and negatives, misclassifying or failing to identify similar shapes. (Yiu Ip et al., 2002)

2.4. Matching 3D Models with Shape Distributions

The paper by Osada et al. (2002) introduces a method for matching 3D models using shape distributions. By representing shapes as probability distributions of geometric properties (e.g., distances between random points on a model's surface), this approach simplifies shape matching and avoids complex traditional methods. The method is robust, efficient, and invariant to transformations and noise, showing high accuracy in distinguishing and classifying 3D shapes in experiments. This technique is useful for applications in shape-based recognition and retrieval systems (Osada et al., 2001).

2.5. Developing an engineering shape benchmark for CAD models

The paper by Jayanti et al. (2006) presents a benchmark database for the evaluation of shape-based search methods in mechanical engineering and discusses the effectiveness of twelve different shape representations for engineering parts. The results of the evaluation, which used average precision and precision-recall curves, showed that 2D view-based techniques routinely outperformed others. In general, histogram-based techniques outperformed feature vector-based techniques (Jayanti et al., 2006).

2.6. Comparing 3D CAD Assembly Models

Kim et al. (2017) propose a novel methodology for comparing 3D CAD assembly models that evaluates both part shape and assembly relationship dissimilarity. This comprehensive approach addresses the limitations of existing methods that focus solely on shape dissimilarity, and provides a more robust comparison framework for 3D models. (Kim et al., 2017)

2.7. Geometric Dataset for Deep Learning

The creation of the ABC dataset by Koch et al. (2019) represents a milestone in geometric deep learning, providing over one million CAD models characterised by parametrically defined curves and surfaces. This dataset facilitates the development and benchmarking of geometric learning algorithms, including surface normal estimation and shape reconstruction, and sets a new standard for algorithm comparison and evaluation (Koch et al., 2019).

2.8. Automated Classification of Mechanical Parts

The automated part classification methodology presented by Rucco et al. (2019) uses supervised machine learning to develop a multilayer artificial neural network. This approach, which is characterized by the assessment of system accuracy through receiver operating characteristic (ROC) curves and similarity coefficients, marks a significant improvement over traditional descriptors, such as the light field descriptor, for the identification of 3D mechanical objects. This contribution is critical to improving the efficiency and accuracy of automated classification in the mechanical and manufacturing industries (Rucco et al., 2019).

3. Solution Concept

Building on the literature reviewed in Chapter 2, this chapter presents our solution, which aims to revolutionise the identification and retrieval of mechanical parts in large datasets. Previous methods such as geometric histograms, shape distributions and statistical approaches have shown varying degrees of effectiveness in 3D model comparison. Evans et al (1993) introduced geometric histograms for robust recognition under difficult conditions. Ankerst et al. (1999) demonstrated the effectiveness of shape distributions in discriminating groups of models. Yiu Ip et al. (2002) highlighted the robustness and limitations of shape distributions for 3D solid models, while Osada et al. (2002) simplified 3D model matching by focusing on efficiency and transformation invariance.

Our solution addresses the gaps identified in previous studies by integrating geometric analysis with machine learning and database management for an accurate and scalable part retrieval process. The core of our solution is an SQL database-centric system that organises part features, including attributes and histogram values, to facilitate similarity searches.

Key steps in our approach include:

1. **STEP to STL Conversion**: Converting CAD models from STEP to STL format for detailed geometric representation.
2. **Histogram Generator**: Creating histograms that capture shape distributions for robust comparison.
3. **Bounding Box Dimensions and Ratios**: Using bounding box dimensions for comprehensive part comparisons, ensuring rotation invariance and efficient pre-filtering.
4. **Pre-filtering**: Efficiently narrowing down the search space based on bounding box ratios before detailed histogram comparison.
5. **Histogram Comparison**: Comparing histograms to identify similar parts.
6. **Visualized Results**: Presenting the results of the similarity search in an intuitive user interface.

We use the open-source ABC dataset (Koch et al., 2019) to ensure reproducibility. Our solution builds on the shape distribution-based approach of Kim et al. (2017), which does not rely on labelled data, eliminating the need for time- and resource-intensive labelling.

