CLUSTERING OF SEQUENTIAL CAD MODELLING DATA

Šklebar, Jelena;
Martinec, Tomislav;
Perišić, Marija Majda;
Štorga, Mario

University of Zagreb

ABSTRACT

Automating modelling activities in computer-aided design (CAD) systems is no exception within design automation, one of the current research endeavours aiming to use and transform design-related data in design decision-making processes and the generation and evaluation facilitation of new design solutions. The paper explores the differences between CAD models based on their feature-based CAD modelling sequences that lead to the final models' design. The dataset collected and structured for the study contains more than 1400 CAD models clustered on two levels by using an unsupervised K-means clustering algorithm. The algorithm is performed on the number (total and unique) and the first-order Markov model transition matrices of the CAD modelling operations and their sequential order, respectively. Therefore, three and ten groups (clusters) of CAD models are obtained regarding the level of clustering. The results show that most of the obtained groups are specified by the dominant transition between particular modelling operations. In addition, the study also provides insight into the potential of using feature-based CAD modelling operations' sequences as a first step toward automating the user interaction with the CAD system.

Keywords: Computer Aided Design (CAD), Computational design methods, Big data, Cluster analysis, CAD modelling sequences

Contact:
Šklebar, Jelena
University of Zagreb
Croatia
jelena.sklebar@fsb.hr

Cite this article: Šklebar, J., Martinec, T., Perišić, M. M., Štorga, M. (2023) 'Clustering of Sequential CAD Modelling Data', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.94

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

1 INTRODUCTION

Design automation (DA) is one of the current research endeavours in using and transforming design-related data to suit design decision-making processes and extract knowledge valuable for the generation and evaluation of new design solutions (Cantamessa et al., 2020). Data mining techniques (e.g., cluster analysis, conjoint analysis), artificial intelligence (AI) tools (e.g., neural networks, generic algorithms) and machine learning algorithms (supervised and unsupervised learning) are thereat widely applied (Tao et al., 2018). Computer-Aided Design (CAD) systems are no exception here, intending to automate several CAD modelling activities identified as routine-like. Hence, among many, researchers have increasingly acknowledged DA as an approach toward reducing the repetitiveness of mundane design tasks in CAD systems, with the constant aim to improve the efficacy and precision of user interaction with the tool (Machchhar and Bertoni, 2021). Given their expanded support for the design process as facilitators of model-based engineering, CAD systems are now generally recognized in various industries. In general, CAD systems allow the user to perform the activity of solid modelling, a field that covers a wide area of activity directed toward the solid physical objects' representation and basic operations on them (Hoffmann, 1989). Therefore, the fundamental representation of designs within a CAD system is the solid object or model, with its geometric and topological representations (Regli, 1995). A solid object or model can be built using design features, a CAD paradigm known as design-by-features or feature-based modelling (Salomons et al., 1993). In a most modern CAD systems, the latter offers the user a library of predefined features that can be instantiated to build a CAD model (Pedley, 1997). Features are parts of components that constitute the CAD model (Salomons et al., 1993), or generic shapes with characteristics defined by attributes that define the geometry of a model and have knowledge associated with them. When interacting with the CAD tool, users choose the features from the library while creating the model, simultaneously recording CAD modelling operations or the history of chosen features that lead to the final model (Regli, 1995). Regarding CAD systems and DA, data mining tools and machine learning is integrated into the design research to predict geometries and reconstruct the CAD models. However, the great potential of automating the CAD modelling process is the prediction of the next user's feature-based operation, inspired by text editors' predictions, e.g., their anticipation and suggestions of words or phrases while typing. This research is motivated by the desire to automate and predict users' CAD feature-based modelling operations. It is, therefore, important to determine the different groups of feature-based CAD models and the main differences in their modelling operations transitions or changes from one feature to another while performing feature-based modelling. Thus, the overall objective of this study to categorize CAD models based on numbers and transitions between CAD modelling operations as one of the very first steps toward CAD modelling automation and predictions, motivated by the following two research questions:

- What are the groups of feature-based CAD models regarding the total and unique number of CAD modelling operations?
- What are the main differences within the groups from the first research question in the domain of transitions of feature-based CAD modelling operations?

This paper first discusses the related work employing CAD modelling sequences and clustering the CAD models and databases. That is followed by the methodology description, including the dataset collection and structuring, as well as the clustering methods applied in the study. Statistics of the dataset and the results of clustering CAD models are then presented to understand the differences among sequences of CAD modelling operations within obtained CAD models' clusters, followed by a discussion of whether the valuable insights can be inferred from the CAD models' clustering analysis.

