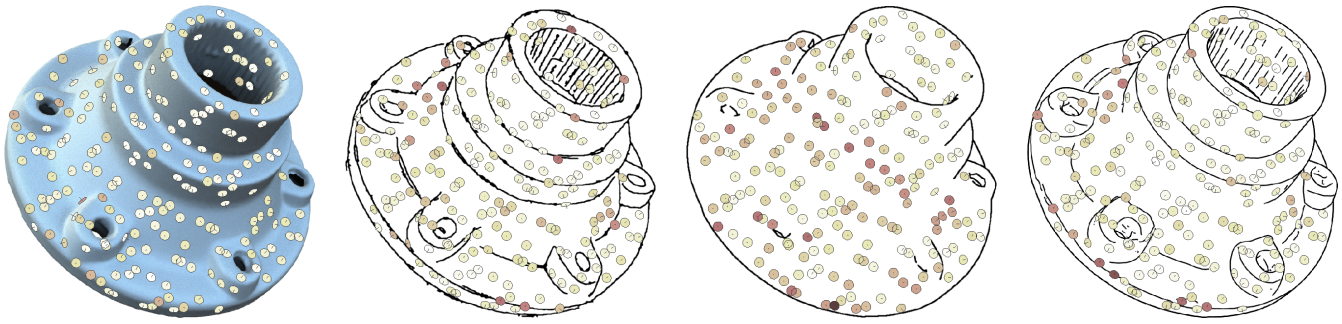
How Well Do Line Drawings Depict Shape?

Forrester Cole[1] Kevin Sanik[2] Doug DeCarlo[2]
Adam Finkelstein[1] Thomas Funkhouser[1] Szymon Rusinkiewicz[1,3] Manish Singh[2]
[1]Princeton University [2]Rutgers University [3]Adobe Systems



(a) shaded image (b) human drawing (c) contours (d) apparent ridges

Figure 1: *Gauge figure results. In this study, people were shown one of six different renderings of a shape: (a) a shaded image, (b) a line drawing made from the shaded image by a person, (c) contours, (d) apparent ridges, and (shown in Figure 7) ridges/valleys and suggestive contours. Overlaid are representative "gauges" (discs revealing the surface normal) oriented on the images by people in the study, colored by how far they deviate from the ground truth.*

Abstract

This paper investigates the ability of sparse line drawings to depict 3D shape. We perform a study in which people are shown an image of one of twelve 3D objects depicted with one of six styles and asked to orient a gauge to coincide with the surface normal at many positions on the object's surface. The normal estimates are compared with each other and with ground truth data provided by a registered 3D surface model to analyze accuracy and precision. The paper describes the design decisions made in collecting a large data set (275,000 gauge measurements) and provides analysis to answer questions about how well people interpret shapes from drawings. Our findings suggest that people interpret certain shapes almost as well from a line drawing as from a shaded image, that current computer graphics line drawing techniques can effectively depict shape and even match the effectiveness of artist's drawings, and that errors in depiction are often localized and can be traced to particular properties of the lines used. The data collected for this study will become a publicly available resource for further studies of this type.

Keywords: non-photorealism, line drawings, shape perception

1 Introduction

Line drawings are used for a variety of applications because they offer a "minimal" visual representation of a scene with little visual clutter, they can focus attention on critical areas, and they reproduce well over a broad range of scales and media. Common experience tells us that line drawings can be made for a wide range of shapes such that people generally agree about the shape they see; that some line drawings are more effective than others at doing this; and that some shapes are difficult to draw effectively. However, there is little scientific evidence in the literature for these observations. Moreover, while a recent thread of the computer graphics literature devoted to automatic algorithms for line drawings is flourishing, to date researchers have had no objective way to evaluate the effectiveness of such algorithms in depicting shape.

In this paper, we investigate how accurately people interpret shapes depicted by line drawings. At first, this goal seems difficult to achieve. Aside from asking sculptors to craft the shape they see, how can we know what shape is in a person's mind? Koenderink et al. [1992] have proposed several strategies for experimentally measuring perceived geometry, based on collecting the results of many simple questions. This paper describes such a study, based on the gauge figure protocol, in which the subject is asked to rotate a *gauge* (see Figure 3) until it appears to be tangent to the surface, providing a perceived surface normal. Previous studies using gauge figures and other related methodologies have used photorealistic images of shiny and diffusely shaded objects. With photographs, researchers find that people perceive shape accurately (up to a family of shapes related by the *bas-relief ambiguity*).

This paper offers substantial evidence that people can interpret shapes accurately when looking at *drawings*, and shows this for drawings made by both artists and computer graphics algorithms. Not all drawings are equally effective in this regard. We offer evidence that viewers interpret individual lines as conveying particular kinds of shape features (e.g. ridges, valleys, or inflections). Where different line drawing algorithms place different lines, the algorithms may be more or less effective at conveying the underlying shape. This paper offers both statistical and anecdotal data regarding the performance of various algorithms and drawings created by hand, with the goal of informing future development of computer graphics algorithms.

The results presented herein have been made possible by two resources unavailable in earlier research. Last year, Cole et al. [2008] published a dataset containing line drawings made by artists. This dataset provides us with a set of drawings for a variety of shapes, along with registered 3D models and viewpoints that serve as ground truth. The other enabling resource for this study is the Amazon Mechanical Turk, which has allowed us to distribute the

potentially tedious gauge figure task out to more than 500 subjects. Via this service we have collected more than 275K gauge samples distributed over 70 images. While this paper begins to analyze this data, there is more to learn from it. We therefore make the data available to other researchers for further analysis.

This paper makes the following contributions:

- We show that different people interpret line drawings roughly as similarly as they interpret shaded images.
- We demonstrate that line drawings can be as effective as photorealistic renderings at depicting shape, but that not all line drawings are equally effective in this regard.
- We provide new evidence that mathematical descriptions of surface features are appropriate tools to derive lines to convey shape in drawings.
- We offer a publicly available data set of gauge figures placed over a variety of drawings of 3D shapes; we believe this to be the largest gauge figure data set recorded to date.

2 Related Work

We use the data of [Cole et al. 2008] in a perceptual study of line drawings, and draw upon a range of work in computer graphics, computer vision, and the psychophysics of shape perception in our methods and analyses.

2.1 Lines and Interpretation

A range of lines on the surface that convey shape can be defined in terms of the geometry of the surface and viewer. The two fundamental lines of this type are discontinuities in depth—occluding contours [Koenderink 1984; Markosian et al. 1997; Hertzmann and Zorin 2000], and discontinuities in surface orientation—sharp creases [Markosian et al. 1997; Saito and Takahashi 1990]. Both are classical elements in line drawings [Willats 1997], and are commonplace in non-photorealistic rendering systems. In this paper, we study line drawings which contain contours along with one of the following three line types: (1) suggestive contours, (2) ridges and valleys, and (3) apparent ridges. Suggestive contours [DeCarlo et al. 2003; DeCarlo et al. 2004] are places where occluding contours appear with a minimal change in viewpoint. Lines along ridges and valleys are formed by local extrema of surface curvature along one of the principal curvature directions [Interrante et al. 1995; Thirion and Gourdon 1996; Pauly et al. 2003; Ohtake et al. 2004] and might be considered a generalization of sharp creases to smooth surfaces. Apparent ridges [Judd et al. 2007] are a variant of ridges and valleys, which are extrema of a view-dependent curvature that takes foreshortening into account. In future work, we hope to investigate other types of lines, including demarcating curves drawn in black [Kolomenkin et al. 2008], and highlight lines drawn in white [Lee et al. 2007; DeCarlo and Rusinkiewicz 2007]. Note that for the current study we specifically exclude hatching or texture lines, focusing instead on "sparse," shape conveying lines.