Figure 1 illustrates the concept of the application. The data source is an SQL database from which mechanical part features are derived and stored in tabular form. These features include part attributes and histogram values used for similarity searches. The calculation of histogram values and other components of the application are explained in the following sections.

3.1. STEP2STL conversion

Converting CAD models from STEP to STL format is the first step. An STL file uses triangular facets to directly represent the surface of the part, which is more practical for our purposes than the STEP format, which focuses on the operations used to create the CAD model. This conversion process has been automated to handle large datasets efficiently, ensuring accurate capture of surface geometry for subsequent analysis.

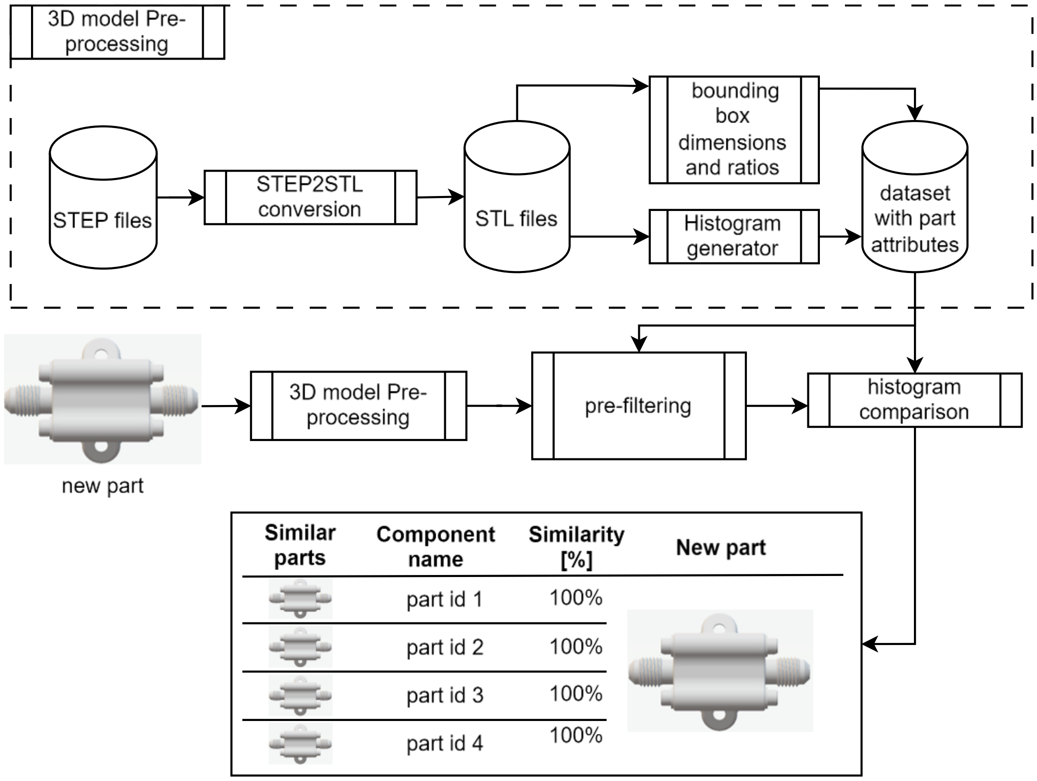
3.2. Histogram generator

We used the D2 method to generate histograms, focusing on scalability for large datasets and industrial applications. Part geometries are characterised by capturing the geometry-based shape distribution, represented by histograms. To compute the histogram, 1024^2 random point pairs are selected on the part surface and the Euclidean distances between these pairs are calculated. These distances are then represented in a histogram of 1024 bins, effectively creating a 1024-dimensional vector. The attributes and histograms are stored in an SQL database to allow quick inference. The number of selected point pairs balances computational efficiency and histogram quality, with 1024^2 proving to be a reasonable compromise in practice. This ensures that the system can effectively handle large amounts of data, providing accurate and efficient similarity searches.