2 RELATED WORK

2.1 CAD action sequences

Recording designer actions in CAD systems is a common design analysis approach for design behaviour and decision-making research. Thus many researchers extensively obtain log or time series data sequences to better understand CAD modelling patterns from various perspectives, such as studying iterative CAD modelling cycles with the aim of data-driven decision making (Rahman et al., 2019), inferring behaviour differences of designers' with different CAD skill levels (Chen et al., 2021),

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

or differentiate designers' CAD actions on individual and team level (Celjak et al., 2022) . However, CAD systems also offer valuable information on the CAD modelling design history. It is beneficial since CAD models have become the standard shape representation in almost every industrial production sector (Wu et al., 2021). Therefore, more recent work on CAD modelling sequences could be classified into three categories: (a) works on 3D shape generation, (b) CAD models reconstruction and (c) automating the CAD modelling process. All the categories are challengeable due to the diverse ways of CAD models' design. Current work on the learning-based 3D shape generation uses CAD datasets and machine learning to predict the 3D geometries from various geometry representations such as boundary representations (B-reps) (Jayaraman et al., 2022), 2D sketches (Li et al., 2020), point-clouds (Uy et al., 2021) or triangle meshes. Furthermore, CAD reconstruction (e.g., feature recognition) is the detection of geometric primitives and the correspondence of primitives holistically modelled as a modelling sequence. The learning-based approaches are also mainly applied to potentially learn the rules that can automate scenarios requiring user input and generalize when confronted with unfamiliar geometry (Willis *et al.*, 2021), both being compared with the traditional methods that cannot completely rebuild the model operation sequences, such as rule-based or grammar-based. Thus, Fusion 360 Gallery (Willis et al., 2021), Zone Graphs (Xu et al., n.d.), DeepCAD (Wu et al., 2021), and CSGNet (Sharma et al., 2017) are the recent research learning-based attempts to reconstruct the 3D models from CAD modelling sequences generated from B-reps or Constructive Solid Geometry (CSG) representations.

Additionally, datasets of CAD models used in these studies use large-scale synthetic CAD modelling data, whereas only Fusion 360 Galley provides approximately 8600 human-designed CAD modelling sequences. Further, the feature-based CAD modelling sequences can undoubtedly be useful data for machine-learning applications to predict and suggest the following user modelling operation. The recent work on predicting the hole feature by using association rule learning focuses exactly on automating the CAD modelling process (Vasantha et al., 2021). In contrast to all these works and despite the research progress in using CAD data for learning-based purposes, only a limited number of research focuses on sequences of CAD modelling operations. Hence, we focus our research on understanding how the CAD model's design came about by using a fuller array of human-designed CAD modelling operations beyond a sketch and extrude (Willis et al., 2021; Wu et al., 2021), sweep (Li et al., 2020) or boolean operations (Xu et al., 2021). As a first step in automating user interaction with CAD systems, the characteristic transitions between feature-based operation needs to be understood. Additionally, CAD moved to the cloud, thus enabling access to a large amount of user-generated CAD modelling data. Such data is subject to various learning-based and statistical methods, including clustering, which could be used to determine groups that could answer the study's research questions.

2.2 CAD models clustering

Clustering is an unsupervised machine learning technique aiming to group unlabelled data objects (Omran et al., 2007). The grouped data called clusters are represented of objects similar to each other concerning the surrounding or other clusters. It is mainly used for pattern recognition. Therefore, the clustering algorithms are also used for exploratory data mining of CAD models. The existing approaches of clustering CAD models can be classified into two categories regarding their geometry and topology: (1) segmentation of CAD models and (2) grouping CAD models from databases. Many CAD model segmentation research focuses on mesh or point cloud and solid CAD models. For the former, hierarchical clustering is widely used for decomposing meshes of 3D objects via finding the meaningful components of the mesh and generating the exact boundaries between the components (Katz and Tal, 2003; Xiao et al., 2011) or for partitioning a surface of the 3D object into a hierarchy of disjoint face clusters (Garland et al., 2001). However, clustering solid 3D models for their segmentation or CAD models database partitioning research works is rare. Most of the available work on clustering 3D CAD models uses a B-rep descriptor, a form of an internal representation of a 3D CAD model that essentially consists of a set of edges and a set of faces used by designer to describe the shape of the model in 3-dimensional space. B-reps are, for the clustering analysis, transformed into two-dimensional coordinate points corresponding to the nodes of the attributed adjacency graph (AAG). For example, Yuan et al. and Li et al. use single-body CAD parts in the form of STEP files to divide point sets into several groups to achieve the model's segmentation using k-means, k-medoids and spectral clustering techniques. In contrast, Han et al. and Bonino et al. perform spectral clustering for CAD assembly segmentation. The latter is defined as connected sets of parts sharing some

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

characteristics (Bonino et al., 2021). CAD assembly model is also described by an AAG where the node and its attribute represent respectively part assembled and attributes' information of the part, edges, their connection relationship, and assembly constraints between parts (Han et al., 2019). Partitioning the CAD repository into subspaces of similar CAD models is another research direction with the overall objective of facilitation and automation of design retrieval and reuse. Thus, a model signature graph (MSG) and AAG of CAD model parts used as a solid model representation constructed from B-Rep are used as input data into the clustering algorithm for CAD models recognition (Peabody et al., 2001; Roj et al. H.-B., 2015).
The related work regarding CAD clustering shows a lack of research on clustering the CAD modelling history data. Hence, performing the clustering methods on sequences of CAD modelling operations is the overall objective of this study. To get the insight into the clusters' most frequent feature-based transitions, the clustering methods are performed on transition probability matrices of features-based CAD modelling history.

