Erroneous Answers Categorization for Sketching Questions in Spatial Visualization Training

Tiffany Wenting Li
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801
wenting7@illinois.edu

Luc Paquette
Department of Curriculum & Instruction
University of Illinois at Urbana-Champaign
Champaign, Illinois 61820
lpaq@illinois.edu

ABSTRACT

Spatial visualization skills are essential and fundamental to studying STEM subjects, and sketching is an effective way to practice those skills. One significant challenge of supporting practice using sketching questions is the vast number of possible mistakes, making it time-consuming for instructors to provide customized and actionable feedback to students. The same challenge persists for computer programs as well. This paper introduces a clustering model designed to categorize sketching answers based on the severity and characteristics of their mistakes. The model is designed to be used by a computer-based training platform to provide customized, actionable formative feedback to students in real-time. The promising results also suggest a new and comprehensive set of evaluation criteria to assess a student's performance on sketching questions. As a broader contribution, our work is a proof-of-concept for a modeling approach to automatically evaluate and provide formative feedback on complex free-hand sketches using abstract features that may be generalized to a variety of disciplines that involve the creation of technical drawings.

Keywords

Automatic grading, Sketching, Clustering, Spatial Visualization, Formative feedback

1. INTRODUCTION

Spatial visualization is the ability to represent and mentally manipulate two-dimensional and three-dimensional objects [11]. A body of research has shown that good spatial visualization skills help students succeed in STEM education [39, 3, 13, 25, 27, 32, 41, 44]. It is encouraging that existing research also demonstrates that spatial visualization skills are malleable and can be trained and improved, for example, via forms of workshops and seminars [42]. There have been successes in increasing the retention rates of STEM freshmen students with spatial visualization skills training in recent years, especially for minority groups such as female students [39, 23].

Besides multiple-choice questions that are traditionally used in spatial visualization training, free-hand sketching on grid paper is an effective type of practice question [38]. Sketching questions can imitate the sketching tasks required in many engineering disciplines, which is particularly helpful since sketching is a fundamental skill for engineering designs [22]. In the training process, since students gain from learning from their mistakes instead of failing in the first try and giving up based on the immediate-feedback assessment technique [26], students can benefit from having a second chance on a practice problem. However, providing formative feedback while not giving away the answer, which is known to support self-regulated learning [28], on free-hand sketching can be challenging due to the wide variety of possible incorrect answers on such activities.

While human instructors possess the capability to analyze an erroneous free-hand sketch, identify the source of potential errors and provide formative feedback, it is a time-consuming process and providing such feedback to a large student population would require prohibitive efforts that would likely prevent the feedback from being provided in a timely fashion [2]. Computer-based systems able to provide timely formative feedback can be considered as an alternative to address this limitation. However, one significant challenge to automatically providing immediate customized feedback for sketching questions is the need for a computer-based system to be able to recognize and understand how much an answer is different from the answer key and the types of mistakes students are making.

On the one hand, sketching questions have an enormous number of possible incorrect answers, which are often specific to a unique problem, making it difficult, if not impossible, to identify every possible error and to prepare unique feedback for each one. As an alternative, a computer-based system could be designed to recognize categories of answers based on the severity or characteristics of their errors and provide feedback relevant to each one. However, to the best of our knowledge, there is no existing research that categorizes answers to complex sketching questions based on their errors, either conceptually or computationally. The lack of solution motivated us to identify patterns that exist in students' erroneous sketching answers and create a computer-based algorithm that can categorize them in real-time.

Tiffany Wenting Li and Luc Paquette "Erroneous Answers Categorization for Sketching Questions in Spatial Visualization Training" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020),* Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 148 - 158

Due to the lack of existing categories of erroneous answers in free-hand sketching problems, we propose the use of a clustering approach to identifying such categories. Our research questions are the following:

RQ1 What categories exist in students' sketching answers based on the severity and characteristics of their errors?

RQ2 How meaningful are the identified erroneous answer categories, and what actionable feedback can be provided for each category?

We constructed a list of features that can be used to characterize students' erroneous sketching answers. Using a k-mean clustering approach, we discovered six common answer categories for incorrect sketches that are distinct from one another according to the severity and characteristics of the errors. Our clustering results suggest a new set of evaluation criteria for complex free-hand sketching answers that is more interpretable and generalizable than those in prior work [7, 43, 5]. Also, we provide initial suggestions for the kinds of formative feedback appropriate for each answer category without giving away the answer [36].

To the best of our knowledge, our study is the first to identify categories of erroneous sketches, both computationally and conceptually, in spatial visualization sketching problems using abstract features. Our approach also has the potential to be generalized to other subject areas that require sketching practices, mostly technical drawings in various Engineering and Science subjects, such as circuit diagrams in Electrical Engineering, engine models in Mechanical Engineering, building plans in Architecture, and structural formula in Organic Chemistry.

2. RELATED WORK

2.1 Spatial Visualization Skills and Sketching

Spatial visualization skills were estimated to play an important role in 84 careers [37], most of which are STEM-related. A longitudinal study showed that psychometrically-assessed spatial ability predicts career in STEM fields after accounting for Math and Verbal aptitudes [45].

Spatial visualization skills are applied in various STEM areas. Research shows that students with better spatial visualization skills perform better in Chemistry [32, 6]. In Organic Chemistry, for example, students with strong spatial visualization skills draw preliminary figures more often. Hence they use figures to gain a better understanding of the questions and are more likely to answer them correctly [32]. Another body of research revealed the connection between spatial skills and Geoscience [17, 30]. In particular, students with strong visual penetration ability, e.g., imagining cross-sections, perform better in Geology [17]. Furthermore, understanding cross-sectioning is a basic skill in many other engineering subjects [9, 12]. Spatial visualization is also found to be tightly related to performance in Anatomy in Biology [34], Radiology in Medicine [16].