3.3. Bounding box dimensions and ratios

Using bounding box dimensions and ratios as key properties of a 3D model allows the identification of similar parts that may differ in volume and surface area but have the same structure. Sorting each part's bounding box dimensions by size provides a rotation-invariant view, so that parts that are identical but rotated in space can be considered similar. The sorted dimensions (A, B, C) are used to create ratios between the lengths, resulting in three additional ratios for each component.

Rotational invariance is crucial for identifying scaled parts and those with different origins and orientations in the coordinate system. The calculation of bounding box dimensions and ratios ensures accurate comparisons between different parts, meeting the objective of identifying geometrically similar parts scaled by a single factor. Figure 2 illustrates the calculation of these ratios based on the bounding box dimensions.



Similar parts	Component name	Similarity [%]	New part
	part id 1	100%	
	part id 2	100%	
	part id 3	100%	
	part id 4	100%	

Figure 1. Illustration of part finder solution concept and respective building blocks

3.4. Pre-filtering

The pre-processing steps for a new CAD part are the same as for historical parts. Pre-filtering the dataset ensures that the subsequent search for similar parts is more specific and efficient. After calculating the three dimensions of the bounding box for the new part, the ratios of these dimensions are calculated in the same way as for historical parts in the database. These newly calculated ratios are then compared with those of the parts in the database. For each ratio of the new part, a limit is set based on a percentage deviation from the ratios of the new part. Parts within these limits are retained, while others are discarded.

The new part is assigned to the appropriate subset defined during the pre-filtering step. The histogram of the new part is then compared with the pre-filtered set of historical parts selected on the basis of their bounding box dimensional ratios. Pre-filtering using bounding box ratios rather than absolute dimensions allows the identification of scaled 3D models (e.g. scaled by a factor of 2) to meet a user-specific requirement. This approach is particularly effective for large datasets, such as the multi-million 3D model dataset used in this study.

Conversion from STEP to STL results in some loss of information due to surface approximation with triangles. To maintain accuracy, we set a threshold of 1% deviation from the bounding box ratios of the reference parts to allow for slight variations due to approximation. The effectiveness of this pre-filtering step is demonstrated in section 4.2, which shows its impact on the overall system performance.

3.5. Histogram comparison

The histogram acts as a unique fingerprint of the 3D model, making it easy to process and store. By comparing different histograms, we can determine the similarity or dissimilarity between parts. The final step is to output possible similarities based on these comparisons. Various distance measures such as Euclidean distance, Wasserstein distance and Canberra distance can be used to compare two histograms. The comparison metrics are summarised in Table 1, with each comparison yielding a value indicating the degree of similarity, allowing the output components to be sorted accordingly. The details of the comparison metrics used and their performance evaluation are covered in Section 4.5, which provides an in-depth analysis of the effectiveness of different metrics in comparing histograms to identify similar parts.

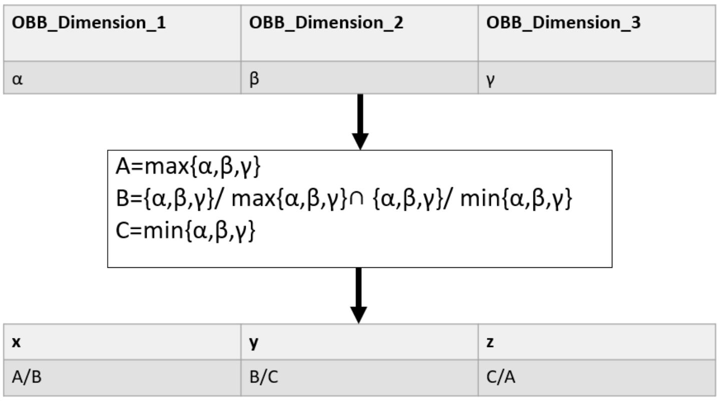


Figure 2. Calculated bounding box ratios to allow comparability between the new component and the existing component set for the solution space under consideration

3.6. Visualized results

The last step is the visualisation of the similarity search. If available, identical or similar parts are presented to the user. The result is a list of similar parts ranked according to the selected comparison value, starting with the most similar part. In the context of this project, a user interface was developed that also displays the reference part and the currently selected similar part, as shown in Figure 1.