3 METHODOLOGY

This section presents the approach for clustering CAD models performed on CAD design history in terms of feature-based modelling operations and their sequential order. Thereat, 3D CAD models and their sequences of modelling operations are used as the main raw dataset. Its collection, cleaning, structuring and clustering is described in this section.

3.1 Dataset collection

The main aim of data collection is to retrieve 3D CAD model parts and their feature-based design history to extract the CAD modelling sequences, which were further used to perform unsupervised clustering methods. Generally, a CAD model design history lists the feature-based operations used during 3D CAD modelling in the order of their creation. Therefore, the dataset used in this study consists of 1737 single-body Onshape Part Studio 3D models, of which 1273 were created by the 3rd year mechanical engineering students at the University of Zagreb. During a design project-based course, students have created, by using the feature-based design approach, 3D CAD representation of clamping device assemblies and their parts. The rest of the database was obtained by scraping the Onshape public database. Once the dataset as collected, sequences of CAD modelling operations were sourced using the Get Feature List API described in the official Onshape Developer Documentation (Onshape, n.d.).

3.2 Dataset cleaning

Data extracted in the form of CAD modelling operations sequences were cleaned and structured based on the following criteria. Although a sketch is the backbone of the solid 3D model, the sketch items were removed from the modelling sequences to get insight into transitions exclusively used for the 3D CAD shape representation. Also, certain features imply the existence of sketch features, thus implying the sequences between 3D features and sketch. Similarly, to use the transition matrices of CAD modelling operations, CAD models that consisted of zero (53 models) or only one (256 models) operations were removed from the dataset. Thus the final dataset is reduced to 1419 CAD models consisting of, cumulatively, 13 unique features. In addition, to reduce the number of feature-based operations in the dataset, circular pattern, linear pattern, and mirror features were replaced with an operation named "composite feature". In the same manner, the chamfer and fillet features were replaced with "edge-cut feature". In conclusion, the following are ten CAD modelling operations, both for addition and removal of 3D representation material, considered for this analysis: extrude, revolve, sweep, loft, edge cut, composite feature, hole, draft, helix, and shell (Hoffmann, 1989).

3.3 Dataset structuring

Once CAD models’ operations and their sequential order were collected and cleaned, the following information was extracted:

- Total number of CAD modelling operations per CAD model,
- Number of unique CAD modelling operations per CAD model, and
- Transition probability matrix for every CAD modelling sequence.

Regarding numbers, total number corresponds to sequence operations' items lead to the final CAD model, whereas the unique number counts only the distinct values of feature-based operations. Further,

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

the first-order Markov chain describes transitions from one CAD modelling operation state to another. A Markov chain describes a sequence of states where the probability of transitioning from states depends only on the current state. Those probabilities are summarized in the transition matrix, an approach used in this study to describe the sequences of CAD modelling operations considering ten operations in total. Thus, a final transition matrix for each 3D model from the dataset was a square (10x10) matrix whose rows and columns correspond to 10 unique feature-based CAD modelling operations sorted alphabetically. In the transition matrix:

- Rows represent the first state,
- Columns represent the final state and
- Entry (i, j) is the conditional probability that first state = j, given that final state = i: the probability of going from state i to state j.

To prepare data for the clustering as the next analysis step, data information was converted to vectors. Since the clustering was performed on two levels, two types of vectors were extracted. For the first-level clustering, each CAD modelling sequence is converted to a 1x2 vector, representing a point in 2-dimensional space whose elements correspond to the numbers of total and unique operations in a CAD modelling sequence. Similarly, the second clustering level requires the conversion of 10x10 transition matrix to a 1x100 vector, representing a point in 100-dimensional space, whose columns correspond to the states from the feature-based CAD modelling transition matrix. Furthermore, since the clustering analysis is carried out on the rows of an array or matrix (Murtagh & Contreras, 2011), 1419 CAD models were formed into 1419x2 and 1419x100 matrices, respectively. Driven by the research objective, both matrices were used for clustering methods, each corresponding to the order of research questions.

3.4 Clustering

Clustering, as method for underlying patterns recognition (Hamerly & Elkan, 2002) is an unsupervised algorithm used for grouping data into clusters based on some similarity measure (e.g., Euclidean distance) (Oyelade *et al.*, 2019). Furthermore, hierarchical and partitional clustering are the techniques on which most clustering algorithms are based. This study adopts hierarchical and K-means clustering (as a partitional clustering algorithm) methods for defining cluster numbers and clusters within the dataset, respectively.

The K-means clustering algorithm requires the number of clusters to be specified in advance. Thus, hierarchical clustering (the clustering method that doesn't require the cluster number input) and the elbow plot method were applied to find the suitable number of clusters. A hierarchical clustering method is a bottom-up approach where the first algorithm step corresponds to placing each data object in a separate cluster. After that, it finds the nearest data object and merges them, repeatedly executing the algorithm until one big cluster is formed (Oyelade et al., 2019). The number of clusters is then determined based on the hierarchical tree-like structure of the dendrogram that is created after algorithm execution. On the other hand, the method commonly used to determine the suitable number of clusters is the elbow plot. The method performs K-means clustering for an arbitrary range of cluster numbers (K). In addition, it calculates the distortion score, e.g., the sum of squared distances between the data objects in generated clusters and their centroids (Shi et al., 2021). The suitable number of clusters suggested by the elbow plot is the inflexion point on the curve.

