Clustering Free-Form Sketch Scenes through Perceptual Similarity

by

Şerike Çakmak

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Science and Engineering

July 14, 2016

Clustering Free-Form Sketch Scenes through Perceptual Similarity

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Şerike Çakmak

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. T. Metin Sezgin (Koç University)(Advisor)

_____

Assoc. Prof. Engin Erzin (Koç University)

_____

Assist. Prof. Ayşe Küçükyılmaz (Yeditepe University)

Date: _____

To my family

ABSTRACT

Similarity is a fundamental concept in psychology that underlies object recognition, classification and clustering. Psychologists have theorized many explanations for human mind's ability to perceive similarity. Yet, there is still no agreed upon theory, partly because similarity is quite subjective and varies with the features attended by the subject. In this thesis, we explored the concept of similarity for hand-drawn sketches, and address the problem of building a gold standard for assessing similarity and clustering free-form sketch scenes through perceptual similarity. Toward this end, we collected a large dataset consisting of 2400 hand-drawn scenes. We further designed a table-grouping protocol for obtaining a measure of similarity through similarity ratings of human assessors. We verified the validity of the constructed gold standard through inter-rater agreement. We evaluated the performance of the clustering system by measuring the degree of agreement with the constructed gold standard. We obtained high agreement scores, showing that the clustering system operates very similar to human way of grouping sketch scenes.

ÖZETÇE

Benzerlik; nesne tanıma, sınıflandırma ve gruplama kavramlarının altında yatan psikolojideki temel kavramlardan biridir. Psikologlar, insan beyninin benzerliği algılama yetisini açıklamak için bir çok teori ortaya atmışlardır. Fakat, kısmen benzerliğin öznel olması ve öznenin dikkat ettiği özelliklere göre değişmesi yüzünden, henüz üzerinde anlaşma sağlanan genelgeçer bir teori yoktur. Bu tezde, benzerlik kavramını elle çizilen çizimler için inceleyip, benzerlik ölçümü için bir altın standart oluşturma ve serbest formlu çizim sahnelerini algısal benzerlikleri üzerinden gruplama problemini ele aldık. Bu amaçla, elle çizilmiş 2400 çizim sahnesinden oluşan büyük bir veri kümesi derledik. Ayrıca, uzman insanların oluşturduğu gruplardan benzerlik değerlerini elde edebileceğimiz bir masa gruplama protokolü tasarladık. Oluşturduğumuz altın standardın doğruluğunu, kullanıcılar arasındaki uyumu ölçerek kanıtladık. Gruplama sisteminin performansını, oluşturulan altın standart ile uyum derecesini ölçerek değerlendirdik. Gruplama sisteminin insanların çizim sahnelerini gruplama şekline çok benzer çalıştığını gösteren, yüksek uyum skorları elde ettik.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Assoc. Prof. T. Metin Sezgin for the valuable guidance and support that he provided throughout my studies.

I am also thankful to Assoc. Prof. Engin Erzin and Assist. Prof. Ayşe Küçükyılmaz for agreeing to be in my thesis committee and contributing with their comments.

Most specially, I would like to express my gratitude to my beloved parents Fahri Çakmak and Emine Çakmak for always believing in me during this long process. I would specifically thank my dearest Hafize Çakmak, for being my little joyful sister. Not only did she provided such a great motivation to me, but she also helped me with my experimental studies.

Seperately, I would also like to present my warmest thanks to my childhood friends Zeynep Öztürk and Cemile Kuytu, for always being available when I needed them and helping me in all situations. Besides, I thank Buket Yüksel with all my heart for being such a sincere friend with all her emotional support.

Finally, I owe many thanks to all members of IUI Lab: Kemal T. Yeşilbek, B. Berker Türker, Özem Kalay, Banuçiçek Gürcüoğlu, Ezgi Emgin, Ozan C. Altıok, Kurmanbek Kaiyrbekov, Ferhat Çağan, Emre Karaman for listening to my research presentations countless times and providing valuable feedback. In particular, I would like to thank Cansu Şen for her great friendship and support, Neşe Alyüz and Çağla Çığ for their valuable helps during my time here in Koç University.

This thesis is funded by TUBITAK under grant number 113E059.

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE

IDM	Image Deformation Model
AUC	Area Under Curve
STD	Standard Deviation
CSV	Comma-Separated Values
QR code	Quadratic Residue Code
UML	Unified Modelling Language

Chapter 1

INTRODUCTION

Similarity is one of the central problems in psychology, which is defined as the perceptual or conceptual relationship holding between two objects. In psychology, similarity plays a fundamental role in understanding human's way of classifying objects, making generalizations and forming concepts. It is the basic principle that underlies object recognition and classification in computer vision literature.

Psychologists have suggested various theoretical approaches for defining similarity, namely common elements approach, template approach, geometric approach and feature approach [Blough, 2001]. In common elements approach, the proportion of common elements determines similarity of the objects. Template models define similarity as a point to point correspondence between image-like representations of objects. Geometric models represent the objects as points on the coordinate space, and similarity between these objects is defined as the inversely proportional distances between these points. Tversky defines feature approach as the representation of two objects each of which contains its own unique features and also contains common features [Tversky, 1977]. Despite the whole research in the area, deciding the best approach is still controversial. The most appropriate method changes according to the nature of the problem. Among all these approaches, we prefer feature model by considering its suitability to sketch recognition domain.

Providing a natural way of communication, free-hand sketching is being more and more popular with the availability of pen-based computers. So, the use of pen-based tablets and computers is increasing in our daily lives. With the ease they provide, pen-based tablets are candidate for taking the place of traditional pen and paper in

schools. Instructors prefer using educational tools like tablets and boards for teaching the lectures and testing students. Students answer the questions on their tablets by drawing diagrams, figures, namely free-form sketches. The current popularity of pen-based tablets in modern education life brings a huge interest in computational methods for processing and understanding sketches. Developing sketch recognition based intelligent systems has several advantages of helping teachers such as simplifying the question preparation process and shortening the duration of exam grading task. However, research conducted to develop intelligent sketch interpretation algorithms so far mainly focused on domain specific, template based approaches with isolated symbols.

In this thesis, we explore the similarity perception from sketch recognition aspect and address the problem building a gold standard for assessing similarity and clustering free-form sketch scenes through perceptual similarity of corresponding to the answers of specific questions which are drawn by students. Traditional methods would try to solve this problem by recognizing a predefined list of sketch symbols which would require domain specific knowledge. Instead, we offer a domain independent clustering approach that can measure the similarity between free-form sketch scenes and group them accordingly.

To better understand how people would solve the problem we studied famous Gestalt theory [Wertheimer, 1938]. This theory asserts that human mind perceives concepts as a global whole instead of individual parts. Gestalt theory is established on grouping principles including law of proximity, law of similarity, law of closure, law of symmetry and law of continuity [Sternberg and Sternberg, 2016]. In order to build a gold standard for our clustering system we organized perceptual sketch grouping experiments with human assessors. During the experiments, we observed that people instinctively follow the principles introduced by Gestalt theorists.

Our approach is inspired from human perception. We used classic sketch features to measure the similarity among the scenes. We further clustered the large set of hand-drawn sketch scenes by using the agglomerative clustering algorithm. Our

methodology is then confirmed by measuring the agreement degree of our system's resulting clusters with the human created gold standard clusters.

Chapter 2 explores the literature to review the research performed in the area. Chapter 3 describes the data collection process in detail. Chapter 4 provides information about the carefully designed sketch grouping experiment to build a gold standard for performance evaluation. Chapter 5 describes our methodology of developed algorithms for feature extraction and hierarchical clustering. Chapter 6 summarizes the performance evaluation results by following two different metrics. Chapter 7 concludes the thesis and presents possible future work about this research problem.

Chapter 2

RELATED WORK

Our work concentrates on discovering perceptual similarity among sketch scenes. Therefore, we summarize the related work from these two aspects. We first present the recent findings from sketch recognition domain in human computer interaction literature. Then, we discuss different approaches for measuring human perception of similarity.

Sketch recognition is applied to many different domains like UML diagrams (in software engineering), military action diagrams, architectural drawings and hand-drawn chemical diagrams. Works by [Hammond et al., 2010] and [Ouyang and Davis, 2007] represent state of the art in this subject. However, the research problems focused in the literature are all dependent on the predefined specific class information of sketch symbols. In reality, it is not possible to define all the symbol classes that are going to be used in sketching applications. Our thesis work does not use any context information or predefined symbol classes. On that sense, we differ from the works existent in the literature.

Classroom Presenter is the first system in the area of pen-based educational systems that provides applications for students to answer questions by free-form drawings [Anderson et al., 2007]. Following this, the application called Classroom Learning Partner was developed by [Jóźwiak, 2011]. These two systems revealed the problem of interpreting many sketches at once in a restricted amount of time for teachers. Then, a system that applies sketch recognition to detect special symbols like arrows and boxes from the scenes was developed by [Smith, 2006]. This system provided a way of aggregating student answers. Yet, these works did not solve the problem of our interest, specifically the case when teachers desire to ask questions that require

drawing free-form answers which do not match any templates.

Clustering algorithms are frequently preferred when automatic grouping of similar objects is desired. Vast amount of clustering algorithm types exists, including, but not limited to, K-means clustering, aggloremative hierarchical clustering, mean-shift clustering, spectral clustering and affinity propagation [Sculley, 2010], [Defays, 1977], [Comaniciu and Meer, 2002], [Ng et al., 2002], [Frey and Dueck, 2007]. Among those algorithms, we preferred agglomerative hierarchical clustering where individual sketch scenes were initialized in their own clusters in the beginning and similarity groups were formed as the output of the algorithm. Being suitable for large number of samples and large number of clusters, being adaptable to any pairwise distance metric and not requiring any kind of random initialization lead this clustering algorithm to be the perfect choice for our problem statement.

The most similar works to ours in terms of utilizing clustering methods for sketch similarity assessment is the two systems developed for pen-based interactive boards that assist users with their selection [Lindlbauer et al., 2013], [Perteneder et al., 2015]. The clustered items in these two work include isolated sketch objects, not the complex sketch scenes. Another similar system to ours clusters sketches collected via pen-based tablets from students and assists instructors with the grading job [Hatfield, 2011]. Our motivation is parallel to theirs, but they fail to handle free-form sketch scenes. Instead, their work focuses on hand-written answers that require domain knowledge. Another usage area for clustering similar items is image clustering. Similar to our proposed system, the auto album creates a hierarchy of albums [Platt, 2000]. Unlike our perceptual grouping system, it uses the time and order of photo creation together with the color information.

We analyzed different approaches for measuring human perception with the purpose of evaluating clustering results. We performed an extensive search on this area to make use of the most appropriate methods in the sketch grouping experiment. The aim was to select the methodology that provides similarity ratings for the large sketch scene dataset within a doable amount of time. An efficient technique called

similarity via spatial arrangement was introduced by [Goldstone, 1994]. The spatial arrangement task was later used with the purpose of learning perceptual kernels for visualization design [Demiralp et al., 2014]. Subjects are given a random set of objects distributed on a computer screen and required to arrange them spatially in such a way that the distances among those objects correspond to dissimilarities. A similar technique called multi-arrangement was proposed by [Kriegeskorte and Mur., 2012]. With this method, subjects arrange subsets of objects spatially on the screen and dissimilarity information is inferred from multiple arrangements. These two methods actually fit into our purpose of obtaining similarity information from humans, but they are not suitable in terms of the experiment duration and screen size of the computers. To eliminate the area restriction of the computer screens, we followed a similar approach, called table scaling, which was originally developed for measuring perceptual image similarity [Rogowitz et al., 1998]. For our sketch grouping experiment, we followed Rogowitz et al.'s approach and extended Gurcuolgu's experiment setup of acquiring messiness information of sketches to meet our needs[Gurcuoglu, 2014].

Chapter 3

DATA COLLECTION

Since our problem is to build a gold standard for perceptual similarity of hand-drawn sketch scenes and group them accordingly, we needed a large dataset of free-form sketches. We designed a pretty comprehensive scenario of data collection for this purpose.

We had two major constraints to satisfy in the data collection task. First constraint was to provide a free environment for the participants where they do not feel under pressure while drawing and the observer does not guide them unintentionally. Considering this constraint, carrying out the data collection task in a fully controlled environment would not be preferable at all. Therefore, we preferred a comforting environment for the participants. Second constraint of ours was to create a realistic scenario that would supply perceptually similar free-form sketch scenes for our research problem. Desired sketch scenes consisted of multiple sketch words that were freely decided by the drawer. Number of sketch words was not specified, scenes did not fit into any template and most importantly participants were not a given any instruction on what to draw, all of which meet with our purpose of collecting free-form sketch scenes. The only instruction provided to the participants was that they had to answer a question by drawing within their knowledge store.

Our realistic data collection scenario satisfying the explained constraints was based on a diverse set of primary school questions from mathematics, physics and biology domains, that are proper to be answered by hand-drawn sketches. We asked teachers to prepare 8 questions from those domains and directed these questions to students. This process required us developing appropriate applications both for students and teachers of concern. In order to ensure natural interaction with the device, we devel-

oped our applications for pen-based tablets having active stylus for drawing purposes.

3.1 Development of Student and Teacher Applications

We developed android applications for the data collection task. Two separate applications for students and teachers were implemented on Samsung Galaxy Note GT-N8005 16GB 10.1" tablet which has Android 4.1 (Ice Cream Sandwich) operating system. Later, the applications installed on 15 android tablets owned by IUI Laboratory at Koç University. Free-form sketch scenes were collected using these tablets together with the developed applications.

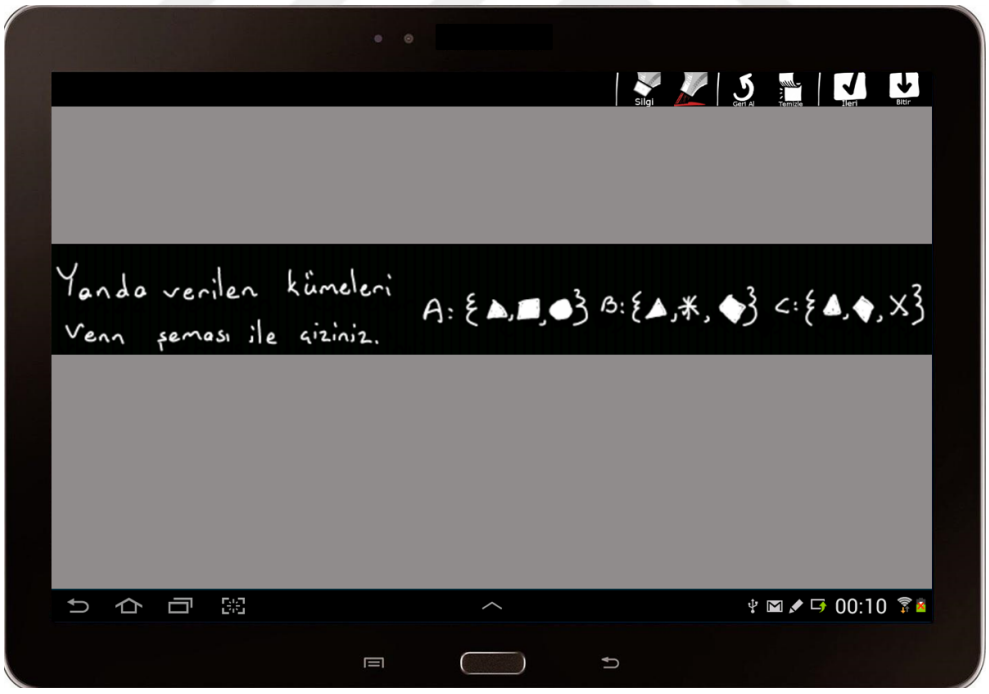


Figure 3.1: Teacher application - question preparation

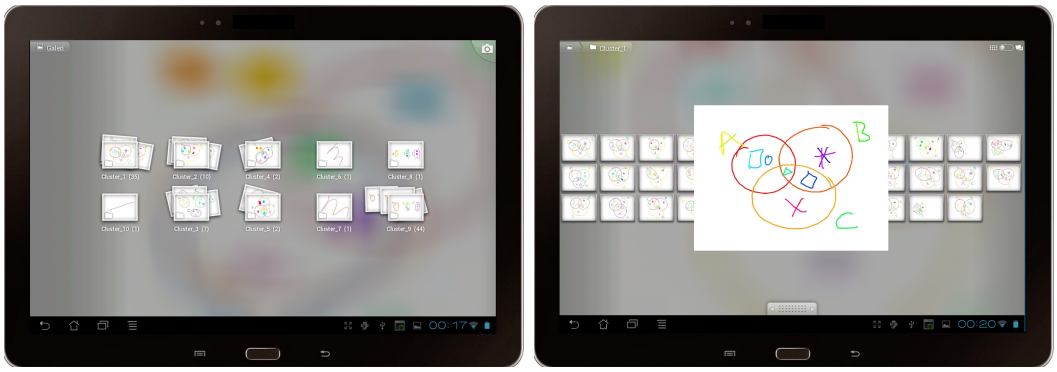
3.1.1 Teacher Application

Teacher application provides functions for preparing a question, directing it to the students' tablets and viewing the solutions coming from the students in a grouped

manner. Figure 3.1 illustrates the interface of the teacher application's question preparation option. Teacher is able to write the question on the provided area and he/she is able to draw any shape that he/she needs for explaining the question.