Table 1. Comparison metric used as a measure of similarity of the compared histograms

Squared chord	$\sum_{i=0}^{1023} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2$
Euclidean distance	$\sqrt{\sum_{i=0}^{1023} (a_i - b_i)^2}$
Wasserstein distance	$\inf_{\pi \epsilon \Gamma(u,v)} \int_{R \times R} \lvert x - y \rvert d\pi(x,y)$
Canberra distance	$\sum_{i=0}^{1023} \frac{\lvert a_i - b_i \rvert}{\lvert a_i \rvert + \lvert b_i \rvert}$

4. Evaluation

In this evaluation section, we perform a rigorous evaluation of the proposed solution concept to determine its effectiveness and functional capability. To achieve this, we focus on key evaluation parameters, including precision, reproducibility, comparison of different algorithms, the efficiency of the pre-filtering process, and quality of predictions concerning manually labelled data. By systematically analysing these critical aspects, we aim to comprehensively understand the solution's performance and potential impact on the identified problem.

4.1. Underlying dataset

Data selection and preparation are critical components of the evaluation process. In this study, the proposed solution is evaluated using the ABC open dataset, as mentioned in Section 2.2. This dataset, obtained from an open repository, is a valuable resource for evaluating the solution's performance. A subset of 1000 files was specifically selected for analysis to ensure a comprehensive evaluation.

While the dataset provides a substantial collection of CAD parts, it requires manual labelling to facilitate accurate evaluation. Although the ABC dataset contains labels, the labels were found to be too generic to distinguish between two parts and accurately assess their identity. As a result, a meticulous manual labelling process was performed on the 1000 CAD parts, increasing the granularity of the data and enabling a more accurate evaluation of the solution's capabilities. Figure 3 below shows the five reference CAD models used to evaluate the proposed solution.

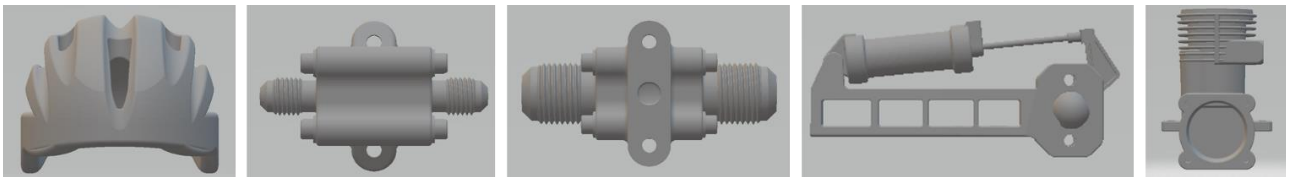


Figure 3. The CAD models used from left to right 00000011_e909f412cda24521865fac0f_trimesh_000.stl, 00000148_d9a2aa6d24764b809c265460_trimesh_001.stl, 00000210_33bd159d563f438fbbebd9fa_trimesh_002.stl, 00000572_5a4cef68211d4706b1ec8586_trimesh_001.stl, 00000473_db2f9eab292d47fa8304bcf9_trimesh_000.stl

4.2. Capability

The evaluation of the performance of the proposed solution involves a careful study of its computational efficiency, a key advantage highlighted in this study. Specifically, this evaluation compares two variants with and without pre-filtering. To facilitate the pre-filtering process, a criterion has been established that the histogram comparison is performed only on parts with a similarly oriented bounding box, with a tolerance of plus or minus 1% in each dimension. This strict criterion ensures that only parts with comparable dimensions are considered, eliminating the possibility of non-identical parts being identified as similar, drastically reducing the number of possible matches and, therefore, the computation time. The evaluation uses micro-benchmarking to measure and test the solution's performance (Poggi, 2019). By using this methodology, we aim to provide a comprehensive analysis of the proposed solution. This approach allows us to focus on individual functions or algorithms within the solution's computational efficiency, allowing us to understand its performance in different scenarios.