A wide variety of empirical research has shown that spatial visualization skills are malleable. Interventions designed to improve spatial visualization skills reach, on average, a medium effect size of 0.47 [42]. A well-known training developed by Sorby (2009) showed significant post-test improvement for each class of college students over a 6-years-long study. In particular, Sorby found that the training significantly improved female students' retention rate but not that of male students [39]. The finding suggested the critical role of spatial visualization skills training in increasing the diversity of STEM field students.

Sketching ability is fundamental to engineering design [22] and highly correlates with many STEM subjects [35]. To improve spatial visualization skills, sketching is one of the most effective approaches [38]. Electronic sketching has also demonstrated potential in training spatial visualization skills [8, 47]. Thus, the application of sketching practice is worth studying for better improving spatial visualization skills.

2.2 Computer-based Evaluation and Formative Feedback for Sketches

To the best of our knowledge, there is no prior work on the evaluation of sketches in spatial visualization training, both conceptually or computationally. The use of computer-based formative feedback for spatial visualization sketching has not been studied either. There is a body of research on computer-based evaluation and formative feedback for other types of sketches [5, 7, 43, 40, 15, 18, 19, 20]. However, some of them are too simple or too domain-specific to be generalized to a complicated case as in spatial visualization sketches. Others' evaluation methods cannot provide actionable or easy-to-interpret formative feedback.

For free-hand sketching that is evaluated mostly based on the shape and structure, there are a few existing evaluation approaches in domains other than spatial visualization training. Bhat (2017) developed Skechography, a river-sketching auto-grading tool for Geology [5]. This tool could perform sketch recognition and compare the river's shape similarity using the Shape Context algorithm, the distances of start points and endpoints between a student's answer and the answer key. Based on the degree of similarity and distances, the tool provided a score that was a weighted sum of these three features. Skechography evaluated a river, which had only one line with specific features of a start point, an endpoint, and the shape of the line. The simplicity of this application has a weak external validity and cannot be used in evaluating spatial visualization sketches.

The work by Chandan et al. (2018) [7], on the other hand, worked on a complicated case of free-hand drawing of objects of specific categories, e.g., a bee, an airplane, etc. They applied a Convolutional Neural Network approach for object categorization and a Scale Invariant Feature Transform approach to check the similarity between a given sketch and the "standard" sketch. As feedback, the tool showed the percentage of similarity to various categories of objects. The use of deep learning methods made the interpretation of results challenging. Hence, this approach is limited in its capability to generate specific and actionable feedback to help students improve their answers.

Mechanix, a sketch-based tutoring system for learning forces applied on a truss, could provide specific feedback to free-

hand sketching of forces [43]. In this case, the errors that could occur were known and clearly defined on an arrow-basis. Given the small number of arrows, it is relatively easy to cater specific and actionable feedback to each error. In the case of spatial visualization sketches, a sketch contains far more number of lines, making it infeasible to provide a piece of feedback for each line.

There exists another body of work that focused on the recognition of East Asian characters, which are similar to a simple sketch [40]. However, these solutions applied an "all or nothing" approach to recognize the structure of a character, which was not helpful in providing specific formative feedback. A few other works aimed to evaluate and provide feedback on the quality or aesthetics of a sketch, but not on the correctness in terms of the structure of shape [15, 18]. There is also an evaluation approach for computer-aided design solid models specifically, using criteria related to parameters set in the computer-aided model, which does not apply to free-hand sketching because the concept of parameters is not intuitive in free-hand sketching [19, 20].

Overall, there is limited work on a computer-based evaluation of complex free-hand sketching based on structural correctness that can generate specific and actionable formative feedback. Our work aims to fill in this gap.

2.3 Answers Categorization in Content-based Automated Evaluation

In evaluating constructed response automatically from a content-based perspective, there is a rich body of work in evaluating short answer questions for a variety of subjects and domains [24]. However, except for the studies mentioned in the last section, there is very few existing literature related to the content-based evaluation of complex free-hand sketching. Therefore, we draw our inspiration from the existing research in evaluating short answer questions and apply it to complex free-hand sketches, a different type of constructed response.

Answer categorization is one of the most frequently used approaches to perform a content-based evaluation of short answers. In most cases, supervised learning is applied using a manually labeled training set based on pre-defined rubrics [21, 33, 1, 10, 29]. For example, c-rater applied NLP techniques that determined whether an answer contained each key concept and was widely applied on short answer questions in Biology, Psychology, Math, and Reading, to not only grade but to provide specific real-time feedback [21, 1]. Pulman and Sukkarieh (2005) experimented with Inductive Logic Programming, Decision Tree and Naive Bayes to classify short answers into the desired category for Biology [33].

In our case, however, there are neither pre-existing robust rubrics as the evaluation standard for spatial visualization sketches nor known categories of error. This brought difficulties to label a training set manually accurately. Also, most content-based evaluation approaches only provided up to three levels of scoring. Some exceptions that provided more than three levels of scoring were either unclear about the definition of the levels or the levels were only mechanical composition of the correct answer [24]. As an alternative,

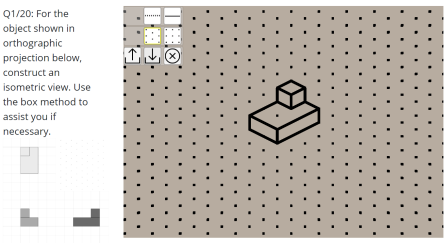

Figure 1: Free-hand sketching tool for isometric sketching on the online spatial visualization training platform

we turned to unsupervised learning to perform answer categorization to identify categories that were as granular yet meaningful as possible. Clustering is an often-used unsupervised learning approach in short-answer grading, especially in the case of answering open-ended questions. Previous work [4, 48] has shown that clustering could group answers that are similar in text characteristics, semantics, and topics. Our work aims to leverage this method to categorize complex sketches in spatial visualization training.