The clustering algorithm further used in this study is K-means, which is widely used for unlabelled datasets. It partitions the dataset into K clusters represented by its randomly selected or apriori-derived centroid values. Then, each data object in the given clusters is assigned to the closest centroid in the iterative centroid recalculation process until convergence is achieved (Omran et al., 2007).

In this study, the K-means clustering was performed on two research levels:

1. Grouping CAD models according to the total and unique numbers of CAD modelling operations
2. Grouping the clusters' items from the above statement according to their transition matrices

The number of clusters was determined using the elbow plot method and substantiated by plotting the hierarchical dendrogram for each clustering level. The expected clustering output, thus, was to get the insight to:

1. Average transitions of CAD modelling operations for the clusters of CAD models grouped based on the numbers (total and unique) modelling sequences
2. Average transitions of the CAD modelling operations for the second-level subclusters

Clusters on both clustering levels are then visualized in 2D scatter plots. Clusters' items from the second level represent points in 100-dimensional space. Thus the Principal Component Analysis

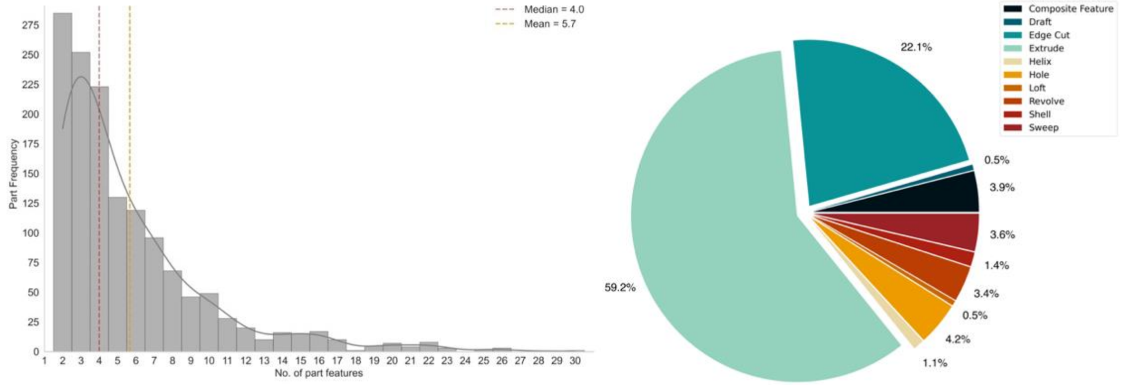
https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

(PCA) is used to reduce the number of dimensions without much loss of information (Jollife and Cadima, 2016).

4 RESULTS

4.1 Dataset characterisation

After dataset cleaning and structuring, CAD models used for the cluster analysis contain a total of 8039 feature-based CAD modelling operations and 6620 operation transitions. As shown in Figure 1(a)., the distribution of the total number of CAD modelling operations per single CAD model is positively skewed. That means most CAD models contain between 2 and 6 feature-based operations with a mean value of 5.7 and a median of 4 operations per CAD model, with two operations being the shortest feature-based design sequence in the dataset (285 or 20% of analysed CAD models). In contrast, the models with the longest feature-based CAD modelling sequence has 30 features. Concurrently, seven is the maximum number of unique operations per CAD model, whereas one unique operation is a minimum. The average and median are 2.3 and 2.0, respectively. Furthermore, the most common operation in the dataset is extrude, which appears in 59.2% of CAD models, followed by the edge cut operation, whose share is 22.1%. Conversely, the dataset's draft and loft operations are the least represented. The dataset operations proportion is shown as a pie chart in Figure 1(b). Furthermore, given the transition between operations in CAD modelling sequences, out of 6620 CAD modelling operation transitions, 74.5% or 4952 transitions falls on transitions between extrude and edge cut (both directions). Additionally, 10% is the share of transitions from extrude to the remaining eight features (composite feature, draft, helix, hole, loft, revolve, shell and sweep, respectively). In comparison, the transitions from the 8 previously mentioned operations to extrude have a share of 8.8%.



(a) (b)

Figure 1. Distribution of CAD modeling operations (a) and their proportion (b)

4.2 Clustering

4.2.1 First-level clustering

Firstly, from the hierarchical agglomerative clustering dendrogram, the elbow plot method was performed on the range number of clusters from k=1 to k=9. Hence, for the given data, the optimal number of clusters obtained from the elbow plot method is three, annotated with a dashed line and shown in Figure 2(b). Also, three clusters are presented in the dendrogram in Figure 2(a). Furthermore, the first-level clusters of CAD models correspond to the range of CAD modelling operation numbers (total and unique).