The performance results show significant differences between the histogram comparison with and without pre-filtering. Specifically, the execution time for the histogram comparison with pre-filtering for 1000 parts was 0.00234 seconds, while the execution time for the same task without pre-filtering was significantly longer at 0.5699 seconds. These results indicate that the implementation of pre-filtering significantly improves the computational efficiency of the histogram comparison process, resulting in a significant reduction in the overall execution time.

4.3. Reproducibility

The assessment of reproducibility in this study involves an evaluation of the predicted outcomes to ensure their consistency and reproducibility. The evaluation uses the relative standard deviation (RSD) as a statistical measure to assess the precision and reproducibility of the predicted outcomes (Gao et al., 2013). In the histogram generation process, the extraction of point pairs on the part surface is random, while every other function within the process is deterministic. A reproducibility evaluation is performed to evaluate the potential impact of randomness on the histogram generation and subsequent results. To verify the reproducibility of the results, the study uses a test protocol that repeats the point selection and histogram generation 100 times on the same part. The results of the reproducibility evaluation are illustrated through visual displays in Figure 4 and statistical analysis. The left side of the analysis shows the overlay of 100 runs, showing 100 histograms with small variances between them. This visual representation demonstrates the consistency of the generated histograms over multiple iterations. The analysis shows a maximum RSD value of 0.558%, indicating minimal variance within the generated vectors. In addition, most of the histograms have an RSD of less than 0.1%, indicating a remarkably small variation across the data points.

4.4. Quality

The quality of the proposed solution was assessed through a performance evaluation using two datasets, one provided by our industry partner, CLAAS Engineering, and the ABC dataset. The CLAAS dataset, derived from real engineering processes, provides a valuable benchmark for evaluating the solution's performance under real-world conditions. The ABC dataset was used to demonstrate the general capability of the solution applied to an open-source dataset. Within this CLAAS dataset, 11 CAD models were processed, including seven different types of CAD models. Two types of CAD models in this dataset contained identical duplicates. All 11 CAD models were manually labelled to ensure accuracy and reliability. The results of the performance evaluation were promising. The proposed solution accurately identified all 11 labelled CAD models in testing, demonstrating exceptional precision and accuracy. While this dataset may not fully represent the diverse challenges of predicting the quality of engineering components, it stresses the solution's capability to handle real-world data, a critical factor for practical implementation. Evaluation of the ABC dataset based on 1000 labelled models resulted in 100% correct differentiation of similar and dissimilar parts. The results of the quality evaluation are shown in Table 2 below.

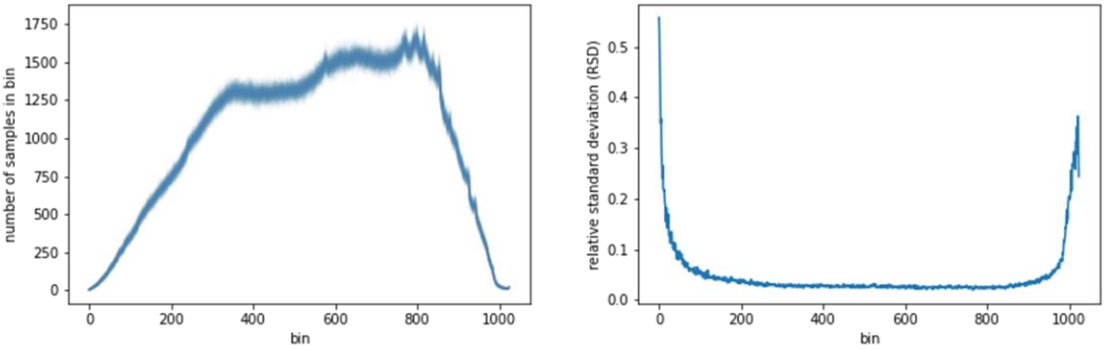


Figure 4. Plot the number of samples over the bins of the histogram (left) & the relative standard deviation for each bin (right)