Research in computer vision offers a number of precedents for conceptualizing the process of understanding line drawings [Waltz 1975; Malik 1987]. This research emphasizes that reconstructing a 3D scene accurately involves recognizing qualitatively what each line depicts as well as inverting the geometry of line formation. In polyhedral scenes, each line segment has a consistent labeling across its entire length: it is either a convex or concave edge, or an occluding edge. Waltz [1975] showed how the set of possible globally consistent labelings could be inferred efficiently by constraint satisfaction from local possibilities to label junctions where lines meet. Malik [1987] extends this approach to drawings of piecewise smooth surfaces that contain occluding contours or creases, where labels are not necessarily consistent along the length of each line. The interpretations these algorithms find seem reasonable to people, but we do not know how the human perceptual system solves this problem. Even so, these ideas have found their way into effective interfaces for sketch-based modeling [Kaplan and Cohen 2006].

2.2 Psychophysical Measurements of Shape

In order to understand the shape perceived when people look at a line drawing, we rely on the *gauge figure* technique from visual psychophysics to obtain local estimates of surface orientation at a large number of points spread over a picture [Koenderink et al. 1992; Koenderink et al. 2001]. A gauge figure is simply a circle and a line in 3D, parameterized by slant (orientation in depth) and tilt (orientation in the image plane). When properly adjusted, it resembles a "thumb tack" sitting on the surface: its base sits in the tangent plane, and its "pin" is aligned with the outward-pointing surface normal. See Figure 3 for an example. Gauge studies can document not only shape interpretation but also the priors, bias, and information used by the human visual system. For instance, the direction of illumination affects the perceived shape [O'Shea et al. 2008; Caniard and Fleming 2007], and specular reflections can improve the perception of shape [Fleming et al. 2004]. Most gauge figure studies consider diffuse imagery. The only precedent for gauge figure study with line drawings is [Koenderink et al. 1996], who presented a single shape rendered as a silhouette, a hand-crafted line drawing, and a shaded picture, and found that the percept was better from the line drawing than the silhouette, and nearly as good as the illuminated version.

To interpret our results, we draw on the larger context of psychophysical research. For example, since people can successfully and consistently locate ridges and valleys on diffusely rendered surfaces in stereoscopic views [Phillips et al. 2003], it seems likely that the visual system represents these features explicitly. Also, perceived shape is likely an interaction between the global organization of line drawings and inherent biases of the visual system [Langer and Bülthoff 2001; Mamassian and Landy 1998], such as preferences for convexity over concavity in certain kinds of imagery.

A final wrinkle concerns the inherent underdetermination of depth from imagery. Given a shaded image, the 3D surface is determined only up to an affine transformation: individual points on the surface can slide along visual rays as long as planarity is preserved. This is the *bas-relief ambiguity* [Belhumeur et al. 1999]. Thus, to show the veridicality of human percepts of diffuse images, it is necessary to correct for this ambiguity [Koenderink et al. 2001]. In cases where human perception is highly noisy or inaccurate, however, a bas-relief correction is problematic. Because the correction uses the best fit between subjects' percepts and the true scene geometry, it can produce a misleading result when a good fit cannot be found.

2.3 Evaluation of Effectiveness in NPR

The psychophysical results we present offer fine-grained information about the effectiveness of NPR displays. Previous perceptual studies in NPR have largely focused on three areas: user preference, cognitive effort and task performance. User preferences for particular visual styles [Isenberg et al. 2006] show for example that users can find computer-generated imagery appealing even when it does not look like what a human artist could produce. To assess cognitive effort, experiments can use methodologies such as eye-tracking to confirm that people's attention is drawn to locations where detail is placed in a meaningful way [Santella and DeCarlo 2004; Cole et al. 2006]. Finally, researchers assess the overall effectiveness of stylized renderings by measuring performance on tasks like facial expression recognition [Wallraven et al. 2007] and recognition of facial caricatures [Gooch et al. 2004]. Stylization typically comes with a boost in performance (recognition speed), provided the information remains clear in the picture. As measured by the ability of subjects to identify a set of objects, one can determine that shading

leads to the best performance, followed by contours, followed by texture [Winnemöller et al. 2007]. One can also measure the effectiveness of cognitive models of assembly instructions and rendering style [Agrawala et al. 2003], by assessing people performing a real assembly task.

Previous evaluations let designers of systems make decisions about rendering style to achieve certain effects, recognition rates and user performance. For instance, in the illustrative rendering style used in the video game *Team Fortress 2* [Mitchell et al. 2007], characters are rendered to make individuals and team membership more recognizable. By contrast, our study is designed to tease out general relationships between algorithms for depiction and interpretation of shape.

3 Study

The study is designed to determine how viewers interpret line drawings and shaded images of the same shapes, and how consistent those interpretations are. The study is also designed to be broad enough to allow general conclusions about the effectiveness of line drawings. To achieve these goals, several issues must be decided: what images and drawings to show, how to sample each image with gauges, and how to structure the presentation to each participant so that a large amount of quality data can be gathered.

3.1 Subject Matter

One of the chief subjects of interest for this study is the effectiveness of line drawings by artists compared to computer generated line drawings. To compare human and computer generated drawings we use the dataset collected by Cole et al. [2008].

Besides offering artists' drawings registered to models, the dataset includes a useful range of 3D shapes. Intuitively, it seems that the usefulness of a line drawing varies with the type of shape depicted. For example, boxy shapes with hard edges are easy to depict with lines, while smoothly varying shapes are more difficult. We test this intuition by using the 12 models from the Cole dataset.

The study includes six different rendering styles: fully shaded, occluding contours only, apparent ridges [Judd et al. 2007], geometric ridges and valleys [Ohtake et al. 2004], suggestive contours [DeCarlo et al. 2003], and a binarized human artist's drawing. Contours are included in all the computer-generated drawings. The shaded image was created using global illumination and a completely diffuse BRDF with the eucalyptus grove HDR environment map [Debevec 1998]. For the cubehole model, the suggestive contour and apparent ridge styles are identical to the contour only and ridge and valley images, respectively, and are therefore omitted.

We endeavored to represent each drawing style as kindly as possible. For computer generated drawings, we smoothed the model and chose thresholds to produce clean, smooth, continuous lines. For the human drawings, we chose the subjectively best drawing available for the model (Figure 6). Our choice of view for each model was dictated by the view of the best drawing. Example computer-generated and human drawings are shown in Figure 2.

