Do These Students Have Similar Strategies? Clustering Math Work in Uploaded Images on an Online Learning Platform

BLINDED AUTHOR, Blinded Institute, blinded
BLINDED AUTHOR 2, Blinded Institute, blinded
BLINDED AUTHOR 3, Blinded Institute, blinded
BLINDED AUTHOR 4, Blinded Institute, blinded
BLINDED AUTHOR 5, Blinded Institute, blinded
BLINDED AUTHOR 6, Blinded Institute, blinded
BLINDED AUTHOR 7, Blinded Institute, blinded

This exploratory study delves into the complex challenge of analyzing and interpreting student responses to mathematical problems, typically conveyed through image formats within online learning platforms. The main goal of this research is to identify and differentiate various student strategies within a dataset comprising image-based mathematical work. A comprehensive approach is implemented, including various image representation, preprocessing, and clustering techniques, each evaluated to fulfill the study's objectives. The exploration spans several methods for enhanced image representation, extending from conventional pixel-based approaches to the innovative deployment of CLIP embeddings. Given the prevalent noise and variability in our dataset, an ablation study is conducted to meticulously evaluate the impact of various preprocessing steps, assessing their potency in eradicating extraneous backgrounds and noise to more precisely isolate relevant mathematical content. Two clustering approaches—k-means and hierarchical clustering—are employed to categorize images based on student strategies that underlies their responses. Preliminary results underscore the hierarchical clustering method could distinguish between student strategies effectively. Our study lays down a robust framework for characterizing and understanding student strategies in online mathematics problem-solving, paving the way for future research into scalable and precise analytical methodologies while introducing a novel open-source image dataset for the learning analytics research community.

CCS Concepts: • **Computing methodologies → Image processing**; **Image representations**; **Cluster analysis**.

Additional Key Words and Phrases: Open-ended questions, Image responses, Embeddings, Clustering, Mathematics

ACM Reference Format:
Blinded Author, Blinded Author 2, Blinded Author 3, Blinded Author 4, Blinded Author 5, Blinded Author 6, and Blinded Author 7. 2018. Do These Students Have Similar Strategies? Clustering Math Work in Uploaded Images on an Online Learning Platform. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, online learning platforms have witnessed substantial growth, accelerated by factors such as globalization, advancements in technology, and more recently, global challenges like the COVID-19 pandemic [30]. This transition to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
Manuscript submitted to ACM

digital platforms has led to an unprecedented influx of diverse student data, including mathematics education. Among the various data types about student learning captured through this mathematics education platform, image-based submissions — capturing handwritten equations, sketches, and diagrams — are particularly noteworthy.

Such image submissions, often termed as 'visual artifacts' of learning, provide an unparalleled window into students' thought processes, their conceptual understanding, and their problem-solving strategies [25]. They transcend the limitations of traditional text-based responses, enabling educators to decipher nuances like hesitation in strokes, the sequence of problem-solving, or even errors and corrections made during the process [3, 27]. This level of granularity can be pivotal in understanding not just the 'what' but the 'how' students are learning, allowing educators to provide precise feedback and tailored instruction based on students' strategies and reasoning reflected in their responses. Moreover, analyzing images can aid in automatically identifying common misconceptions, patterns of thought, and even predicting potential hurdles a student might face in the future. For instance, the way a student sketches a parabola or labels a geometric figure might give hints about their comprehension of underlying concepts [8]. Such insights can be instrumental in the timely remediation of learning gaps and fostering a more supportive and efficient mathematics learning environment.

However, the richness and complexity of these image-based submissions also pose distinctive challenges in their analysis and interpretation. Traditional analytic techniques, designed primarily for textual or numeric data, fall short when applied to images, necessitating the development of innovative methods attuned to the nuances of visual data [1]. Some pioneering efforts have been made to analyze hand-drawn diagrams or sketches using image recognition techniques to provide instant feedback in domains like engineering and physics [8].

In the context of mathematics, earlier studies have often relied on simplistic pattern recognition methods to classify hand-written equations and geometrical sketches [26]. Nevertheless, the diverse nature of student strategies, especially when conveyed through images, calls for a more holistic and nuanced approach. It is this gap in the literature that our study seeks to address, integrating advanced embedding techniques and sophisticated clustering algorithms to delve deeper into the world of image-based student responses to evaluate students' underlying strategies. Our study embarks on a mission to decipher image-based student submissions by addressing three pivotal research questions:

(1) **RQ1**- Does the incorporation of embeddings enhance our capacity to differentiate between distinct categories of students' mathematical reasoning and strategies depicted in images?
(2) **RQ2**- How does the choice of preprocessing method impact the differentiation process?
(3) **RQ3**- To what degree does the utilization of different clustering techniques enhance our ability to distinguish between various students' responses?

In navigating these questions, we aim to enhance the empirical evidence through detailed, step-by-step comparisons, evaluating whether students' response strategies and reasoning in mathematical problems can be assessed using a sample of image data derived from two sample math problems. Recognizing the scarcity of specialized datasets in this domain, we are releasing our meticulously curated, image-based dataset to the broader research community. This open-source resource will not only serve as a valuable foundation for further studies in the realm of image-based learning analytics but also stimulate the development of novel analytical methods specifically tailored for such data.

2 RELATED WORKS

2.1 Online Learning Platforms in Math Education

The digital transition in education has witnessed the rise of online platforms explicitly tailored for various subjects, with mathematics being a prominent area of focus. This shift towards online math platforms has been catalyzed by the increasing need for flexible, accessible, and interactive learning environments [15, 16]. The digital transition in education has ushered in the rise of online platforms, explicitly tailored for various subjects, mathematics being a notably prominent focus. This shift towards online math platforms has been catalyzed by an ever-increasing need for flexible, accessible, and interactive learning environments [15, 16].