Therefore, the first cluster (Cluster 1) corresponds to CAD models containing 2 to 5, while the second cluster (Cluster 2) groups CAD models with 6 to 12 operations, thus implying the third cluster (Cluster 3) of CAD models with equal to and more than 13 CAD modelling operations in a sequence. Furthermore, CAD models had been modelled on average using 3.2, 7.9 and 17.4 feature-based operations, whereas using only 1.9, 2.8 and 3.2 unique operations for cluster order, respectively. The number of CAD models in each cluster is 890, 426 and 103.

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

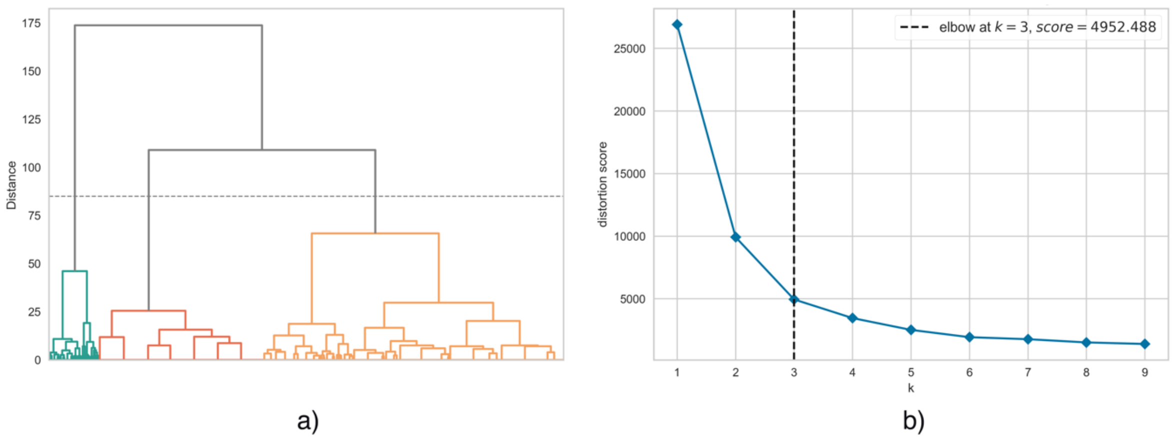


Figure 2. Dendrogram (a) and elbow plot for k=1 to k=9 (b)

To further analyse the patterns of CAD modelling sequences, the average first-order Markov chain transition matrices for each cluster are calculated and visualized as transition diagrams in Figure 3. Each cluster shares the highest probability between extrude operations. Thus, the transition probability between extrude operations is 0.47, 0.55, and 0.62, respectively, for the cluster number order.

Although diagrams reveal a significant equivalence level among clusters by sharing a high probability of transitions between extrude and edge-cut operations, this transition probability also increases as the cluster number increases. Moreover, as the number of operations appearing in the CAD model sequences increases, the average transitions between operations other than exclusively extrude and edge-cut increases. Thus, the following highest probabilities are 0.08 from extrude to hole operation (Cluster 1), 0.13 from sweep to extrude operation (Cluster 2) and 0.2 for the transition from composite feature to extrude (Cluster 3).