3.2 Methodology

We follow the gauge placement protocol described by Koenderink et al. [1992]. Participants are shown a series of gauges in random order and asked to place each gauge correctly before moving on to the next. The participants have no control over the position of the gauge, only its orientation. Each gauge is drawn as a small ellipse representing a disc and a single line indicating the normal of the disc. The gauges are superimposed over the drawing images and colored green for visibility (Figure 3). To avoid cueing the participants to shape, the gauges do not penetrate or interact with the 3D model at all. The initial orientations of the gauges are random.

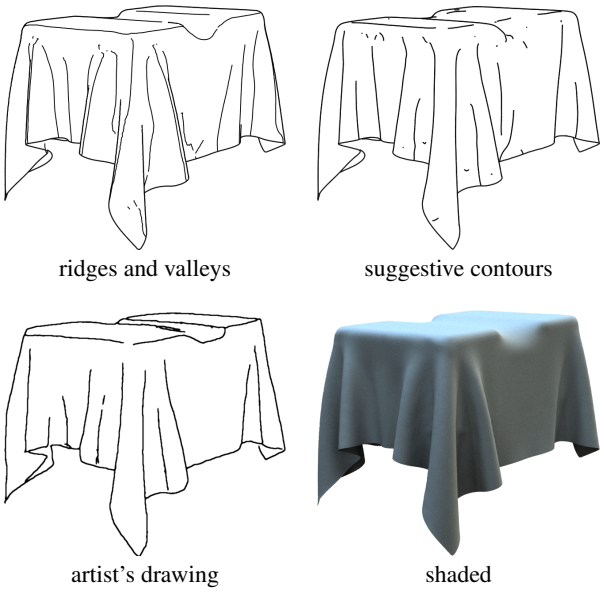


Figure 2: Example prompts. *Each style was represented by a clean, sparse, drawing, or a full-color shaded image.*

Participants were shown a simple example shape that is not included in our dataset in the instructions. The shape had examples of good and bad placement (Figure 3). Each time the participant started a session, they were allowed to practice orienting gauges on the example shape before moving on to the actual task. Participants were asked to orient the gauge by dragging with the mouse, and to advance to the next gauge by pressing the space bar. Participants were shown the total number of gauges to place in the session and the number they had placed so far.

The placement of gauges for each shape is determined in advance and is the same across the different rendering styles. We place gauges in two ways: evenly across the entire model, and in tight strings across areas of particular interest.

The evenly spaced positions are generated by drawing positions from a quasi-random Halton sequence across the entire rectangular image, and rejecting samples that fall outside the silhouette of the shape. All 12 models have at least 90 quasi-random gauge positions. Four representative models – the flange, screwdriver, twoboxcloth, and vertebra – have 180 positions in order to better understand how error is localized.

Four densely sampled lines of gauges (*gauge strings*) are also included in the study, one each on the flange, screwdriver, twoboxcloth, and vertebra. The gauge strings consist of 15 gauges spaced by 5 pixels along a straight line in screen space.

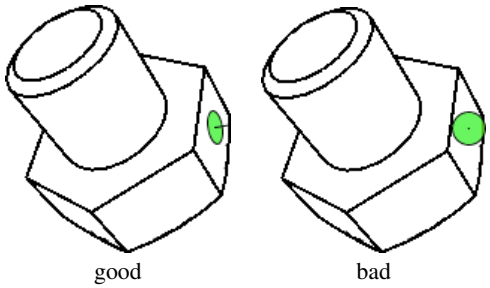


Figure 3: Instructional example. *Images such as these were included in the instructions for the study. Left: good setting. Right: bad setting.*

3.3 Data Collection

Previous studies of this type have included small numbers of highly motivated subjects. Each subject in the 2001 study by Koenderink et al. [2001], for example, was asked to perform approximately twelve hours of work, an impractical amount of time for any but a tiny number of subjects. In previous studies, the authors often provided all the data themselves. Our study has many subjects, but each is asked only a small number of questions. Rather than relying on the motivation of our subjects, we rely upon the robust statistics of many participants.

We used the Amazon Mechanical Turk as the source of participants in our study. The Mechanical Turk is a internet service that allows requesters (such as researchers) to create small, web-based tasks that may be performed by anonymous workers. Each worker is typically paid between $0.05 and $0.25 to complete a task. The number of workers on the service is such that particularly attractive tasks are usually completed within minutes. Unfortunately, workers on the Mechanical Turk are generally interested in work that takes around 5-10 minutes. This restriction dictates the number of gauges that we can present to a single participant. We found empirically that workers completed tasks at a satisfactory rate for up to 60 gauges in a session, but with more gauges workers became likely to give up on our task without completing it.

We asked each worker to set each gauge twice in order to estimate their precision. Setting reliability provides a measure of the perceptual naturalness or vividness of the observer's shape interpretation [Fulvio et al. 2006]. If the observer clearly and consistently perceives a specific shape in a line-drawing, then setting reliability will be high. If the percept is weak, and the observer has to "guess" to some extent, then reliability will be low. We showed each worker two sets of the same 30 gauges (60 gauges total), shuffled randomly and presented back-to-back. For simplicity, the sets of 30 are defined as consecutive sets in the Halton sequence for the 90 or 180 evenly spaced gauges. The statistics of each set of 30 are thus approximately equal.

To avoid training effects, each participant is allowed to see only a single style for each model. A participant is allowed repeat the study until they have performed the task once for each set of gauges. There are 52 distinct sets of 30 gauges across the 12 models, so a participant could perform the study up to 52 times.

Each worker is randomly assigned a stimulus from the available images each time they begin a new session. To favor more uniform sampling across the dataset, we weight more heavily the probability of receiving stimuli for which there is already little data, as follows. Out of the set of available stimuli we select randomly with probability proportional to $1/k_i + 1$ where k_i is the number of participants who have already successfully completed stimulus i.

Participants were told that the estimated time for completion was 5-10 minutes. The maximum time allowed for each session was one hour. The average time spent per session was 4 minutes.

In total, 560 people participated in our study and positioned a total of 275K gauges. The most active 20% of workers (115 people) account for approximately 75% of all data. The median number of sessions an individual performed was 4, but the median for the most active 20% was 28. The average time all individuals spent setting a gauge was 4 seconds.

3.4 Data Verification

Since we have no way of ensuring workers do a good job, we have to be careful to filter obviously bad data. Since we are asking for information about interpretation, however, there are no definitively wrong answers. We therefore base our rejection criteria on the reliability with which the worker positions each gauge.

We assume that if a worker is making a good faith effort, each duplicate set of gauges will be aligned in roughly the same orientation. Since each gauge is presented in a random orientation, a worker who simply speeds through our task will provide gauge settings distributed randomly across the hemisphere. Therefore, a worker's data is rejected if fewer than 70% of the duplicate gauges are set within 30 degrees of each other. These numbers were found empirically during pilot testing and remove grossly negligent workers without removing many good faith workers.

During pilot testing, we also noticed that through guile or misunderstanding some workers oriented all gauges in a single direction (usually directly towards the screen). This data passes our consistency check, but is useless. We therefore add an additional check, whereby a worker's data is discarded if the standard deviation of all gauge positions in a session is less than 5 degrees.