The COVID-19 pandemic further expedited the transition to online learning, leading to an augmented application of online learning platforms in K-12 mathematics classrooms [32]. Online mathematics learning platforms provide various advantages, making mathematics learning more accessible and personalized. These platforms enable personalized and self-paced learning, facilitating students' engagement with mathematical concepts and practices at their convenience [14]. They also provide interactive learning and assessment resources, which cater to students' individual needs and aid in establishing more effective and efficient learning environments. Moreover, these platforms allow for potentially instant feedback and progress monitoring through online assessments. The adoption of automated grading of student responses [4, 6, 23], analysis of students' writing patterns and discourse [2], and generation of teacher feedback [12, 17], have been rigorously demonstrated in the previous literature.

2.2 Automated Scoring in Online Math Assessment

Automated scoring systems in mathematics education have predominately focused on evaluating students' computational skills [11], problem-solving strategies [5], and, occasionally, the procedural steps undertaken during problem-solving [29]. The recent introduction of transformer-based models into the scoring systems remarkably extends their capabilities, particularly in improving scoring accuracy [4, 35], ensuring scoring consistency and fairness [10], and expanding these models for the generation of timely feedback [6]. Despite such advancements, one major focus area remains the incorporation of new response formats, such as image-based responses.

2.2.1 Automated Scoring of Image-based Math Responses. Image-based responses require students to create a visual representation of their work using a traditional paper and pencil approach or using digital media and upload their work to online learning platforms. GeoGebra[13] and Desmos[9] are some examples of computer-based applications that allow students to interact with graphs and algebraic expressions. While these kind of tools and support for visual representation of answers exists in online learning platforms, some teacher still prefer the traditional approach of paper and pencil and some others use a blend of both in their classrooms.

While previous automated assessment methods have relied heavily on text-based constructed responses, where students type an answer directly into the online learning platforms [4, 11, 35], recent works have explored a diverging type of student responses, particularly, image-based responses. Baral et al.[3] proposed methods to auto-score open-ended mathematics questions containing text and image responses [3]. Using optical character recognition and deep learning models like CLIP reduced scoring errors for mixed text and image responses over models that only handled text. As online assessments expand the types of responses they allow, automated scoring techniques must evolve to handle multimedia response formats.

2.3 Approaches in Image Processing and Analysis

2.3.1 Embeddings for Image Representation. One of the primary roles of embeddings in image analysis is to facilitate the understanding and representation of visual content within them. In recent years, image embeddings have emerged as a transformative technique in image processing and analysis, offering powerful ways to represent and understand visual content. The features used in image analysis are often of high dimension, thus necessitating techniques like feature extraction for handling multi-modal features for image classification tasks. [3, 24]. Luo et al. [24] have explored the domain of multi-modal multi-task feature extraction, highlighting the advantages of leveraging multiple modalities in such scenarios.

The CLIP (Contrastive Language-Image Pretraining) model, introduced by Radford et al. [31], is an image classification model based on transformer architecture, commonly used in natural language processing tasks. CLIP learns joint image-text embeddings, allowing images to be encoded into a shared space with natural language descriptions. This versatile representation has found applications in various image analysis tasks, from image classification to zero-shot learning. In our analyses, we explore the use of image embeddings from a pre-trained CLIP model as a method of image representation and further investigate the ability of such a pre-trained model in the domain of mathematics education.

2.3.2 Data Clustering Techniques.

3 DATA COLLECTION AND PARTICIPANTS

In this study, we utilize a dataset of student responses to open-ended mathematics questions taken from an online learning platform. The main goal of this study is to analyze and compare various clustering methods with image processing techniques to identify and distinguish various student approaches to solving a math problem within the student-uploaded works. As such we mainly look into the response from students that are images. This dataset was collected from a BLINDED Online learning platform, from a middle school mathematics classroom. The students were assigned mathematics assignments using the BLINDED learning platform; which consisted of both close-ended and open-ended problems. For open-ended problems, the students were allowed to provide either a textual response or they had the option to write their answer and upload the image of their work directly to the learning platform.

The dataset, in addition to the image responses from students, consists of a numeric assessment score given by a teacher. The scores for these responses are on an ordinal 5-point scale ranging from 0 to 4.

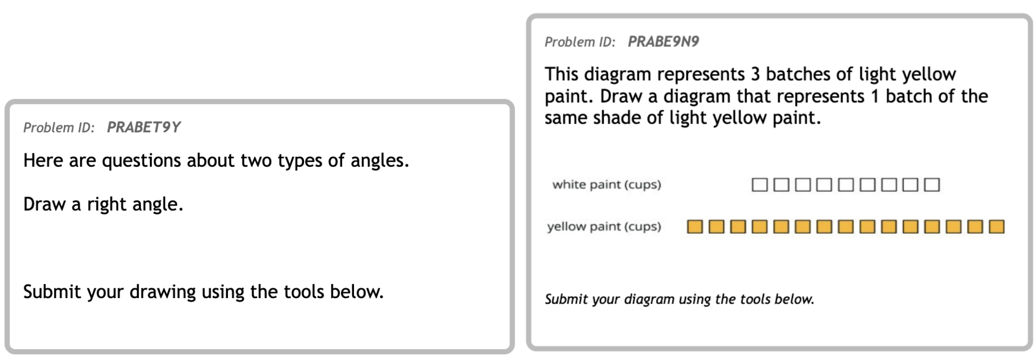

Fig. 1. The two open-response mathematics problems used for the analyses in this study, Problem 1 (on the left), and Problem 2 (on the right)

For this study, we selected two specific mathematics problems which had mostly image-based responses from students. Figure 1 shows these two math problems. The first problem which we call "Problem 1" throughout this study is a 4th-grade problem based on the Relationship of Angles, while the second problem "Problem 2" is a 6th-grade math based on Defining Equivalent Ratios.