Teachers' application has the rubber option, which can be used to delete any part of the drawing. The application also has the undo and clear options. Undo option is used for deleting the last drawn stroke when the button with the label 'Geri Al' is clicked. Clear option is used for deleting everything drawn on the screen when 'Temizle' button is clicked. In order to back switch to the writing mode, teachers can use the button having an image of pen. When they were done writing the current question, in order to prepare one more question, they could click the next button on the top right, labeled 'İleri'. When teacher is completely finished with preparing the question set, then she/he can use the finalize button, on the very top right corner of the screen, labeled as 'Bitir'.

The options for directing the question to the students and viewing solutions is available on the home screen of the teacher application. Teachers can simply use the relevant buttons for directing the questions to the students and viewing the grouped free-form answers given by the students. Figure3.2 shows the screen-shot of the teacher application for displaying grouped answers.



(a) View the groups (b) Zoom in to groups

Figure 3.2: Teacher application - viewing grouped sketch scenes

3.1.2 Student Application

Student application provides only one function for answering the question sets prepared by the teacher. Figure 3.3 illustrates the interface of the student application's question answering option. Student is able to draw the answer on the provided area. He/she can use the whole area as they wish. Drawing area is intendedly kept so large for the students to enable them freely use the area without any restriction while drawing the answers.

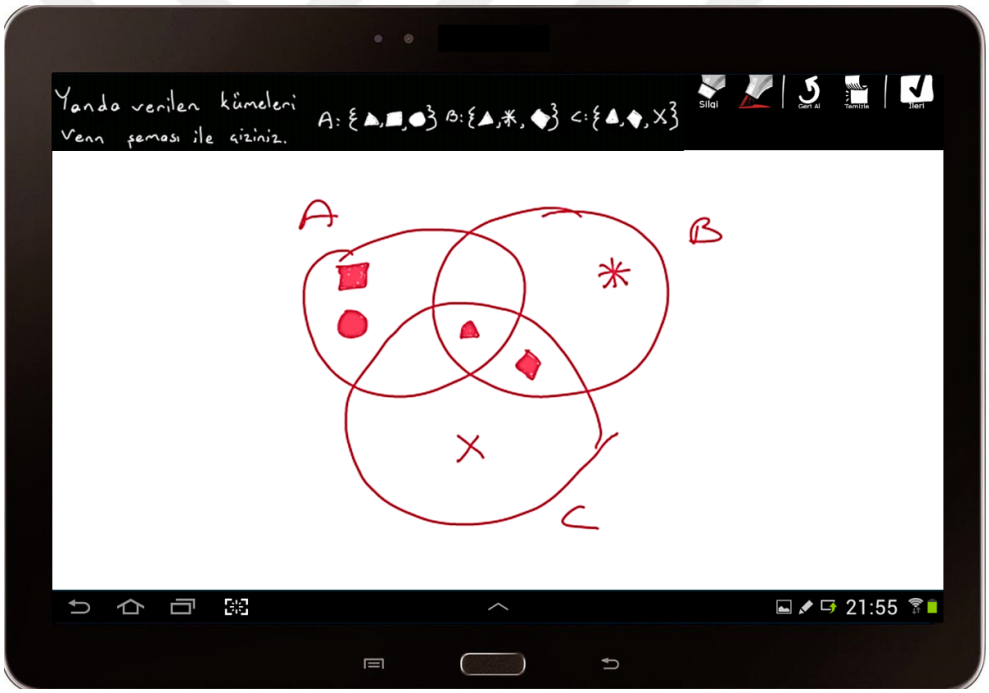


Figure 3.3: Student application - question answering

Students' application has the rubber option, which can be used to delete any part on the screen in order to correct the mistakes while answering the question. The application also has the undo and clear options for the same purpose. Undo option is used for deleting the last drawn stroke when the button with the label 'Geri Al' is clicked. Clear option is used for deleting everything drawn on the screen when 'Temizle' button is clicked. In order to back switch to the writing mode, students can use the button having an image of pen. When they are done drawing the answer of

the current question, in order to proceed to the next question, they can click the next button on the top right corner, which is labeled as 'İleri'.

3.2 Dataset Construction

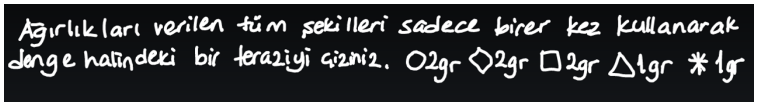
The free-form sketch dataset was constructed as a result of the collaborative work with teachers and students from 12 schools (secondary school and high school) located in Istanbul. We had legal permission from the ministry of education in order to perform data collection in those schools. We requested the teachers to prepare questions from science and mathematics domain and organized a workshop for that purpose. Teacher application was used for preparing the questions. Among those, we identified 8 proper questions from different categories and collected answers for that categories of questions. Questions were posed to the students from that 12 schools. Students told that the task was not an exam and they would not be graded based on their answers. Teachers did not observe students while drawing the answers, they solved the questions as self exercise. The constructed dataset consists of 2400 free-form sketch scenes in total, having 300 free-form answer for each question category. Figure 3.4 shows the set of questions that were used in the data collection task.

3.3 Annotation

The constructed dataset of free-form sketch scenes was annotated via a sketch annotation tool developed by IUI Laboratory at Koç University. Thanks to the annotation tool, we labeled the class names of the sketch words in 2400 scenes.

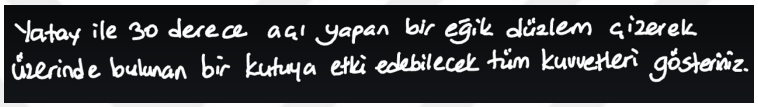
Figure 3.5 shows the sketch word annotation process of an answer given to the flower question. This tool uses Douglas–Peucker algorithm for point detection on a curve by is approximating the curve as a series of points [Douglas and Peucker, 1973].

Related sketch words for this question are flower, leaf, root, arrows and written words. In order to label one of the sketch words in the scene, points of interest should be selected by performing free sketch on the screen. For this purpose, annotator first circles the interested sketch word with the use of a stylus pen and clicks select

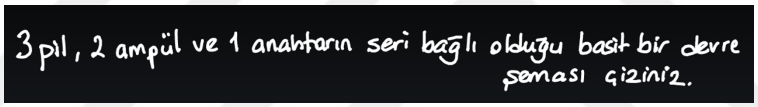


(a) Balance Question

(b) Flower Question



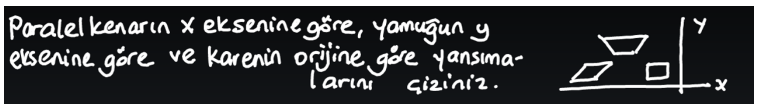
(c) Force Question



(d) Lamp Question

(e) Money Question

(f) Pattern Question



(g) Reflect Question

(h) Set Question

Figure 3.4: Question texts belonging to 8 categories from mathematics and science domain

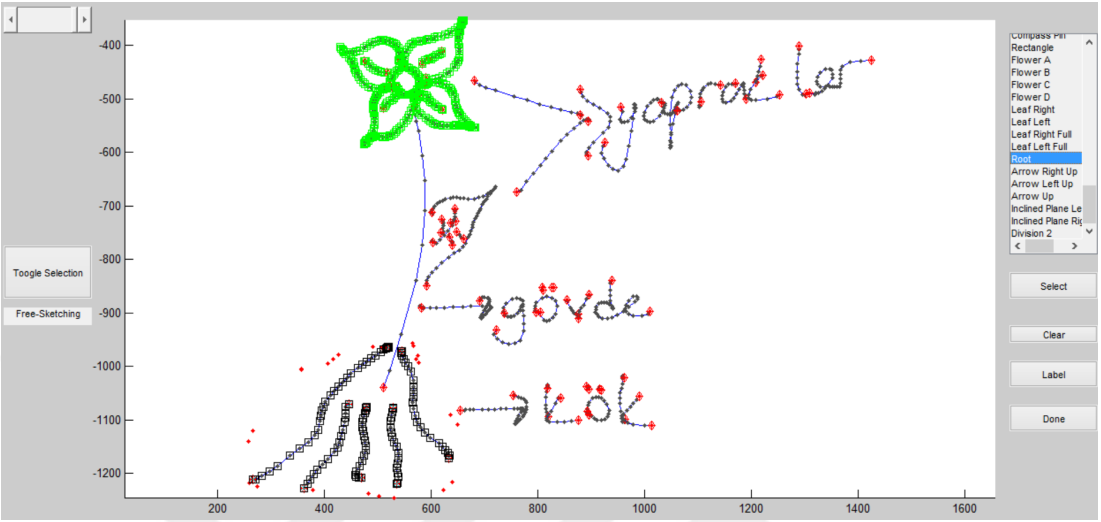


Figure 3.5: Annotation tool

button. Then, the tool shows the selected part by highlighting that part in black color. Selection process is illustrated on the root part of the flower in Figure 3.5. If selection is not correct, clear button is used to reselect the correct piece of sketch. For labeling step, there is a drop-down list showing all the possible labels on the right side. Selected sketch word is labeled by clicking the correct option from that list. When labeling is done system highlights the labeled sketch word in green and saves that sketch word together with its label information automatically.

For each question category, there are 300 free-form answers drawn by students. Since the question categories have different characteristics, answers provided for that questions are also variant in terms of the number of sketch words used. Total of 20750 sketch words labeled in 2400 scenes. Figure 3.6 shows the distribution of sketch words among question categories. As seen from the figure, set question was answered with the use of most sketch words and pattern question was answered by least number of sketch words.

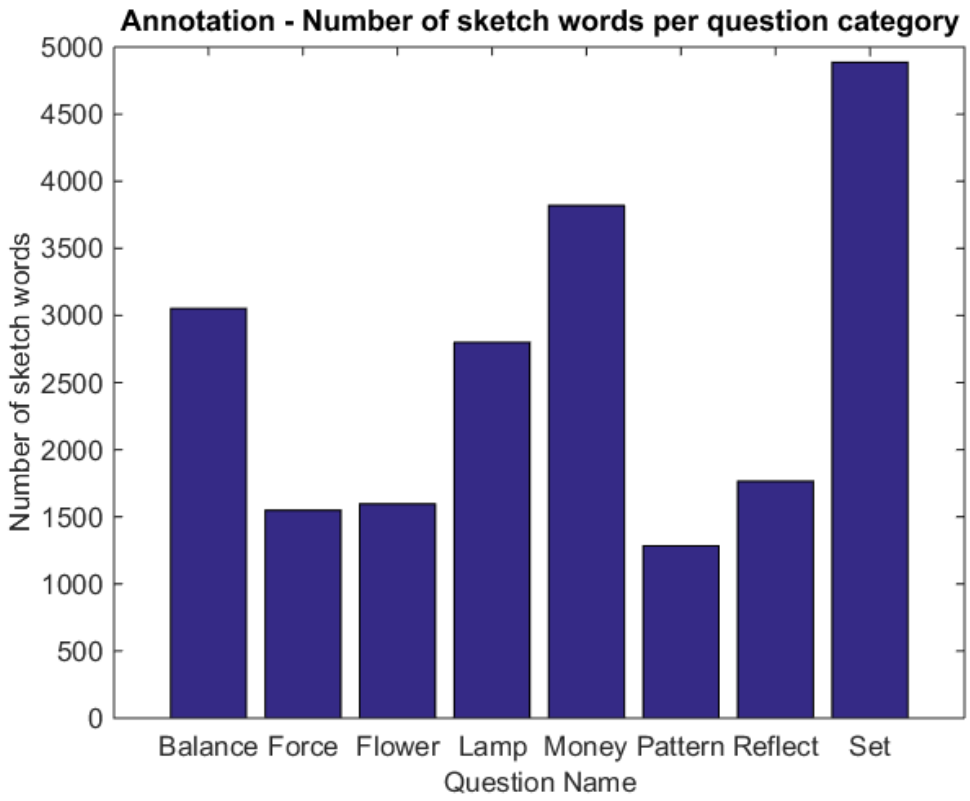


Figure 3.6: Number of sketch words per question category

Chapter 4

GOLD STANDARD

We build a gold standard of perceptually similar sketch scenes to serve as the ground truth clustering. In order to construct the desired gold standard for measuring the similarity of sketches, we carefully designed an experiment setup.

Our purpose in this experiment is to examine the issue of sketch similarity from the perspective of the human observers in order to build a gold standard. We aim to use this gold standard in the evaluation phase of the sketch clustering methods that we develop. Different behavioral methods for acquiring similarities are available in the literature. The most preferred methods are pairwise similarity judgments, perceptual confusion tasks, free sorting, single arrangement, and multi arrangement [Kriegeskorte and Mur., 2012].

In pairwise similarity judgments, each pair of items is presented in isolation and the subject rates the similarity on a scale. In confusion tasks, subjects are presented with two similar items and asked whether the items are the same or different. The probability of the confusion between these two items is measured simultaneously. These two methods are very slow since $(n^2 - n)/_2$ separate judgments are required, where n is the number of items. In free sorting, subjects are instructed to place items into groups. This method actually suffers from graded similarity estimates for individuals. Single arrangement method has been proposed to overcome this problem, where the subject arranges items in $2D$ by considering the distances between the items which reflect dissimilarities [Goldstone, 1994]. However, spatially arranging 300 sketch cards at the same time is not possible both on a table and on a computer screen in our case. Even worse, if we chose the multi arrangement method, multiple item subsets would be arranged iteratively, causing the experiment to last for hours.

Due to the time complexity of other methods, we decided to follow free sorting method in the experiment. Free sorting method is also consistent with the famous Gestalt principle of perceptual organization, which states that similar things will tend to be grouped together by humans [Köhler, 1970]. In order to overcome the lack of graded similarity estimates problem, we averaged the scores over all participants and acquired the graded similarity estimates. Throughout the experiment, we assumed that the frequency with which two items are placed in the same group is proportional to their perceptual similarity.

Another important issue while designing the experiment was to decide the experiment environment. Due to the restricted size of any possible display, a computer interface would be a limitation for the participants. Rather than designing a computer interface for the experiment, we prepared a more realistic and a more interactive environment. Our experiment design was inspired from table scaling experiment, which was previously used in the perceptual image similarity context[Rogowitz et al., 1998]. We took Gurcuoglu's sketch ranking experiment as an example and modified it to meet our needs [Gurcuoglu, 2014].

4.1 Sketch Grouping Experiment

We carefully prepared an experiment in which the participants were requested to group the similar sketches. In order to make the experiment conditions more realistic, we conducted the experiment with the use of printed versions of sketches.