3. METHODS

3.1 Data Collection

We collected data from students solving free-hand sketching problems in a 100-level engineering course called "Spatial Visualization" that utilized an online training platform over half a semester in Fall 2019 at our home institution, a large public university in the Midwestern United States. The online training platform was previously developed as a computer-based spatial visualization training platform [47] to enable practicing at scale using online exercise and automatic grading. Previous work has shown a significant improvement in spatial visualization skills for those who completed the exercises on the platform [47].

Students in the course met once a week in-person for an hour, and the majority part of the course was working through practice problems on the platform on their own as their weekly assignment, given the instructions. The focus of practice questions each week was different, depending on the particular set of skills that were being trained, such as mental rotation, cross-sectioning, and coded plan. The platform supports both multiple-choice questions and sketching questions. Figure 1 and Figure 2 show the free-hand sketching tool on the platform that allows students to sketch out their answers on the computer. Students can draw and erase lines on the grid paper freely. Students could also save their sketch when they leave the platform and load what they saved when they come back. In the course, students were given a maximum of two attempts for each sketching question, i.e., they were given a second chance if they answered incorrectly in the first attempt. All the sketching questions were graded with an "all or nothing" approach.

The collected dataset includes 370 incorrect sketches from 14 students in the course that covers five types of sketching questions and 61 unique questions. We excluded correct sketches in the categorization because they would naturally

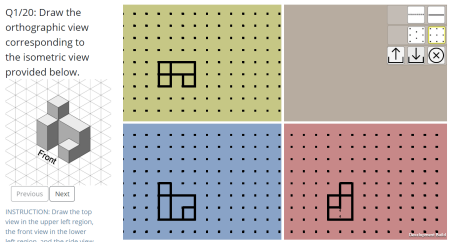

Figure 2: Free-hand sketching tool for orthographic sketching on the online spatial visualization training platform

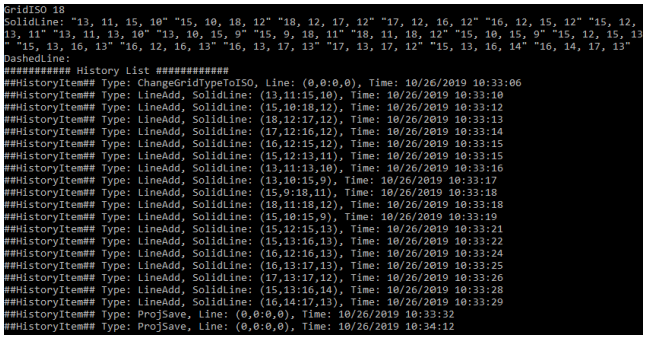

Figure 3: An example of a raw log file generated from sketching questions on the online spatial visualization training platform

be in one category by mapping exactly to the answer key. Examples of the types of sketching questions include drawing the orthographic view of a 3D object given the isometric view or vice versa, and drawing the resulting 3D object after rotating a given 3D object with a certain degree in a given direction. Each type of sketching questions contained a series of different questions with 3D objects of various shapes. On average, each sketch contains approximately 30 to 80 lines of unit length.

Each submission of an attempt to answer a question produced a raw log describing their answer. In the raw log, two major types of information were recorded. First, it contained the set of lines in the final submitted sketch. Second, it recorded the history of all the timestamped steps a student took of adding or deleting a line, clearing, or loading the sketch for that question (Figure. 3). In this paper, we focused on the final submitted sketch only since the goal is to categorize the final answer instead of analyzing students' process of solving a free-hand sketching problem.

Each final submitted sketch is represented by the X-Y coordinates of a list of lines. The lines are further denoted by the type of the lines, either solid line or dashed line, which are the two standard types of lines used in the sketching exercise for different purposes. A sketch is mostly made up of solid lines, but a dashed line should be used instead of a solid line to represent a hidden edge from a particular perspective.

Another data point in the raw log is the type of grid paper used for a sketch. There are two types of grid paper in the sketching exercises: an isometric grid for isometric drawing, and a dot grid for orthographic drawing. A sketch is considered as correct only if the shape and the size of the object match with those of the answer key, and uses the correct type of grid paper. The position of where a sketch is drawn on the grid paper is flexible.

We performed two steps of data standardization on the raw log before feature extraction. First, we aligned both the student's answer and the answer key to the lower-left corner of the sketch-pad. Second, all the lines were broken down into unit length and de-duplicated so that lines that overlapped with each other would only be counted once. We conducted these two steps for the ease of comparing student's answers against the answer key.

3.2 Feature Extraction

We developed a total of 8 features to use as input for our clustering model. We performed feature engineering manually after observing a small subset of the data to get an idea of what information human instructors might use when interpreting incorrect answers. In order to get a preliminary view of possible errors that would be as comprehensive as possible, we selected three questions that had the highest number of incorrect answers and observed the errors made by students on those problems. Based on our preliminary observation, we created three categories of features that represent different characteristics of the observed errors.

The first group of features uses a unit-length line as its basic unit, i.e., a line connecting adjacent points, and represents the number of lines that are wrong compared to the answer key. We observed from the subset of mistakes that the number of incorrect lines involved in a sketch varied widely, from only one wrong line to over 80% of lines being wrong. The number of incorrect lines is a straightforward way to quantify the degree to which a sketch was incorrect. We considered three scenarios in which a line is wrong.

1. **An extra line:** a line is in the student's answer, but there is no line at the same position in the answer key.
2. **A missing line**: a line is in the answer key, but there is no line at the same position in the student's answer.
3. **A line with incorrect type:** two lines with the same position in the student's answer and the answer key are of different types, i.e., solid line vs. dashed line.

To normalize the number of incorrect lines against the complexity of the sketch, we adopted the percentage of wrong lines instead of the absolute number, i.e., dividing by the total number of lines in a sketch. The three features in this group are Percentage of Extra Lines, Percentage of Missing Lines, and Percentage of Lines with Correct Position but Incorrect Type.