Problem 1, had 159 scored image responses, while Problem 2 had about 269 scored images in total. For Problem 2 we randomly sampled 159 images to balance out the dataset for the purpose of the study. The final dataset includes 318 image-based responses from 318 unique students who answered the 2 math problems. This dataset of images was scored by 51 different teachers.

Utilizing this dataset [1] , we perform an exploratory analysis through the application of various clustering techniques in order to distinguish different approaches taken by students for solving math problems. We discuss the methods taken in detail in the following sections.

4 METHODS

4.1 OpenCV Template Match vs CLIP

In this subsection of the methods, we perform a comparison of different image representation techniques to identify the best approach for differentiating and grouping various students' math works. We compare the raw pixel matching using OpenCV's template match techniques, with context-rich encoding of images from a popular deep-learning method called "CLIP" for clustering images.

4.1.1 OpenCV Template Match. OpenCV, a popular computer vision library, offers a technique known as template matching, for comparing different images. We leverage this method to identify similarities in our dataset of images and group these using K-means clustering. Template matching is typically used for finding instances of a template image (a small image) within a larger target image. The goal of this is to find regions in the target image that closely match the template. In our case, we adapt this technique to compare and group images by using a bidirectional approach. We use the "TM_CCORR_NORMED" method for template matching. This method calculates the cross-correlation between the images with the highest value indicating the best match.

Initially, we apply template matching by matching Image A onto Image B. This means we treat Image A as the template and try to find the best matches in Image B. Next, we reverse the process and match Image B onto Image A. Now, Image B serves as the template, and we look for matches in Image A. After performing template matching in both directions, we obtain match scores for Image A matched onto Image B and Image B matched onto Image A. To determine the overall similarity between the two images, we consider the minimum of these match scores. This is done to account for cases where one image may match well with the other, but the reverse might not be true. For example, if image A is a perfectly drawn right-angled triangle and image 2 is a plain sheet of graph paper then image A matches on to image B well but not vice versa. We perform this for each of the images in our dataset. The resulting match scores for pairs of images used as a distance metric are then utilized as input for k-means clustering, a common method for grouping similar data points.

4.1.2 CLIP Embeddings. CLIP (Contrastive Language-Image Pre-training)[31] is an image classification model based on transformer architecture that offers a versatile and context-rich means for representing visual content. Introduced

[1] A curated dataset of these images with cropped background is shared through this URL: https://osf.io/9a8xq/?view_only=f0cd8d45acfd49f3be8aa1f0c1eb375b. For review purposes, we only share the cropped image data, but a more comprehensive dataset will be shared for the final version.

by OpenAI, this model harnesses the power of a vision-language transformer architecture and is able to encode both natural languages (text) and images in the same vector space by using a multi-modal pre-training approach. While the CLIP model was initially designed for the combination of text and images, its embeddings can be effectively used for image representation tasks independently of textual information.

In this study we use the *"clip-vit-large-patch14-336"* version of the CLIP model, to generate an embedding vector representation for each of the images of students' math works. These image embeddings encapsulate both the visual characteristics and semantic content of the images, effectively allowing us to understand not only what the images contain but also what they mean making them potential for various image-related tasks. To assess the similarity between images encoded by CLIP, we use an angle-based metric. For this particular method, instead of relying on the traditional Euclidean distance, we measure the angle between the CLIP embedding vectors and further use these as input to the K-means clustering. The number of clusters is chosen based on the elbow plot for the accuracy scores with cluster size k ranging from 2 to 12.

4.2 Ablation Study on the Impact of Preprocessing

The original dataset of images posed a challenge due to the presence of background noise, including elements like students' faces, backgrounds, and other non-relevant content alongside the mathematical content. In addition to this, there are differences in the images aside from the mathematical content and background, coming from the use of different types of papers (like graph, math, or plain paper, the use of digital media vs. conventional pencil and paper, etc. As such, the image representations may pick up on the non-relevant content, tampering with the results of clustering. For this reason, we apply the following preprocessing steps and conduct an ablation study of these different methods to identify the best-suited preprocessing method for the image clustering tasks.

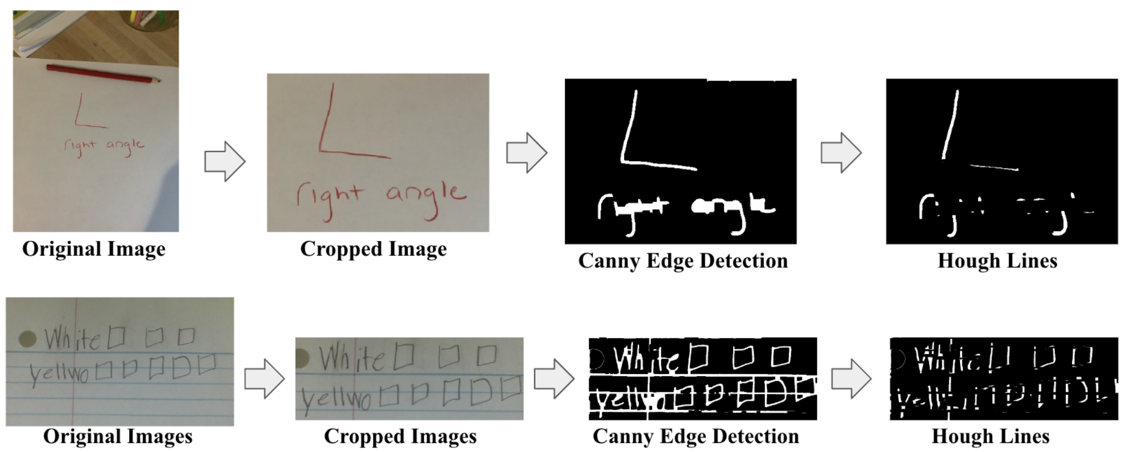

Fig. 2. Example image response from Problem 1(top) and Problem 2(bottom) with the applied preprocessing steps.