In all, approximately 80% of the data that we gather passes our two criteria and is included in the dataset. Each gauge in the study had an average of 15 opinions (two measurements of each). The minimum number of opinions was 9, and the maximum was 29.

3.5 Perspective Compensation

The dataset of Cole et al. [2008] includes drawings made from a camera with a vertical field of view of 30 degrees. The gauges, however, are drawn with an orthographic projection, to avoid any possible cues to the underlying shape. In order to compare the workers' gauge settings to ground truth, we must compensate for this difference by computing the ground truth gauge settings in the orthographic camera space.

The compensation method is as follows: create gauges for the ground truth normals, project the ellipses of those gauges by the camera projection matrix, reconstruct slant and tilt values from the projected ellipses, and finally reconstruct a normal from slant and tilt in the same way as the gauges set by the participants. Our comparisons against ground truth are made against these projected ground truth normals.

3.6 Compensation for Ambiguity

Koenderink et al. [2001] found that different subjects' perceptions of photographs of shape were related by bas-relief transformations [Belhumeur et al. 1999]. As Koenderink did, we can factor out this transformation separately for each person before making comparisons between different peoples' responses.

The bas-relief transform for an orthographic camera looking down the z-axis maps the surface point $[x, y, f(x,y)]$ to $[x, y, \lambda f(x,y) + \mu x + \nu y]$, given parameters for depth scaling (λ) and shearing (μ, and ν). To determine the best fit for a subject's set of gauges, we need to find values of μ, ν and λ that best map the subject's settings to the ground truth.

Given a set of bas-relief parameters, we can transform a set of normal vectors (re-normalizing them after the transformation). Thus, using a non-linear optimization procedure (which we initialize to the identity, $\lambda = 1$ and $\mu = \nu = 0$), we find the bas-relief parameters that minimize the L^1 norm of angular differences between the (normalized) transformed normals and the ground truth. We found the use of the L^1 norm to be robust to spurious gauges.

4 Results

Our data allows us to investigate several broad questions. First, how closely do people's interpretations of the stimuli match the ground truth shape? Second, how similar are two different peoples' interpretations of the same stimulus? Third, when compared with a shaded image, how effective are the line drawings in depicting shape? Fourth, are there particular areas for each model that cause errors in interpretation, and if so, can we describe them?

4.1 How well do people interpret shaded objects?

Before examining the results for line drawings, we investigate how well our participants were able to interpret the objects under full

shading. We expect that people will perform most accurately on the shaded prompts, so the results for these prompts provide a natural way to determine how successfully the average Mechanical Turk worker performed our task.

Across all the shaded prompts, the mean error from ground truth is 24 degrees, with a standard deviation of 17 degrees. After bas-relief fitting, mean error is 21 degrees, with a standard deviation of 16 degrees. Histograms of the errors for each style before and after bas-relief fitting are shown in Figure 4a and c. For comparison, a worker placing gauges randomly in each of the same tasks would have a mean error of 66 degrees, with a standard deviation of 31 degrees. After bas-relief fitting, random data would have a mean of 42 degrees with a standard deviation of 19 degrees. A worker rotating all gauges to face the camera (a situation we mark as bad data) would have a mean error of 42 degrees before bas-relief fitting, and 40 degrees after.

The reliability or precision with which the participants positioned gauges can be measured by comparing against the median vector for that gauge. If a gauge has n settings v_i, the median vector is the vector v_k that minimizes the total angular distance to every other v_i. Given the median vectors for each gauge, we can plot the error from the median (Figure 4b and d). In the case of the shaded prompts, the mean error from the median vector was 16 degrees, with standard deviation 14 degrees. These numbers do not change significantly with bas-relief fitting.

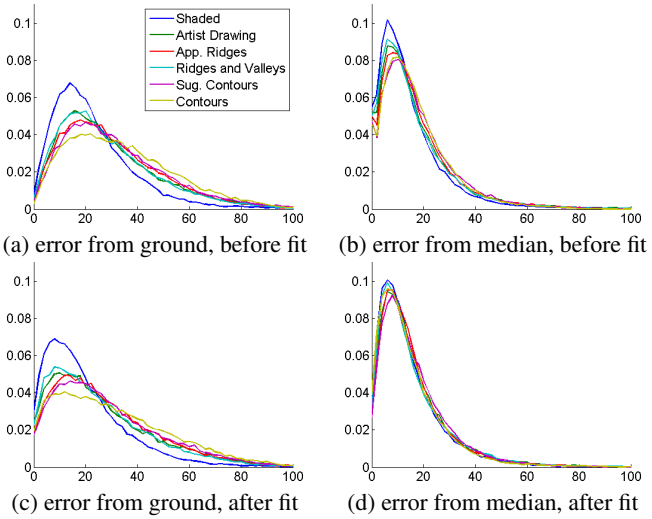
The scatter plots in Figure 5 give an alternate visualization of the distribution of errors for the shaded prompts. The orientations are shown using an equal-area azimuthal projection, so the area of each ring matches the area of the corresponding slice of the hemisphere. If the participants were placing gauges randomly, the points would appear uniformly distributed across the disk. The participants' settings, however, are clustered towards the center of each plot: for error from ground, 75% of the samples are within 31 degrees, or 14% of the area of the hemisphere, while for error from the median, 75% are inside 23 degrees, or 8% of the area of the hemisphere.

There is variation in the accuracy and precision with which workers placed gauges when seeing the shaded models, suggesting that some models were more difficult to interpret than others even under shading (Tables 1 and 2). The models for which the viewers had most difficulty are the smooth, blobby shapes such as the lumpcloth and the vertebra. For the obvious shapes such as the cubehole, however, the viewers interpreted the shape very closely to ground truth, lending confidence that viewers were able to perform the task successfully and that errors from ground truth in the line drawings are not simply the effect of negligent or confused participants.

4.2 How similarly do people interpret line drawings?

A simple way to examine how similarly our participants interpreted the line drawings is to compare statistics of the distributions around the median vectors at each gauge. We find that when seeing the line drawings, our participants set their gauges nearly as close to the other participants' as when seeing the shaded image (Figure 4b and d, Table 2). This result suggests that the participants all had roughly similar interpretations of each line drawing, and positioned their gauges accordingly.

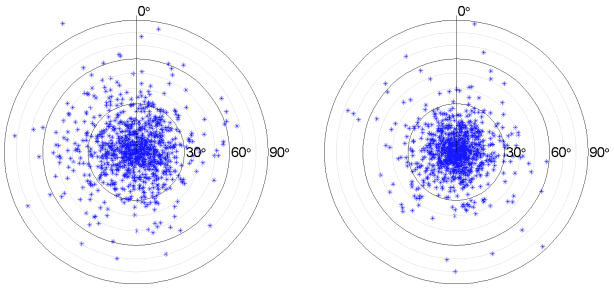
To get a more complete picture of the differences between the error distributions, we perform the Wilcoxon / Mann-Whitney non-parametric test for comparing the medians of two samples. This test yields a p-value. The full pair-wise comparisons of all error distributions are visualized in the inset of Figure 6. The colors match the colors in the legend of Figure 4. Black indicates a p-value > 0.05, meaning that we can not disprove the null hypothesis that those two distributions are the same. We find a statistically significant difference in error between most pairs of line drawings. This result suggests that while the interpretation of each drawing was similar across viewers, the viewers usually had different interpretations for each drawing.