Table 2. Confusion Matrix Example Data Industry Use Case and ABC Test Dataset

		CLAAS dataset		00000011_e909f412cda24521865fac0f_trimesh_000.stl		00000148_d9a2aa6d24764b809c265460_trimesh_001.stl		00000210_33bd159d563f438fbbebd9fa_trimesh_002.stl		00000572_5a4cef68211d4706b1ec8586_trimesh_001.stl		00000473_db2f9eab292d47fa8304bcf9_trimesh_000.stl	
		Predicted condition		Predicted condition		Predicted condition		Predicted condition		Predicted condition		Predicted condition	
		P	N	P	N	P	N	P	N	P	N	P	N
Actual condition	Positive (P)	7	0	6	0	8	0	8	0	4	0	12	0
	Negative (N)	0	48	0	994	0	992	0	992	0	996	0	988

4.5. Benchmarking

The evaluation section includes a comparison of various comparison metrics, shown in Table 1, to determine their effectiveness within the proposed solution. Given the number of available histogram comparison metrics, this study seeks to identify the most effective metric for the intended purpose. To achieve this, the identical computation is performed on a predetermined dataset, with the variation replacing the comparison metric function. By systematically replacing the comparison metric function, the study facilitates an evaluation of the performance of each metric within the solution. The comparison focuses on identifying the parts distinguished by each metric and then evaluating the intersection between the different metrics. In addition, the analysis considers the decrease in the calculated comparative value for "dissimilar" parts, highlighting the differentiation in the calculation process for the selected metrics. Given the different calculation methods based on the selected metric, the resulting graphs are expected to show dissimilar patterns. This comparison method allows the identification of the most effective comparison metric.

Quantifying a similarity score to distinguish between similar and dissimilar parts is challenging. One approach is to evaluate the decay of similarity values within a given dataset, where similar parts typically yield high similarity scores.

On the other hand, the first dissimilar part yields a significantly lower similarity score, resulting in a significant decline, as shown in Figure 5. The slope of this decrease can be used to distinguish between similar and dissimilar parts.

The Canberra distance shows superior performance when applied to multiple reference parts in the ABC dataset. In contrast, the Wasserstein distance, while promising, has an inherent flaw. It incorrectly identifies fully inverted histograms as identical histograms, ignoring the different distribution of distances between pairs of points. This may not be a problem in small datasets, but it can lead to the misidentification of similar parts in large and diverse datasets.