4.2.1 Image cropping. In this step, a meticulous inspection was performed on each image within the dataset, to identify the relevant math content. The objective was to crop and isolate the core math content within the images while removing any extraneous noise or background elements. This step not only enhanced the clarity of the images but also facilitated more precise template matching and CLIP-based analysis.

4.2.2 Edge Detection. To further refine image preprocessing, Canny edge detection [7] was employed. This technique identifies prominent edges within the images. By emphasizing edges and contours, this method enhances the ability to detect and differentiate key mathematical elements within the images.

4.2.3 Hough Lines. The Probabilistic Hough lines algorithm [21] was used in conjunction with the Canny edge detection to address the specific challenges related to graph and math paper lines within the images. While Canny edge detection effectively identifies edges, it may also detect lines originating from the underlying graph paper or grid, which are unrelated to the mathematical content. The Hough lines algorithm is used to identify and remove these extraneous lines if they occur repeatedly and are parallel to each other.

The example of student responses in the dataset with applied preprocessing steps are shown in Figure 2.

4.3 K-means vs Hierarchical Clustering

To unravel latent structures and patterns, which are associated with students' mathematical response strategies, within the dataset, we conducted two clustering techniques, including K-means [22] and hierarchical clustering [28] as our primary analytical approaches. Euclidean distance was chosen as the distance metric to measure the similarity between CLIP embeddings.

Hierarchical clustering allowed us to construct a dendrogram that organized the image responses into a hierarchy of clusters, with each node representing a group of similar responses. Determining the optimal number of clusters is crucial for obtaining meaningful and interpretable clustering results. We addressed this challenge by employing silhouette analysis, a widely-used technique for evaluating the quality of clustering. By systematically varying the number of clusters from 2 to 10 based on the baseline threshold (t=2), which were identified from the dendrogram, and computing the silhouette score for each configuration, we identified the optimal number of clusters that maximized the cohesion within clusters and separation between them. This process ensured that our clustering was both statistically robust and reflective of the underlying patterns in students' problem-solving approaches.

4.4 Evaluation Metrics

We used three commonly adopted cluster evaluation metrics: the Gini score (or Gini index) [18], the purity score [33], and the silhouette score. The Gini score, also known as the Gini index or Gini coefficient, quantifies the inequality or impurity within a cluster. Often applied in hierarchical clustering or decision tree algorithms, it measures how mixed or heterogeneous the elements within a cluster are. In clustering analysis, a Gini score of 0 indicates perfect purity, signifying all elements in the cluster belong to the same class or category—in our case, the same mathematical strategy or reasoning to solve a problem.

The purity score is another metric used to evaluate clustering quality, especially in unsupervised learning and clustering algorithms. It gauges how closely the elements within a cluster relate to the same class or category. In clustering analysis, a purity score of 1 signifies perfect purity, indicating all elements in the cluster pertain to a singular class or category. A diminished purity score implies the elements within the cluster are diverse and may affiliate with multiple classes or categories.

However, both the Gini and purity scores can be susceptible to the effects of increasing cluster sizes, potentially leading to skewed evaluations. To counteract this limitation, we also employed the silhouette score. The silhouette score measures an object's similarity to its own cluster in contrast to other clusters. Its values lie between -1 and 1. A high silhouette score indicates the object aligns well with its own cluster and poorly with neighboring clusters. Conversely, a low silhouette score suggests potential misclustering. If most objects boast high silhouette scores, the clustering configuration is deemed appropriate. Yet, if many objects present low or negative scores, the clustering may encompass too many or too few clusters.

5 RESULTS

Our final results indicate that both K-means and Hierarchical clustering analyses performed comparably in identifying and clustering images based on their underlying mathematical reasoning. A total of seven clusters were retrieved from K-means clustering, and based on the distance threshold of 2 according to the ward linkage, Hierarchical clustering with the cluster size of 4 to 11 were compared for further evaluation.

5.1 OpenCV Template Match vs. CLIP

Comparing the GINI index and the purity score between OpenCV Template Match and CLIP embedding yielded distinct results as shown in Table 1. The CLIP embedding, when combined with K-means clustering, consistently showcased an enhancement in the clustering outcomes, marked by a lower GINI index and a higher purity score.

5.2 Ablation Study on the Impact of Preprocessing

The most optimal GINI index was obtained using the cropped image, registering at 0.157, paired with a notably high purity score of 0.894 as seen in Table 1. The original image also achieved an equivalent purity score of 0.894. The influence of other image processing techniques seemed marginal in augmenting the clustering outcomes. Specifically, the images processed with edge and hough line techniques recorded the peak GINI index of 0.396 and the lowest purity score of 0.676. Based on these findings, we juxtaposed the clustering methodologies, namely K-means and Hierarchical clustering, using the original and cropped images. This comparison was intended to further assess improvements in image clustering predicated on students' mathematical reasoning.