(a) error from ground, before fit (b) error from median, before fit

(c) error from ground, after fit (d) error from median, after fit

Figure 4: Angular error distributions across all shapes (angle / frequency). *Errors from ground truth and from the median are shown before bas-relief fitting (a, b) and after (c, d). Note that the errors for the shaded prompts were considerably lower on average. Compared with errors from ground truth, the deviations from the medians are consistent across styles.*



error from ground error from median

Figure 5: Distribution visualization for shaded prompts. *Errors for 1000 randomly sampled settings. Left: errors from ground truth (blue distribution in Figure 4c), right: errors from median (blue distribution in Figure 4d). Radial distance is magnitude of error, compass direction is direction on the screen. Errors are roughly uniform in all directions, and errors from ground truth are larger than errors from the median.*

4.3 Do line drawing interpretations match ground truth?

Unlike precision, the accuracy with which our participants interpreted shape from the line drawings varies considerably with the type of drawing and the model (Figure 6, Table 1). In general, the performance of the occluding contours alone was considerably worse than the other drawing types, while the performance of the other drawing types (apparent ridges, ridges and valleys, suggestive contours, and the human drawing) were roughly comparable, though still often statistically significantly different.

The types of surfaces in the model have a large effect on the accuracy of interpretation. For the cubehole, which is largely made up of flat surfaces joined with ridges, the error from ground truth for all but the contours drawing is small: approximately 15 degrees on average. For the vertebra and cervical models, which are smooth and not obviously good candidates for line drawings, the errors for the best drawings are much larger: 35-40 degrees on average.

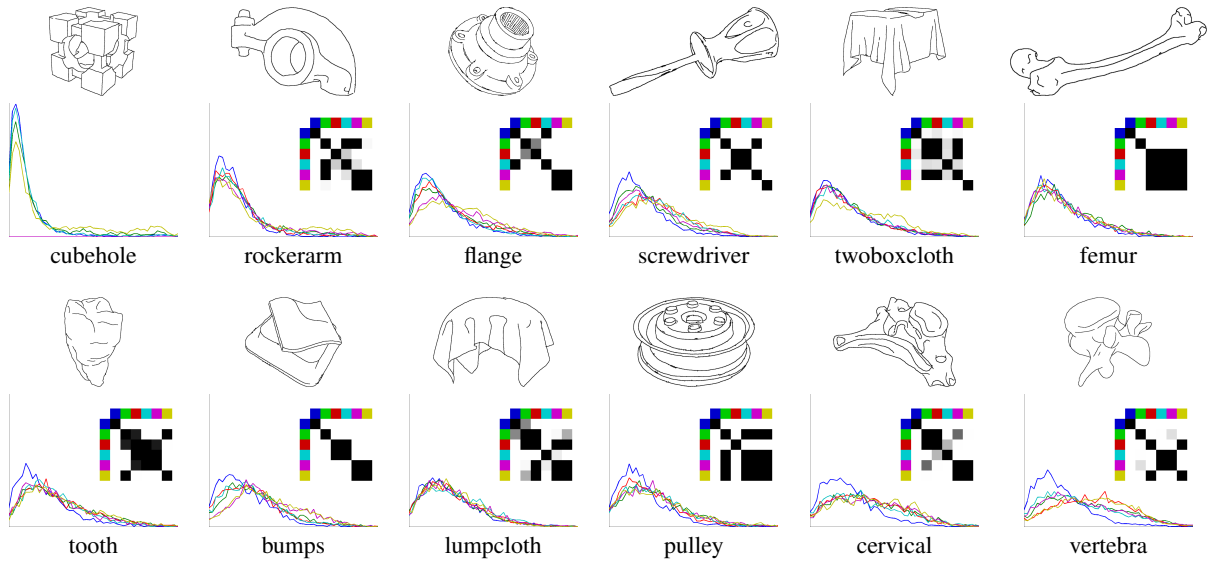


Figure 6: Distributions of angular errors from ground truth for all models. *Colors and axes are as in Figure 4 (x-axis: error,* $0 - 100$ *degrees, y-axis: frequency). Above the graphs are the human artists' drawings used for the models. Inset in each graph is a visualization of the p-values for significance (black: p-value* > 0.05*) of difference between distributions, where the colors correspond to the styles in the histogram. The table for the cubehole is incomplete and therefore omitted. Images are ordered by the mean error for the human artist's drawing.*

In these cases, even the human artists were unable to effectively convey the shape with only lines.

Examining the statistical significance between distributions, in almost all cases we find that the lit image did provide a statistically significant improvement over any line drawing, suggesting that some information is lost in the translation from model to line drawing. A notable exception is the flange model, for which the errors in the shaded and ridges and valleys versions are not reliably different (for a visualization, see Figure 1).

4.4 Is error from ground truth localized?

Beyond the aggregate statistics for each model, we can inspect the individual gauges to immediately determine if error is localized in a few important places, or if it is evenly spread across the model. If it is highly localized, then it may be interesting to examine high error areas in detail and attempt to form theories for why the errors occurred. In order to answer this question convincingly, we chose four representative models and scattered 180 (rather than 90) gauges on their surfaces.

Figure 7 shows gauges for the four representative models: the flange, screwdriver, twoboxcloth, and vertebra. It is immediately apparent from the plots that the errors from ground truth are not uniformly spread across the models, but rather exist in hotspots that vary with the style of drawing. For example, on the flange model we see heavy error in the suggestive contours drawing around the outer rim and the neck of the model. For the screwdriver, error is localized in the handle area. Error in the twoboxcloth model is localized around the folds near the front of the shape, whether lines are drawn there (suggestive contours, upper image) or not (apparent ridges, lower image). Error in the vertebra is large almost everywhere, even for the human artist's drawing, but relatively low in the flat area near the back of the model.

4.5 How can the local errors be described?

Once we have established that error is often localized in particular areas of each model, we can closely examine these areas by placing gauge strings. We chose four areas of interest, one on each of the four representative models in Figure 7. The median vectors for each gauge string, colored by error from ground truth, are visualized in the left and middle columns of Figure 8 (the images

Model	S	H	AR	RV	SC	C
cubehole	12	18	-	14	-	26
rockerarm	15	21	19	21	23	26
screwdriver	20	25	31	29	27	34
flange	21	26	25	22	32	32
pulley	21	29	27	29	29	30
bumps	22	29	27	27	36	36
femur	22	28	25	25	26	25
tooth	22	32	30	28	29	32
twoboxcloth	23	25	25	26	26	32
vertebra	24	38	42	35	37	42
cervical	25	37	35	35	37	38
lumpcloth	26	27	28	29	28	27
average	21	28	29	27	30	32

Table 1: Mean error in degrees from ground truth for each model and style. *Values shown are after bas-relief fitting. Rows are ordered by the mean error of the shaded prompt. Columns correspond to styles: S, shaded, H, human drawing, AR, apparent ridges, RV, ridges and valleys, SC, suggestive contours, C, contours only.*