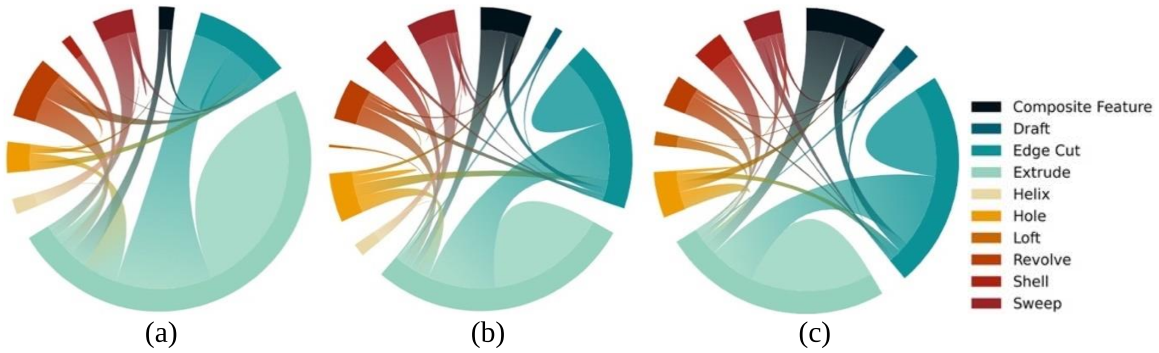
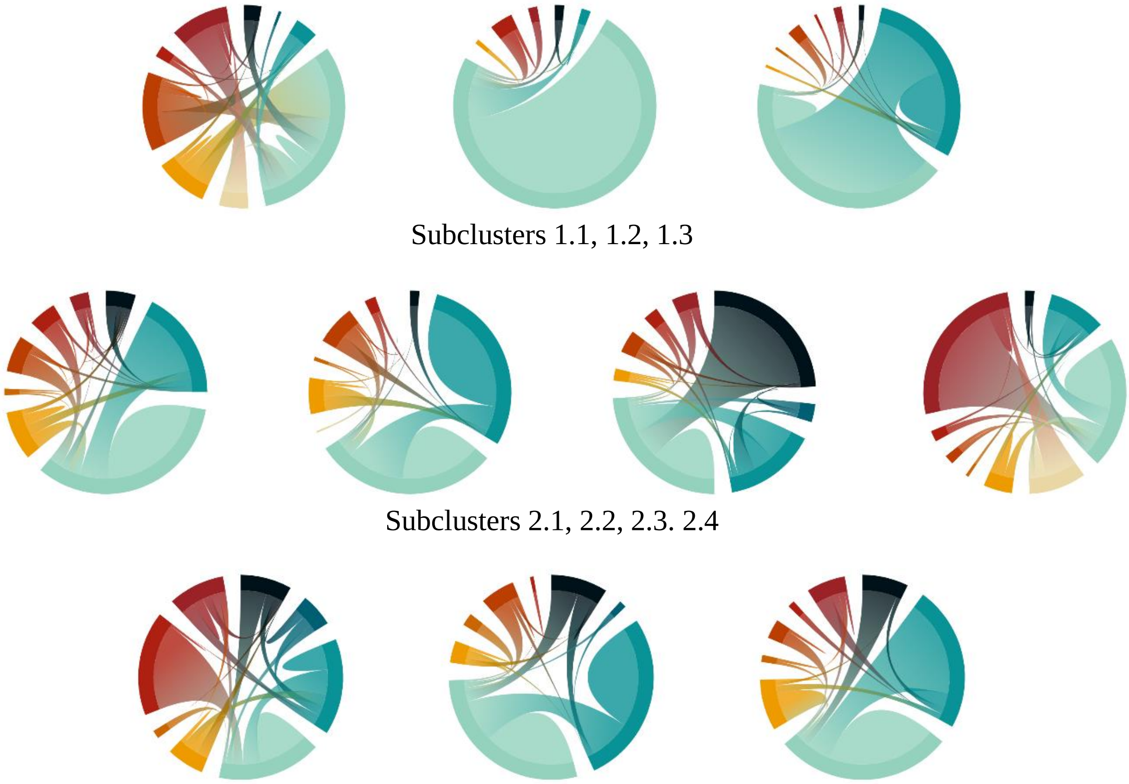


Figure 3. Characteristic transitions between operations for clusters 1 (a), 2 (b) and 3 (c)

4.2.2 Second-level clustering

Further clustering was performed on the transition matrices for the three datasets corresponding to the clusters obtained from the first-level K-means clustering analysis. For the optimal number of subclusters, the elbow plot method and hierarchical dendrogram were used: 3 clusters for Cluster 1 and 3, respectively, and 4 clusters for Cluster 2. The resulting subclusters shown in Figure 4. represent the final groups of CAD models obtained by unsupervised clustering based on transitions of feature-based CAD modelling operations. From the average transition matrices for each subcluster, it can be inferred that clusters share similarities among subclusters but significant differences as well. Regarding the similarity between subclusters, each has groups of CAD modelling sequences with dominant transitions between extrude and combinations of extrude and edge-cut operations (subclusters 1.2, 1.3, 2.1, 2.2, 3.2 and 3.3). On the other hand, regarding differences, there are subclusters with a high value for the average transitions from composite feature to extrude (0.96, subcluster 2.3) and transitions from sweep to extrude (0.84, subcluster 2.4), respectively. Similarly, subcluster 3.1 has a dominant transition between shell and extrude (0.97). Interestingly, the transitions between extrude features follow the mentined transitions according to their value. Finally, in subcluster 1.1, no dominations in terms of transitions of CAD modelling operations can be inferred.

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press



Subclusters 1.1, 1.2, 1.3

Subclusters 2.1, 2.2, 2.3. 2.4

Subclusters 3.1, 3.2, 3.3

Figure 4. K-means clustering plot for ten subclusters (clusters 1, 2 and 3)

5 DISCUSSION

The presented results provide insights into the differences between the characteristic transitions of CAD modelling operations for the obtained clusters of 3D models.

With the help of clustering algorithms, the answer to the first research question has been provided. Namely, three groups of CAD models are obtained using a clustering unsupervised learning algorithm based on the number of operations in CAD modelling sequences. Results suggest that regardless of the modelling sequence length, only a few unique CAD modelling operations are used for building most of the feature-based 3D CAD models. This finding is in line with the statement of Willis et al. that the sketch and extrude are the two most common modelling operations used in approximately 80% of CAD models from their dataset of CAD models created by human designers. Indeed, the here presented data show that if only models with at least two CAD modelling operations are considered, the proportion of the models containing the extrusion feature increases to approximately 88%. They also claim the modelling operations the most used after the sketch and extrude are - fillet and chamfer (edge cut operations) (Willis et al., 2021), which our study has confirmed with the share of 26.8% of the total number of operations used in the dataset.