Table 1. Clustering Evaluation Metrics

Processing and Clustering Methods	Cluster size	GINI index	Purity score
OpenCV Template Match			
Original image	7	0.254	0.787
CLIP and K-means Clustering			
Original image	7	0.169	0.896
Cropping	7	**0.157**	**0.894**
Cropping and Edge	7	0.203	0.864
Edge and Hough Lines	7	0.386	0.676
Cropping, Edge, and Hough Lines	7	0.297	0.789
CLIP and Hierarchical Clustering			
Original image	10	0.11	0.917
Cropping	10	**0.088**	**0.943**
Cropping and Edge	10	0.221	0.837
Edge and Hough Lines	10	0.238	0.831
Cropping, Edge, and Hough Lines	10	0.227	0.839

5.3 K-means vs. Hierarchical Clustering

Hierarchical clustering was analyzed across a range of cluster sizes from 4 to 11. This range was determined based on a threshold of 2 when inspecting the dendrogram using the ward linkage method (see Figure 3). The results highlighted that a cluster size of 4 yielded the highest silhouette score for both original and cropped images. On the other hand, a cluster size of 10 resulted in the lowest GINI index and the highest purity score. It's important to note that the silhouette score for cluster size 10 was notably better than for cluster size 11. To further scrutinize the clustering methods' efficiency in unveiling students' intrinsic mathematical reasoning depicted in their image responses, additional comparisons between the K-means and hierarchical clustering outcomes were undertaken. The final clustering analyses were performed comparing the *cropped* images clustered into 7 groups using K-means, against the total of 10 clusters derived from hierarchical clustering.

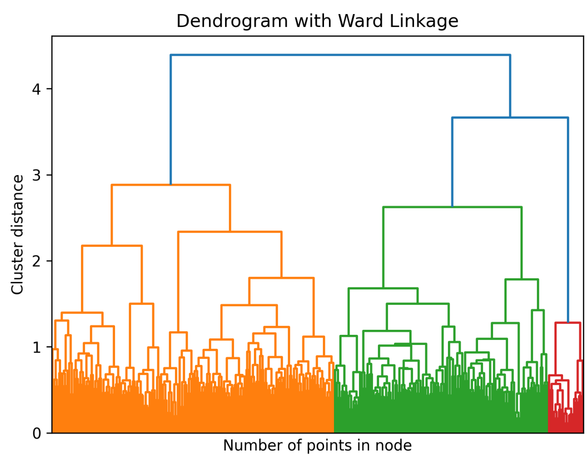

Fig. 3. Dendrogram for the hierarchical cluster with ward linkage method to identify the baseline threshold

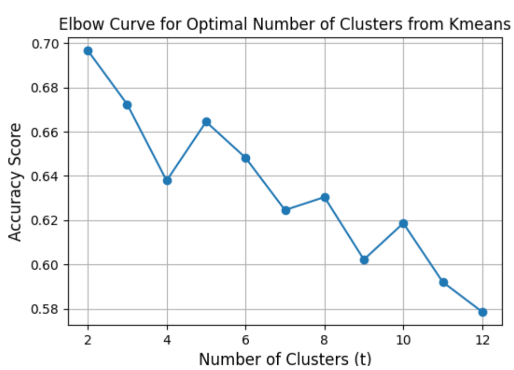

Fig. 4. Elbow plot with accuracy scores of the k-means clusters

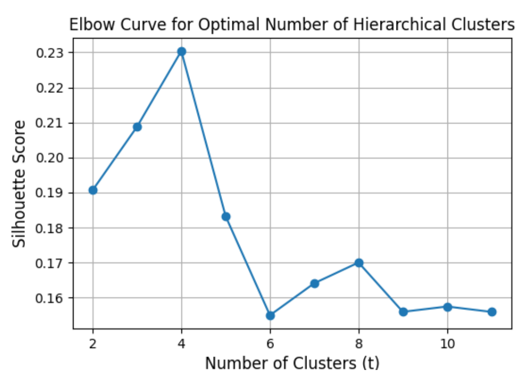

Fig. 5. Elbow plot of silhouette scores of the hierarchical clusters

Table 2. CLIP and Hierarchical Clustering Results with the Ward linkage method

	Original Image			Cropped Image		
Cluster size	*GINI index*	*Purity score*	*Silhouette score*	*GINI index*	*Purity score*	*Silhouette score*
4	0.179	0.876	**0.230**	0.152	0.895	**0.205**
5	0.144	0.901	0.183	0.138	0.908	0.190
6	0.153	0.898	0.155	0.116	0.922	0.175
7	0.131	0.913	0.164	0.100	0.933	0.167
8	0.138	0.896	0.170	0.105	0.931	0.170
9	0.123	0.907	0.160	0.096	0.937	0.165
10	0.110	0.917	0.157	0.088	0.943	0.142
11	**0.100**	**0.924**	0.156	**0.087**	**0.944**	0.128

Table 3. Distribution of Image Responses across Clusters

Cluster	Total Images	Problem 1	Problem 2
K-means with Cropped images (t=7)			
3	56	0	**56**
2	58	1	**57**
4	18	1	**17**
7	51	**51**	0
6	50	**47**	3
5	23	**16**	7
1	62	**43**	19
Hierarchical Clustering with Cropped images (t=10)			
1	33	0	**33**
2	38	0	**38**
3	39	1	**38**
7	38	2	**36**
4	10	**10**	0
8	23	**23**	0
10	35	**35**	0
6	37	**36**	1
5	44	**38**	6
9	21	**14**	7

5.3.1 Clustering Performance Accuracy. The clusters acquired from the two methods of clustering with cropped images were compared based on their accuracy to retrieve the underlying students' mathematical reasoning represented in their image responses. Table 3 provides the final performance accuracy of the two clustering methods, identified by the distribution of the image responses that originated from the two math problems (i.e., Math Problem 1, Math Problem 2) across the clusters. These two math problems as shown in Figure 1 required students to approach and solve the problems with two distinctive mathematical reasoning.

