ORIGINAL ARTICLE

Perceptually grounded quantification of 2D shape complexity

Dena Bazazian[1] · Bonnie Magland[2] · Cindy Grimm[3] · Erin Chambers[4] · Kathryn Leonard[5]

Abstract

The importance of measuring the complexity of shapes can be seen by the wide range of its application such as computer vision, robotics, cognitive studies, eye tracking, and psychology. However, it is very challenging to define an accurate and precise metric to measure the complexity of the shapes. In this paper, we explore different notions of shape complexity, drawing from established work in mathematics, computer science, and computer vision. We integrate results from user studies with quantitative analyses to identify three measures that capture important axes of shape complexity, out of a list of almost 300 measures previously considered in the literature. We then explore the connection between specific measures and the types of complexity that each one can elucidate. Finally, we contribute a dataset of both abstract and meaningful shapes with designated complexity levels both to support our findings and to share with other researchers.

Keywords Shape complexity · Complexity measures · 2D shapes

1 Introduction

The notion of shape complexity is a fundamental one, which has been investigated in many different areas of computer vision, computer science, mathematics and psychology [3,16,26,34,36]. Definitions of complexity of shape vary widely, sometimes depending on an application domain, and sometimes depending on a particular theoretical framing, but rarely are these definitions constructed in conjunction with human perception of complexity.

In this paper, we build on prior work that attempts to quantify various aspects of complexity by determining quantitative measures that agree with human evaluation of complexity, and then relating those measures to different categories of complexity. In doing so, we develop a theoretical foundation for shape complexity that is rooted in human perception.

In [10], we identify categories of complexification—adding parts to a shape, creating indentations, adding noise to a shape boundary, and disrupting symmetry—and conclude that no single quantitative measure is likely to capture the full range of shape complexity. Instead, we propose aggregating measures and explore an extensive list of possible measures grouped by whether they are local measures on the boundary of the shape, local measures on the region of the shape, measures based on the Blum medial axis of the shape [7], measures that capture self-similarity, or global shape measures. The complexity clusters we obtain using k-medoids clustering on the groupings of measures, and on all measures together, do not indicate that shapes of similar perceived complexity are necessarily closer to each other than they are to shapes of differing complexity in the respective embedding spaces. In this paper, we build on that prior work by conducting user studies of human perception of complexity and applying the results to guide and refine our understanding

✉ Kathryn Leonard
  kleonard.ci@gmail.com

  Dena Bazazian
  dena.bazazian@bristol.ac.uk

  Bonnie Magland
  bonnie.magland@gmail.com

  Cindy Grimm
  grimmc@oregonstate.edu

  Erin Chambers
  erin.chambers@gmail.com

1 Department of Computer Science, University of Bristol, Bristol, UK

2 BYU, Provo, USA

3 Oregon State University, Corvallis, USA

4 Department of Computer Science, St. Louis University, St. Louis, USA
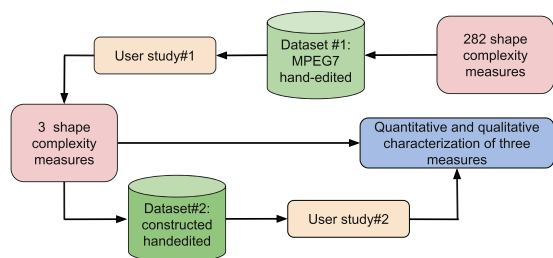
5 Occidental College, Los Angeles, USA

Fig. 1 An overview of our approach. We apply 282 baseline shape complexity measures from [10] to a dataset and then apply a user study to identify three measures that are well-correlated with human perception of complexity. We then apply those three measures to a small dataset of constructed shapes and apply another user study in order to evaluate which aspects of complexity the three measures are capturing

of those quantitative measures, and to obtain measures that better capture *perceived* complexity.

Contributions: This paper makes three main contributions. First, in Sect. 4, we apply results from a forced-choice user study to identify which of the 282 measures from [10] correlate most strongly with human perception of complexity. We then apply those significant measures to three small datasets of constructed shapes created to have predetermined complexity levels to see how well the selected measures distinguish between the predetermined levels. Finally, we apply results from two additional user studies to determine if human perception of complexity matches with raw and user-weighted rankings from our selected measures, and to identify the relationship between types of complexity and values of specific measures. An overview of our approach is depicted in Fig. 1.

2 Related work

Shape complexity is studied across several fields such as computer vision [26], design [16,36] and psychology [3]. In the context of 3D shapes, the topic of shape complexity has the potential to be useful in shape retrieval [1,4], in measuring neurological development and disorders [18,25], in determining the processes and costs involved for manufacturing products [16,35], and in robotics for learning where to grasp objects [9]. 2D shape complexity also has a wide range of applications in cognitive studies and eye tracking [23]. The relationship between eye-tracking metrics and the psychological factors explored in [15] is used to obtain the physiological and psychological indicators of the visual complexity of art images from the perspective of visual cognition. Complexity has also been used in image understanding, such as in [34] where visual complexity is defined as an image attribute that humans can subjectively evaluate based on the level of details in the image. The authors then link attributes to deep intermediate-layer features of neural networks. Shape

complexity measures are also applied in writer verification techniques to analyze handwritten text in [5,6]. In [6], the authors determine if two samples have been written by the same person by evaluating the similarity of the two most complex shapes extracted from each word. In [5], the authors explore different notions of shape complexity by applying them to a library of shapes using k-medoids clustering, and then use the results to solve the handwriting similarity problem as a particular case of a shape matching problem.

Human perception has been applied in various aspects of computer graphics in addition to shape complexity studies [30,32]. The features from a human visual system (HVS) are applied for incorporating perception-based computer graphics approaches as a computational model [32]. ICTree [30] introduces an automated system for realism assessment of the tree model based on their perception. PTRM [32] introduces Perceived Terrain Realism Metrics that assigns a normalized value of perceived realism to a terrain represented as a digital elevation model.

In [10], the authors explore a wide range of measures of shape complexity arising from information theory [13], computer vision [28], computational geometry [12], and curve analysis [14,20,27,33], and introduce new notions of complexity based on measurements taken along Blum medial axis [7] and persistence of certain features under down-sampling. We discuss these in more detail below. The authors apply k-medoids clustering to values of those measures extracted from shapes from the MPEG-7 database [8], providing an initial understanding of complexity neighborhoods based on the selected measures. Evaluating the clusters subjectively, the authors conclude that no single measure successfully captures complexity but rather that an aggregation of measures is most likely to produce results consistent with our human perception [10].