Furthermore, second-level clustering analysis has answered the second research question. Among the high similarity among groups generated when addressing the first research question, the second-level clustering provided a more thorough analysis that facilitated the differentiation of groups of CAD models with dominant transitions that could not have been noticed in three main clusters. For example, the subcluster with a dominant transition from sweep to extrude operation has emerged (subcluster 2.4), suggesting groups of CAD models with curved geometry segments. Furthermore, transitions from composite feature to extrude (subcluster 2.3) and shell to extrude (subcluster 3.1) also emerged after the second-level clustering. Moreover, the second-level clustering analysis shows a significant difference comparing the transitions of CAD modelling operations between CAD models with a small and larger number of operations. While the subclusters (1.2 and 1.3, respectively) of CAD models with a small number of operations are entirely dominated by transitions from extrude to extrude and edge cut, in the subclusters of CAD models containing more than five operations per model, new transition patterns have emerged. Thus, subcluster 2.3 is dominated by transitions from composite feature to extrude, whereas the transition from sweep operation to extrusion is dominant in subcluster 2.4.

In contrast, despite not dominating any subcluster, the composite feature is exclusively followed by extrude or edge cut operations, suggesting that CAD models within these transitions are created using individual extrude or edge cut operations and then multiplying or mirroring them. By further analysing CAD models with the largest number of operations, the domination of transition from shell operation to extrude is specific for cluster 3 only (subcluster 3.1). Additionally, the specificity of subcluster 3.1., but also the other subclusters, albeit in a smaller proportion, is domination of transition from edge cut to extrude. Simply put, CAD users create the whole model, and then modify the necessary filleted and bevelled edges. Furthermore, the helix to sweep transitions exclusively have the mentioned direction, suggesting that subclusters 1.1 and 2.4 have the CAD models representing spring elements. To summarise the analysis of the transitions, it can be concluded that, although some of the transitions among operations have been expected or assumed (extrude to extrude, extrude to edge cut or helix to sweep), the study has empirically confirmed the assumptions on a larger number of CAD models.

Furthermore, the study has also provided insight into the potential of using sequences of feature-based CAD modelling operations. The insights for using CAD modelling operations beyond extrude, sketch, and sweep is scant; hence, this study has proven that the fuller array of operations can be used for research work that aims to automate the CAD modelling process. Compared to the work using the association rule learning to predict the hole feature by Vasantha et al. (Vasantha et al. 2021), performing the clustering unsupervised learning methods on the first-order Markov model transition matrices of the CAD modelling operations has also shown the potential of being the first step toward automating the user interaction with the CAD system. Indeed, the three identified groups (clusters) of models with a different number of total and unique modelling operations, as well as the ten identified subgroups (subclusters) show that there exist characteristic sets of sequences for different target geometries when 3D modelling. The sequences imply rules on which the automation of the user interaction with CAD can be based.

6 CONCLUSION

The study has attempted to provide insight to different groups of 1419 CAD models containing ten unique modelling operations by using the unsupervised learning technique of clustering. The CAD models have been clustered on two levels, performing the algorithm on a total and a unique number of CAD modelling operations and first-order Markov model transition probability matrices for sequences of CAD modelling operation. As a result, three CAD groups (clusters) emerged in the first clustering level, corresponding to the number of CAD modelling sequences, whereas the second clustering level refined the clusters and provided ten subclusters. The obtained clusters and subclusters have given insight into the characteristic average transitions between CAD modelling operations.

However, the next steps need to be defined due to the study's limitations, which are also guidelines for future work. First, despite the fuller array of feature-based operations used in the study analysis, sketch features and the removal and addition of the 3D shape representation material need to be included in future analyses. In addition, the first-order Markov model states depend only on the present state but not the preceding states in the sequence of CAD modelling operations. Thus, the further implementation of statistical models used to describe the evolution of observable events depending on internal factors (e.g., the Hidden Markov model or deep learning techniques) needs to be considered as the study's extension. Furthermore, despite more than 1000 CAD models humans have created, we aim to broaden their number and scope (e.g., using data provided by the professionals) with the overall objective of accurately predicting the next CAD user's step while performing feature-based CAD modelling.

REFERENCE

Bonino, B., Raffaeli, R., Monti, M. and Giannini, F. (2021), “A heuristic approach to detect CAD assembly clusters”, Procedia CIRP, Vol. 100, Elsevier B.V., pp. 463–468.

Cantamessa, M., Montagna, F., Altavilla, S. and Casagrande-Seretti, A. (2020), “Data-driven design: The new challenges of digitalization on product design and development”, Design Science, Cambridge University Press

Celjak, R., Horvat, N. and Skec, S. (2022), “Exploring the Potential of Tracking CAD Actions in Project-based Courses”, CAD’22 Proceedings, CAD Solutions, LLC, pp. 302–307.

Garland, M., Willmott, A. and Heckbert, P.S. (2001), “Hierarchical Face Clustering on Polygonal Surfaces”, Proceedings of the 2001 Symposium on Interactive 3D Graphics - SI3D ’01, pp. 49–58.

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press

Hamerly, G. and Elkan, C. (2002), “Alternatives to the k-means algorithm that find better clusterings”, Proceedings of the Eleventh International Conference on Information and Knowledge Management - CIKM ’02, pp. 600–607.

