Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing

Forsythe, Alex; Mulhern, Gerry; Sawey, Martin

Published in:
Behavior Research Methods

DOI:
10.3758/BRM.40.1.116

Publication date:
2008

Citation for published version (APA):
Forsythe, A., Mulhern, G., & Sawey, M. (2008). Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior Research Methods*, *40*(1), 116-129. https://doi.org/10.3758/BRM.40.1.116

Behavior Research Methods
2008, 40 (1), 116-129
doi: 10.3758/BRM.40.1.116

Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing

Alex Forsythe
John Moores University, Liverpool, England

and

Gerry Mulhern and Martin Sawey
Queen's University, Belfast, Northern Ireland

Complexity is conventionally defined as the level of detail or intricacy contained within a picture. The study of complexity has received relatively little attention—in part, because of the absence of an acceptable metric. Traditionally, normative ratings of complexity have been based on human judgments. However, this study demonstrates that published norms for visual complexity are biased. Familiarity and learning influence the subjective complexity scores for nonsense shapes, with a significant training × familiarity interaction [$F(1,52) = 17.53$, $p < .05$]. Several image-processing techniques were explored as alternative measures of picture and image complexity. A perimeter detection measure correlates strongly with human judgments of the complexity of line drawings of real-world objects and nonsense shapes and captures some of the processes important in judgments of subjective complexity, while removing the bias due to familiarity effects.

Pictorial complexity refers to the degree of detail or intricacy in a picture (Snodgrass & Vanderwart [S&V], 1980). Complexity is one of several subjective image characteristics frequently collected by researchers in normalization studies. Subjective ratings have long been used to provide normative data for the characteristics of visual stimuli for use in studies of object recognition, memory, naming, and semantic priming in normal populations and in those suffering neurological deficits. Proctor and Vu (1999) have indexed some 142 normative studies published by the Psychonomic Society since 1960, covering picture categories such as imagery, concreteness, familiarity, age of acquisition, naming times, and complexity. The origin of this approach lies in the work of Paivio, Yuille, & Madigan (1968), who published normative ratings of the concreteness, imagery, and meaningfulness of words. Prior research had sometimes relied on "unspecified judgments by the experimenter alone" (Paivio et al., 1968, p. 2). When S&V produced a set of norms for pictures, their motivation was similar to that of Paivio et al. Of particular concern was the extent to which picture sets created by researchers represented the intended picture characteristics and the degree to which it was possible to generalize the findings of experiments, using unstandardized pictures.

Since Proctor and Vu (1999) published their index of studies, others have developed and continue to develop new population norms—for example, in English (Barry, Morrison, & Ellis, 1997; Vitkovitch & Tyrrell, 1995), Icelandic (Pind, Jonsdottir, Trggvadottir, & Jonsson, 2000), or Italian (Dell'Acqua, Lotto, & Job, 2000)—and new sets of pictures for concepts not previously represented (Bonin, Peereman, Malardier, Méot, & Chalard, 2003).

Complexity and Its Influence on Reaction Time

S&V felt it likely that increased complexity would influence the speed at which pictures are categorized. Anthropogenic objects (simpler) would be categorized most quickly, and naturalistic complex images, such as *insects* more slowly. S&V suggested how, in episodic memory tasks, complexity is likely to influence stimulus recognition: The extra detail depicted in an object may give an image added novelty, and this novelty may slow the recognition process. In support of this idea, Rossion and Pourtois (R&P; 2005) reported some categorical reaction time advantage—that is, some categories tend to be responded to more quickly than others—although this seemed to be mainly a function of diagnostic color in categories, such as fruits/vegetables versus animals, rather than a function of complexity. Bonin et al. (2003) also found no significant relationship between visual complexity (VC) and naming times. In the icon/symbol literature, complexity does seem to influence response latency (rather than naming latency); whereas *concreteness* or how *real world* an image appeared determined accuracy, VC determined the speed at which users could search and respond (McDougall, de Bruijn, & Curry, 2000).

A. Forsythe, a.m.forsythe@ljmu.ac.uk

Measures of Complexity Affected by Familiarity: Studies With Adults

Close examination of the published norms pertaining to complexity points to the presence of a possible confound that was not taken into account in the original analysis and interpretation of the data. This is the relationship between familiarity and VC. S&V (1980) asked raters to consider familiarity as "how unusual the object is in your realm of experience"; this was defined as "the degree to which you come into contact with or think about the concept."

S&V (1980) reported a complexity–familiarity correlation of $r_s = -.46$, $p < .01$; likewise, when standardizing the S&V pictures in French, Alario and Ferrand (1999) report a correlation of $r = -.39$, $p < .01$. Alario and Ferrand argued that this correlation arises because visually complex pictures *tend* to be unfamiliar and more novel. Close inspection of their data does not support this explanation. In their study, pictures were scored on a 5-point scale. If 2.5 is taken as the midpoint on the scale, one can identify 63 pictures that are both unfamiliar and complex (i.e., familiarity < 2.5 and complexity > 2.5). There are 111 pictures that are both familiar and complex (i.e., familiar > 2.5 and complex > 2.5). The S&V picture set contains proportionally more complex familiar pictures than complex unfamiliar ones, and, as such, Alario and Ferrand's explanation for the inverse correlation does not hold.