Sketch images were prepared automatically by a matlab script. By this script, sketch data was retrieved from stored xml files and {sketch id,related question name} information was embedded to the image. For the purpose of embedding that information into the sketch cards, we developed a QR code generator. With the help of this QR code generator script, identificatory information for the sketch was placed on the lower right corner of the image. Sketch scenes were centralized and the size of the images was standardized. Prepared sketch images were printed on rigid cardboard and they were cut out according to a fixed sized the template. A sample sketch card

used in the experiment is presented in Figure 4.1.

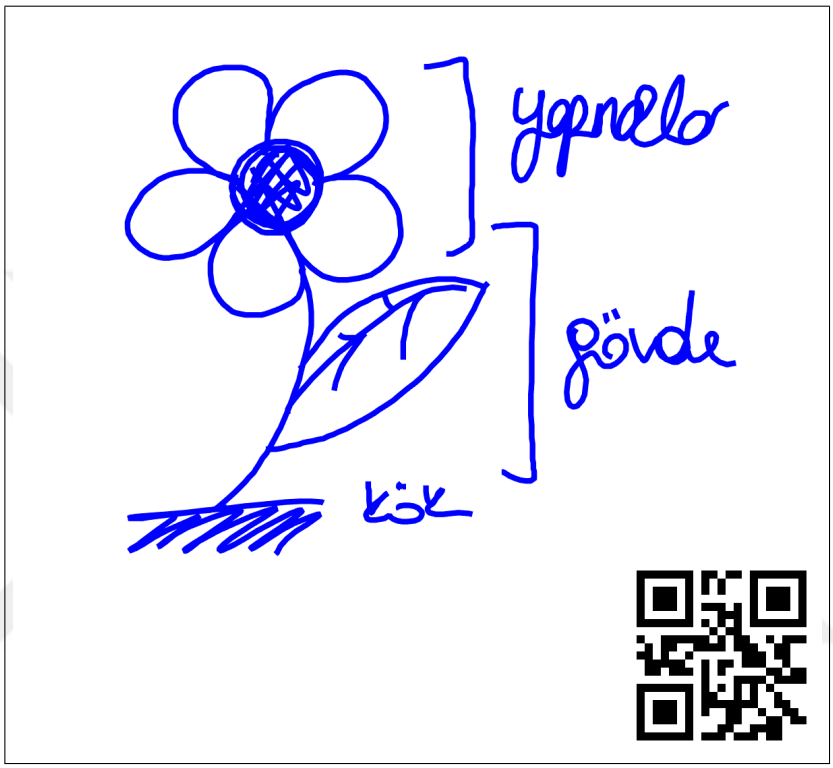


Figure 4.1: A sample sketch card

Experiment was conducted in separate sessions and two participants attended each session at the same time, grouping 300 sketch cards each. The meeting room (ENG 208 at Koç University) for the experiment, where a suitable table was present to spread the printed sketch cards, was prepared by the conductor beforehand. The experiment setting can be seen in Figure 4.2.

Each human assessor participating to the experiment, attended a training session before starting the grouping task. In the training session, we told the participants that these 300 printed sketch cards are all answers to the very same question given by primary or high school students. We also informed the participants that the context of the question would not be shared with them. So, they were asked to group 300 sketches just by considering visual and perceptual similarities. We warned the participants that the sketch scenes with similar visual properties should be placed in

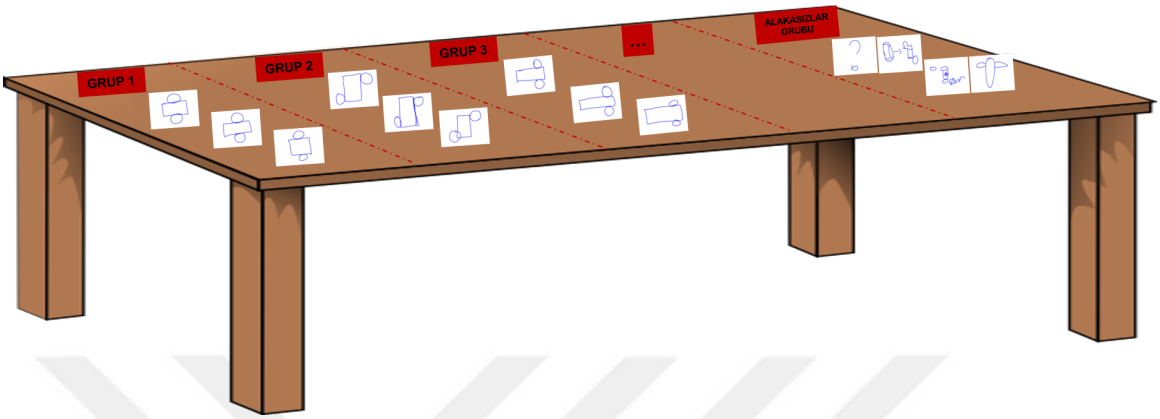


Figure 4.2: Sketch grouping experiment setup

the same cluster, while the ones with different visual properties should not be in the same cluster. Since this specification had an important role for the experiment, they were provided with various examples illustrating the case.

Number of maximum groups was specified as 15 for all types of questions and table was prepared accordingly beforehand. However, the assessors had the right to define the exact number of groups that is proper for the current question. They were free to construct as many clusters as they wish, provided that they do not exceed the limit 15. If the assessors choose to use less number of groups than 15, then they had to rearrange the table by easily sliding the group separators.

One of the groups was always reserved for irrelevant answers. The irrelevant group was used only for the answers that has no common property with any of the other groups. If the assessors could not find any suitable group for a sketch, then they were allowed to include that sketch into the irrelevant group. It was pointed out for all participants that, irrelevant group had to be used correctly.

After the training session, we invited the participants to complete a quick test session in order to verify that they have fully understood the task. We asked some questions to confirm their understanding and explained the points if they were unclear.

Participants completing the training and test phases were taken to the meeting room for the experiment. We provided the printed sketch cards containing the free-

Figure 4.3: A snapshot from the sketch grouping experiment: Two assessors grouping perceptually similar sketches

form answers to one type of question as a deck to the participants. Each participant had a different type of question. In order to begin the grouping, one needs to quickly scan all 300 sketch scenes for gaining an insight. In the preliminary experiments performed by the conductor, this scanning phase took a long time. Therefore, for each question a video was prepared to preview the answers. Participants asked to try estimating possible groups to be formed, while watching the preview video. Later, human assessors were provided with the deck of 300 sketch cards and they were asked to gather perceptually similar sketches into groups within a limited time period. While forming the groups, they were free to change the group of sketch cards as they wish unless the task was finished. A snapshot taken during the experiment is shown in

Figure 4.3.

After the assessors finished the grouping task, experiment conductor recorded the group information of the sketch cards with the help of the QR code scanner program, QuickMark [Qui, 2013]. QuickMark provides continuous scanning functionality which eases the job for scanning 300 sketch cards. It also provides the functionality of exporting scanned data as CVS format.

Figure 4.4 illustrates a finalized sketch grouping task. All sketch cards has its own identificatory QR code on the lower right corner. Group names are also encoded with QR codes and printed on top of the group separators. Scanning job begins by recording the QR codes of the group names. After scanning the QR code for a group, all the cards belonging to that group are scanned consecutively. Remaining groups and sketches are scanned similarly. When the scanning job is finished, scanned history of CVS files are read. For this purpose, another script was written for reading the scanned data and instantly constructing that grouping on a computer with the proper structure. By this means, experiment conductor checked the correctness of the job immediately after scanning all the 300 sketch cards and revised the records if something went wrong during the scanning task.

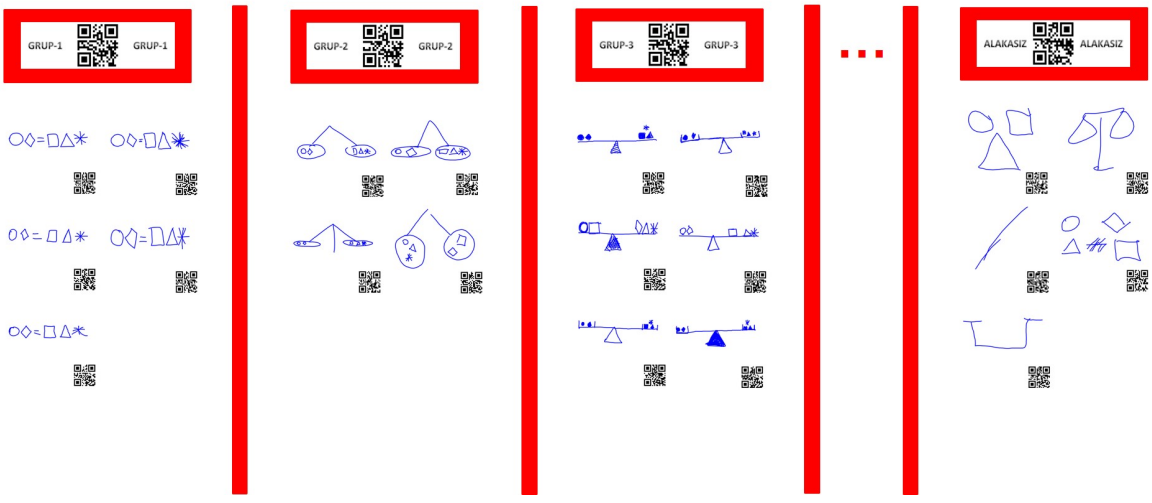


Figure 4.4: A sample scene for a set of grouped sketches

Each assessor participated in the experiment for 8 different categories of questions on different days. Each question category was grouped by 9 different assessors. Question categories consist of mathematics, physics or biology related concepts. For each question category, 300 free-form answers were randomly chosen from the database. Representative answers for each category are presented in Figure 4.5.

Among 9 participants, there were 4 female and 5 male assessors participating in the experiment, whose ages vary between 16-24. Each grouping session took approximately 40-45 minutes. After the grouping task, data was scanned within 15 minutes. In total, perceptual similarity data for 8x300 = 2400 sketch scenes were efficiently obtained.

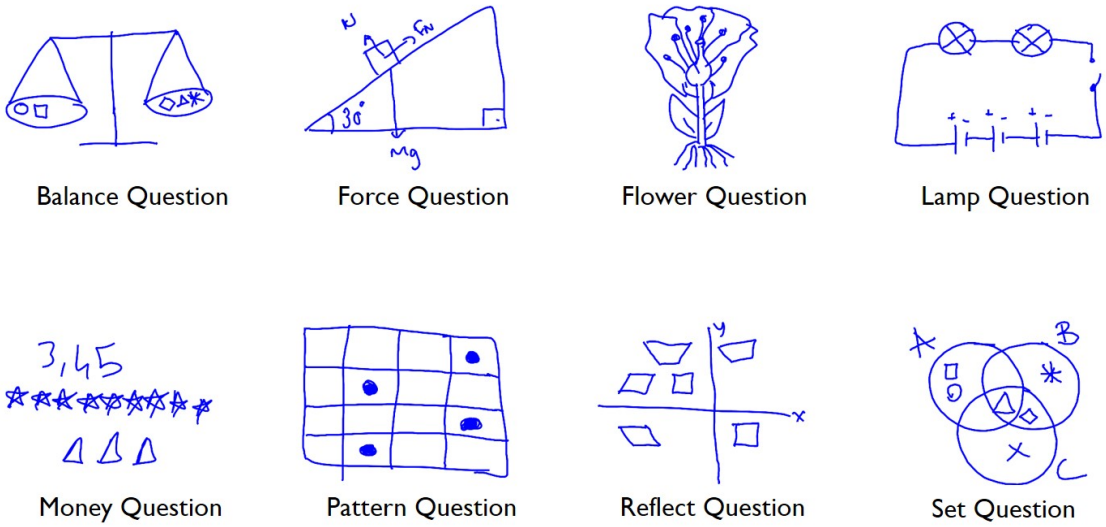


Figure 4.5: Representative sketches for 8 different categories of questions used in the experiment

4.2 Verification of the Sketch Grouping Experiment Data

Verification of the data acquired through user studies is a common concern and hard to achieve. In order to validate the grouping data we collected, we performed various analyses from different aspects to see how the subjects agree.

4.2.1 Participant Data Analysis based on Question Categories

Grouping data was analyzed extensively in terms of all question categories in order to see both the similar and dissimilar aspects of the participants.

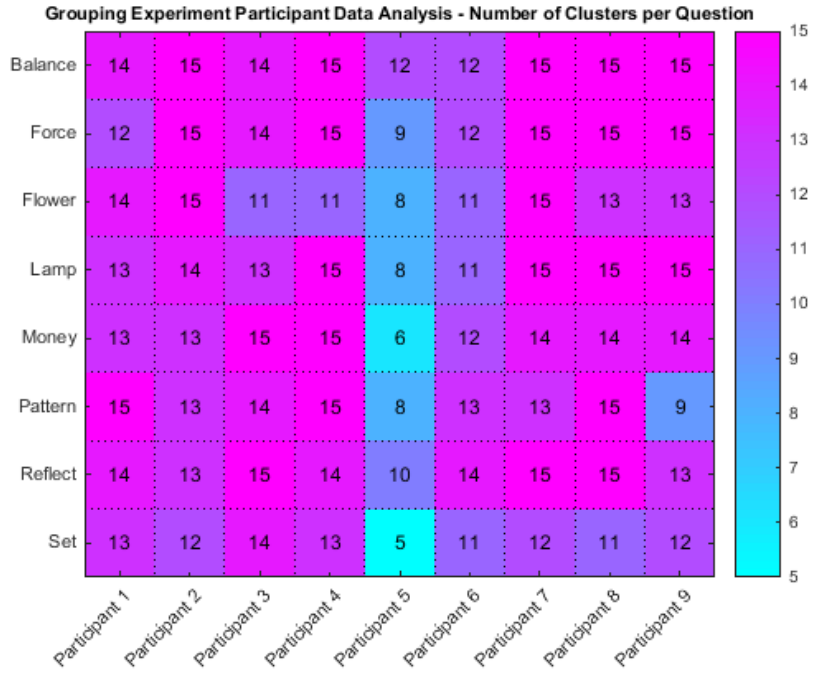


Figure 4.6: Number of clusters formed in the experiment in terms of participants vs. question categories

In the training phase of the sketch grouping experiment, participants informed that they are allowed to form at most 15 clusters for each question category. Yet, they were free to use less number of clusters. We summarized number of formed clusters per question category by each participant. Figure 4.6 shows participants' tendency to use more or less number of clusters. As seen, participant 5 prefers forming less clusters. Other than that, everyone else forms more clusters for all question types. This means that, participant 5 does not pay attention to details too much. General visual properties of the scene are considered as more important in that case. Participant 5 might seem as outlier from the figure, yet it just the result of a subjective preference. That participant's data was examined and it was observed that he/she

forms the groups by considering more explicit similarities of visual properties.

Figure 4.7 shows number of clusters formed in the experiment averaged for 9 participants. It can be seen that, participants prefer the number of clusters to be in the range between 11-15. This both shows that maximum limit is sufficient and groups were formed with adequate attention. Average number of the formed groups slightly differs among question categories. Set question is clustered with the least number of groups and balance question is clustered with the most number of groups.