The second category of features represents the groupings of the incorrect lines based on their location in a sketch. In our preliminary observation, we found that, between two

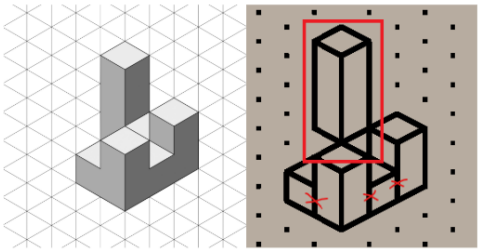

Figure 4: An example of a sketch (on the right) with four error components, i.e., four sites of mistakes. The sketch on the left is the answer key.

sketches with a similar number of incorrect lines, the incorrect lines may be inter-connected and concentrated in one place in a sketch while being scattered in multiple spots in another sketch. These two cases represented the mistakes of different natures.

Based on the assumption that incorrect lines that are connected are more likely caused by the same mistake, we treated all the incorrect lines as an undirected graph and defined each component in the graph as one "site" of mistake. A component here has the same definition of a component in an undirected graph, a subgraph in which any two vertices are connected by paths, and which is connected to no additional vertices in the supergraph [46]. As an example, in Figure 4, there are a total of four error components in the sketch, three extra lines in different locations, and a disconnected taller stack separated from the bottom of the object.

We constructed three features in this category. The first feature is the number of components in the graph made of incorrect lines, which is a representation of the number of mistake sites in a sketch. Since the size of a component represents how severe a mistake is, the second feature is the average size of all the error components in a sketch. The larger the average component size is, the more severe the mistakes are on average. The last feature is the maximum size difference among all error components, which reflects the range of severity across multiple mistake sites in a sketch.

The last set of features describes the general characteristics of the sketch. One feature is whether the student uses the same type of sketching grid as the answer key. Another feature is whether the sketch is empty. If it is empty, it indicates either the student did not attempt the question or accidentally skipped the question.

3.3 Model Construction

As there was no prior framework or knowledge on how to categorize the erroneous sketches, it was not possible to obtain labels (ground truth) describing each answer. As such, we used an unsupervised clustering algorithm to identify categories of erroneous answers from existing data. Based on prior observation of the data, we hypothesized that the features of each cluster should have a sphere-like shape. Therefore, we used k-means clustering with squared Euclidean distance. The algorithm aims to assign all the data points into a specified number of clusters such that every data point is in the cluster with the nearest mean. Ideally, data points that have similar values across all the features are grouped in one cluster.

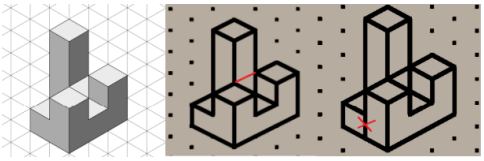

Figure 5: Examples of mistakes in Cluster 0, having one minor mistake. The sketch on the left is the answer key.

After feature extraction, we performed further data normalization as the first step of model construction. Since the k-means clustering algorithm is sensitive to the scale of the features, we normalized each of the three features (Number of Components, Average Size of Components, and Maximum Difference between Size of All Components) into the unit interval respectively across all data, so that they were on the same scale as the other features that were either in percentages or in a boolean format.

We performed parameter tuning to decide on the optimal number of cluster k. We started with two clusters and repeatedly increased the number of clusters by one. We evaluated the choice of k using two criteria. The main criterion we used to evaluate the quality of the clustering results was how interpretable a new cluster was and whether it could help us provide more specific and actionable feedback. Another complementary criterion for evaluation was the Silhouette score, measuring the quality of the clusters based on the cohesion of the separation of the identified clusters (Silhouette score ranges from -1 to 1). We valued the interpretability of a cluster over a higher Silhouette score. Therefore, as long as the Silhouette score remained at an acceptable level, we increased k until the interpretation of the newly generated cluster did not make sense or did not differ much from the existing clusters.

4. RESULTS

Our clustering approach identified a set of six clusters related to categories of erroneous answers in free-hand sketching problems, as listed in Table 1. The 6 clusters are ordered based on the severity of the errors in the table. The clustering model yields a Silhouette score of 0.6659, which is a reasonable value.

Cluster 0 is the most common cluster in the dataset. From the centroid value, we can see that the sketches in this cluster only have one mistake (Number of Component = 1) with about two incorrect lines (Avg Component Size = 1.89). The centroid values suggested that a large portion of the errors had only one minor mistake, which was most likely due to drawing errors such as forgetting an edge at the corner, or drawing an extra edge on a plane (see examples in Fig 5).

Cluster 1, the second-largest cluster in the dataset, differs from Cluster 0 mainly by the number of mistakes in the sketch. On average, there are 2.21 mistake components in

Cluster ID	Cluster Size	Interpretation	Perc Missing	Perc Extra	Perc Type	Num Comp	Avg Comp Size	Max Size Diff	Same Grid?	Empty?
0	218	Have one minor mistake	2.39%	2.16%	0.04%	1.00	1.89	0.00	1.00	0.00
1	65	Have more than one minor mistakes	4.13%	10.46%	0.11%	2.14	2.88	1.43	1.00	0.00
2	30	Have both major and minor mistakes, mostly minor mistakes	20.61%	32.14%	0.29%	3.70	5.13	5.63	1.00	0.00
3	15	Have both major and minor mistakes, mostly major mistakes	37.82%	22.46%	0.77%	2.80	10.69	15.73	1.00	0.00
4	39	More than half of the sketch as a whole is completely wrong	80.08%	67.04%	0.00%	1.05	45.35	0.08	1.00	0.00
5	3	Empty sketch	100.00%	0.00%	0.00%	1.67	29.78	1.00	0.33	1.00

Table 1: Clustering Results Summary Table: The size, interpretation and centroid of each cluster are shown in the table. The centroid values are transformed back to its original scale if unit normalization was performed. Values are color-coded with different shades of red, representing low values to high values)

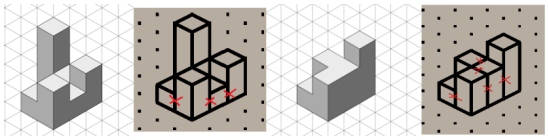

Figure 6: Examples of mistakes in Cluster 1, having multiple minor mistakes. The sketches with a white background are the answer keys.