A few measures have been proposed since [10]. Authors of [29] analyze the geometric basis of spatial complexity. An index of total absolute curvature proposed by [24] reflects the amount of concavity on a curved surface as an index of the quantification of "complexity" as defined by the cumulative area on the spherical surface indicated by the Gauss map on the curved surface. In 3D, an investigation of shape complexity measures performed in [2] introduces a 3D dataset and evaluates the performances of the methods by computing Kendall rank correlation coefficients both between the orders produced by each complexity measure and the ground truth and between the pair of orders produced by each pair of complexity measures.

3 Background

The 282 measures explored in [10] group naturally into three categories: boundary-based, regional, and skeletal. Some

Fig. 2 Image of down-sampled boundary with 500, 100, 50, 25, and 8 vertices

measures are global, and some are based on persistence during down-sampling, which we denote sampling based. We summarize these measures below, but refer to [10] for full details.

3.1 Boundary-based measures

The boundary-based measures include ratio of peri-meter to area, total curvature, and a number of sampling-based measures: the ratio of down-sampled boundary to length of original boundary, ratio of area enclosed by the down-sampled boundary to area of original boundary, L^2 norm on the approximation error produced by down-sampling, Hausdorff norm on the approximation error produced by down-sampling, distribution of errors between down-sampled boundary and original, distributions of curvature at each sampling level, distribution of tangent angles at each sampling level, distribution of change in tangent angles at each sampling level, distribution of edge lengths in the Voronoi diagram at each sampling level, distribution of triangle areas in the Delaunay triangulation at each sampling level, and percentage of Voronoi cell centers that lie inside the shape versus outside at each sampling level. Note that all sampling-based measures are normalized by corresponding values in the full shape, and that values computed locally are stored as histograms. Boundaries are down-sampled until convex. The boundary is linearly approximated using a steadily decreasing number of points at five levels—500, 100, 50, 25 and 8 points–using arclength sampling at shifted starting points (Fig. 2).

3.2 Regional measures

The regional measures include the ratio of the area of the down-sampled shape to the area of the original shape, and the histogram of percentage of fill for pixels at the 100% resolution after down-sampling. Areas are down-sampled by scan-converting the original boundary curve into a 256×256 image I with 16 pixels of padding on all sides. Regions are down-sampled by placing a grid with an n-pixel neighborhood (for $n \in [2, 4, 8, 16]$) on top of I, resulting in four levels.

3.3 Skeletal measures

The skeletal measures are derived from the Blum medial axis [7] computed for each shape using circumcircles of the Delaunay triangulation. Centers and radii of the circumcircles give skeletal points and radii for the Blum axis. Following [19,22], the Extended Distance Function (EDF), Weighted Extended Distance Function (WEDF), Erosion Thickness (ET) and Shape Tubularity (ST) are computed for all skeletal points. EDF computes the geodesic depth of a skeletal point within a shape measured along the skeleton. WEDF computes the area-based depth of a skeletal point by taking the area of the shape part subtended by a given skeletal point. ET captures the local blobbiness of a shape, while ST capture the local tubiness. Together, they capture the fundamental geometric properties of a given shape part [21]. Histograms of each of these measures are computed point-wise along the skeleton, and also for the subset of skeleton points that are branch points and neighbors of branch points, where two parts of the shape join together.

3.4 Rank support vector machine (SVM)

Rank SVM applies the framework for linear SVM classification to ranking problems [11,31]. Training data for the system consists of items embedded into an n-dimensional feature space, and ground truth pairwise comparisons of ranks of those items—each item is either ranked higher or lower than each other item, but the full ordered ranking is not required. With a loss function meant to optimize Kendall's tau rank correlation [17], rank SVM produces a vector of weights in the feature space, $\mathbf{w} = \{w_i\}_{i=1}^{n}$, so that projection of items onto $\mathbf{w}$ results in a ranking that is as close as possible to the ground truth rank information. Originally developed for search result evaluations, where web pages in response to a search query are the items to be ranked and ground truth ranking is inferred from user click behavior, we apply it here to images of shapes with ground truth ranking provided by a forced-choice user study. Our feature space is defined by the complexity measures.

4 Identifying human-linked measures: Forced-choice user study

Given the perceptually unsatisfying complexity clusters presented in [10], we design and implement a small user study to determine which of those 282 complexity measures correlate best with human perception of complexity, and in which settings. Because the average user may not have a well-developed notion of shape complexity, we pose four different questions to address four potentially different aspects of complexity.
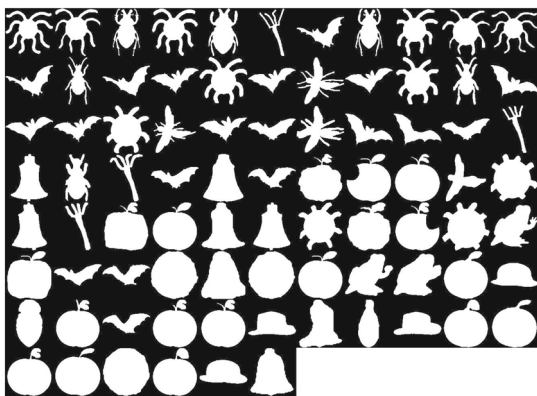
Fig. 3 Hand-edited images, ranked by the aggregation of user responses in order of complexity (all questions combined, all measures)

4.1 Methods for forced-choice study

We perform an initial forced-choice study with a small subset of images from the MPEG-7 images [8]). This generates, for each image pair, a complexity rank comparison based on majority vote of the users who compared those two images. These rankings can then be converted to a ranking scheme with associated weights on the measures using rank SVM [11], as described in Sect. 3.4.

For a follow-up study, we use a subset of this original dataset (see Sect. 4.1.2) and augment it with hand-edited images that represent specific complexity edits, such as removing a detail.

4.1.1 Initial forced-choice ranking

We use 69 images, one randomly selected from each category of the MPEG-7 images. We ask four questions, each capturing a slightly different notion of complexity (familiarity of shape, smoothness of boundary, complexity of boundary): (1) Which shape is more complex? (2) Which would be harder to draw from memory? (3) Which would be harder to cut out with scissors? (4) Which would take longer to trace?

The questions are always presented simultaneously in the same order and each participant answers all four questions for each image pair. We do not randomize these questions to avoid adding to the cognitive load the participant. Each participant sees every image paired at random with another image from the 69 images. All images are the same size, with a white shape on a black background and no interior features. Using Mechanical Turk, we gathered 242 responses, resulting in approximately four responses per ordered image pair. Average completion time was around five minutes, ranging from 3 to 10 minutes. Note that our image question arrangement results in *both* image orderings being present in the survey, preventing left-right image bias.

Data validity was checked using comparisons of the simple shapes (e.g., square) against complex shapes (e.g., insect or animal). No evidence of unreliable users was detected using these checks. See Fig. 4 for user agreement by question type.