Model	S	H	AR	RV	SC	C
cubehole	11	15	-	12	-	17
rockerarm	11	13	13	14	13	8
lumpcloth	14	16	16	13	15	14
femur	15	17	16	17	16	15
pulley	15	16	15	16	15	15
flange	16	18	19	16	21	20
screwdriver	16	17	16	15	17	13
bumps	17	18	16	18	14	13
twoboxcloth	17	16	18	18	18	20
tooth	18	22	21	21	21	21
cervical	19	17	18	18	13	12
vertebra	19	20	18	22	20	18
average	16	17	17	17	17	16

Table 2: Mean error in degrees from median orientations for each model and style. *Values shown are after bas-relief fitting. Rows are ordered by the mean error of the shaded prompt. Columns are same as Table 1. Note that, unlike the errors from ground, errors from median are sometimes* lowest *for the occluding contours drawing.*

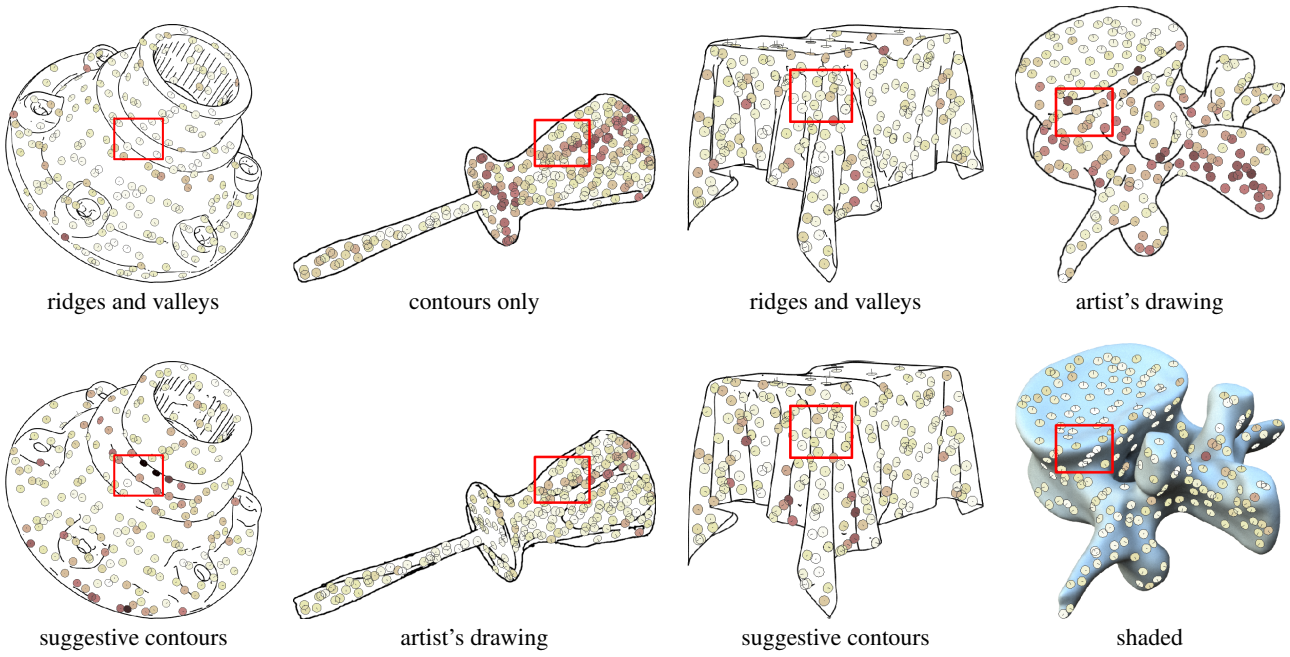


Figure 7: Plots of error of median vector from ground truth for four representative models. *Each pair shows a different finding from our study. Flange: lines are sometimes interpreted as ridges or valleys regardless of position, leading to errors when lines lie on suggestive contours. Screwdriver: without additional lines, the screwdriver appears as a cylindrical solid (contours only), but a skilled artist can depict the inflection points in the handles (artist's drawing). Twoboxcloth: errors in the folds appear both with lines (ridges and valleys) and without (suggestive contours). Vertebra: some shapes are difficult for even a skilled artist to depict. In such cases, the shaded image is significantly superior (though not perfect). Red box: area of interest shown in detail in Figure 8.*

are shown magnified for clarity, but the prompts were the same as for the random grids). Surface curvature can be estimated by differentiating along the string and is shown in the right column of Figure 8. Because our model of bas-relief ambiguity is global, we do not apply a bas-relief transformation to the gauge string data. Global fitting applied only to a small patch of local data is not well constrained, and can erroneously flatten the surface (set $\lambda = 0$) if the gauge settings are inaccurate.

Looking at the gauge strings in detail, we can conjecture what types of shape interpretation errors occur with each drawing style.

Flange: The errors on the flange model suggest that lines can be interpreted as representing particular geometic features, regardless of their exact location. The neck area of the flange is roughly a surface of revolution and includes a ridge and valley across the circumference of the shape. When presented with the ridge and valley drawing (Figure 8b), viewers interpreted the shape about as accurately as the shaded version. They were also quite consistent with each other, except where the gauge string crosses the valley line. When presented with the suggestive contour version (Figure 8a), however, viewers did poorly. It appears that viewers often interpreted the two suggestive contour lines as a ridge and a valley. The median absolute deviation for the suggestive contour gauge string is between 10-30 degrees, however, suggesting that the viewers held strongly differing opinions.

Screwdriver: The gauge string on the screwdriver lies across an inflection point in the surface. Without a line to depict the inflection point (Figure 8d), the change in curvature is lost – the surface appears as an area of uniform positive curvature, similar to a cylinder. The human artist managed to depict the inflection point effectively (Figure 8c). Both these interpretations are relatively reliable: median absolute deviation for each gauge in both drawings is 10 degrees or less.

Twoboxcloth: The fold on the front of the twoboxcloth provides a counterexample to the misidentification effect on the neck of the flange. Here, viewers interpreted both the the suggestive contour drawing (Figure 8e) and the ridge and valley drawing (Figure 8f) roughly correctly, though they performed better on the suggestive contour drawing. The median orientations indicate that viewers intepreted the ridge and valley drawing roughly as a triangle, with gauges oriented similarly on either side of the middle ridge. In the suggestive contour example, the viewers interpreted the lines correctly as inflection points, leading to a more accurately rounded shape. Viewers were roughly equally reliable in both of these interpretations.

Vertebra: The string on the vertebra is an example where the artist appears to have included a ridge line and an inflection point line (Figure 8h), but the depiction is not successful. Viewers interpreted the shaded image (Figure 8g) roughly correctly, but the drawing is completely misinterpreted. Viewers were also relatively unreliable when interpreting the artist's drawing: the median absolute deviation for the drawing gauges is between 10-20 degrees, approximately double the value for the shaded image.

5 Conclusion

This study allows us to comment on several issues in line drawings and depiction that have previously been speculative. In particular, we now have a partial answer to our original question: how well do line drawings depict shape? Further work is necessary, however, to answer other pertinent questions, such as how best to place lines for shape depiction, and the relationship between the aesthetic quality of a drawing and its effectiveness in depiction.