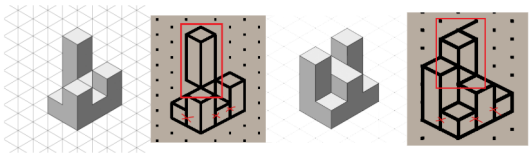

Figure 7: Examples of mistakes in Cluster 2, having multiple minor mistakes and a small number of major mistakes. The sketches with a white background are the answer keys.

the sketch. The average size of 3.04 lines of the components suggests that these are still minor mistakes with three incorrect lines on average. It is reasonable to interpret Cluster 1 as sketches that have several minor mistakes. Examples of this category are shown in the examples in Fig 6. Even though both Cluster 0 and Cluster 1 contain minor errors, they are different enough because students in Cluster 0 make one small mistake likely due to being careless. In contrast, those in Cluster 1 may have misconceptions that are causing a series of mistakes.

Cluster 2 and 3 are quite different from Cluster 0 and Cluster 1. Both of them have a much higher Percentage of Missing Lines and Percentage of Extra lines compared to Cluster 0 and 1, suggesting more severe mistakes in the sketch. More severe errors are more likely to be due to an incorrect structure at specific parts of the sketch rather than careless mistakes. These two clusters both have a high number of components (3.70 and 2.80 for Cluster 2 and 3 respectively), suggesting a series of mistakes across the sketch. Cluster 2 and 3 are different in two perspectives. First, Cluster 2's average component size is small (5.13), while Cluster 3's average component size is a lot bigger (10.69). Second, Cluster 3 has a massive difference in size across the different components (15.73), while Cluster 2 has a medium difference of 5.63. These differences suggest that within the series of mistakes in a sketch in Cluster 2, more of them are minor, and there is only a small proportion of major mistakes, as shown in Figure 7. On the other hand, a sketch in Cluster 3 has mainly major mistakes and fewer minor mistakes, as shown in Figure 8. The major mistakes in Cluster 3 are also more severe than those in Cluster 2 on average.

Cluster 4 has 80% of the lines missing and 67% extra lines, a lot higher than the previous clusters. Interestingly, most of the sketches in this cluster have only one component in their mistake (1.05 components on average), with an average size of 45.35 lines. These features suggest that there is one substantial mistake that spans over half of the sketch, which is often due to either an utterly wrong structure or a wrong orientation. For example, both examples in Fig 9 have the correct structure but wrong orientations.

Lastly, Cluster 5 contains empty answers, either due to the student not attempting a question or accidentally skipping it. Even though the cluster size is small, with only 3 data points due to the low number of empty answers, it is distinct enough from all the other clusters to be on its own.

Overall, we considered the erroneous answer categories detected to be intuitive and well-defined. They are distinct in the severity and characteristics of the mistakes. Being able

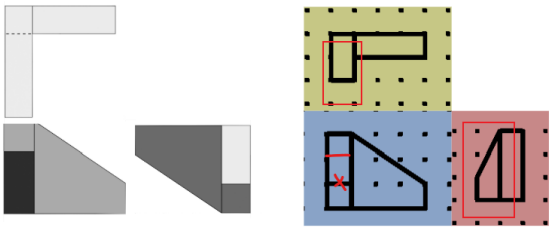

Figure 8: An example of mistake in Cluster 3, having multiple major and minor mistakes, but mainly major mistakes. The sketch on the left is the answer key.

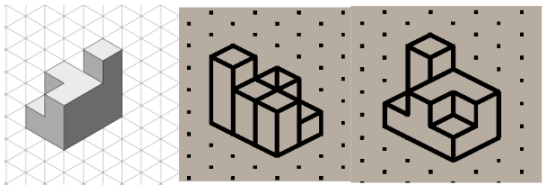

Figure 9: Examples of mistakes in Cluster 4, having one huge cluster of mistake. The sketch with a white background is the answer key.

to automatically identify six categories of erroneous answers demonstrated the potential advantage of using an unsupervised approach in answer categorization than a supervised learning approach that tries to align the model capability with human judgment of the answer categories, which could often only yield up to three clearly defined categories [24]. Additionally, we did not observe any significant difference between the frequency distribution of the error categories across the different types of questions in our dataset, i.e., the frequency of each answer category did not differ significantly across all five types of sketching questions, suggesting the generalizability of the error categories to more variety of questions.

5. DISCUSSION

5.1 Evaluation Criteria for Sketching

Due to the lack of prior work on erroneous answer categories in complex free-hand sketching problems, there is no currently available set of criteria to evaluate the degree of correctness of a complicated sketching answer. In multiple offerings of the spatial visualization training in the past in our school, an instructor either used an "all or nothing" evaluation approach, or used a subjective standard on one or two dimensions to judge a sketch, e.g., taking off 0.5 point for each missing or extra line up to a maximum of 1 point, taking off 1.5 points any time when not all features of the top, front, and right sides are correct. These evaluation schemes are too coarse to reflect the degree of correctness of a sketch accurately. The results of our clustering analysis provide promising results towards the development of a more comprehensive view on how to evaluate a sketch using a scale of multiple levels.