We use the raw pairwise ranking data from user responses and rank SVM to create five weightings of the measures: one for each question and one for the combined answers to all questions.

4.1.2 Expanded study

Using the hypothesized editing operations from the prior work [10], we next create a set of hand-edited images that represent specific changes to the image: removing detail, thinning structures, editing the curvature of the boundary, adding noise to the boundary. We apply these changes to the image categories where there is a natural edit—apples (2), bats (4), beetles (2), bells (1), fly, fork, frog, hat, octopus. The complete set of edited images (sorted by the rank SVM output ranking for all questions combined) is shown in Fig. 3. We also hand-assign a ground truth rank-edit measure for each hand-edit by increasing (or decreasing) the rank by 1 for each major edit (e.g., removing all detail), and 0.5 for a minor edit (shortening the legs by one-half). All objects start with a score of 1 before editing.

Finally, we extract a curated set of image pairs based on object type and rank proximity. We hand-label the entire MPEG-7 dataset with type of object (manmade, abstract, animal) and whether or not that image was oriented "correctly" (several of the images are simply rotated copies of other images). For each of these reduced sets, we sort the images by their user rank from the initial study (using one of the 5 possible rankings) and then randomly select two images that were within plus or minus 10 ranking positions of each other, according to the users, in an attempt to capture subtler complexity shifts. We generate 10 pairs for each combination, for a total of 80 curated image combinations. We do not include both orderings of the images, but we do balance whether the left image or the right is higher ranked.

The expanded study used the same question format as the initial study, with all possible same object-type pairs from the hand-edited images plus the 80 image combinations from the curated image set for a total of 265 image comparisons. Each participant saw 50 questions and there were 50 participants, producing an average of four responses for each question.

4.2 Results from forced-choice study

We present analysis and results related to how consistently participants ranked the images, which measures were more correlated to user rankings for different question types, and how well the measures weighted by the first study predicted
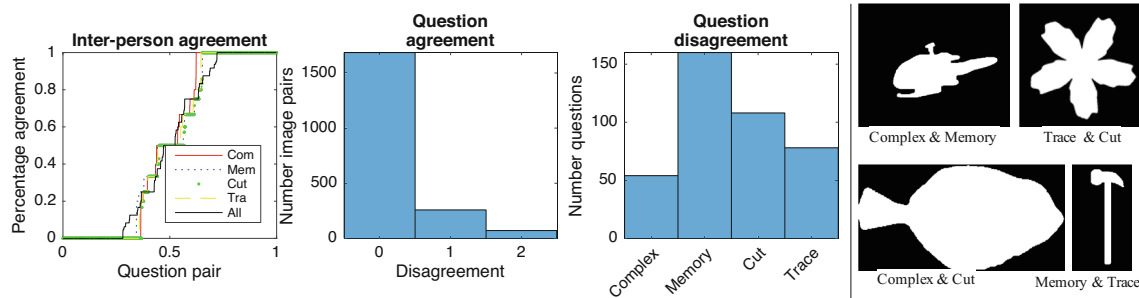
Fig. 4 From left to right (first three images): Inter-person agreement by question type, for image pairs with a preference, number of other questions that disagreed with that preference, distribution of that disagreement, and two examples of that disagreement. The four shapes to the right- Top left: The helicopter was ranked as more complex and more difficult to draw from memory, while the flower was more difficult to trace and cut out. Bottom left: The fish (more complex, difficult to cut) versus the hammer (more difficult to draw from memory and trace). See Sect. 4.2.1 for full details

the results from the expanded study. Note that user rankings of hand-edited images can be seen in the ordering of the images in Fig. 3. See also Table 2, top row, for Kendall's tau rank correlation between the induced ranking resulting from rank SVM computed on our images and the users' pairwise rank comparisons. We also note that the *distribution* of rankings of all images is more uniform for the complexity and memory questions than the scissor and trace questions.

4.2.1 Question consistency

We analyzed the responses both for when the participants differed in their responses (inter-subject) and for when they differed by response for one of the four specific complexity questions (intra-subject).

To make the inter-subject plot (left of Fig. 4) for each image pair, we recorded the number of votes for the first image versus the second for each of the four question types ("all" is the combined votes) and then sorted the values. Approximately 20% of the image pairs had disagreement amongst participants, with <1% having a roughly equal vote (between 0.4 and 0.6 percent agreement).

To measure question agreement, we counted the number of times one of the four study questions had *more* votes for the first image for one question and *more* votes for the second image for a different question. There were 606 image-pair questions (of the over 9,000 total) for which the votes were equal; they are not included in the plots shown. Middle left of Fig. 4 shows the number of image pairs for which the questions were in agreement (0), one question was different (1), or two of the four questions were different (2). The memory question was the most likely to vary from the other questions, with the complex question the least likely. Two examples of image pairs that differ in two questions are shown on the right of Fig. 4.

4.2.2 Measure importance

Analyzing the overall importance of measures in our large measures set as determined by the magnitude of their corresponding rank SVM weights in the user study, we find remarkable consensus among the top ten measures for the four questions and the grouped questions. We find fifteen unique measures in the top ten. See Table 1.

The top measures, those in the top 10 for all questions, cover a range of shape qualities. The first bin of the histogram of sampling-based boundary error at the highest two levels of down-sampling captures large-scale features that persist (i.e., have low error) for the coarse boundary samplings. The first bin of histograms of WEDF values for skeleton branch points and their neighbors captures the proportion of shape parts (where a part is defined by a branch in the medial axis) with the smallest areas—the smallest details of the shape. The middle bin of the sampling-based percent-filled histogram at the highest level of down-sampling captures the proportion of pixels at the original pixelation level that are half-filled by the image at the coarsest sampling level. This again gives information about persistence of boundary features in down-sampling, since a half-filled pixel is necessarily one that contains a portion of the boundary of the shape region. Finally, the middle bin of the curvature histogram at the two coarsest sampling levels captures the mid-range curvature features that persist. Of these high-ranking measures, we believe only curvature has been extensively studied as a complexity measure.