The results indicate that both K-means and hierarchical clustering methods are able to differentiate the two distinct mathematical reasoning coming from the dataset of two different math problems as seen by the proportions of the images in each of the resulting clusters. Overall, both methods could clearly separated the math problems based into different cluster categories – such as Clusters 2, 3, 4, 6, and 7 in K-means and Clusters 1, 2, 3, 4, 6, 7, 8, and 10 in hierarchical clustering – with few exceptions.

In terms of the clusters that failed to clearly separate the two method problems, in K-means, out of the 7 resulting clusters, 2 clusters (Clusters 1 and 5) represented a considerable mix of image responses from both problems 1 and 2. Cluster 1 and 5 both had 70% of images from Problem 1 and 30% images from Problem 2. Similarly, for the hierarchical clustering, out of the 10 clusters, two of the clusters (Cluster 5 and 9) represented a mix of the image responses from both Problems 1 and 2. Overall, the performance accuracy to separate the responses based on the math problems, indicate that our hierarchical clustering approach could show slightly improved clustering results compared to the K-means with less clusters with a mix of image responses from the different problem categories.

5.3.2 Illustrative Examples from Hierarchical Clustering Results. We further qualitatively assessed the clusters to understand the characteristics of the clusters from the hierarchical clustering results. First, we evaluated the clusters (e.g., Clusters 4, 10, 3 and 7) that demonstrated clear separation between the two problems. Figure 6 presents some examples from the clusters 4, 10, 3 and 7, identified through hierarchical clustering using the cropped images. Clusters 4 and 10, present responses from Problem 1 in the dataset. All the responses in Cluster 4 "right angle" text written as a label to the drawn right angle. While cluster 10, groups the perfectly drawn right angles with "90°" marking as the label. Clusters 3 and 7, present responses solely from Problem 2. Cluster 3 picks up mostly on the textual format of responses, whereas Cluster 7 presents digital images which are the screenshots of the question with markings for the answer.

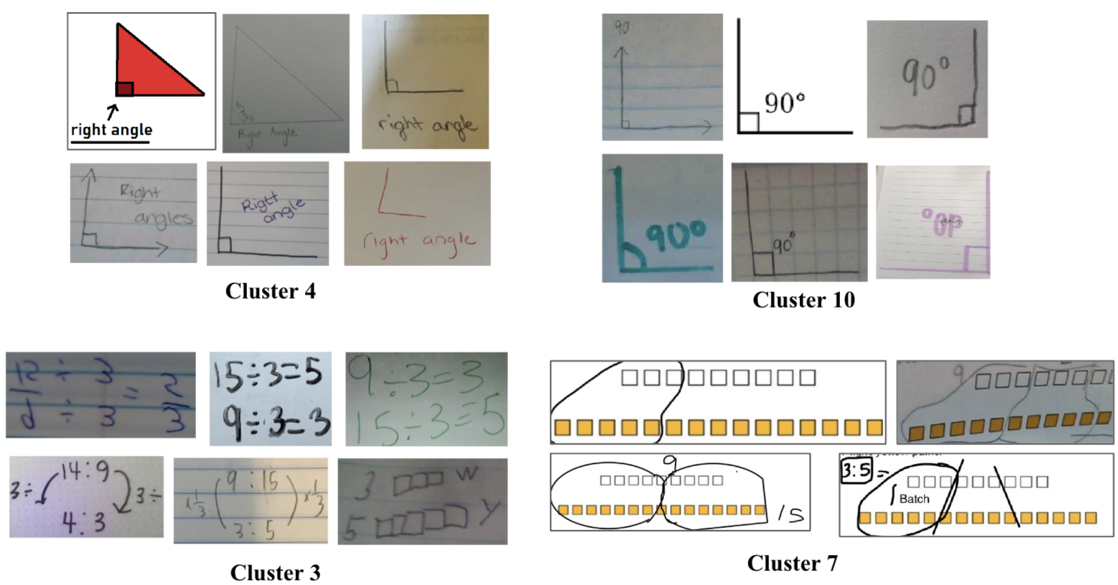

Fig. 6. Example images from the clusters with clear separations (Clusters 4, 10, 3, & 7) from the Hierarchical clustering method with cropped images.

Second, we evaluated the clusters (e.g., Clusters 5, 9) that showed a less clear separation of students' approaches seen in the images. The clusters presented a mix of images from both math problems grouped together into the same clusters. Figure 7 displays selected images from clusters 5 and 9, which were identified through hierarchical clustering. Cluster 5 predominantly consisted of images from Problem 1 (86%) with a minor portion from Problem 2 (13%). All images in this cluster were of handwritten math work on paper. A notable similarity among these images was the type of paper

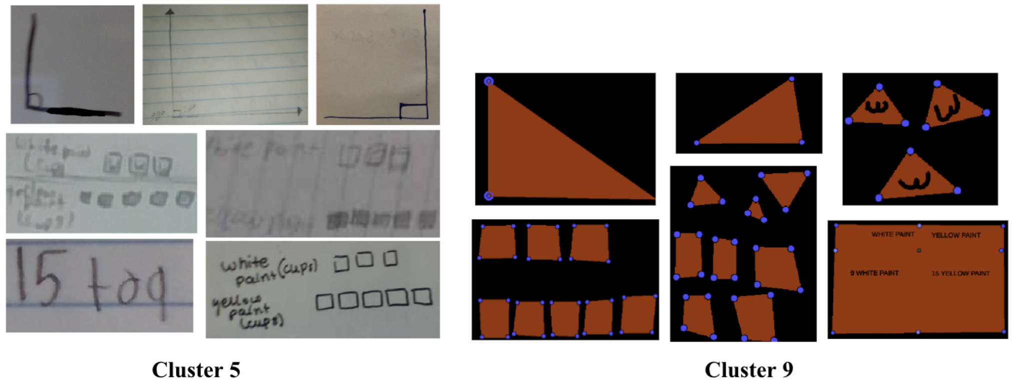

Fig. 7. Example images from the clusters with *less* clear separations (Clusters 5 & 9) from the Hierarchical clustering method with cropped images.

used, as illustrated in Figure 7. On the other hand, Cluster 9 had a more balanced distribution with 66% of images from Problem 1 and 33% from Problem 2. Images in this cluster primarily represented digital submissions from both problems. The study's dataset featured images with a distinctive black background, adorned with shapes like triangles and squares filled in brown. Cluster 9 captured these characteristic images from both problems, as can be seen in Figure 7.