Our model demonstrated that more than one dimension is needed concurrently to provide a nuanced interpretation of the state of a sketch. In our model, the percentages of missing, extra lines or lines with the wrong type, the number of mistakes sites, the average size of the mistakes, and how different the various mistakes sites are in a sketch are used in combination with one another to determine the degree of correctness and the type of errors. For example, a distinction between Cluster 2 and 3 suggests that with a similar percentage of incorrect lines, the number of mistakes components and the average size of the components brings additional insights into whether a sketch contains a large number of minor mistakes or a small number of major mistakes. As another example, even though Cluster 0 and Cluster 1 have a similar average size of mistakes, the number of mistake sites suggests that students in Cluster 1 may have a more systematic misconception than those in Cluster 0 who likely commit a mistake due to carelessness.

Our approach could also be used to define minor mistakes versus major mistakes in a sketch for a group of sketching questions with similar size and complexity. Without a systematic review of all the mistakes in a group of sketching questions, it is hard for an instructor to draw an objective line between an error that is significant and one that is not. As a result, the evaluation criteria may be overly strict or overly generous. The clustering model computationally categorizes what it considers as minor and major mistakes based on the optimal separation principle. Its outcome can serve as analytical support for an instructor's grading decision.

5.2 Potential Intervention

Since one of the motivations to construct this model is to provide real-time, customized, and actionable formative feedback, we propose potential customized intervention messages for each erroneous answer category. Based on the best practices of offering formative feedback [36], each of the messages follow a similar structure of (1) first letting the student know how far they are from the correct answer, (2) describing what types of mistake there are, and (3) suggesting ways for the student to approach solving the errors. A summary of the interventions is provided in Table 2.

Students having answers that fall into Cluster 0 or Cluster 1, which consist of having one or more minor errors, understand what the object should look like structure-wise. When the system tells them that they are wrong, they may find it confusing since they are likely confident in their answer. Hence, the feedback message could first assure the students that they have got the general structure of the object correct. Then, the system could let the students know that they have X number of minor mistakes, where X is the feature Number of Components. The feedback may also include whether they have some missing lines, extra lines, or lines of the wrong type. Lastly, the feedback message would suggest the students check for details of their drawing by listing out the common reasons for such errors, such as extra edges on a flat plane, missing edges at a corner.

If the answer falls within Cluster 2 or Cluster 3, the feedback message should be different from that for Cluster 0 and 1 because there is at least one major mistake in the answer,

Cluster ID	Cluster Size	Interpretation	Potential Intervention
0	218	Have one minor mistake	• Encourage students that they get the general structure correct • Inform students the number of minor mistake sites they have • Suggest students to check for detail errors and list the common reasons for such errors, e.g. extra edges on a flat plane, missing edges at a corner
1	65	Have more than one minor mistakes	
2	30	Have some major and minor mistakes, mostly minor mistakes	• Encourage students that they are heading towards the right direction • Inform students the number of minor and major mistake sites they have • Suggest students to revisit some parts of the structure • Suggest students to carefully check for drawing errors and list the common reasons for such errors, e.g. extra edges on a flat plane, missing edges at a corner
3	15	Have some major and minor mistakes, mostly major mistakes	

4	39	More than half of the sketch as a whole is completely wrong	• If students have the correct structure but a wrong orientation:
			  • Encourage students that they get the general structure correct
			  • Inform them that they may have drawn it in an incorrect orientation
			• If students have an incorrect structure:
			  • Let students know that they have the wrong idea for the structure
			  • Suggest students to rethink about the structure from the beginning
			  • Provide hints for the students if available

5	3	Empty sketch	• If students did not make an effort, encourage them to attempt the question • If students forgot to submit a sketch, remind them to submit in the next attempt

Table 2: Interventions Summary Table

likely due to a structural error. The students in these two cases are mostly on the right track in terms of the general structure of the sketch. Hence, the feedback message could first encourage them that they are heading in the right direction. The system could then say that the sketch has X minor mistakes and Y major mistakes, where X is the Number of Components with a size smaller than the Average Component Size of the cluster centroid, and Y is the Number of Components with a size larger than the average. Finally, the intervention message could suggest the student first revisit the structure in detail to identify the major mistake, and then carefully check for drawing errors referring to a list of common minor mistakes.

For a student that falls into Cluster 4, it is likely that the student is either on the wrong track entirely or uses a wrong orientation. The system can perform a further check to compare the student's answer to other possible orientations and see if it belongs to the case of having a wrong orientation. If it is, the feedback message will remind the student that the structure of the sketch is mostly correct, but the orientation is incorrect. If it is not the case of having a wrong orientation, the feedback message will remind the students that they may have the wrong idea for the sketch, and they should reconsider the question from the beginning. The system could consider providing hints to the students as well in this case.

Lastly, if a student submits an empty sketch, the system can check the time spent on the question to determine whether the student did not attempt the question at all or forgot to click the submit button. If the student did not attempt the question, the system would encourage the student to make an effort in attempting to solve the problem. If the student forgot to submit the answer, the feedback message would remind them to submit in the next attempt.

5.3 Generalizability of the Proof-of-concept Approach

Our clustering model is more than a single model that works only in a specific scenario. It is a proof-of-concept approach for the evaluation of a complex free-hand sketch based on abstract features. Our contributions to the evaluation scheme of sketching answers have the potential to be generalized from spatial visualization training to more fields that involve free-hand technical drawings in various Engineering and Science subjects, such as circuit diagrams in Electrical Engineering, engine models in Mechanical Engineering, building plans in Architecture, and structural formula in Organic Chemistry. Technical drawing is similar to spatial visualization sketching in the sense that they both follow strict rules of sketching and are often drawn on grid paper to ensure a consistent proportion and orientation. Technical drawings in these fields usually start from a fundamental practice of drawing and modeling using practice problems that have a limited number of correct answers. With the presence of answer keys, our unsupervised clustering approach is flexible and easy to be retrained on new datasets to adapt to new types of sketches, even with additional features developed based on the learning goal of the type of sketches.