4.2.3 User ranking versus our ranking for hand-edited shapes

For our hand-edited images, we compare our hand-assigned rank score to the user study results for the edited pairs. We also compare the expanded user-study rankings to the rankings from the initial user study for *all* pairs of images. Our

Table 1 Importance of measures for each complexity question in user study described in Sect. 4, as determined by weight magnitude for rank SVM output weights

Measures/questions	Complex	Draw	Cut	Trace	All
Boundary error, bin 1, level 4	1	1	1	1	1
Boundary error, bin 1, level 5	2	2	3	3	2
Neighbor WEDF, bin 1	3	4	5	2	4
Percent area filled, bin 6, level 4	4	3	10	9	3
Curvature, bin 5, level 5	6	8	6	4	6
Curvature, bin 5, level 4	8	10	9	5	9
WEDF, bin1	7	5	17	6	8
Curvature, bin 5, level 3	10	9	11	8	12
Boundary error, bin 1, level 3	9	13	14	10	7
Boundary length change, level 4	11	11	7	30	5
Boundary length change, level 5	5	53	2	7	10
WEDF, bin5	43	6	58	110	123
Percent area filled, bin 3, level 4	20	7	37	57	133
EDF, bin 4	16	28	8	19	30
Mean area, level 5	44	58	4	25	25

Values in the table are the ordered positions of the absolute values of the corresponding weights in **w** for those measures. Note that the top measure is top for all questions. Note also that most questions share the same measures in their top rankings. The two questions capturing a different understanding of complexity are the questions about cutting and drawing, whose top measures are outside the top 50 for the other questions

hand-ranking was in 70% agreement with the user study rankings, with most of the disagreement arising from the edits that changed shape but did not remove detail. For example, our hand ranking marked greater complexity given operations such as making the bell asymmetric or fattening a stem on the apple, which was not reflected in the user study. The user rankings agreed with our explicit editing: removing detail reduced complexity, as did shortening legs, whereas adding noise or curvature increased complexity, see Fig. 3). In general, bending, thinning, or curving shapes also increased complexity. The ground truth rankings on the hand-edited images were in 80% agreement with the initial user study user-weighted rankings. As before, we compare inter-person agreement (see left of Fig. 5). There was more disagreement within this dataset than the original (4% versus 1% between 0.4 and 0.6).

Results on image rankings for the full expanded study data showed, on average, 90% agreement with the results on image rankings of the initial user study (by question type: 68/75, 65/75, 66/75, 77/75). Unsurprisingly, because the images were chosen to be "close" to each other in ranking, the inter-person disagreement is larger than for the initial study (see right of Fig. 5) with 7% versus 1% being between 0.4 and 0.6.
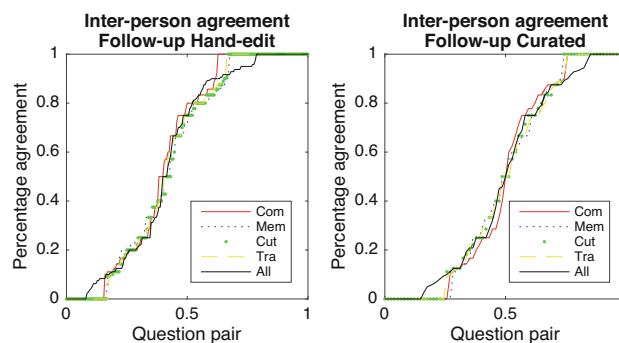

Fig. 5 Question disagreement for follow-up study

5 Constructed shapes with controlled complexity

Given the strong support by users for a small number of complexity measures from the user study, as shown in Table 1, we next explore the capacity of a subset of those top measures to capture complexification in shapes with a known complexity level.

5.1 Methods for constructed shapes

We select three of the high-ranking measures from the user study that are likely to be unrelated: the first bin out of 10 bins of the squared boundary error histogram after four levels of down-sampling (the top measure), the first bin out of 5 bins of the histogram of the WEDF values of branch points of the

medial axis and their neighbors (the second top measure), and the fifth bin out of 10 bins of the point-wise curvatures of the boundary at the highest level of down-sampling (ranked in the top 5 and based on a heavily studied complexity measure). The total magnitude of user-based weights from rank SVM for this subset of measures is 0.415, contributing almost half of the total unit vector **w**. With 282 total weights comprising a unit-length weight vector, these three measures carry a substantial proportion of the complexity ranking information, and their values are highly uncorrelated. We will refer to these in what follows as the boundary measure, the neighbor WEDF measure, and the curvature measure. We note here that we explored but discarded a fourth measure, the middle bin of the percent-filled measure at the top level of down-sampling, because it was both correlated with the curvature and boundary measures and also did not consistently identify any specific form of complexification in our experiments.

We then design two additional constructed shape sets with controlled complexity to augment our hand-edited set from Sect. 4. The two new shape sets are abstract in form so that we may clearly separate geometric complexity from perception of shape complexity due to semantic interpretation.

5.1.1 Top measures

Boundary Our boundary measure is generated from down-sampling, which reduces the number of vertices used to create the shape and erodes long protrusions first (Fig. 2). At the highest level of down-sampling, the shape becomes convex and loses most protrusions. The boundary measure, the first bin of the histogram at the next-to-highest level of down-sampling of distances between down-sampled boundary to the original boundary, measures the proportion of the original shape boundary points that remain after significant down-sampling and therefore have a small distance to the down-sampled boundaries.

Curvature The curvature measure is also a sampling measure. Taking the middle bin at the highest level of down-sampling gives the proportion of down-sampled boundary points in the middle of the curvature distribution. This measures persistence of mid-scale shape features. A circle, which has constant curvature and is not thought of as complex, would have a value of 0. A square, with curvatures of 0 and ∞ (or a discrete approximation thereof), would also have a value of 0 since all its curvature measures would fall into the first and tenth bins. A shape with some sharp corners, some straight regions, and some variability in between would have a nonzero value in the middle bin.

Neighbor WEDF The interior Blum medial axis gives a skeleton of a shape where branch points on the skeleton indicate shape parts connecting. The WEDF at a point measures the volume of the shape part supported by that point, which is the volume of the part that would be lost if the shape were truncated at that point. The first bin of the neighbor WEDF histogram gives the proportion of branch point neighbors that are supporting very small shape parts. This value will be small for very smooth shapes and shapes with primarily large parts, and will be close to one for simple shapes with a large amount of noise on the boundary creating multiple small volume parts.

To further support the effectiveness of these measures, we repeat a rank SVM calculation using *just* these three measures, see Table 2. Kendall–Tau rank correlations with the raw user-rankings, with the top three measures for each user study question, and with the full measure rank SVM are shown in Table 2. Although the correlation values drop a bit from the full measure set, they are still high for the 3-measure set, indicating that these three measures capture a considerable proportion of the information in user rankings.

5.1.2 Constructed shape data

We generate two constructed shape datasets with controlled complexity of different types. The first dataset is a blob dataset, meant to generate shapes with salient shape parts. We use interpolation between a set of random points to create a smooth closed shape. This results in blob-like shapes with large protrusions. Starting with a larger set of points for interpolation generally produces more protrusions in the resulting shape, which typically leads to increased complexity. See Fig. 6.