6 DISCUSSION

This research presents and compares different methods of image representation methods, preprocessing steps, and clustering techniques to identify and distinguish different types of student approaches seen in image-based responses. This study is a preliminary analysis conducted using a small dataset of 318 image responses from two mathematics problems, that sheds light on the application of unsupervised methods to distinguish different student approaches seen in image-based responses. The results from this exploratory analysis presented valuable insights into the use of suitable image representation methods along with preprocessing and clustering techniques for the analyses of image-based responses in mathematics.

The findings from the analyses indicate that CLIP embeddings provide a powerful means of representing and analyzing image-based student responses. Further, the conducted ablation study on the impact of the preprocessing step suggested that removing background noise and irrelevant features with the cropping of images enhanced the accuracy of both the clustering methods. However, it is noteworthy that the other two preprocessing steps – edge detection and the subsequent application of the Hough lines algorithm, had detrimental effects on the clustering outcomes. Edge detection method, while valuable for simplifying the representation of images, in the context of mathematical image responses seemed to oversimplify the data representation, potentially discarding some of the crucial information within these images. Moreover, the Hough lines algorithm, intended to remove extraneous grid lines, might have inadvertently interfered with the interpretation of certain mathematical components within the images. The Hough lines algorithm can be especially useful for finding prominent linear features, such as grid lines coming from graph paper. In our context, this method seems to capture and remove the horizontal line drawing in most of the right-angled triangle problems, as seen in the example in Figure 2. Further, it also tampered with some of the text in the images as seen in Figure 2.

In terms of two different clustering methods – K-means and Hierarchical – hierarchical clustering yielded better results in distinguishing various student approaches, potentially due to the varying factors addressed in previous literature [19, 20]. First, the robustness to outliers in the hierarchical clustering approach may have helped improve performance accuracy in our context. The presence of outliers can heavily affect centroid selection in K-means clustering approaches, where the varying sizes, shapes, and noisy features that frequently appear in image data could have had a negative effect. Second, hierarchical clustering can generally accommodate more flexible dissimilarity and similarity comparison measures through the linkage method, which can aid in constructing clusters that are flexible in both size and shape. This is particularly advantageous in cases where distance measures are computed from a large dimension of embedding, where the computation of similarity or dissimilarity becomes a less clear, and challenging problem [34].

7 LIMITATIONS AND FUTURE WORK

Despite the promising findings, our study encounters several limitations. Primarily, the dataset used for this analysis was relatively small, comprising 318 responses based on only two mathematics problems. This preliminary analysis aimed to compare various methods of image representation and clustering, seeking to identify the optimal unsupervised method. Expanding this dataset to include a broader variety of student responses in mathematics would enhance the generalizability of our approach.

Moreover, the dataset exhibited an uneven distribution of scores, with over 80% of the image responses receiving a full score of 4 from teachers. Scores, serving as labels, can be pivotal in analyzing the ability of image clustering methods to distinguish between correct and incorrect responses. Nevertheless, a dataset with an uneven distribution of correct versus incorrect responses presents a challenge in developing and evaluating these methods in a nuanced manner. Furthermore, the study utilized a manual cropping process for the images. Future work could explore developing automated methods for background removal.

Beyond the constraints and limitations inherent in the methodologies and techniques explored, analyzing image-based student work presents its own set of formidable challenges. These stem from the inherent variations and complexity found in the dataset of image-based responses. Factors such as variability in writing styles, different types of handwriting, the use of various symbolic notations, and the unstructured format of the responses all contribute to the complexity of these analysis methods.Additionally, the limited availability of a comprehensive dataset in the educational domain poses a significant challenge in constructing improved methods of analysis and support for these images.

8 CONCLUSION

In conclusion, this research offers valuable insights into the differentiation of student approaches in image-based responses. By leveraging advanced techniques of image representation, optimizing preprocessing steps, and conducting systematic analyses employing image clustering, we have taken significant steps toward this goal. Future work will continue to refine and expand these methods to further enhance educational outcomes and facilitate personalized learning experiences in the online education landscape.

ACKNOWLEDGMENTS

BLINDED FOR REVIEW

REFERENCES

[1] Jessica Andrews-Todd, Jonathan Steinberg, Michael Flor, and Carolyn M Forsyth. 2022. Exploring automated classification approaches to advance the assessment of collaborative problem solving skills. *Journal of Intelligence* 10, 3 (2022), 39.

[2] Michelle Banawan, Jinnie Shin, Renu Balyan, Walter L Leite, and Danielle S McNamara. 2022. Math Discourse Linguistic Components (Cohesive Cues within a Math Discussion Board Discourse). In *Proceedings of the Ninth ACM Conference on Learning@ Scale*. 389–394.

[3] Sami Baral, Anthony Botelho, Abhishek Santhanam, Ashish Gurung, Li Cheng, and Neil Heffernan. 2023. Auto-scoring Student Responses with Images in Mathematics. In *The Proceedings of the 16th International Conference on Educational Data Mining*.