On the other hand, for technical drawing that involves a creative component or pure creative drawing, it may be harder to apply our approach directly. In evaluating creative drawing that does not have a limited number of correct answers, a mistake may be more subjective, and the evaluation may extend beyond getting a sketch correct to being functional, optimal, creative or aesthetic. The clustering approach based on abstract features of a sketch, however, may be used for

other purposes in this case. For example, our approach could be used to group sketches with similar characteristics together for the convenience of human graders, especially in a large course with limited human resources, such as Massive Open Online Courses. Reconsideration in feature engineering would be needed to achieve the new goals.

6. LIMITATIONS AND FUTURE WORK

The current erroneous answer categories do not take into account specific reasons that lead to a particular error in an answer. There may be multiple reasons for a student to end up with mistakes in the same category. To the best of our knowledge, there is neither prior work that studies the common misconceptions in spatial visualization sketching, nor cognitive models that describe the process of this task. The closest available work in cognitive models for spatial ability focuses on how people solve multiple choice spatial visualization questions, i.e., when candidate solutions are provided [14, 11, 31]. These models do not cover the process of generating a spatial object from scratch, which is what sets spatial visualization sketching apart from the traditional spatial ability tests. Hence, our proposed model is unable to distinguish the errors by their causes. Future research conducting qualitative interviews with students to understand the reasons why an error occur could provide valuable insights towards identifying not only broad categories of erroneous answer, but also the causes behind various error categories. It would also be beneficial to create cognitive models to understand systematically the strategies students used to solve these problems. These information would be valuable in further developing other features that could distinguish errors according to their underlying cause, for example, by leveraging the temporal sequence of actions executed by the student leading to their error. Improving current models to include information about the most probable cause of an error would be beneficial in generating formative feedback that goes beyond providing information about the nature of the students' error, and integrates conceptual information to support students in addressing misconceptions.

The current training data for the model only involved 14 students, which is a relatively small sample. As such, the current model can be seen as a proof-of-concept for the feasibility of erroneous answer categorization. Applying the same approach to a larger population of students will be necessary to validate the stability of the model and ensure that there are no additional answer categories that may not have been included in our current dataset. Future studies can re-train and test the model on a larger population to confirm the existence of the answer categories identified within the current study. Since the training process of the model is simple, re-training the model based on another dataset would be straightforward.

Another next step for this research is to deploy the model in an online training platform and conduct user testing to examine the effectiveness and accuracy of the categorization and intervention. Last but not least, the method proposed in this study is designed to be flexible and be applied to other disciplines. Future work in other disciplines, such as evaluating circuit diagrams in Electrical Engineering, engine models in Mechanical Engineering, building plans in Architecture, and structural formula in Organic Chemistry, will need to be conducted to evaluate the extent to which the proposed method generalizes to new topics.

7. CONCLUSION

In conclusion, this paper presents a clustering model as a solution to categorize erroneous answers in complex freehand sketching questions in spatial visualization training. Eight abstract features were developed and proven to be effective in the categorization of erroneous answers, including percentages of various types of incorrect lines, number of mistake components, and metrics of the size of the components. The clustering model detected six answer categories based on the severity and scale of the mistakes. With these detected categories, an online training platform will be able to present customized and actionable formative feedback in real-time. Moreover, our approach suggested a new and comprehensive set of evaluation criteria to assess a sketch, which could potentially be generalized to other disciplines that require sketching practices.

8. REFERENCES

[1] Y. Attali and D. Powers. Effect of immediate feedback and revision on psychometric properties of open-ended gre® subject test items. *ETS Research Report Series*, 2008(1):i–23, 2008.

[2] H. Ault and A. Fraser. A comparison of manual vs. online grading for solid models. In *Proceedings of the 2013 ASEE Annual Conference, Atlanta, Georgia, June 23*, volume 26, 2013.

[3] H.-D. Barke and T. Engida. Structural chemistry and spatial ability in different cultures. *Chemistry Education Research and Practice*, 2(3):227–239, 2001.

[4] S. Basu, C. Jacobs, and L. Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402, 2013.

[5] A. N. Bhat. *Sketchography-Automatic Grading of Map Sketches for Geography Education*. PhD thesis, 2017.

[6] C. S. Carter, M. A. Larussa, and G. M. Bodner. A study of two measures of spatial ability as predictors of success in different levels of general chemistry. *Journal of research in science teaching*, 24(7):645–657, 1987.

[7] C. Chandan, M. Deepika, S. Suraksha, and H. Mamatha. Identification and grading of freehand sketches using deep learning techniques. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1475–1480. IEEE, 2018.

[8] M. Contero, F. Naya, P. Company, J. L. Saorin, and J. Conesa. Improving visualization skills in engineering education. *IEEE Computer Graphics and Applications*, 25(5):24–31, 2005.

[9] R. T. Duesbury et al. Effect of type of practice in a computer-aided design environment in visualizing three-dimensional objects from two-dimensional orthographic projections. *Journal of Applied Psychology*, 81(3):249, 1996.

[10] M. O. Dzikovska, J. D. Moore, N. Steinhauser, G. Campbell, E. Farrow, and C. B. Callaway. Beetle ii: a system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010*

System Demonstrations, pages 13–18. Association for Computational Linguistics, 2010.

[11] D. E. Egan. Testing based on understanding: Implications from studies of spatial ability. *Intelligence*, 3(1):1–15, 1979.

[12] H. B. Gerson, S. A. Sorby, A. Wysocki, and B. J. Baartmans. The development and assessment of multimedia software for improving 3-d spatial visualization skills. *Computer Applications in Engineering Education*, 9(2):105–113, 2001.

[13] B. J. Gimmestad. Gender differences in spatial visualization and predictors of success in an engineering design course. In *Proceedings of the National Conference on Women in Mathematics and the Sciences*, number 801, pages 133–136, 1990.

[14] J. Gluck and S. Fitting. Spatial strategy selection: Interesting incremental information. *International Journal of Testing*, 3(3):293–308, 2003.