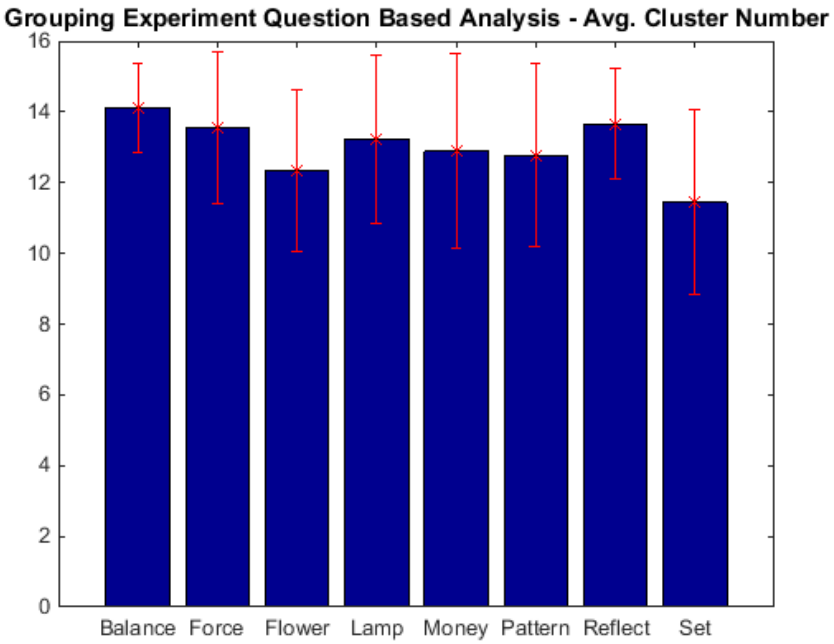


Figure 4.7: Average number of clusters formed in the experiment

Another important issue was to balance the number of instances in the formed groups. For example, if a participant forms 15 clusters in total but use just 2 of them heavily and ignores the remaining clusters, then this situation leads to a faulty gold standard in the end. Figure 4.8 proves that mean number of sketches per cluster is both balanced and consistent.

Last analysis of participants' grouping data is related to the correct use of irrelevant group. Use of irrelevant group is critical since a participant could easily prefer

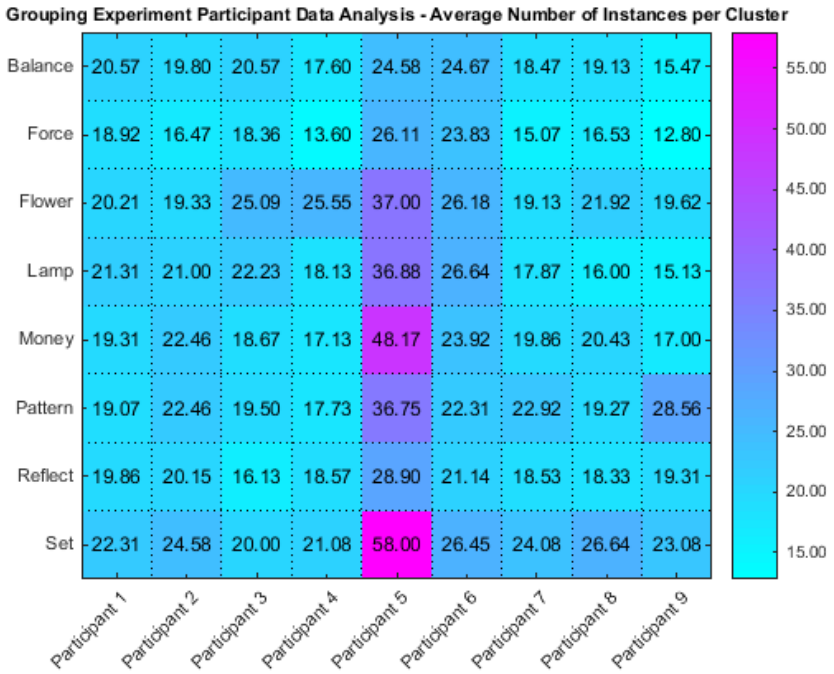


Figure 4.8: Average number of instances per cluster in terms of participants vs. question categories

using this group frequently due to the boredom or rush. During the grouping task, we do not manipulate participants by any warning. Therefore, validating the use of this special group gains significance for the ground truth to be constructed.

Figure 4.9 shows number of sketch cards placed into the irrelevant group. As can be seen in the figure, amount of outliers is not excessive except for the force question. The extraordinary situation with the force question is probably due to the difficulty level of the question. We saw that, it really contains lots of irrelevant, wrong and dissimilar answers, proving that the students answering that question did not understand the subject. Moreover, participant 9 uses irrelevant group more frequently than other users. When we reviewed that participant's grouping solution, we noticed that he/she paid much more attention to the cleanness of other formed groups. So, we validated that the use of special group of irrelevant answers is accurate.

We also calculated the mean number of outliers separately for each question cate-

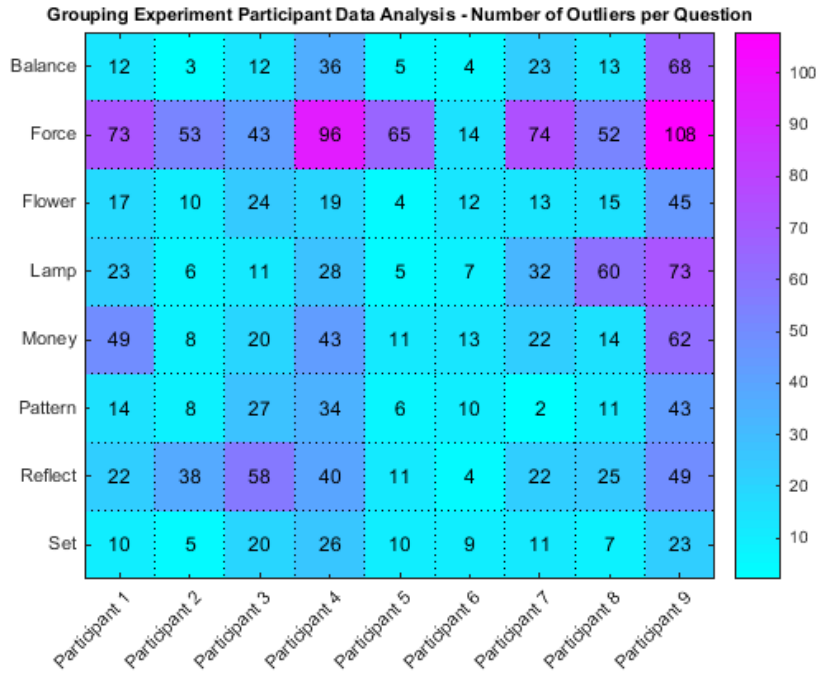


Figure 4.9: Number of outliers in terms of participants vs. question categories

gory. Figure 4.10 illustrates the situation for all question categories. Not surprisingly, force question has the most outliers. Other question types are generally grouped with the use of less outliers, set question being the least of all. We can conclude from here that, irrelevant group was used properly and use of this group is somehow related to the difficulty level of the question.

4.2.2 Inter-rater Agreement

After analyzing the sketch grouping experiment data for 8 different question categories and 9 participants, we validated the consistency of assessors within themselves to make sure that the acquired data is not coherent by chance. For this purpose, we need an evaluation metric for comparing clusterings.

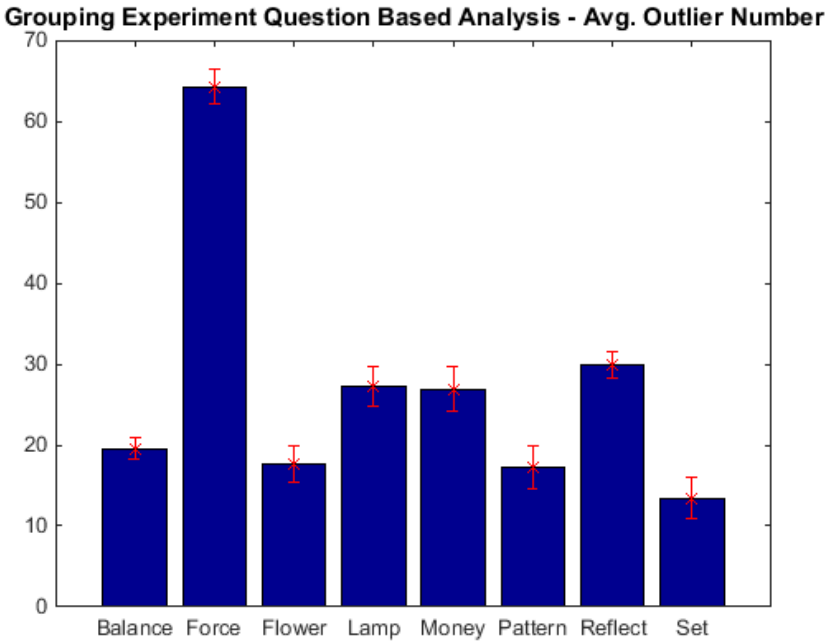


Figure 4.10: Average number of outliers per question category

Comparing Clusterings

There exists two kinds of evaluation metrics, either intrinsic or extrinsic. Intrinsic metrics measure how distant an element from elements in other clusters, and how close elements from one cluster are to each other. On the other hand, extrinsic metrics are based on comparisons between the output of the system and a gold standard which is usually built by human assessors. We focus on extrinsic measures to compare clustering solutions with each other. When doing extrinsic evaluation, determining the distance between clustering solutions, the system output and the gold standard, is complex and it is still subject to discussion. Many different evaluation metrics have been proposed for this purpose, such as Purity and Inverse Purity, clusters and class entropy, VI measure, Q_0, V-measure, Rand Statistic, Jaccard Coefficient, Mutual Information etc. [Amigó et al., 2009]

There are four basic constraints covering all necessary quality aspects that any evaluation metric should satisfy according to Amigó et al. First formal constraint is

cluster homogeneity. This constraint is a very basic restriction which states that the clusters must be homogeneous, meaning that they should not mix items belonging to different categories. The second constraint, called as cluster completeness, states that items belonging to the same category should be grouped in the same cluster. In other words, different clusters should contain items belonging to different categories. Remaining two constraints are the ones that some of the most popular evaluation metrics can not satisfy. One of them is called rag bag. This constraint supports having a "rag bag" of items ("unclassified", "other" or "miscellaneous" cluster). Rag bag constraint states that introducing disorder to into an already disordered cluster is less harmful than introducing disorder into a clean cluster. Last formal constraint that a successful evaluation metric should satisfy is called cluster vs. size quantity. According to this constraint, a small error in a big cluster is preferable to the case where a large number of small errors exist small clusters.

There are various clustering evaluation metrics in the literature, namely metrics based on set matching, metrics based on counting pairs, metrics based on entropy, metrics based on edit distance. These metrics were tested against the four formal constraints mentioned above and it was discovered that none of the existing metrics can satisfy all formal constraints at the same time. In [Amigó et al., 2009], a new mixed family of metrics called BCubed is proposed. BCubed precision and recall metrics satisfy all four constraints. BCubed precision of an item represents how many items in the same cluster belong to its category. Conversely, BCubed recall of an item represents how many items form its category appear in its cluster. Figure 4.11 shows how the BCubed precision and recall of an item is computed. The overall BCubed precision and recall of the clustering solution is simply the averaged precision and recall of all items in the distribution.

BCubed precision and recall are combined into a single evaluation metric (BCubed FScore) by Van Rijsbergen's F [Van Rijsbergen, 1974]. It is computed with the use of following equation:

$$F(P, R) = \frac{1}{\alpha(\frac{1}{P})+(1-\alpha)(\frac{1}{R})}$$

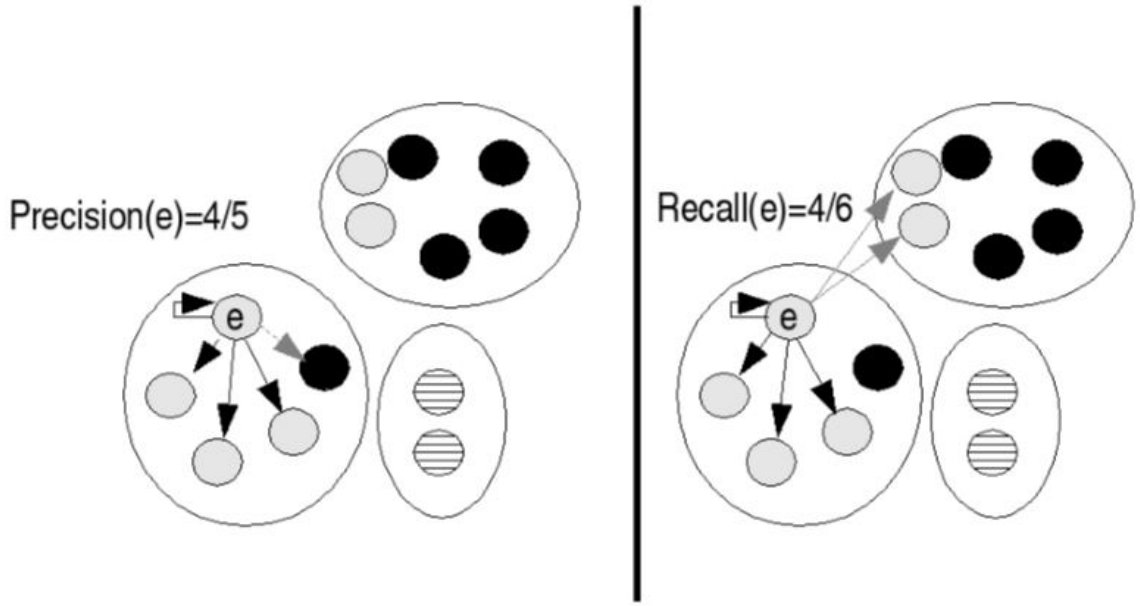


Figure 4.11: Computing the BCubed precision and recall for one item [Amigó et al., 2009]

In the equation, P represents BCubed precision and R represents BCubed recall. Relative weights of these two metrics are represented by α and $(1 - \alpha)$. If α is taken as 0.5, then BCubed FScore equals to the harmonic mean of precision and recall.

Inter-rater Agreement Analysis (BCubed FScore)

We calculated the consistency among assessors with the use of BCubed extrinsic clustering comparing metric and reported BCubed recall, BCubed precision and BCubed FScore values in Figure 4.12.

High BCubed FScore values represent high agreement between the clustering solutions being compared in terms of the constraints; homogeneity, completeness, rag bag and size vs. quantity. High BCubed precision guarantees that noisy items do not occur in the same cluster, while high BCubed recall implies that most of the related items can be found without leaving the cluster.

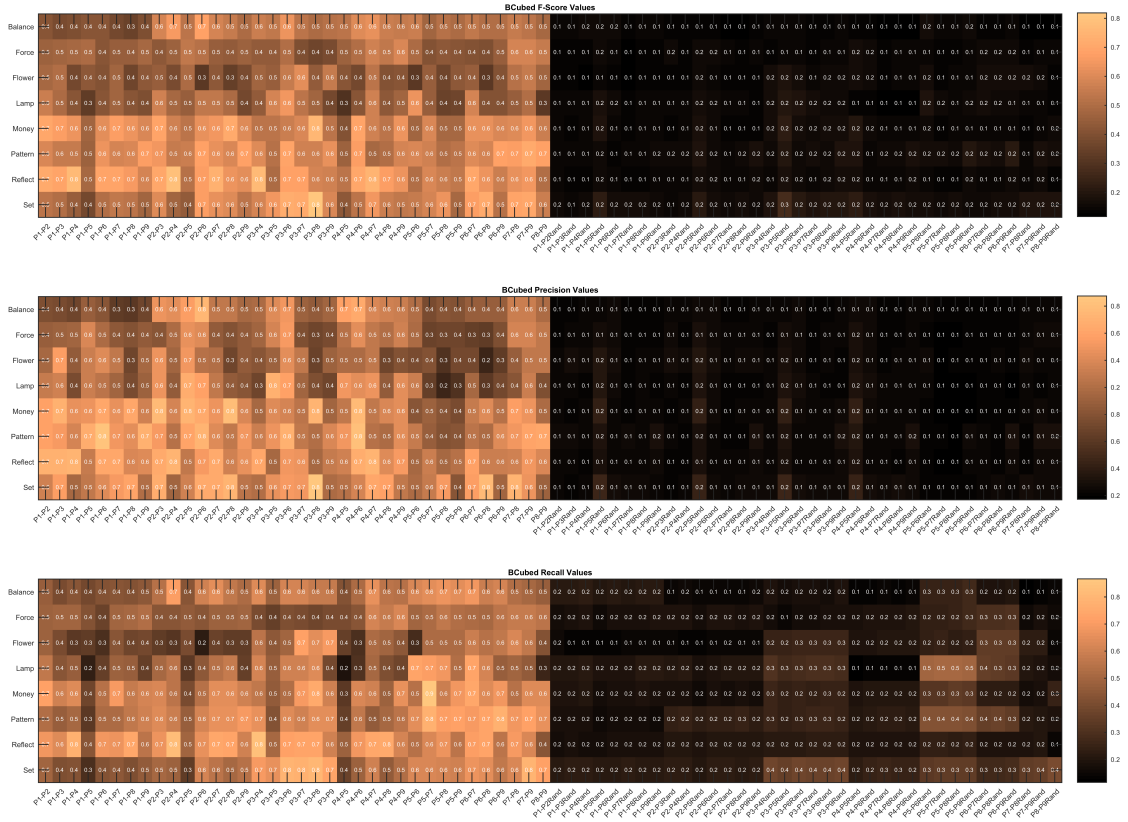


Figure 4.12: Agreement of the clustering solutions provided by users, calculated FScores are participant to participant vs. participant to random

Clustering solutions provided by the participants are compared in pairs. Each participant is tested against all the other participants. For each question type, any two participants' clustering solution is compared and agreement is represented by BCubed scores. Participant to participant comparisons are shown on the left half of Figure 4.12. On the right half, each participant's grouping solution is compared with a random clustering generated by the computer. While generating the random clustering, number of clusters is specified as the number used by the other participant from the left side.