[4] Sami Baral, Anthony F Botelho, John A Erickson, Priyanka Benachamardi, and Neil T Heffernan. 2021. Improving Automated Scoring of Student Open Responses in Mathematics. *International Educational Data Mining Society* (2021).

[5] Randy Elliot Bennett. 2011. Automated scoring of constructed-response literacy and mathematics items. *Retrieved April* 14 (2011), 2011.

[6] Anthony Botelho, Sami Baral, John A. Erickson, Priyanka Benachamardi, and Neil T. Heffernan. 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning* 39, 3 (2023), 823 – 840.

[7] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

[8] Melanie M Cooper, Mike Stieff, and Dane DeSutter. 2017. Sketching the invisible to predict the visible: From drawing to modeling in chemistry. *Topics in cognitive science* 9, 4 (2017), 902–920.

[9] David Ebert. 2014. Graphing projects with Desmos. *The Mathematics Teacher* 108, 5 (2014), 388–391.

[10] John A Erickson and Anthony Botelho. 2021. Is it fair? Automated open response grading. In *International Conference on Educational Data Mining*.

[11] John A Erickson, Anthony F Botelho, Steven McAteer, Ashvini Varatharaj, and Neil T Heffernan. 2020. The automated grading of student open responses in mathematics. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 615–624.

[12] Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph J Williams, and Sharad Goel. 2019. MathBot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence* (2019).

[13] Markus Hohenwarter and Markus Hohenwarter. 2002. GeoGebra. *Available on-line at http://www. geogebra. org/cms/en* (2002).

[14] Mahmood H Hussein, Siew Hock Ow, Monther M Elaish, and Erik O Jensen. 2022. Digital game-based learning in K-12 mathematics education: a systematic literature review. *Education and Information Technologies* (2022), 1–33.

[15] Gwo-Jen Hwang, Sheng-Yuan Wang, and Chiu-Lin Lai. 2021. Effects of a social regulation-based online learning framework on students' learning achievements and behaviors in mathematics. *Computers & Education* 160 (2021), 104031.

[16] Muhammad Irfan, Betty Kusumaningrum, Yuyun Yulia, and Sri Adi Widodo. 2020. Challenges during the pandemic: use of e-learning in mathematics learning in higher education. *Infinity Journal* 9, 2 (2020), 147–158.

[17] Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education* 112 (2022), 103631.

[18] Lan Jiang, Huidong Chen, Luca Pinello, and Guo-Cheng Yuan. 2016. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome biology* 17, 1 (2016), 1–13.

[19] B Karthikeyan, Dipu Jo George, G Manikandan, and Tony Thomas. 2020. A comparative study on k-means clustering and agglomerative hierarchical clustering. *International Journal of Emerging Trends in Engineering Research* 8, 5 (2020).

[20] Manju Kaushik and Bhawana Mathur. 2014. Comparative study of K-means and hierarchical clustering techniques. *Int. J. Softw. Hardw. Res. Eng* 2, 6 (2014), 93–98.

[21] Nahum Kiryati, Yuval Eldar, and Alfred M Bruckstein. 1991. A probabilistic Hough transform. *Pattern recognition* 24, 4 (1991), 303–316.

[22] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36, 2 (2003), 451–461.

[23] Nava L Livne, Oren E Livne, and Charles A Wight. 2007. Can automated scoring surpass hand grading of students' constructed responses and error patterns in mathematics. *MERLOT Journal of Online Learning and Teaching* 3, 3 (2007), 295–306.

[24] Yong Luo, Yonggang Wen, Dacheng Tao, Jie Gui, and Chao Xu. 2015. Large margin multi-modal multi-task feature extraction for image classification. *IEEE Transactions on Image Processing* 25, 1 (2015), 414–427.

[25] Peter Maclaren. 2014. The new chalkboard: the role of digital pen technologies in tertiary mathematics teaching. *Teaching Mathematics and Its Applications: International Journal of the IMA* 33, 1 (2014), 16–26.

[26] Mahshad Mahdavi, Richard Zanibbi, Harold Mouchere, Christian Viard-Gaudin, and Utpal Garain. 2019. ICDAR 2019 CROHME+ TFD: Competition on recognition of handwritten mathematical expressions and typeset formula detection. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1533–1538.

[27] Filip Moons, Ellen Vandervieren, and Jozef Colpaert. 2022. Atomic, reusable feedback: a semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers. *Computers and Education Open* 3 (2022), 100086.

[28] Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378* (2011).

[29] Ryosuke NAKAMOTO, Brendan Flanagan, Taisei Yamauchi, Dai Yilling, Kyosuke Takami, and Horoaki Ogata. 2023. Enhancing Automated Scoring of Math Self-Explanation Quality using LLM-Generated Datasets: A Semi-Supervised Approach. (2023).

[30] Sumitra Pokhrel and Roshan Chhetri. 2021. A literature review on impact of COVID-19 pandemic on teaching and learning. *Higher education for the future* 8, 1 (2021), 133–141.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[32] İpek SARALAR-ARAS. 2022. Countries' responses to Covid-19 pandemic in K-12 education: Toward a digital education in mathematics. *Uluslararası Sosyal Bilimler Eğitimi Dergisi* 8, 2 (2022), 450–478.

[33] Satya Chaitanya Sripada and M Sreenivasa Rao. 2011. Comparison of purity and entropy of k-means clustering and fuzzy c means clustering. *Indian journal of computer science and engineering* 2, 3 (2011), 343–346.

[34] Michael Steinbach, Levent Ertöz, and Vipin Kumar. 2004. The challenges of clustering high dimensional data. In *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition*. Springer, 273–309.

[35] Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic Short Math Answer Grading via In-context Meta-learning. *arXiv preprint arXiv:2205.15219* (2022).

Received 2 October 2023