The second dataset, noisy circles, is designed to capture noisy small protrusions rather than the salient large protrusions of the blob dataset. We begin with a circle, a shape that both our intuition and the constructive model of complexity outlined in [10] consider to be the simplest shape. We then add noise by running a pseudo-random walk on each point on the boundary of a binary image of the circle and adding a pixel at each step of the walk. The walk moves in a random cardinal direction and for a random number of steps for each point on the boundary. We add levels of complexity by increasing the number of times that the circle cycles through every point on the random walk. We generate eight levels of complexity with ten shapes per level. See Fig. 7.

In addition, we extract all sequences of the hand-edited shapes where parts are gradually added or lengthened. See Fig. 12.

Fig. 6 Examples of blob constructed shapes, with four levels of complexity increasing from left to right

Table 2 Kendall's tau rank correlation between user rankings in the first user study, our full-measure rank SVM induced ranking, and our 3-measure rank SVM induced ranking

	Complex	Draw	Cut	Trace	All Qs
Full-measure to user rank	0.856	0.849	0.878	0.886	0.900
3-Measure user rank	0.763	0.664	0.803	0.815	0.776
Full-measure to 3-measure	0.851	0.766	0.863	0.880	0.838

Note that our 3-measure ranking performs almost as well as the full-measure ranking for Questions 1 (complex) and 3 (cut). For the other two questions, we are not capturing the user rankings as well. Neither the full- nor the 3-measure rankings perform well for the combined questions
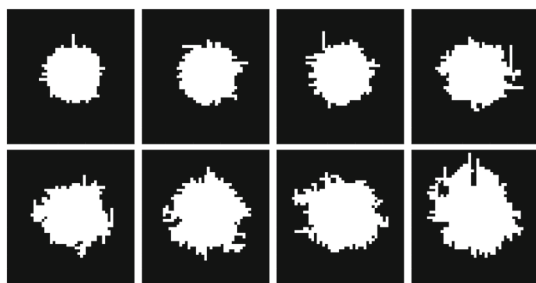
Fig. 7 Example sequence of the eight levels of complexity for noisy circles. Top row, L to R: complexity levels 1–4. Bottom row, L to R: complexity levels 5–8
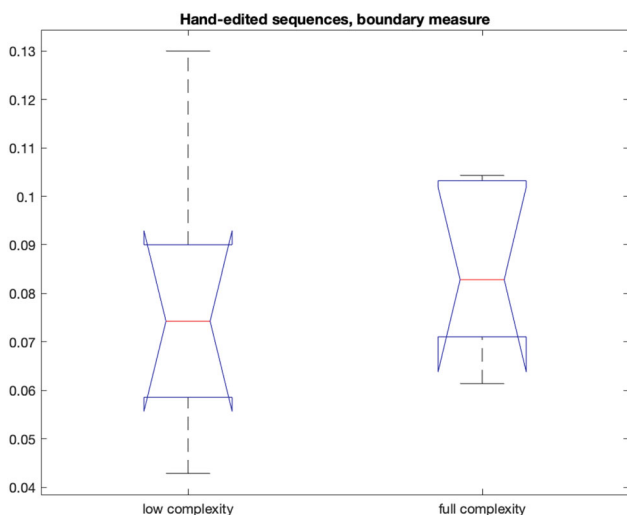
Fig. 8 Boxplots of the boundary measure values for the simplified hand-edited shapes (L) and the full complexity versions of the same shapes (R)

Fig. 9 Boxplots of the curvature measure values of the blob datasets at increasing levels of shape complexity (L to R)

5.2 Quantitative results for constructed shapes

We find from considering our constructed dataset that the three complexity measures are sensitive to different forms of complexification.

The boundary measure distinguishes well between the most simplified and most complex of the hand-edited shapes. See Fig. 8. The measure is not sensitive enough, however, to correctly identify the intermediate complexity stages, nor to identify the complexity increases in the noisy circles.
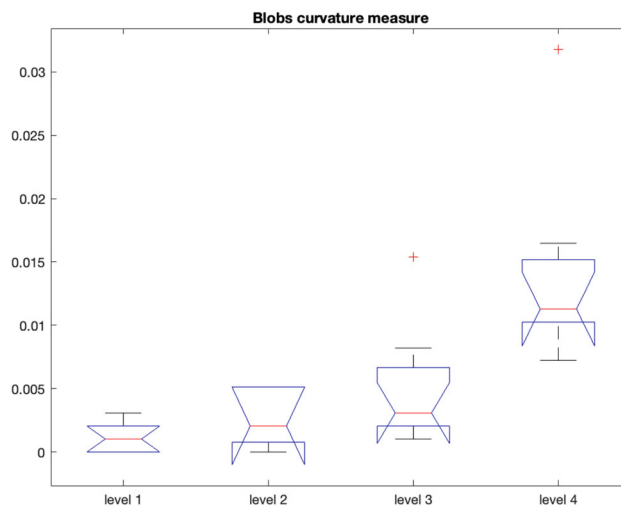
The curvature measure captures larger-scale complexification such as that found in the blob shape set. Figure 9 shows the box plots for the blobs at each of the four levels of increasing complexity and the corresponding increases in the curvature measure. Because the curvature histograms of the noisy circles are fairly consistent across the complexity levels, the curvature measure joins the boundary measure in failing to distinguish complexification in that dataset.

The neighbor WEDF measure distinguishes extremely well between the small scale complexity of the noisy circles and the larger-scale complexity of the blobs. See Fig. 10. But again, the measure does not distinguish between complexity levels within the noisy circles dataset.

Using the weighted rankings resulting from the three-measure rank SVM applied to the dataset from Sect. 4, we obtain a ranking that distinguishes well between the three shape categories of blobs, circles, and hand-edited shapes, but not so well within each category. See Fig. 11. In particular, the weighted ranking again does not distinguish between complexity levels within the noisy circles.