BCubed FScore, precision and recall values are much higher among participants for all sketch categories. This proves that the assessors are more consistent within

themselves when compared with lower values coming from random group assignments. We performed a two-sample Kolmogorov-Smirnov test to compare the distributions of BCubed FScore values at 1% significance level [Pettitt and Stephens, 1977]. The null hypothesis of two populations with the same distribution is rejected with p-value equals $3.2893 * 10^{-128}$, proving that these two distributions is significantly different. To sum up, it is clear that the consistency among assessors is not achieved by any chance.

Building the Gold Standard

Based on the whole analysis and agreement results, experiment data was verified for establishing the gold standard grounding on the common decisions of 9 assessors. For each question category, $\binom{300}{2}$ pairs of sketches were examined and the number of assessors claiming this pair as similar was identified. By this means, a similarity matrix consisting of the perceptual similarity scores per sketch category was constructed. Figure 4.13 illustrates the perceptual similarity matrix. Thanks to this perceptual similarity matrix obtained from the common judgments of 9 assessors, the gold standard was built. In order to build our gold standard for clustering, perceptual similarity matrix was used together with the hierarchical clustering method. We clustered the 300 sketches by their similarity matrices with changing cluster number 2 to 50.

As a result, we built a gold standard clustering based on the common decisions of 9 assessors for all individual question categories. This gold standard will be used while developing the hierarchical clustering algorithm for perceptual sketch grouping and evaluating the algorithm's success on these sketches.

4.2.3 Further Agreement Analysis

Pairwise agreements among participants were reported with the use of BCubed evaluation metric. We now further analyze how an individual assessor is consistent with other 8 assessors. This analysis provides us to see whether an assessors contradicts

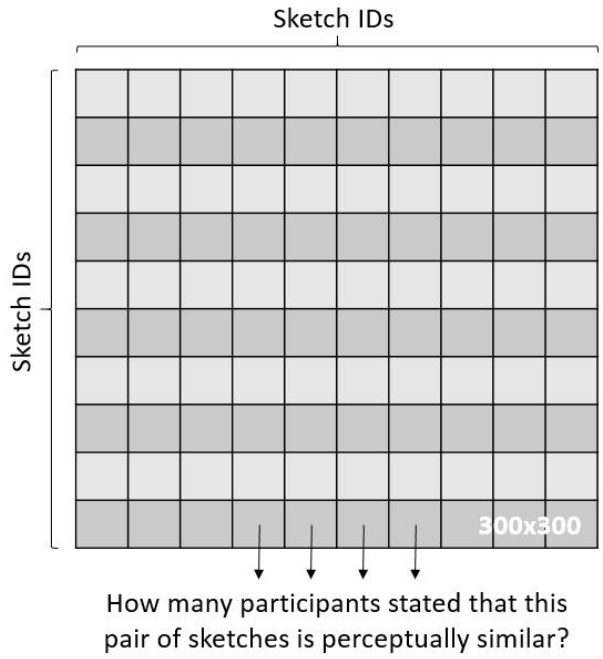
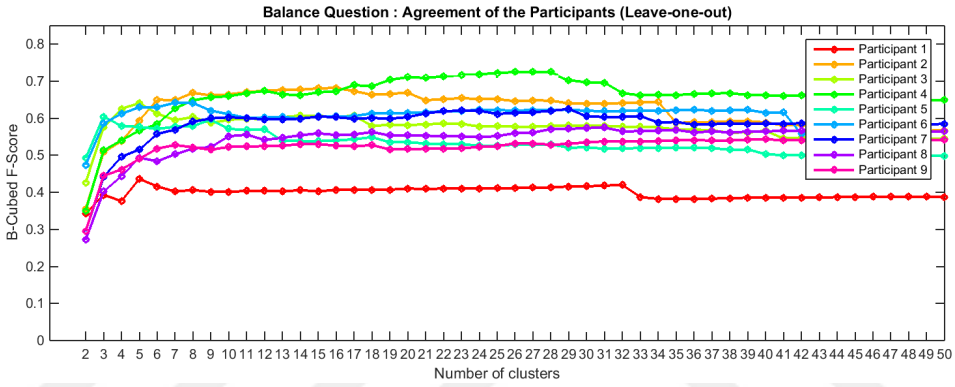


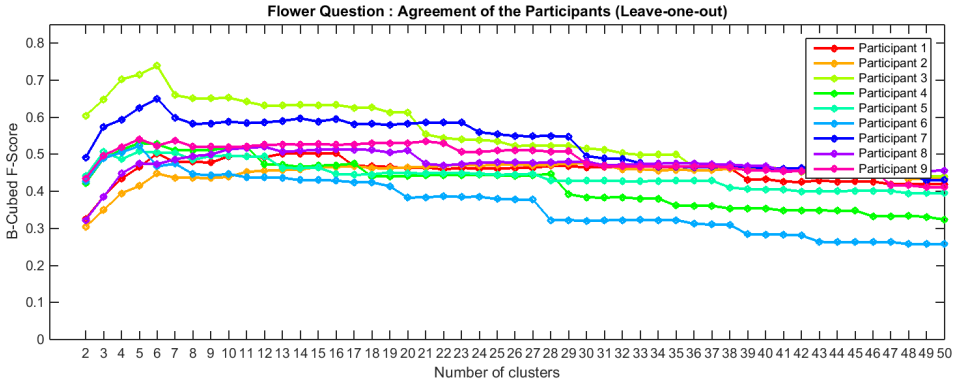
Figure 4.13: Illustration of perceptual similarity matrix

with the common perception.

We measure the BCubed agreements for each question category in a leave-one-out fashion. Each participant is excluded once and gold standard is built with the perceptual similarity matrix of remaining 8 participants. The clustering solution of the excluded participant and the new gold standard clustering built from other 8 participant's data is compared by BCubed metric and BCubed FScores are calculated to measure the agreement. Agreement results are reported separately for each question category. Results show that, there exists no participant whose agreement is very low in all of the question categories. Generally, excluded participants have a high degree of agreement when they are compared with the common clustering solution of other participants. Leave-one-out agreement results are shown in Figure 4.14.

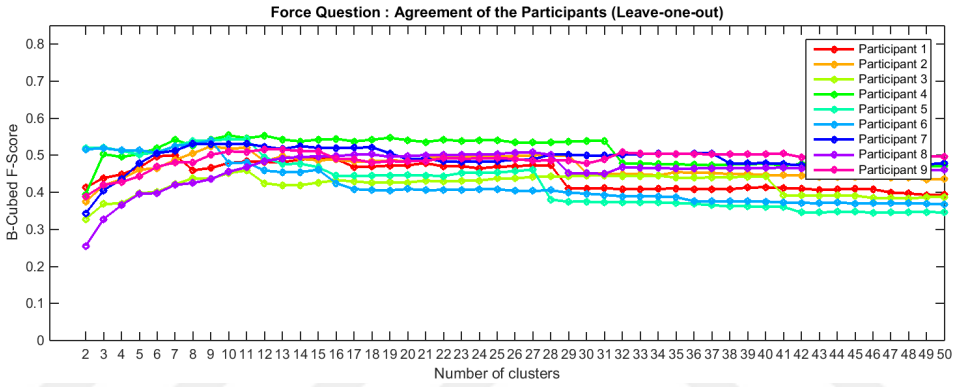


(a) Balance Question

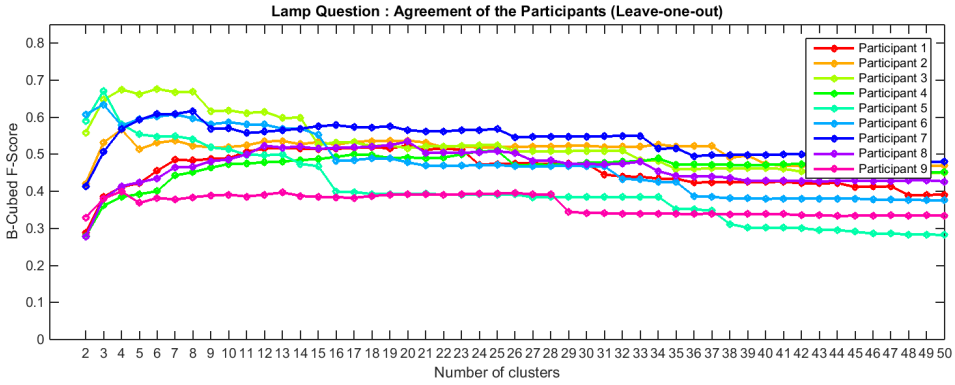


(b) Flower Question

Figure 4.14: Excluded participant's agreement with others

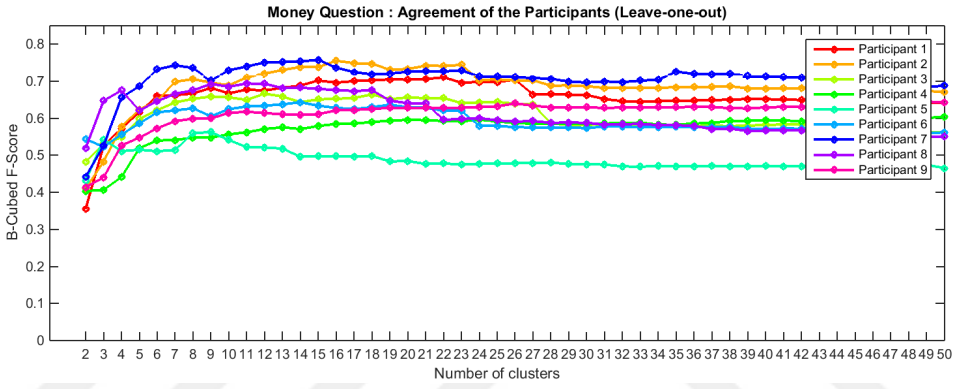


(c) Force Question

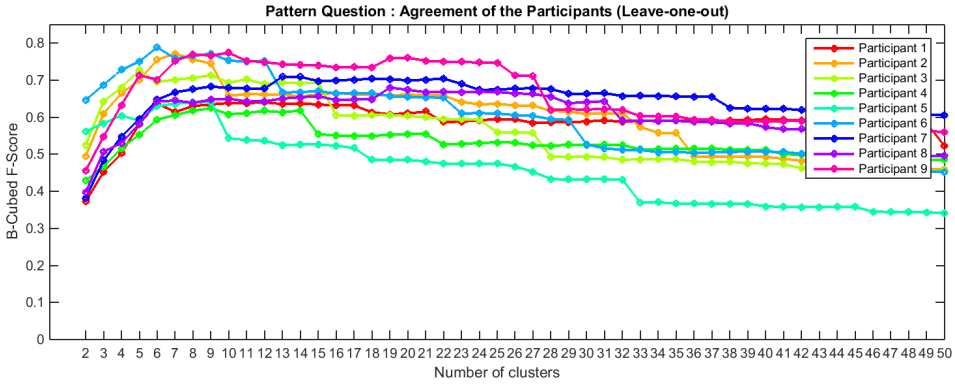


(d) Lamp Question

Figure 4.14: Excluded participant's agreement with others

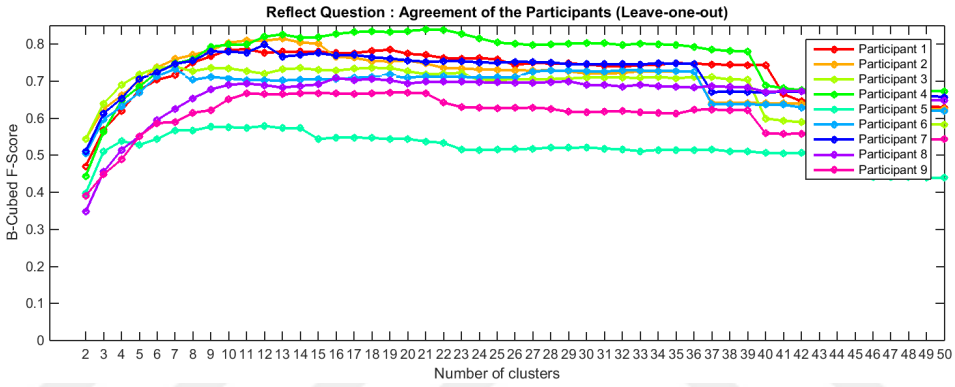


(e) Money Question

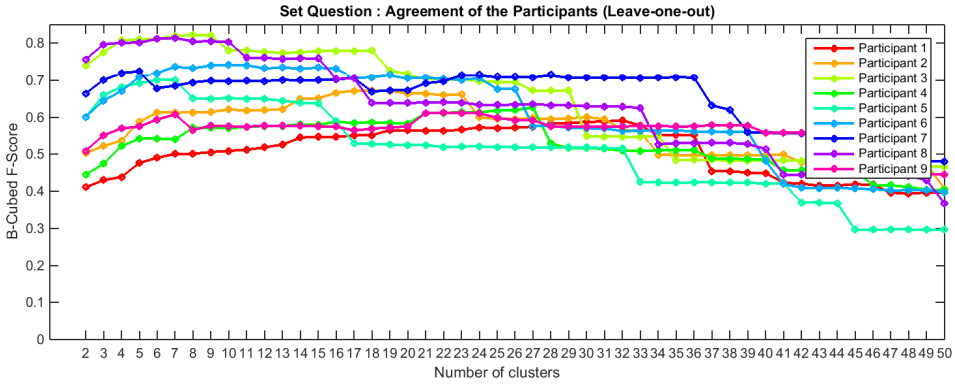


(f) Pattern Question

Figure 4.14: Excluded participant's agreement with others



(g) Reflect Question



(h) Set Question

Figure 4.14: Excluded participant's agreement with others

To see the average performance of the excluded participant by a single numeric value in terms of all question categories, we calculated area under the curves representing BCubed FScore values in Figure 4.14. We demonstrate AUC values of the excluded participant for each question category. In Figure 4.15, we symbolize participants as P1, P2, etc. and show the agreement scores of each participant for each question category.