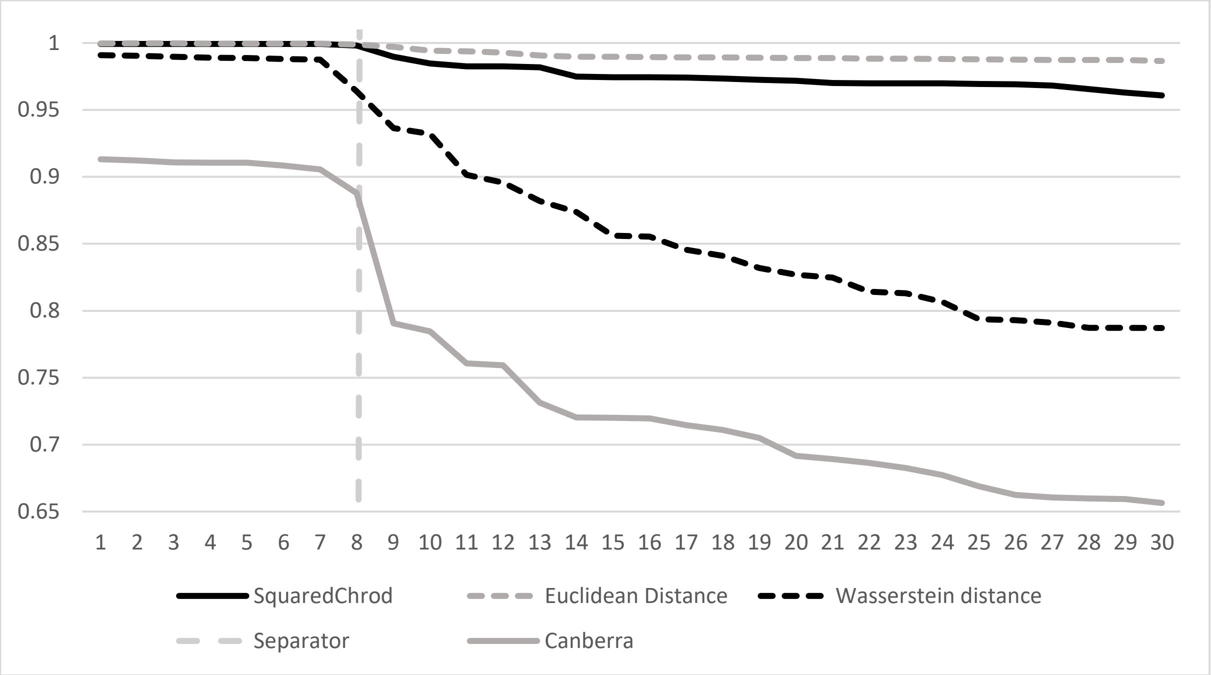


Figure 5. Plot benchmarking different comparison metrics

4.6. Discussion

The evaluation of the solution's capabilities is limited by the use of the available small dataset of only 1000 CAD parts, which may not fully represent the challenges faced by companies managing large databases of millions of CAD parts. While the solution is designed for multi-million CAD datasets, the manual labelling required for evaluation is unrealistic for such large datasets. Despite this limitation, the solution shows promising results in identifying similar parts within the smaller dataset, highlighting its potential effectiveness in practical applications. The remarkably fast computation time further emphasises its ability to efficiently search for similarities even within large datasets. In addition, the low relative standard deviation values confirm the high reproducibility of the results, demonstrating the solution's reliability in producing consistent results.

The solution has been validated by test users at CLAAS Engineering, demonstrating its ability to handle multi-million-part datasets effectively. During testing, the solution demonstrated its ability to identify similar parts in less than 20 seconds. However, it is important to note that the dataset used for testing within the CLAAS database was unlabelled, limiting the ability to definitively assess whether all duplicates were successfully identified. The integration of the application has created a baseline comparison of identical parts that will be used within the company to identify duplicates and duplicate parts within the parts database and to prevent the creation of such duplicates for future development. The successful implementation of this process also replaces manual component searches, reducing repetitive activities. In addition, part matching supports the company's procurement processes mentioned in the introduction for using reference prices for components already used in previous development projects.

We employ the D2 shape distribution methodology alongside random point sampling techniques to assess shape similarity (Osada et al., 2001; Yiu Ip et al., 2002). The selection of the D2 shape distribution is based on its demonstrated efficacy in previous studies and its simplicity. Our findings suggest that further enhancements to the shape distribution do not substantially improve outcomes for our specific use case. However, different scenarios might necessitate more sophisticated similarity assessment methods. For instance, alternative approaches are discussed in (Iyer et al., 2005). Additionally, employing Poisson-disc sampling could ensure a more uniform distribution of sampled points, enhancing the reliability of the results.

5. Conclusion

Effective CAD part management is critical to the success of industrial companies. Existing approaches do not scale with the increasing number of parts due to technological complexity and individualised products. In this paper, we present a novel approach for the automated identification of similar or identical parts based on an efficient comparison of the geometric form of individual parts. We discuss several variation points in the proposed data processing pipeline and validate the solution pipeline on test data from the open-source ABC dataset and with test users from CLAAS. The proposed solution shows promising results, so CLAAS is actively working on rolling out the solution throughout the company. In future work, we plan to evaluate this approach with other companies to better understand the application areas. In addition, we plan to investigate further analytical methods for recommending similar parts already during part design.

6. References

AI Marketplace (2020) *Integration of AI in Computer Aided Design (CAx)* [Online]. Available at https://ki-marktplatz.com/en/claas/ (Accessed 13 November 2023).

Ankerst, M., Kastenmüller, G., Kriegel, H.-P. and Seidl, T. (1999) '3D Shape Histograms for Similarity Search and Classification in Spatial Databases' [Online]. Available at https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=14a9111ecc64a522f4267ff982d095dcc586b19a (Accessed 13 May 2024).