While each measure identifies some forms of complexity, none of these measures successfully distinguishes the complexity levels for the noisy circles. We conjecture that this is
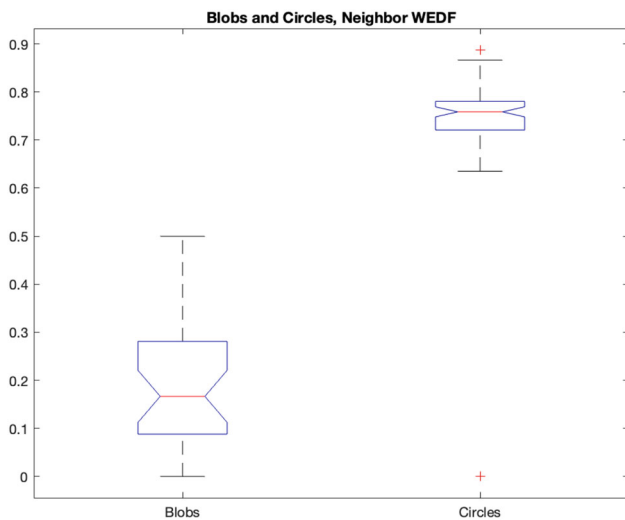
Fig. 10 Boxplots of the neighbor WEDF measure values of the blob (L) and circle (R) datasets at increasing levels of shape complexity
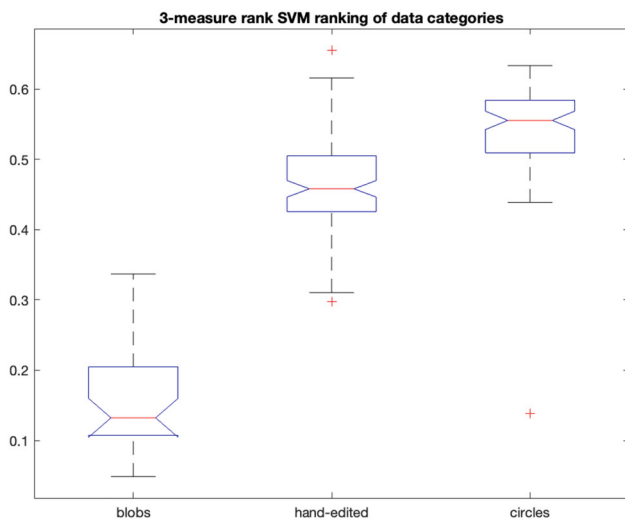

Fig. 11 Boxplots of weighted rankings from the three-measure rank SVM applied to shape categories in the constructed dataset

due to the randomness and more regular sizing of the shape parts in the noisy circles (Fig. 12).

6 Final user studies with constructed shapes

We perform two final user studies to validate our quantitative findings in Sect. 5. The first study aims to identify language non-experts use to describe the different types of hypothesized complexity. The second study uses this language to evaluate the user agreement with our three measures, and link specific words to specific measures if possible. For both studies, we use a curated set of image pairs that exemplify the different types of complexity changes with a mix of the hand-

Fig. 12 Hand-edited shapes with growing complexity

edited shapes, blobs, and circles, as well as a small subset of the original MPEG-7 set.

6.1 Methods for constructed shapes user study

Our first study is an open-answer Mechanical Turk study. Participants are presented with two images and asked to explain the *difference* between the two (natural language response). We use three prompting questions: (1) What words would you use to describe the differences between the two shapes? (2) What features make the shape more complex? (after being asked to pick which was more complex), (3) How would you describe the difference between the two shapes? A subset of pairs of images from the previous study (11 hand-edited, 15 MPEG) plus 4 constructed shapes (blobs and circles), were used. Approximately half the pairs were "recognizable" images. A total of 389 responses were collected from 40 participants, approximately 12 image pairs per participant. The text was analyzed for both anthropomorphic terms and common phrases.

For the second study, we repeat the forced-choice study design of Sect. 4, again using Mechanical Turk, but also ask participants to pick what phrase best-describes the complexity differences and to evaluate how similar the shapes are in terms of complexity (very similar, somewhat similar, not similar). The vocabulary of the phrases is based on our results from the open-answer study: (1) The boundary of the complex shape is more bumpy, (2) The boundary of the complex shape is more curvy, (3) The simpler shape is smoother, (4) The complex shape has more parts that stick out, (5) The simpler shape is more symmetric, (6) The complex shape has more bends in it. Although "parts" were only used once in the first part of this study, named parts *of* things (e.g., legs, petals) were used in all but the most abstract shapes, so we also include it as an option in our questions.

For the second study, we use 49 image pairs (6 low-complexity hand-edited, 9 circle-circle, 21 blob-blob, and 14 circle-blob). Pairs are chosen based on relative intended complexity in the dataset construction. Some pairs are chosen to be the same image type (e.g., both blobs) with either intended complexity level far apart (e.g., level 1 and level 3)
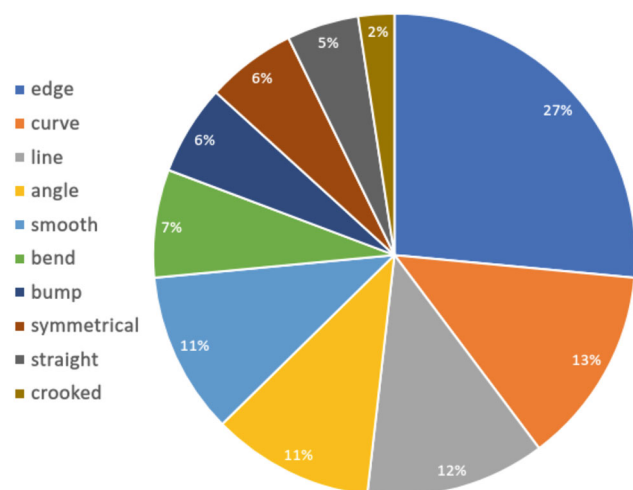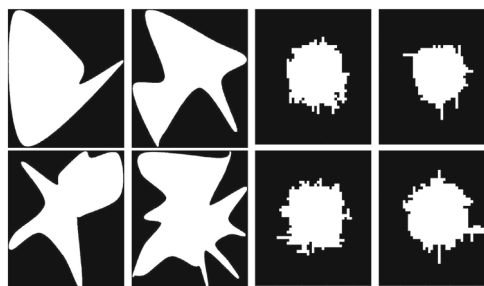
Fig. 14 Examples of shape pairings where our final user study as described in Sect. 6 showed user disagreement with our intended complexity levels. Top row shows shapes of lower complexity, and bottom row, same column, shows the other shape in study pairing that was intended to have higher complexity. Users flipped this ordering in their responses, so users considered the bottom row to be less complex than the top row

Fig. 13 A pie chart of the words provided by users to describe the complex shapes in the first part of the user study in Sect. 6

or closer together (e.g., level 2 and level 3). Some pairs are chosen to be different image types (e.g., blobs and circles) where one of our measures was effective at distinguishing between the types. These questions are designed to determine if our measure ambiguity is reflected in user ambiguity, and if our measure clarity is reflected in user clarity.

Twenty-three participants answered questions for 15 randomly selected image pairs, with an equal sampling from each image pair type, for an average of 5 comparisons for each image pair.

6.2 Results from constructed shape study

Figure 13 shows the most frequent words users in the first part of the study, where users were asked to provide vocabulary to describe why a shape is more complex in comparison with another. For instance, having more "edges" in a shape makes it more complex. The majority of these terms refer to the boundary with only four terms (bend, symmetrical, straight, crooked) referring to the global shape. We choose bumpy, curvy, and smooth for our questions. Although "edges," "lines," and "angle" were commonly used, they are potentially ambiguous and/or grammatically challenging. We also select two overall shape-based terms (bend, symmetrical). Finally, we note that there was a strong correlation between how recognizable the shape was and the use of anthropomorphic/descriptive terms (e.g., has legs, has rays), underscoring the importance of including abstract shape sets in the user studies.