Han, Z., Mo, R. and Hao, L. (2019), “Clustering and retrieval of mechanical CAD assembly models based on multi-source attributes information”, Robotics and Computer-Integrated Manufacturing, Elsevier Ltd, Vol. 58, pp. 220–229.

Hoffmann, C.M. (1989), Geometric and Solid Modeling: An Introduction, 1st ed., Morgan Kaufmann Pub, San Mateo, California.

James Yu-Hsien Chen, by, Olechowski, A. and Yu-Hsien Chen, J. (2021), Development of a Novel Computer-Aided Design Experiment Protocol for Studying Designer Behaviours, University of Toronto.

Jayaraman, P.K., Lambourne, J.G., Desai, N., Willis, K.D.D., Sanghi, A. and Morris, N.J.W. (2022), “SolidGen: An Autoregressive Model for Direct B-rep Synthesis”

Jollife, I.T. and Cadima, J. (2016), “Principal component analysis: A review and recent developments”, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, Royal Society of London, 13 April

Katz, S. and Tal, A. (2003), “Hierarchical Mesh Decomposition using Fuzzy Clustering and Cuts”, SIGGRAPH ’03, pp. 954–961.

Li, C., Pan, H., Bousseau, A. and Mitra, N.J. (2020), “Sketch2CAD: Sequential CAD modeling by sketching in context”, ACM Transactions on Graphics, Association for Computing Machinery, Vol. 39 No. 6

Machchhar, R.J. and Bertoni, A. (2021), “Data-driven design automation for product-service systems design: Framework and lessons learned from empirical studies”, Proceedings of the Design Society, Vol. 1, Cambridge University Press, pp. 841–850.

Murtagh, F. and Contreras, P. (2012), “Algorithms for hierarchical clustering: an overview”, WIREs Data Mining and Knowledge Discovery, Vol. 2 No. 1, pp. 86–97.

Omran, M.G.H., Engelbrecht, A.P. and Salman, A. (2007), “An overview of clustering methods”, Intelligent Data Analysis, IOS Press, Vol. 11 No. 6, pp. 583–605.

Onshape. (n.d.). “Developer Portal API”, available at: https://onshape-public.github.io/docs/ (accessed 4 December 2022).

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghiren, E., et al. (2019), “Data Clustering: Algorithms and Its Applications”, Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019, Institute of Electrical and Electronics Engineers Inc., pp. 71–81.

Peabody, M. and C. Regli, W. (2001), Clustering Techniques for Databases of CAD Models, Philadelphia.

Pedley, A.G. (1997), User Defined Feature Modelling: Representing Extrinsic Form, Dimensions And Tolerances.

Rahman, M.H., Schimpf, C., Xie, C. and Sha, Z. (2019), “A computer-aided design based research platform for design thinking studies”, Journal of Mechanical Design, Transactions of the ASME, American Society of Mechanical Engineers (ASME), Vol. 141 No. 12

Regli, W.C. (1995), Geometric Algorithms for Recognition of Features from Solid Models.

Roj, R. and Woyand, H.-B. (2015), An Examination of Engineering Parts in Large CAD-Databases in Order to Create Adjacency Matrices and Build Clusters.

Salomons, O.W., van Houten, F.J.A.M. and Kals, H.J.J. (1993), Review of Research in Feature-Based Design, Journal of Manufacturing Systems, Vol. 12.

Sharma, G., Goyal, R., Liu, D., Kalogerakis, E. and Maji, S. (2017), “CSGNet: Neural Shape Parser for Constructive Solid Geometry”

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H. and Liu, J. (2021),“A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm”, Eurasip Journal on Wireless Communications and Networking, Springer Science and Business Media Deutschland GmbH, Vol. 2021 No. 1

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H. and Sui, F. (2018), “Digital twin-driven product design, manufacturing and service with big data”, International Journal of Advanced Manufacturing Technology, Springer London, Vol. 94 No. 9–12, pp. 3563–3576.

Uy, M.A., Chang, Y., Sung, M., Goel, P., Lambourne, J., Birdal, T. and Guibas, L. (2021), “Point2Cyl: Reverse Engineering 3D Objects from Point Clouds to Extrusion Cylinders”

Vasantha, G., Purves, D., Quigley, J., Corney, J., Sherlock, A. and Randika, G. (2021), “Common design structures and substitutable feature discovery in CAD databases”, Advanced Engineering Informatics, Elsevier Ltd, Vol. 48

Willis, K.D.D., Pu, Y., Luo, J., Chu, H., Du, T., Lambourne, J.G., Solar-Lezama, A., et al. (2021), “Fusion 360 Gallery: A Dataset and Environment for Programmatic CAD Construction from Human Design Sequences”, ACM Transactions on Graphics, Association for Computing Machinery, Vol. 40 No. 4

Wu, R., Xiao, C. and Zheng, C. (2021), “DeepCAD: A Deep Generative Network for Computer-Aided Design Models”

https://doi.org/10.1017/pds.2023.94 Published online by Cambridge University Press