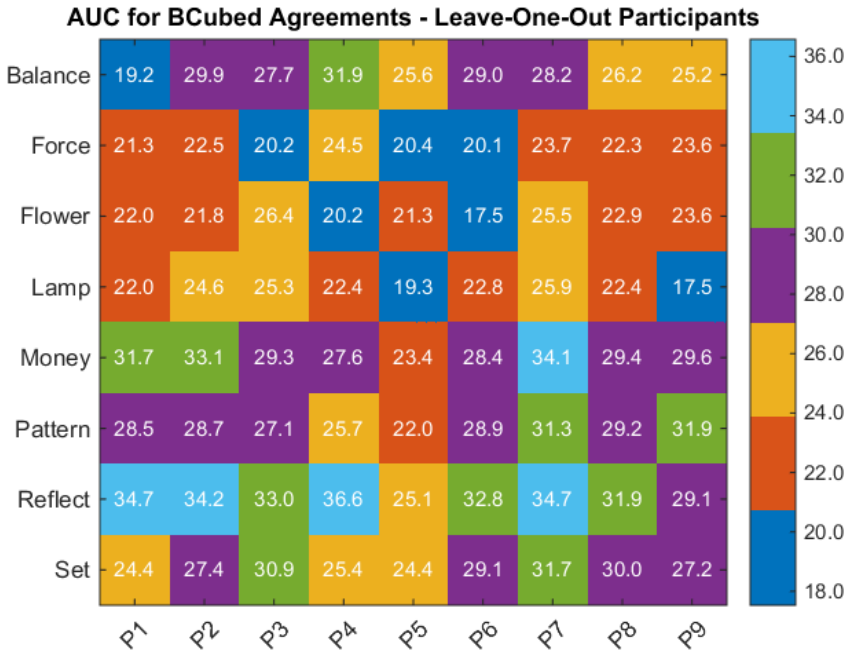


Figure 4.15: AUC for BCubed FScores of agreement (Leave-one-out)

The values varying between 18-36 shows the degree of agreement for a participant with the common opinion of other participants. We divided the whole range into 6 to interpret the meanings of the values falling into those bands.

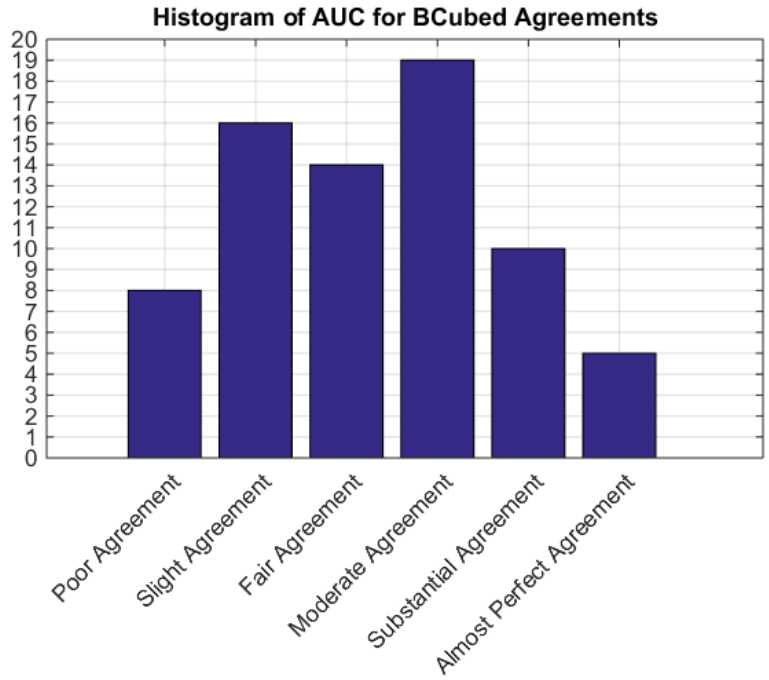


Figure 4.16: Histogram for AUC values of BCubed FScores (Leave-one-out)

Figure 4.16 is the histogram of AUC values of agreement scores. For 8 questions grouped by 9 participants, only 8 cases get poor agreement out of 72. 66% of all the cases get fair, moderate, substantial and almost perfect agreement, which is quite satisfactory considering the fact that inter-rater agreements are generally law due to the subjective nature of the process [Gurcuoglu, 2014]. [Jin et al., 2003] states that, even two users with similar preferences might follow very different rating methods by considering a subjective criteria for similarity.

Chapter 5

METHODOLOGY

Our methodology, proposed for the problem of grouping sketches by considering their perceptual similarity, consists of two main steps, namely feature extraction and agglomerative hierarchical clustering. First of all, sketch scenes are needed to be expressed in terms of meaningful features to be able to measure the degree of similarity among sketch scenes. After the feature extraction step, the degree of similarity between sketch scenes can be calculated according to the distance between these features. Lastly, sketch scenes are clustered based on the distance of similarities.

5.1 Feature Extraction

The issue of how well the extracted features can capture characteristics of the sketches is of great importance, because the system's accuracy is directly affected by these features. Various feature extraction methods from machine vision literature, which work well for hand-drawn data, have been previously adopted to the sketch recognition domain [Tumen et al., 2010]. Our approach makes use of these features prior to clustering phase.

5.1.1 Classic Sketch Features

A number of image-based feature extraction methods have been suggested in the literature. Here we focus on four classic methods widely used in the sketch recognition domain. Among those methods, IDM features obtain highest recognition accuracy in sketch recognition domain [Tumen et al., 2010]. Other three methods were also used in order to compare our results with the state of art methods in terms of accuracy.

Image Deformation Model Features (IDM)

The image-based feature representation method, shortly IDM, was proposed by Ouyang et al. in [Ouyang and Davis, 2009]. In this method, sketches are converted to low density feature images by using the directions and end points of the strokes. Resulting IDM features have 720 entries, and the feature extraction mechanism has three free parameters, k (kernel size), σ (smoothing factor), and r (resampling parameter). We set $k = 4$, $r = 50$ and $\sigma = 10$ to acquire the highest accuracy results as described in [Tumen et al., 2010].

Zernike Moments

Zernike moments are simple and effective feature representations for used for sketch recognition [Hse and Newton, 2004]. Zernike moments feature extraction method has only one free parameter, which is the order of the Zernike moment o. We set the order parameter as $o = 12$ in our experiments.

Shape Context

Shape context is a histogram based local descriptor which captures image intensity statistics in the neighborhood of a reference point. Oltmans et al. have adapted shape context features to sketch recognition domain [Oltmans, 2007]. Algorithm requires the specification of three free parameters, c (the number of concentric circles), s (the number of slices), r (radius of the shape context). We set these free parameters as $c = 3$, $s = 12$ and $r = 50$. Oltmans et al. used a sophisticated matching method to use shape contexts for image segmentation. However, in order to make them fit into our framework, we used a much simpler approach and placed five shape context histograms on the corners and the center of the bounding box of each shape as explained in [Tumen et al., 2010].

Trace Transform

Trace transform is a generalization of Radon transform [Kadyrov and Petrou, 2001]. We followed the same feature extraction process in [Tumen et al., 2010]. As functionals, we used Radon transform and and Fourier functional. Setting the functionals leaves only three free parameters to define: b (the number of bins), θ (the number of discrete angles), σ (the smoothing parameter). We used the following values for these parameters: $b = 20$, $\theta = 30$ and $\sigma = 1.2$.

5.1.2 Pairwise Features

The pairwise features are newly introduced by IUI Laboratory at Koç University [Cakmak et al., 2015]. This feature extraction method is a compact feature representation that basically captures the perceptual relationships in free-form sketch scenes.

5.2 Agglomerative Hierarchical Clustering

Based on the extracted features of the sketch scenes, now the degree of similarity can be measured. To group the similar sketches, we used the teacher application developed within the scope of ASIST project [Cakmak et al., 2015]. Teacher application uses agglomerative clustering algorithm. Agglomerative hierarchical clustering algorithm constructs dendrograms corresponding to the similar groups of sketches. Algorithm consists of three basic steps:

- Find the similarity or dissimilarity between every pair of sketch scene in the data set:
  - $D = pdist(X,'correlation')$ where X is the features for scenes
    Pairwise distance between sketch scenes are calculated. Correlation means one minus the sample correlation between points (treated as sequences of values).
- Group the sketch scenes into a binary, hierarchical cluster:

  - $Z = linkage(D,'complete')$

    Linkage function calculates the distance between clusters. Complete distance means the furthest distance between clusters.

- Determine where to cut the hierarchical tree into clusters:

  - $C = cluster(Z,'maxcluster', 15)$

    Optimal cut point for the hierarchical tree is decided with respect to the inconsistency measures. However, here $'maxcluster' = 15$ is used because of being appropriate for all question categories.

In Figure 5.1, the numbers along the horizontal axis represent the indices of the objects in the original data set, namely the answers given to the balance question. The links between objects are represented as upside-down U-shaped lines. The height of the U indicates the distance between the objects. The height represents the distance linkage computes between clusters. Illustrative hand-drawn answers are shown on the dendrogram belonging to the three clusters with colors yellow, green and purple.

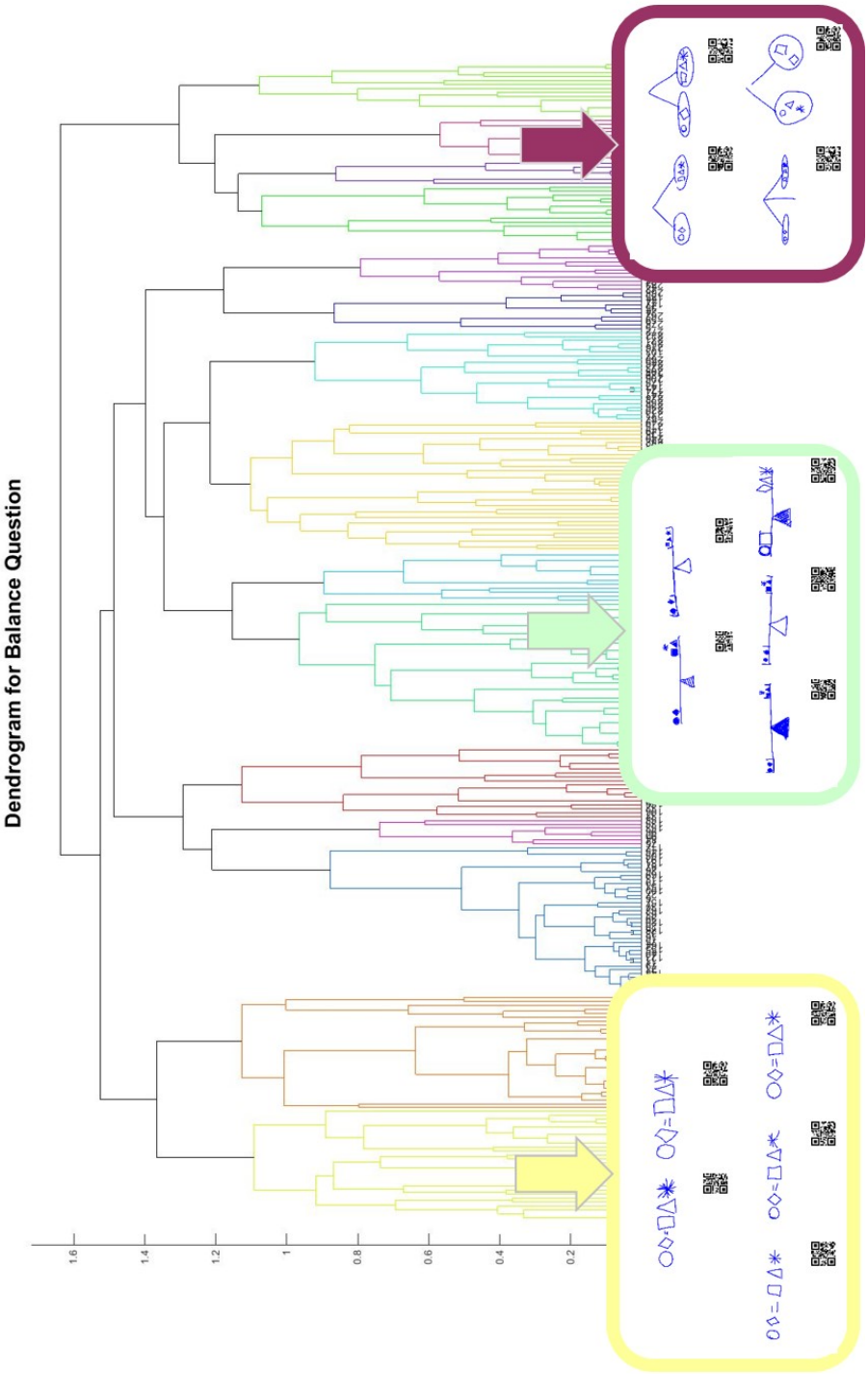


Figure 5.1: Dendrogram for balance question

Chapter 6

EVALUATION

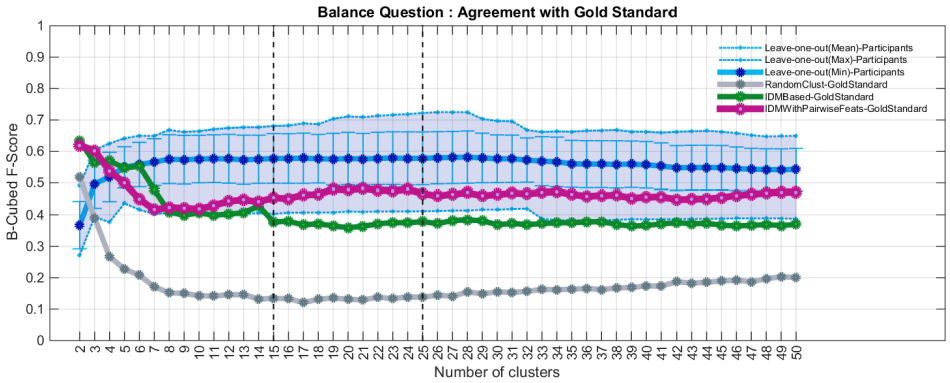
We evaluated the performance of the clustering method from two different aspects. We first measure the agreement with the gold standard and then measure the homogeneity of the constructed clusters in terms of the correctness of the answers in those groups. Evaluation results show that proposed clustering method performs very close to human way of grouping perceptually similar sketch scenes.

6.1 Agreement with Gold Standard

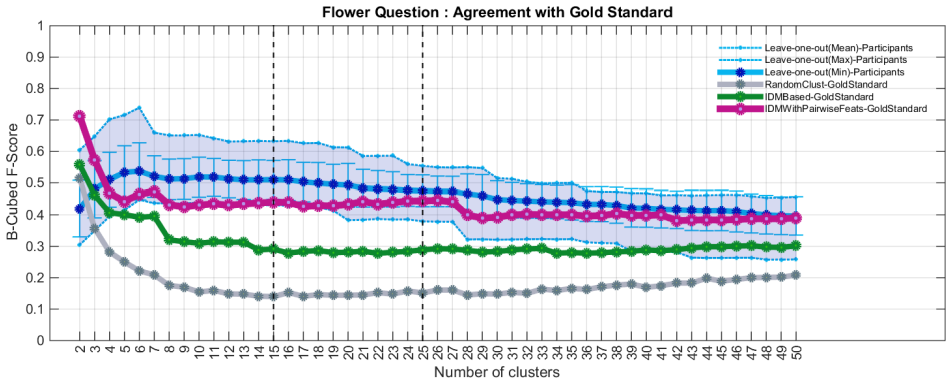
Grouping results of the clustering system was compared with the groups built by human assessors in terms of BCubed FScore values between those two clusterings in order to see whether the two solutions are in good agreement. Agreement results with the gold standard are listed in Figure 6.1.

The first three values in the gold standard agreement figures are 'Leave-one-out mean', 'Leave-one-out max' and 'Leave-one-out min'. These are the agreement results when one of the participants is excluded from the group and common decisions of the remaining participants is assumed to be the system. In this case, BCubed Fscore values are answer to the question 'How would a human perform against a human perception based gold standard?'. Statements 'Leave-one-out min' and 'Leave-one-out max' represent the people who are the most and least agreeable to the public decision. 'Leave-one-out mean' statement is the average of 9 participants, where they are excluded one by one.

'RandomClustering' is the grouping where computer creates the clusters in a randomized manner. Random clustering was also tested against gold standard and it was verified that its agreement to the gold standard is too low as already expected.