In our second part of the study, user vocabulary preferences revealed some complexity category distinctions. By far the most selected phrase for all categories of image pairings was "parts" (approximately 30%), followed by "curvy" and "smoother." "Symmetrical" was only used consistently with

the hand-edited shapes (20% of answers), though the second-most commonly chosen word for the hand-edited shapes was "smoother" (24%). "Bumpy" was rarely selected for the blobs or hand-edited shapes ($< 7\%$), but was the second-most commonly chosen for the circles-blob comparisons (23%, as compared to 27% for "parts"), indicating that it is preferred as a word that distinguishes fine-scale boundary features from the larger-scale features. "Curvy," like "parts," appears to apply to both large and small scales. It was, however, selected more for the blob-blob comparisons than the circle-circles (30% vs 23%) and was the second-most commonly chosen for blobs-blobs. Linking this to our 3-measure results from Sect. 5.2, we might conclude that "bumpiness" is captured by the neighbor WEDF measure, "curviness" is (appropriately) captured by the curvature measure, "smoothness" is captured by the boundary measure, and the 3-measure rank SVM distinguishes between curviness, bumpiness, and smoothness.

The second part of the study also shows that users largely agree with our intended rankings in the constructed shapes, except for in the circles category. For the 37 shape pairings where both shapes were from the same category (e.g., circle-circle rather than circle-blob), 32 pairings (86.5%) showed user agreement with our intended ranking. Of the five pairings with disagreement, two were blob-blob and three were circle-circle. See Fig. 14. Overall, the circle-circle category also showed the most ambiguity in user ratings. For the circle-circle shape pairs with an intended complexity level difference, we consider pairings where average user agreement is close to 50% to be those where users disagreed with each other. Almost half of the circle-circle image pairs showed user disagreement, which may explain why our user-motivated measures are not distinguishing intended complexity well in that class.

7 Conclusion

We have interleaved results from user studies with quantitative analyses in order to identify and evaluate three key measures, out of an initial group of nearly 300, that capture important axes of shape complexity. We find that these three measures distinguish between complexity due to noise and due to salient parts, distinguish complexity levels for abstract shapes with increasing numbers of salient parts, and distinguish between categories of shape complexity. A final user study provides support for these quantitative results, as the users select different words for different categories of shape complexity that our measures are able to detect. We also contribute a database of abstract and meaningful shapes with designated complexity levels for further study.

There are several questions that future work should explore. Why are the top measures for the question "Which shape is harder to draw?" in the first user study so different from the other three questions? What role does semantic shape meaning play for users' interpretation of complexity in user study two? Is there a measure that captures the variation in complexity in the noisy circle database? And, of course, is there a larger set of measures that encompasses a richer understanding of complexity that can improve our comparison to user perception, particularly when comparing shapes across categories, such as comparing a beetle image to a blob? Finally, we note that the data in the user studies can offer up many more insights into human perception of shape complexity than we require in this work.

Data availability statement The data are available upon request.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

1. Arai, K.: Visualization of 3d object shape complexity with wavelet descriptor and its application to image retrievals. J. Visual. **15**(2), 155–166 (2012)
2. Arslan, M.F., Haridis, A., Rosin, P.L., Tari, S., Brassey, C., Gardiner, J.D., Genctav, A., Genctav, M.: Shrec-21: Quantifying shape complexity. Comput. Graph. (2021)
3. Attneave, F.: Physical determinants of the judged complexity of shapes. J. Exp. Psychol. **53**(4), 221 (1957)
4. Backes, A.R., Eler, D.M., Minghim, R., Bruno, O.M.: Characterizing 3d shapes using fractal dimension. In: Iberoamerican Congress on Pattern Recognition, pp. 14–21. Springer, Berlin (2010)
5. Balreira, D.G., Marcondes Filho, D., Walter, M.: Assessing similarity in handwritten texts. Pattern Recogn. Lett. **138**, 447–454 (2020)
6. Bensefia, A.: Arabic writer verification based on shape complexity. In: 2019 7th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6. IEEE (2019)
7. Blum, H.: A transformation for extracting new descriptors of shape. Models for the Perception of Speech and Visual Form, pp. 362–80 (1967)
8. Bober, M.: Mpeg-7 visual shape descriptors. IEEE Trans. Circuits Syst. Video Technol. (2001). https://doi.org/10.1109/76.927426
9. Bohg, J., Kragic, D.: Learning grasping points with shape context. Robot. Autonom. Syst. **58**(4), 362–377 (2010). https://doi.org/10.1016/j.robot.2009.10.003
10. Chambers, E., Emerson, T., Grimm, C., Leonard, K.: Exploring 2d shape complexity. In: Research in Shape Modeling. Springer, Berlin (2018)
11. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with SVMs. Inf. Retr. **13**(3), 201–215 (2010). https://doi.org/10.1007/s10791-009-9109-9
12. Chazelle, B., Incerpi, J.: Triangulation and shape-complexity. ACM Trans. Graph. **3**(2), 135–152 (1984). https://doi.org/10.1145/357337.357340
13. Chen, Y., Sundaram, H.: Estimating the complexity of 2d shapes. In: Proceedings of Multimedia Signal Processing Workshop (2005)
14. Feldman, J.M.: Information along contours and object boundaries. Psychol. Rev. **112**(1), 243–252 (2005). https://doi.org/10.1037/0033-295x.112.1.243
15. Hu, R., Weng, M., Zhang, L., Li, X.: Art image complexity measurement based on visual cognition: Evidence from eye-tracking metrics. In: International Conference on Applied Human Factors and Ergonomics, pp. 127–133. Springer, Berlin (2021)
16. Joshi, D., Ravi, B.: Quantifying the shape complexity of cast parts. Comput. Aided Des. Appl. **7**(5), 685–700 (2010)
17. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
18. Kim, S.H., Lyu, I., Fonov, V.S., Vachet, C., Hazlett, H.C., Smith, R.G., Piven, J., Dager, S.R., Mckinstry, R.C., Pruett, J.R., Jr., et al.: Development of cortical shape in the human brain from 6 to 24 months of age via a novel measure of shape complexity. Neuroimage **135**, 163–176 (2016)
19. Larsson, L.J., Morin, G., Begault, A., Chaine, R., Abiva, J., Hubert, E., Hurdal, M., Li, M., Paniagua, B., Tran, G., et al.: Identifying perceptually salient features on 2d shapes. In: Research in Shape Modeling, pp. 129–153. Springer, Berlin (2015)
20. Leonard, K.: Efficient shape modeling: epsilon-entropy, adaptive coding, and boundary curves -vs- blum's medial axis. Int. J. Comput. Vis. **74**(2), 183–199 (2007). https://doi.org/10.1007/s11263-006-0010-3
21. Leonard, K., Morin, G., Hahmann, S., Carlier, A.: A 2d shape structure for decomposition and part similarity. In: International Conference on Pattern Recognition (2016)
22. Liu, L., Chambers, E.W., Letscher, D., Ju, T.: Extended grassfire transform on medial axes of 2d shapes. Comput. Aided Des. **43**(11), 1496–1505 (2011)
23. Luo, Z., Xue, C., Niu, Y., Wang, X., Shi, B., Qiu, L., Xie, Y.: An evaluation method of the influence of icon shape complexity on visual search based on eye tracking. In: International Conference on Human-Computer Interaction, pp. 44–55. Springer, Berlin (2019)