Gao, Y., Ierapetritou, M. G. and Muzzio, F. J. (2013) 'Determination of the Confidence Interval of the Relative Standard Deviation Using Convolution', *Journal of Pharmaceutical Innovation*, vol. 8, no. 2, pp. 72–82 [Online]. DOI: 10.1007/s12247-012-9144-8.

Hennig, A., Topcu, T. G. and Szajnfarber, Z. (2022) 'So You Think Your System Is Complex?: Why and How Existing Complexity Measures Rarely Agree', *Journal of Mechanical Design*, vol. 144, no. 4.

Iyer, N., Jayanti, S. and Ramani, K. (2005) 'An Engineering Shape Benchmark for 3D Models', *Volume 3: 25th Computers and Information in Engineering Conference, Parts A and B.* Long Beach, California, USA, 24.09.2005 - 28.09.2005, ASMEDC, pp. 501–509.

Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D. and Panozzo, D. (2019) 'ABC: A Big CAD Model Dataset for Geometric Deep Learning', *undefined*, pp. 9593–9603 [Online]. DOI: 10.1109/CVPR.2019.00983.

Lindquist, M. (2023) 'What Is Supply Chain Optimization?', *Oracle*, 4 December [Online]. Available at https://www.oracle.com/scm/supply-chain-optimization/#what (Accessed 11 February 2024).

McKinsey & Company (2001) 'Big data and the supply chain: The big-supply-chain analytics landscape (Part 1)', *McKinsey & Company*, 1 January [Online]. Available at https://www.mckinsey.com/capabilities/operations/our-insights/big-data-and-the-supply-chain-the-big-supply-chain-analytics-landscape-part-1#/ (Accessed 11 February 2024).

McKinsey & Company (2022) 'Cracking the complexity code in embedded systems development', *McKinsey & Company*, 25 March [Online]. Available at https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/cracking-the-complexity-code-in-embedded-systems-development (Accessed 11 February 2024).

Narayan, K. L., RAO, K. M. and Sarcar, M. M. M. (2013) *Computer aided design and manufacturing*, New Delhi, PHI Learning.

Osada, R., Funkhouser, T., Chazelle, B. and Dobkin, D. (2001) 'Matching 3D Models with Shape Distributions' [Online]. Available at https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=923386 (Accessed 13 May 2024).

Parrend, P. and Collet, P. (2020) 'A Review on Complex System Engineering', *Journal of Systems Science and Complexity*, vol. 33, no. 6, pp. 1755–1784.

Poggi, N. (2019) 'Microbenchmark', in Sakr, S. and Zomaya, A. Y. (eds) *Encyclopedia of Big Data Technologies*, Cham, Springer International Publishing, pp. 1143–1152.

Ponn, J., Baumberger, C. and Lindemann, U. (2004) 'GUIDELINES FOR THE DEVELOPMENT OF INDIVIDUALIZED PRODUCTS'.

Sein, Henfridsson, Purao, Rossi and Lindgren (2011) 'Action Design Research', *MIS Quarterly*, vol. 35, no. 1, p. 37.

Stewart, G. (2018) *Master Data Maintenance and How it Impacts Decisions* [Online]. Available at https://nttdata-solutions.com/uk/blog/master-data-maintenance-and-how-it-impacts-decisions/ (Accessed 11 February 2024).

Stroud, I. and Nagy, H. (2011) *Solid Modelling and CAD Systems: How to Survive a CAD System*, London, Springer-Verlag London Limited.

Summers, J. D. and Shah, J. J. (2010) 'Mechanical Engineering Design Complexity Metrics: Size, Coupling, and Solvability', *Journal of Mechanical Design*, vol. 132, no. 2.

Yiu Ip, C., Lapadat, D., Sieger, L. and Regli, W. C. (2002) 'Using Shape Distributions to Compare Solid Models' [Online]. Available at https://dl.acm.org/doi/pdf/10.1145/566282.566322 (Accessed 13 May 2024).

Contact: Fabian Hanke, Fraunhofer IEM, Digital Engineering, Zukunftsmeile 1, Paderborn, Germany, +49 5251 5465289, fabian.hanke@iem.fraunhofer.de