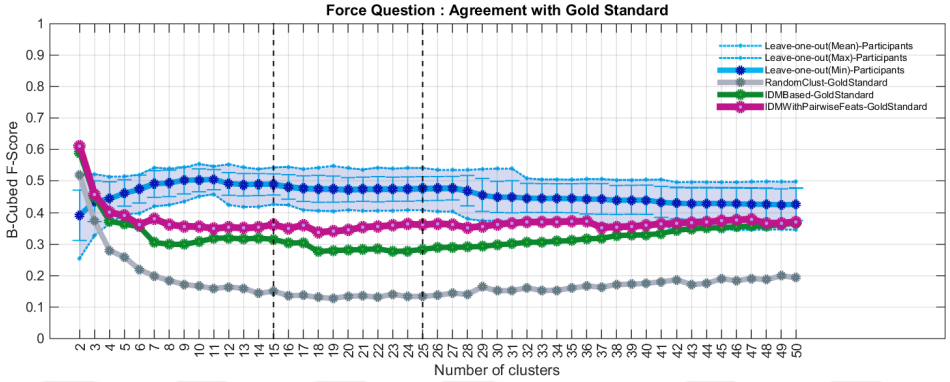


(a) Balance Question

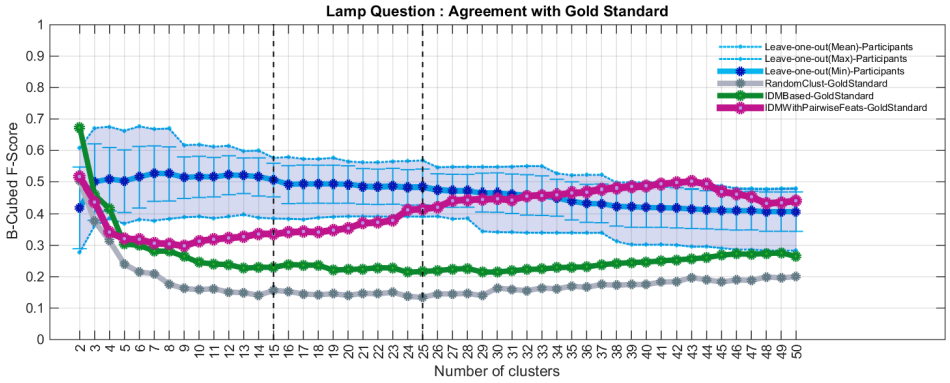


(b) Flower Question

Figure 6.1: Agreement with Gold Standard

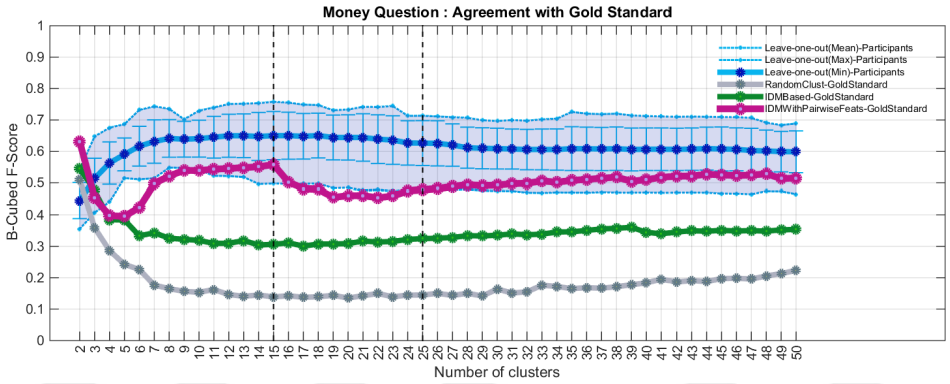


(c) Force Question

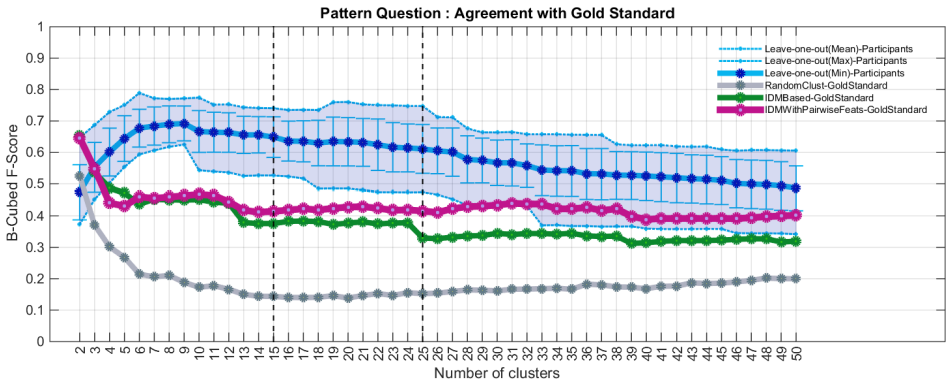


(d) Lamp Question

Figure 6.1: Agreement with Gold Standard

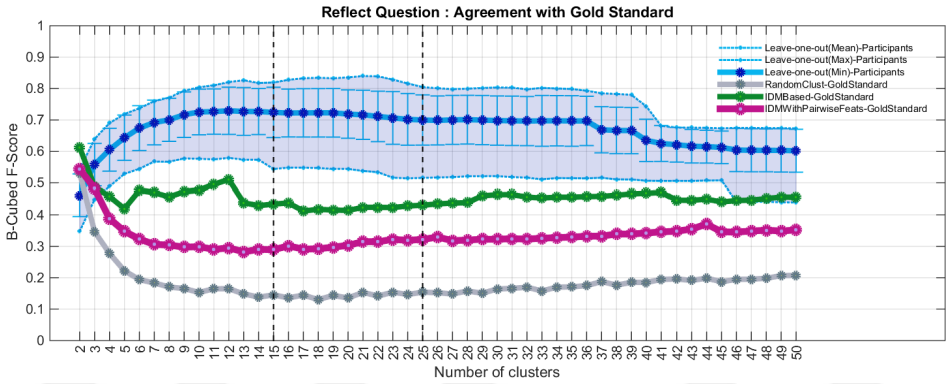


(e) Money Question

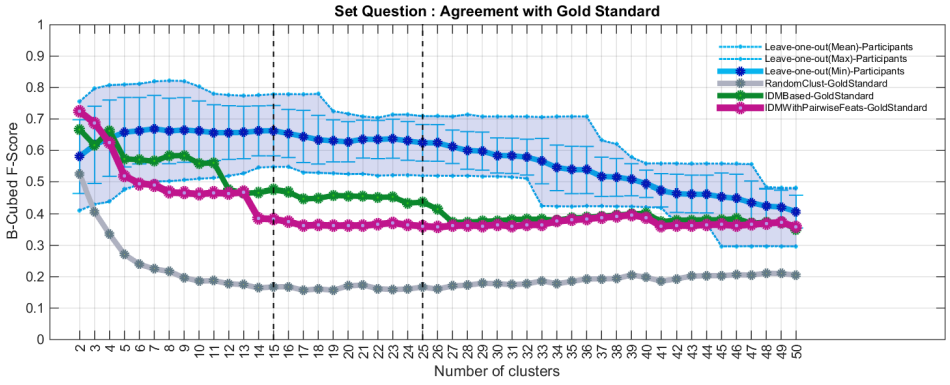


(f) Pattern Question

Figure 6.1: Agreement with Gold Standard



(g) Reflect Question



(h) Set Question

Figure 6.1: Agreement with Gold Standard

Remaining two values in the figures are the agreement scores of our system with the gold standard. First one is the case when classic IDM features are used for hierarchical clustering and second one is the case when pairwise features are used for hierarchical clustering. It was observed that, adding pairwise features to the system increases the agreement with the gold standard. BCubed FScore values rarely decrease, in which cases, the proposed system is not in high agreement with the gold standard.

Summarizing the agreement tests, it is obvious that our system's performance is very close to human performance. Proposed system is almost always performs better than people who have the lowest degree of agreement to the public decision. Our system generally follows the agreement scores of the people's average agreement scores. Even more, our system beats the agreement scores of people with the highest scores in some occasions.

While investigating the cases where our system seems inconsistent with the gold standard, we discovered that people are more likely to group items according to the general appearance without paying much attention to the details. Especially when the number of items is high, 300 or higher, they just have a look at the answers and group them shallowly. Contrarily, the hierarchical clustering method combined with the pairwise features detects the details in the scene and groups items accordingly. Through our analyses, we discovered that the general appearance might not always be sufficient for separating the answers as true/false. Therefore, we evaluated our clustering system from a second aspect.

6.2 Homogeneity Assessment of Clusters

Generally, the entropy of a cluster reflects how the members of the class categories are distributed within each cluster [Amigó et al., 2009]. For further evaluation of the resulting clustering solution of free-form sketch scenes, we investigated the entropy of the groups to assess the homogeneity of the constructed clusters. Entropy of the groups actually depend on the labels of the answers in the clusters and labels of the answers are normally determined by the teacher. In our case, all 2400 answers from 8 question categories has been labeled as true/false by the expert beforehand. Figure 6.2 illustrates the process of average entropy calculation.

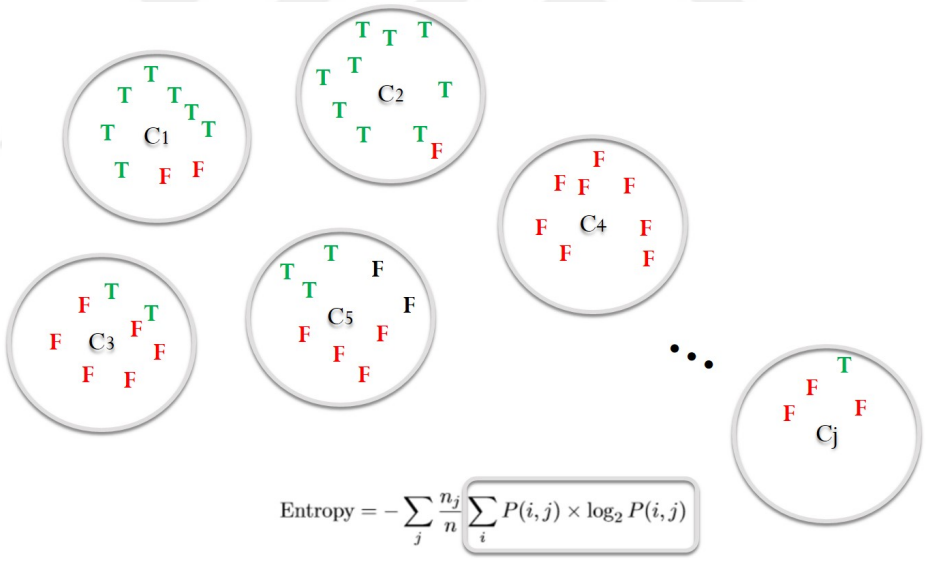
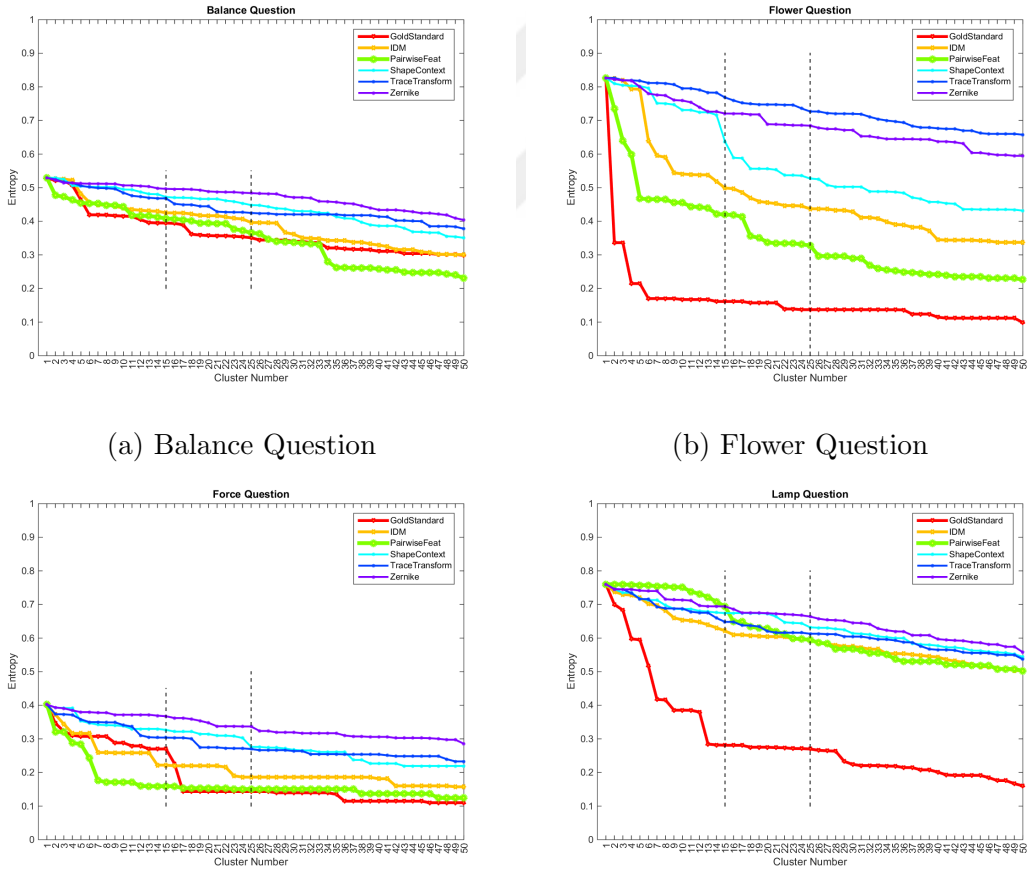


Figure 6.2: Average entropy of clusters

In the figure, $C_1, C_2, C_3, ..., C_j$ are the clusters. Green colored T letters define the true answers in that cluster labeled by the expert. Similarly, red colored F letters represent false answers present in the cluster. In the entropy formula, i index represents the label of the answers and j index represents cluster number. Entropy of a cluster is calculated by the formula:

$\sum_i -P(i,j) * \log_2 P(i,j)$

The global quality measure is computed by averaging the entropy of all clusters: $-\sum_j n_j/n \sum_i P(i,j) * \log_2 P(i,j)$
being $P(i,j)$ the probability of finding an element from the category i in the cluster j, n_j the number of items in cluster j and n the total number of items in the distribution. Using this formula, average entropy of the resulting clusters are calculated and their change with the varying cluster numbers is presented in Figure 6.3 for 8 different question categories. In figures, y axis shows the average entropy value for the corresponding number of clusters shown on x axis.