[15] C.-C. Han, C.-H. Chou, and C.-S. Wu. An interactive grading and learning system for chinese calligraphy. *Machine Vision and Applications*, 19(1):43–55, 2008.

[16] M. Hegarty, M. Keehner, C. Cohen, D. R. Montello, and Y. Lippa. The role of spatial cognition in medicine: Applications for selecting and training professionals. *Applied spatial cognition*, pages 285–315, 2007.

[17] Y. Kali and N. Orion. Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 33(4):369–391, 1996.

[18] S. Keshavabhotla, B. Williford, S. Kumar, E. Hilton, P. Taele, W. Li, J. Linsey, and T. Hammond. Conquering the cube: learning to sketch primitives in perspective with an intelligent tutoring system. In *Proceedings of the Symposium on Sketch-Based Interfaces and Modeling*, pages 1–11, 2017.

[19] S. Kirstukas. A preliminary scheme for automated grading and instantaneous feedback of 3d solid models. In *Proceedings of the midyear conference of engineering design graphics division of the ASEE*, pages 53–58, 2013.

[20] S. J. Kirstukas. Development and evaluation of a computer program to assess student cad models. In *Proceedings of ASEE's 123rd Annual Conference and Exposition, New Orleans, LA*, page 26781, 2016.

[21] C. Leacock and M. Chodorow. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405, 2003.

[22] J. M. Leake and J. L. Borgerson. *Engineering design graphics: sketching, modeling, and visualization*. J Wiley & Sons, 2013.

[23] R. Lehming, J. Gawalt, S. Cohen, and R. Bell. Women, minorities, and persons with disabilities in science and engineering: 2013. *National Science Foundation, Arlington, VA, USA, Rep*, pages 13–304, 2013.

[24] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn. Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33(2):19–28, 2014.

[25] D. Lubinski. Spatial ability and stem: A sleeping giant for talent identification and development. *Personality and Individual Differences*, 49(4):344–351, 2010.

[26] J. D. Merrel, P. F. Cirillo, P. M. Schwartz, and J. Webb. Multiple-choice testing using immediate feedback-assessment technique (if at®) forms: Second-chance guessing vs. second-chance learning? 2015.

[27] N. S. Newcombe and A. Frick. Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, 4(3):102–111, 2010.

[28] D. J. Nicol and D. Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218, 2006.

[29] R. Nielsen, W. Ward, and J. H. Martin. Classification errors in a domain-independent assessment system. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, 2008.

[30] N. Orion, D. Ben-Chaim, and Y. Kali. Relationship between earth-science education and spatial visualization. *Journal of Geoscience Education*, 45(2):129–132, 1997.

[31] S. E. Poltrock and P. Brown. Individual differences in visual imagery and spatial ability. *Intelligence*, 8(2):93–138, 1984.

[32] J. R. Pribyl and G. M. Bodner. Spatial ability and its role in organic chemistry: A study of four organic courses. *Journal of research in science teaching*, 24(3):229–240, 1987.

[33] S. Pulman and J. Sukkarieh. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 9–16, 2005.

[34] K. Rochford. Spatial learning disabilities and underachievement among university anatomy students. *Medical education*, 19(1):13–26, 1985.

[35] J. Roorda. Visual perception, spatial visualization and engineering drawing. *Engineering Design Graphics Journal*, 58(2):12–21, 1994.

[36] V. J. Shute. Focus on formative feedback. *Review of educational research*, 78(1):153–189, 2008.

[37] I. M. Smith. *Spatial ability: Its educational and social significance*. RR Knapp, 1964.

[38] S. A. Sorby. Developing 3-d spatial visualization skills. *Engineering Design Graphics Journal*, 63(2), 2009.

[39] S. A. Sorby. Educational research in developing 3-d spatial skills for engineering students. *International Journal of Science Education*, 31(3):459–480, 2009.

[40] P. Taele and T. Hammond. Boponoto: An intelligent sketch education application for learning zhuyin phonetic script. In *DMS*, pages 101–107, 2015.

[41] D. H. Uttal and C. A. Cohen. Spatial thinking and stem education: When, why, and how? In *Psychology of learning and motivation*, volume 57, pages 147–181. Elsevier, 2012.

[42] D. H. Uttal, N. G. Meadow, E. Tipton, L. L. Hand, A. R. Alden, C. Warren, and N. S. Newcombe. The malleability of spatial skills: A meta-analysis of training studies. *Psychological bulletin*, 139(2):352,

2013.

[43] S. Valentine, R. Lara-Garduno, J. Linsey, and T. Hammond. Mechanix: A sketch-based tutoring system that automatically corrects hand-sketched statics homework. In *The impact of pen and touch technology on education*, pages 91–103. Springer, 2015.

[44] N. L. Veurink and A. Hamlin. Spatial visualization skills: Impact on confidence in an engineering curriculum. In *American Society for Engineering Education*. American Society for Engineering Education, 2011.

[45] J. Wai, D. Lubinski, and C. P. Benbow. Spatial ability for stem domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of educational Psychology*, 101(4):817, 2009.

[46] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[47] Z. Xiao, Y. Yao, C.-H. Yen, S. Dey, H. Wauck, J. M. Leake, B. Woodard, A. Wolters, and W.-T. Fu. A scalable online platform for evaluating and training visuospatial skills of engineering students. In *2017 ASEE Annual Conference & Exposition. ASEE Conferences, Columbus, Ohio. https://peer. asee. org/27509*, 2017.

[48] Y. Zhang, R. Shah, and M. Chi. Deep learning+ student modeling+ clustering: A recipe for effective automatic short answer grading. *International Educational Data Mining Society*, 2016.

9. ACKNOWLEDGMENTS

The authors would like to thank Brian Woodard, Ziang Xiao, and the rest of the SIIP project team at the University of Illinois, Urbana-Champaign for the joint effort in developing the spatial visualization online training platform and offering the Fall 2019 "Spatial Visualization" course.