5.1 Main Findings

For about half the models we examined, the best line drawings depict shape very nearly as well as the shaded image (difference in mean errors < 5 degrees). This is true of the cubehole, rockerarm, flange, twoboxcloth, and femur. In the case of the flange, we did

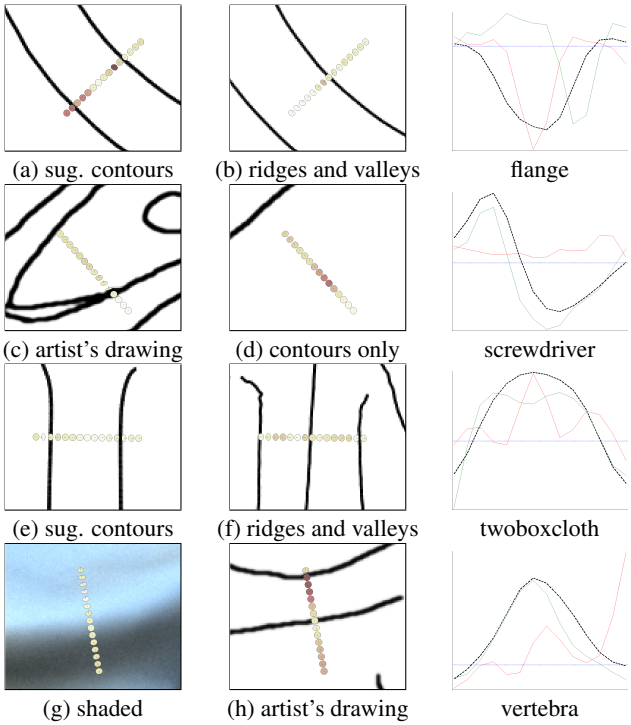


Figure 8: Gauge strings. *Right: curvatures along the string (green: left image. red: right image. dashed: ground truth. dotted: zero curvature). Four points of interest were studied in depth by placing gauge strings. Flange: the curvature valley for suggestive contours (a) is translated along the string. Screwdriver: the artist's drawing (c) depicts inflection points, while the contours drawing (d) does not. Twoboxcloth: suggestive contours (e) depict a smooth shape, ridges and valleys (f) depict a pointy shape. Vertebra: the artist's drawing (h) fails to depict the shape, while the shaded image (g) depicts it closely. Note: bas-relief fitting is* not *used for the strings.*

not find a statistically significant difference in errors between the shaded stimulus and the ridge and valley stimulus, while in the cases of the other models we did find significant difference, but the difference was small. These results suggest that viewers interpreted the shaded and drawn versions very similarly in these cases.

In other cases, such as the cervical and the vertebra, viewers had trouble interpreting the shaded image (mean error 24-25 degrees), but were completely lost on the line drawings (mean error 35 to 42 degrees). Such shapes tended to be smooth, blobby, and relatively unfamiliar to the eye. Viewers could not interpret these shapes accurately even with a drawing made by a skilled human artist. This result supports the intuition that certain shapes are difficult or impossible to effectively depict with a sparse line drawing.

Even when viewers interpreted the shapes inaccurately, however, their interpretations were similar to each other. For some shapes, including the rockerarm, cervical, and vertebra, the errors from ground for the drawings were 50-75% higher than for the shaded prompt, but the errors from median were similar or slightly *lower*. Only for the cubehole model did the average error from median and from ground match (11 and 12, respectively), suggesting that for most prompts, the viewers shared a common interpretation of the shape that differed from ground truth.

This study also indicates that line drawings based on differential properties of the underlying surface can be as effective in depicting shape as the drawings made by human artists. The best computer generated drawing had a slightly lower error than the artist's drawing in every case except the screwdriver. However, different mathematical line definitions can be effective or ineffective depending on context. For example, suggestive contours appear to be confusing in the case of the gauge string on the flange (Figure 8a), but quite effective in the case of the string on the folded cloth (Figure 8d). The human artists drew lines in similar locations to the computer algorithms, but appear capable of selecting the most appropriate lines in each case.

Finally, these results show that the Mechanical Turk can provide useful data for perceptual studies based on gauge figures. The service makes such studies considerably more practical, and should widen the potential areas for future investigation.

5.2 Limitations and Future Work

The gauge orientation task suffers from several limitations. First, there is no way to distinguish between errors due to misunderstanding the orientation of the shape and errors due to misunderstanding the orientation of the gauge. Second, the gauge methodology rewards line drawings that register very closely with the ground truth shape. An artist could depict a shape feature extremely well, but if the feature appeared 10 pixels from its true location, it would have high error. It is impossible, therefore, to distinguish a slightly offset but correct interpretation from a wrong one. Finally, we use only global bas-relief fitting in our implementation of the gauge figure study. Todd et al. [1996] suggests that local bas-relief fitting may give a more accurate model of perception of shape, but we collected too few settings at each gauge for each viewer (two settings) to make this approach practical.

We would also like to extend this study to more drawings and more styles of depiction. Our selection and generation of drawings is subjective, and while we endeavored to make the best drawings we could, we had no knowledge *a priori* what features would be successful. It is possible that different artists' drawings, or slight changes in algorithm parameters, could change our results. Beyond including more drawings, we would like to include wide-ranging depiction styles. It is possible, for example, that a different shading scheme could improve upon even the global illumination renderings, since viewers had trouble interpreting some of the shaded prompts.

As with many studies of this type, the results and analysis we have presented are purely descriptive. A natural area for future work is to investigate prescriptive results: for example, given a new shape, we could attempt to predict what lines will depict the shape most accurately. This study indicates that line definitions such as ridges and valleys and suggestive contours are effective in some cases and not in others, but it does not formalize where each type is effective. Formal rules of this type depend on an interpretation model of lines, which is an important long-range goal that this data may help support. Developing such a model would help us determine *how*, not just *how well*, line drawings depict shape.

Finally, artists create drawings to satisfy both functional and aesthetic goals. A study of this type cannot comment on the latter; it may be that the best drawings for shape depiction are also "ugly." Many factors can contribute to an aesthetic judgment and these factors are difficult to tease apart, but such data would be of tremendous value.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grants CCF-0347427, CCF-0541185, CCF-0702672, CCF-0702580, IIS-0511965, and IIS-0612231, and by Google. Thanks to Andrew Van Sant and John Wilder for early implementation work on the gauge figure study and on data analysis. The models used in the study come from the Aim@Shape, VAKHUN, and Cyberware repositories.

References

AGRAWALA, M., PHAN, D., HEISER, J., HAYMAKER, J., KLINGNER, J., HANRAHAN, P., AND TVERSKY, B. 2003. Designing effective step-by-step assembly instructions. *ACM Trans. Graph. 22*, 3, 828–837.

BELHUMEUR, P. N., KRIEGMAN, D. J., AND YUILLE, A. L. 1999. The bas-relief ambiguity. *Int. Journal of Computer Vision 35*, 1, 33–44.