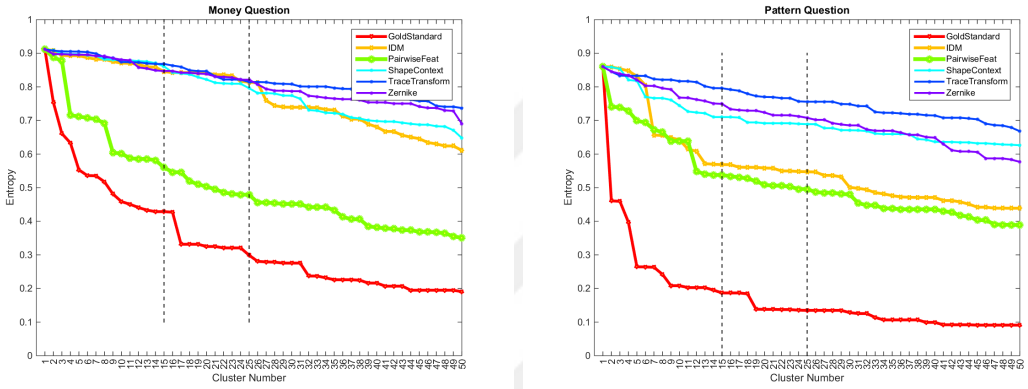
(a) Balance Question

(b) Flower Question

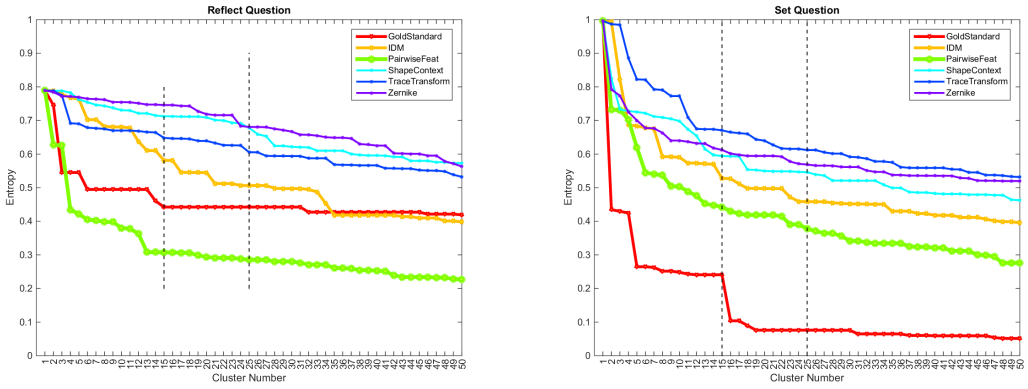
(c) Force Question

(d) Lamp Question

Figure 6.3: Average entropy of clusters in terms of true/false labels of the answers



(e) Money Question (f) Pattern Question



(g) Reflect Question (h) Set Question

Figure 6.3: Average entropy of clusters in terms of true/false labels of the answers

As seen in Figure 6.3, gold standard has the lowest entropy values, meaning that it is the best clustering result. That is, humans are more successful in clustering task when true/false labels of the answers are considered. Our clustering method with pairwise features follow the gold standard in the second place. IDM based hierarchical clustering follows that closely and other classic sketch features based methods are rather far from the gold standard. In Figure 6.4, results for 8 different question categories are summarized by taking their mean values. Our hierarchical clustering method with pairwise features is much more successful than clustering methods with classic sketch features. By using the paired sample t-test, we proved that the improvement obtained via pairwise features rather than IDM is significant ($p = 5.7607 * 10^{-40} < 0.05$). Using IDM features also provides significant benefit towards approaching the gold standard ($p = 1.7879 * 10^{-40} < 0.05$).

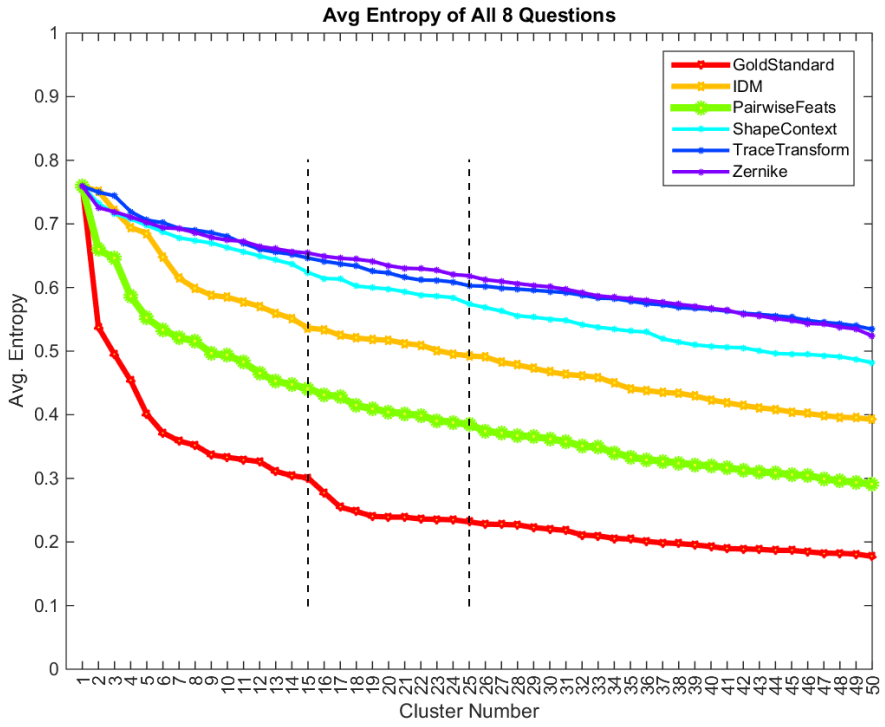


Figure 6.4: Mean of the 8 different question categories

Clustering results with the cluster numbers varying between 15-25 are considered more important than the results with other cluster numbers. The reason behind this is that it would be appropriate to present 300 answers approximately in 15 groups, considering the possible display size of computers. Also, to support our motivation of assisting teachers at the evaluation process, we need to consider cluster numbers below 25. Area under the average entropy curves of 8 question categories is calculated for the cluster numbers between 15-25 and presented in Figure 6.5. In the figure; the smaller the AUC value is, the more successful the corresponding method is.

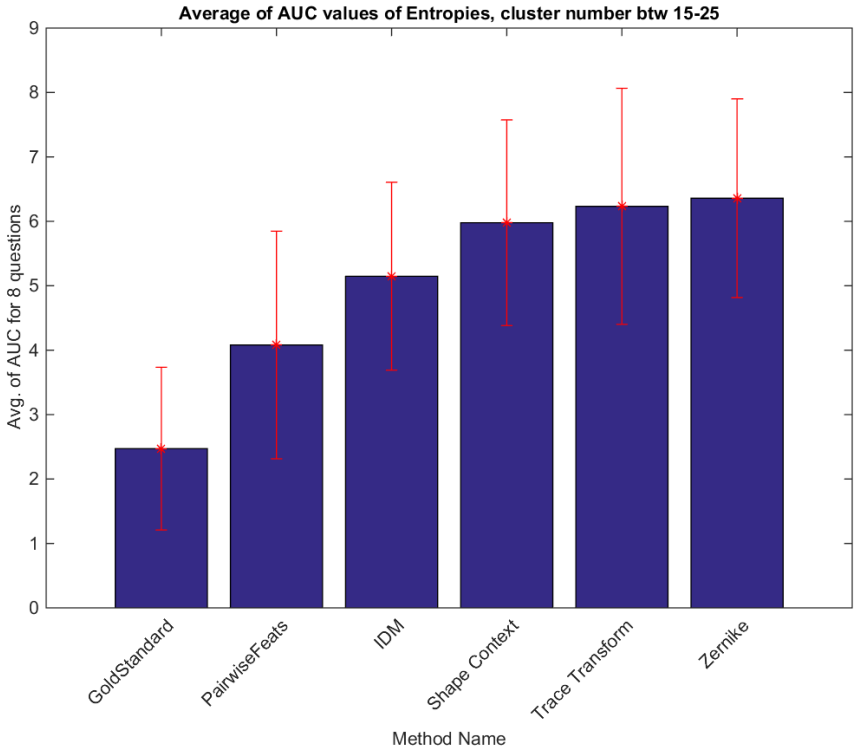


Figure 6.5: AUC for average entropy of 8 question categories, cluster number = [15-25]

Question categories were also individually observed for AUC values of average entropy of cluster numbers varying between 15-25. Figure 6.6 shows how the methods with different feature extraction mechanisms perform for all question categories.

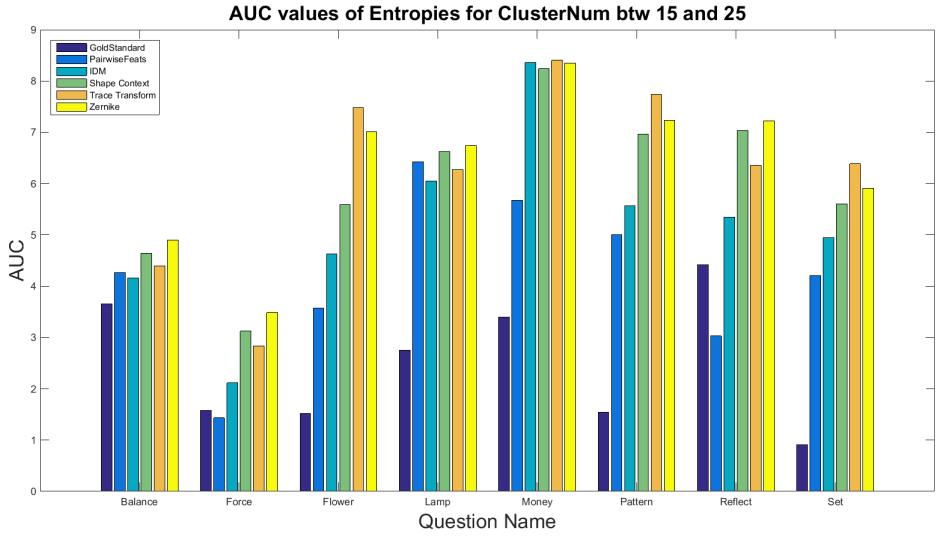


Figure 6.6: AUC values for 8 question categories, cluster number =[15,25]

To summarize, our clustering method with pairwise features is closest to the ground truth, being the most successful method among others. Moreover, in two question categories (namely reflect and force) it performs even better than the ground standard. This proves that general appearance does not always gives us correct clustering solutions. In some cases, the best clustering solution is hidden in the structural details of the answers rather than the general appearance. The proposed method of hierarchical clustering with pairwise features is the best method for capturing structural similarities between the sketch scenes.

Chapter 7

CONCLUSION AND FUTURE WORK

In this thesis, we proposed a system that can assess the similarity of free-form sketch scenes based on the extracted features and construct hierarchically clustered groups of perceptually similar sketches without any use of domain specific knowledge.

Our evaluation is based on the comparison of the proposed system with the human perception based gold standard. We designed a sketch grouping experiment to build the gold standard and validated its reliability with various analyses. Based on the gold standard, we introduced two different metrics for evaluating the system's performance. First of our evaluation metrics includes calculating the agreement scores between two clustering solutions, namely the gold standard and our system's output. Second of our evaluation metrics is the global homogeneity scores based on the entropy calculations of the created clusterings. Results showed that the proposed system performs very close to human way of grouping perceptually similar sketch scenes. Our results obtained with global homogeneity score calculations in some question categories are really promising since they imply the possibility of imitating or even beating human mind's ability of perceiving similarity.

Future work includes extending the proposed framework with more features that are planned to be learned through intelligent machine learning algorithms. We plan to collaborate with experts and learn from their experience about extra attributes which would work globally. In current state of the system, features' benefit varies according to the question categories. By automatically learning the most successful feature for the category of interest, we plan to improve the performance of the proposed system. Another improvement intended to be achieved is the presentation of the created groups in a more visually attractive way. The visual arrangements within

the groups are going to be performed in such a way that similar sketch scenes are positioned next to each other. These visual arrangements could also be extended by additional algorithms of making modifications in the answer scenes belonging to the very same group. All of these improvements are promising since they make the job of interpreting the clustering results of a large dataset much more easier.

BIBLIOGRAPHY

[Qui, 2013] (2013). Quickmark. `http://www.quickmark.cn/En/basic/index.asp`. Accessed: 2016-10-06.

[Amigó et al., 2009] Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

[Anderson et al., 2007] Anderson, R., Anderson, R., Davis, P., Linnell, N., Prince, C., Razmov, V., and Videon, F. (2007). Classroom presenter: Enhancing interactive education with digital ink. *Computer*, 40(9):56–61.

[Blough, 2001] Blough, D. S. (2001). The perception of similarity. *Avian visual cognition*, 6:23–25.

[Cakmak et al., 2015] Cakmak, S., Yesilbek, K. T., and Sezgin, M. T. (2015). Asist project. `http://iui.ku.edu.tr/asist/`. Accessed: 2016-03-07.

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

[Defays, 1977] Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366.

[Demiralp et al., 2014] Demiralp, Ç., Bernstein, M. S., and Heer, J. (2014). Learning perceptual kernels for visualization design. *IEEE transactions on visualization and computer graphics*, 20(12):1933–1942.

[Douglas and Peucker, 1973] Douglas, D. H. and Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122.

[Frey and Dueck, 2007] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

[Goldstone, 1994] Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, and Computers*, 26(4):381–386.

[Gurcuoglu, 2014] Gurcuoglu, B. (2014). Learning people's perception of messiness for hand-drawn sketches. Master's thesis, Koc University, Istanbul, Turkey.

[Hammond et al., 2010] Hammond, T. A., Logsdon, D., Paulson, B., Johnston, J., Peschel, J. M., Wolin, A., and Taele, P. (2010). A sketch recognition system for recognizing free-hand course of action diagrams. In *IAAI*.

[Hatfield, 2011] Hatfield, J. (2011). *Clustering digital ink content to assist with the grading of student work*. PhD thesis, University of Louisville.

[Hse and Newton, 2004] Hse, H. and Newton, A. R. (2004). Sketched symbol recognition using zernike moments. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 367–370. IEEE.

[Jin et al., 2003] Jin, R., Si, L., Zhai, C., and Callan, J. (2003). Collaborative filtering with decoupled models for preferences and ratings. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 309–316. ACM.

[Jóźwiak, 2011] Jóźwiak, M. A. (2011). *Identification and presentation of student answers to in-class exercises in classroom learning partner*. PhD thesis, Massachusetts Institute of Technology.

[Kadyrov and Petrou, 2001] Kadyrov, A. and Petrou, M. (2001). The trace transform and its applications. *IEEE Transactions on pattern analysis and machine intelligence*, 23(8):811–828.

[Köhler, 1970] Köhler, W. (1970). *Gestalt psychology: An introduction to new concepts in modern psychology*. WW Norton & Company.

[Kriegeskorte and Mur., 2012] Kriegeskorte, N. and Mur., M. (2012). Inverse mds: inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3(245):1–13.

[Lindlbauer et al., 2013] Lindlbauer, D., Haller, M., Hancock, M., Scott, S. D., and Stuerzlinger, W. (2013). Perceptual grouping: selection assistance for digital sketching. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*, pages 51–60. ACM.

[Ng et al., 2002] Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

[Oltmans, 2007] Oltmans, M. (2007). *Envisioning sketch recognition: a local feature based approach to recognizing informal sketches*. PhD thesis, Massachusetts Institute of Technology.

[Ouyang and Davis, 2007] Ouyang, T. Y. and Davis, R. (2007). Recognition of hand drawn chemical diagrams. In *AAAI*, volume 7, pages 846–851.

[Ouyang and Davis, 2009] Ouyang, T. Y. and Davis, R. (2009). A visual approach to sketched symbol recognition.

[Perteneder et al., 2015] Perteneder, F., Bresler, M., Grossauer, E.-M., Leong, J., and Haller, M. (2015). cluster: Smart clustering of free-hand sketches on large

interactive surfaces. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 37–46. ACM.

[Pettitt and Stephens, 1977] Pettitt, A. N. and Stephens, M. A. (1977). The kolmogorov-smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2):205–210.

[Platt, 2000] Platt, J. C. (2000). Autoalbum: Clustering digital photographs using probabilistic model merging. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 96–100. IEEE.

[Rogowitz et al., 1998] Rogowitz, B. E., Frese, T., Smith, J. R., Bouman, C. A., and Kalin, E. (1998). Perceptual image similarity experiments. *In Photonics West'98 Electronic Imaging*, pages 576–590.

[Sculley, 2010] Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178. ACM.

[Smith, 2006] Smith, A. C. (2006). *Aggregation of student answers in a classroom setting*. PhD thesis, Massachusetts Institute of Technology.

[Sternberg and Sternberg, 2016] Sternberg, R. and Sternberg, K. (2016). *Cognitive psychology*. Nelson Education.

[Tumen et al., 2010] Tumen, R. S., Acer, M. E., and Sezgin, T. M. (2010). Feature extraction and classifier combination for image-based sketch recognition. In *Proceedings of the Seventh Sketch-Based Interfaces and Modeling Symposium*, pages 63–70. Eurographics Association.

[Tversky, 1977] Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4):327.

[Van Rijsbergen, 1974] Van Rijsbergen, C. J. (1974). Foundation of evaluation. *Journal of Documentation*, 30(4):365–373.

[Wertheimer, 1938] Wertheimer, M. (1938). Laws of organization in perceptual forms.