24. Matsumoto, T., Sato, K., Matsuoka, Y., Kato, T.: Quantification of complexity in curved surface shape using total absolute curvature. Comput. Graph. **78**, 108–115 (2019)
25. Nitzken, M., Casanova, M., Gimel'farb, G., Elnakib, A., Khalifa, F., Switala, A., El-Baz, A.: 3d shape analysis of the brain cortex with application to dyslexia. In: 2011 18th IEEE International Conference on Image Processing, pp. 2657–2660. IEEE (2011)
26. Page, D.L., Koschan, A.F., Sukumar, S.R., Roui-Abidi, B., Abidi, M.A.: Shape analysis algorithm based on information theory. In: Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), vol.1, pp. I–229. IEEE (2003)
27. Page, D.L., Koschan, A.F., Sukumar, S.R., Roui-Abidi, B., Abidi, M.A.: Shape analysis algorithm based on information theory. In: International Conference on Image Processing, pp. 229–232 (2003)
28. Panagiotakis, C., Argyros, A.: Parameter-free modelling of 2d shapes with ellipses. Pattern Recognit. **53**, 259–275 (2016). https://doi.org/10.1016/j.patcog.2015.11.004
29. Papadimitriou, F.: The geometric basis of spatial complexity. In: Spatial Complexity, pp. 39–50. Springer, Berlin (2020)
30. Polasek, T., Hrusa, D., Benes, B., Čadík, M.: ICTree: automatic perceptual metrics for tree models. ACM Trans. Graph. (TOG) **40**(6), 1–15 (2021)
31. Radlinski, F., Joachims, T.: Query chains: Learning to rank from implicit feedback. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, pp. 239–248. Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1081870.1081899
32. Rajasekaran, S.D., Kang, H., Čadík, M., Galin, E., Guérin, E., Peytavie, A., Slavík, P., Benes, B.: Ptrm: Perceived terrain realism metric. ACM Trans. Appl. Percept. (TAP) (2019)
33. Rigau, J., Feixas, M., Sbert, M.: Shape complexity based on mutual information. In: 2005 International Conference on Shape Modeling and Applications (SMI 2005), 15–17 June 2005, Cambridge, MA, USA, pp. 357–362 (2005). https://doi.org/10.1109/SMI.2005.42
34. Saraee, E., Jalal, M., Betke, M.: Visual complexity analysis using deep intermediate-layer features. Comput. Vis. Image Understand. **195**, 102949 (2020)
35. Volarevic, N., Cosic, P.: Shape complexity measure study. In: Annals of DAAAM & Proceedings, pp. 375–377 (2005)
36. Wing, C.K.: On the issue of plan shape complexity: plan shape indices revisited. Constr. Manag. Econ. **17**(4), 473–482 (1999)

Dena Bazazian is a senior research associate at the Visual Information Laboratory of the University of Bristol, UK. Previously, she was a research scientist at CTTC (Centre Tecnològic de Telecomunicacions de Catalunya) and a postdoctoral researcher at the Computer Vision Center (CVC), Universitat Autònoma de Barcelona (UAB) where she accomplished her PhD in 2018. She was working with the Image Processing Group (GPI) at Universitat Politècnica de Catalunya (UPC) between 2013 and 2015. Dena Bazazian was one of the organizers of Deep Learning for Geometric Computing (DLGC) workshops at CVPR2022, ICCV2021, CVPR2020, Women in Computer Vision (WiCV) Workshops at CVPR2018, ECCV2018 and Robust Reading Challenge on Omnidirectional Video at ICDAR2017.
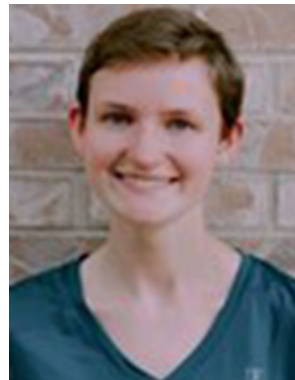
Bonnie Magland is a math PhD student at the University of Washington. They received their Bachelor's in Mathematics in April of 2022 from Brigham Young University with an Applied and Computational Mathematics Emphasis (ACME). There they did research in algebraic combinatorics. Bonnie was a fellow at the Summer Geometry Institute (SGI) at MIT in the summer of 2021. Bonnie is also a strong advocate for increased LGBTQ+ inclusion in academics. While at BYU, they served as secretary for Understanding Sexuality, Gender, and Allyship (USGA) from 2021 to 2022.

Cindy Grimm works in the area of robotic grasping and manipulation for both industry and agriculture, as well as ethics, law and policy related to robotics. Her previous projects include: modeling the developing heart, understanding how the shape of bat ears influences their sonar patterns, 3D sketching, and interfaces for 3D medical image segmentation. She received her PhD from Brown University in 1995 in the area of surface modeling, spent two years working at Microsoft Research on facial animation, then ten years as faculty in Computer Science at Washington University in St. Louis. She is now in the Mechanical Engineering department at Oregon State University.

Dr. Erin Wolf Chambers is a Professor at Saint Louis University in the Department of Computer Science, with a secondary appointment in the Department of Mathematics. Her research focus is on computational topology and geometry, with a more general interest in combinatorics and algorithms. Complementing this work, she is also active in projects to support and improve the culture and climate in computer science and mathematics at all levels. She currently serves as editor for several journals, on the board of trustees for the Society for Computational Geometry, and on the SafeToC organizing committee.

Kathryn Leonard is currently Professor and Founding Chair of Computer Science at Occidental College. Her research interests are in geometric modeling for data science, computer vision, and computer graphics applications, with an emphasis on explainability. She is founding co-Editor-in-Chief of the mathematics research journal, La Matematica, is President of the Association for Women in Mathematics, and has done extensive work to reduce barriers to research in computational sciences for members of minoritized groups.