CANIARD, F., AND FLEMING, R. W. 2007. Distortion in 3d shape estimation with changes in illumination. In *ACM Applied Perception in Graphics and Visualization (APGV) 2007*, 99–105.

COLE, F., DECARLO, D., FINKELSTEIN, A., KIN, K., MORLEY, K., AND SANTELLA, A. 2006. Directing gaze in 3D models with stylized focus. *Eurographics Symposium on Rendering* (June), 377–387.

COLE, F., GOLOVINSKIY, A., LIMPAECHER, A., BARROS, H. S., FINKELSTEIN, A., FUNKHOUSER, T., AND RUSINKIEWICZ, S. 2008. Where do people draw lines? *ACM Trans. Graph. 27*, 3.

DEBEVEC, P. 1998. Rendering synthetic objects into real scenes. In *SIGGRAPH 1998*, 189–198.

DECARLO, D., AND RUSINKIEWICZ, S. 2007. Highlight lines for conveying shape. In *NPAR 2007*.

DECARLO, D., FINKELSTEIN, A., RUSINKIEWICZ, S., AND SANTELLA, A. 2003. Suggestive contours for conveying shape. *ACM Trans. Graph. 22*, 3, 848–855.

DECARLO, D., FINKELSTEIN, A., AND RUSINKIEWICZ, S. 2004. Interactive rendering of suggestive contours with temporal coherence. In *NPAR 2004*, 15–145.

FLEMING, R. W., TORRALBA, A., AND ADELSON, E. H. 2004. Specular reflections and the perception of shape. *Journal of Vision 4*, 9, 798–820.

FULVIO, J. M., SINGH, M., AND MALONEY, L. T. 2006. Combining achromatic and chromatic cues to transparency. *Journal of Vision 6*, 8, 760–776.

GOOCH, B., REINHARD, E., AND GOOCH, A. 2004. Human facial illustrations: Creation and psychophysical evaluation. *ACM Trans. Graph. 23*, 1, 27–44.

HERTZMANN, A., AND ZORIN, D. 2000. Illustrating smooth surfaces. In *Proceedings of SIGGRAPH 2000*, 517–526.

INTERRANTE, V., FUCHS, H., AND PIZER, S. 1995. Enhancing transparent skin surfaces with ridge and valley lines. In *Proceedings of Vis 1995*, IEEE Computer Society, 52.

ISENBERG, T., NEUMANN, P., CARPENDALE, S., SOUSA, M. C., AND JORGE, J. A. 2006. Non-photorealistic rendering in context: an observational study. In *NPAR 2006*, 115–126.

JUDD, T., DURAND, F., AND ADELSON, E. H. 2007. Apparent ridges for line drawing. *ACM Trans. Graph. 26*, 3, 19.

KAPLAN, M., AND COHEN, E. 2006. Producing models from drawings of curved surfaces. In *Eurographics Workshop on Sketch-Based Interfaces and Modeling*, 51–58.

KOENDERINK, J. J., VAN DOORN, A., AND KAPPERS, A. 1992. Surface perception in pictures. *Perception and Psychophysics 52*, 487–496.

KOENDERINK, J. J., VAN DOORN, A., CHRISTOU, C., AND LAPPIN, J. 1996. Shape constancy in pictorial relief. *Perception 25*, 155–164.

KOENDERINK, J. J., VAN DOORN, A. J., KAPPERS, A. M., AND TODD, J. T. 2001. Ambiguity and the 'mental eye' in pictorial relief. *Perception 30*, 431–448.

KOENDERINK, J. J. 1984. What does the occluding contour tell us about solid shape? *Perception 13*, 321–330.

KOLOMENKIN, M., SHIMSHONI, I., AND TAL, A. 2008. Demarcating curves for shape illustration. *ACM Transactions on Graphics 27*, 5 (Dec.), 157:1–157:9.

LANGER, M. S., AND BÜLTHOFF, H. H. 2001. A prior for global convexity in local shape-from-shading. *Perception 30*, 4, 403–410.

LEE, Y., MARKOSIAN, L., LEE, S., AND HUGHES, J. F. 2007. Line drawings via abstracted shading. *ACM Trans. Graph. 26*, 3, 18.

MALIK, J. 1987. Interpreting line drawings of curved objects. *International Journal of Computer Vision 1*, 1, 73–103.

MAMASSIAN, P., AND LANDY, M. S. 1998. Observer biases in the 3d interpretation of line drawings. *Vision Research 38*.

MARKOSIAN, L., KOWALSKI, M. A., GOLDSTEIN, D., TRYCHIN, S. J., HUGHES, J. F., AND BOURDEV, L. D. 1997. Real-time nonphotorealistic rendering. In *Proceedings of SIGGRAPH 1997*, 415–420.

MITCHELL, J. L., FRANCKE, M., AND ENG, D. 2007. Illustrative rendering in Team Fortress 2. In *NPAR 2007*, 19–32.

OHTAKE, Y., BELYAEV, A., AND SEIDEL, H.-P. 2004. Ridge-valley lines on meshes via implicit surface fitting. *ACM Trans. Graph. 23*, 3.

O'SHEA, J. P., BANKS, M. S., AND AGRAWALA, M. 2008. The assumed light direction for perceiving shape from shading. In *ACM Applied Perception in Graphics and Visualization (APGV) 2008*, 135–142.

PAULY, M., KEISER, R., AND GROSS, M. 2003. Multi-scale feature extraction on point-sampled surfaces. *Computer Graphics Forum 22*, 3 (Sept.), 281–290.

PHILLIPS, F., TODD, J. T., KOENDERINK, J. J., AND KAPPERS, A. M. 2003. Perceptual representation of visible surfaces. *Perception and Psychophysics 65*, 5, 747–762.

SAITO, T., AND TAKAHASHI, T. 1990. Comprehensible rendering of 3-d shapes. In *SIGGRAPH 1990*, 197–206.

SANTELLA, A., AND DECARLO, D. 2004. Visual interest and NPR: an evaluation and manifesto. In *NPAR 2004*, 71–78.

THIRION, J.-P., AND GOURDON, A. 1996. The 3d marching lines algorithm. *Graphical Models and Image Processing 58*, 6.

TODD, J. T., KOENDERINK, J. J., VAN DOORN, A. J., AND KAPPERS, A. M. 1996. Effects of changing viewing conditions on the perceived structure of smoothly curved surfaces. *Journal of Experimental Psychology: Human Perception and Performance 22*, 695–706.

WALLRAVEN, C., BÜLTHOFF, H. H., CUNNINGHAM, D. W., FISCHER, J., AND BARTZ, D. 2007. Evaluation of real-world and computer-generated stylized facial expressions. *ACM Trans. Appl. Percept. 4*, 3, 16.

WALTZ, D. L. 1975. Understanding line drawings of scenes with shadows. In *The Psychology of Computer Vision*, P. Winston, Ed. McGraw-Hill, 19–92.

WILLATS, J. 1997. *Art and Representation: New Principles in the Analysis of Pictures*. Princeton University Press.

WINNEMÖLLER, H., FENG, D., GOOCH, B., AND SUZUKI, S. 2007. Using NPR to evaluate perceptual shape cues in dynamic environments. In *NPAR 2007*, 85–92.