To extend the number of standardized pictures available for testing, Bonin et al. (2003) developed a new set of pictures representing concepts not already available. The authors reported one of the smallest correlations between VC and familiarity in the picture-naming literature ($r = -.22$, $p < .01$). In the icon/symbol literature, McDougall, Curry, and de Bruijn (1999) published a relatively small correlation of $r_s = -.30$, $p < .01$.

One of the largest correlations in the literature ($r = -.50$) can be found through an analysis of data reported by R&P (2005). The authors were particularly interested in the differences between picture types, such as color, grayscale, and line drawings. They developed three sets of S&V-like drawings in line, gray shading, and color. R&P found reaction times to be shorter for responses to colorized drawings than for those to line or grayscale drawings. R&P also collected ratings on their S&V-style pictures for naming time, familiarity, complexity, mental image agreement, and color diagnosticity, but correlations among the different ratings were not examined. Subsequent analysis of the raw data (available at www.perceptionweb.com/misc/p5117) indicates the presence of a significant inverse association between complexity and familiarity for line drawings ($r = -.50$, $p < .01$), grayscale drawings ($r = -.41$, $p < .01$), and colorized drawings ($r = -.50$, $p < .01$). The probable cause of these significant correlations is that when judgments of images are elicited against several constructs (e.g., complexity, familiarity, name agreement, etc.), it is normal practice for different groups to be assigned to different image constructs (Bonin et al., 2003; McDougall et al., 1999; S&V, 1980). Participants should not be tested on more than one image construct, for fear that their judgments on one construct will influence the assessments on another. R&P departed from convention and asked the same participants to make judgments of familiarity *and* of complexity. Different groups of people scored items for familiarity *and* complexity, thereby increasing the likelihood that judgments of familiarity and complexity would be confounded. This is problematic, because the authors reported no significant correlations between picture reaction times and complexity or between reaction times and familiarity.

In general, there were few differences between the three sets of pictures (line, gray, and color) rated by R&P (2005). As might be expected, there was little difference in familiarity scores and complexity scores for line drawings, as compared with similar scores for grayscale or color drawings. Adding color or shading did not necessarily increase familiarity or perceived detail and intricacy. Simple correlational analysis of the R&P data by the authors suggests, however, that the relationship between complexity and familiarity is more substantial than was originally reported. There are moderate, negative correlations between reaction time and familiarity for line drawings ($r_s = -.50$, $p < .01$), for grayscale drawings ($r_s = -.45$, $p < .01$), and for color drawings ($r_s = -.46$, $p < .01$). Furthermore, there are small but significant correlations between complexity and reaction time for line drawings ($r_s = .22$, $p < .01$) and color drawings ($r_s = .23$, $p < .01$), suggesting that familiarity may be a stronger mediating factor for reaction times than for complexity.

Measures of Complexity Affected by Familiarity: Studies With Children

When gathering ratings from children, Cycowicz, Friedman, and Rothstein (1997) also tested each child on several picture constructs (name, familiarity, and VC). However, their data seem less problematic, because there was a smaller albeit significant inverse correlation between complexity and familiarity ($r = -.22$, $p < .01$). A possible explanation for this smaller correlation is that rather than using a standard instruction for grading complexity (e.g., "the amount of detail or intricacy of line in the picture") Cycowicz et al. asked the children "how difficult is it to draw or trace this picture." Although the children might have been puzzled by the standard instruction, it is possible that the alternative encouraged them to take into account their own drawing skill and that this partly influenced their judgments of complexity.

Measures of Complexity Unaffected by Familiarity

All of the major studies of picture norms have pointed to the presence of a moderate, statistically significant inverse correlation between complexity and familiarity. None have considered the important implication that the reported norms for picture complexity may be systematically flawed. An unbiased measure would be one in which a judgment of complexity is unaffected by the familiarity of the content. One approach to devising such a measure entails removing human observers and replacing them with an objective, automated metric.

The study of complexity has received relatively little attention—in part, because of the absence of an acceptable metric (Johnson, Paivio, & Clark, 1996). There have been

several attempts to develop rule-based metrics, with varying degrees of success. Geiselman, Landee, and Christen (1982) developed an index of discriminability and identified nine *primitive* attributes—for example, numbers of straight lines, arcs, quasiangles, and blackened-in elements. This metric was applied in an embedded search-and-select task. Participants were required to select symbols from a larger corpus of symbols in which three alternative representations of each concept were present. Using this metric, they found that stimuli with a high discriminability score were selected more quickly than those with low scores.

Garcia, Badre, and Stasko (1994) developed a metric based on a calculation of several image features, including the number of closed and open figures, and horizontal and vertical lines. For example, Figure 1 has a complexity score of 6 (two vertical lines, two horizontal lines, two arrowheads, and one closed figure). Garcia et al. originally intended this metric to be an objective measure of concreteness or how *real world* something appears. They reported that images that are pictorially similar to their real-world counterparts are more likely to be judged as complex. However, McDougall et al. (1999; McDougall et al., 2000) found that this metric was not, in fact, a good measure of icon concreteness but that it was useful for interpreting complexity norms gathered from human observers (McDougall et al., 1999). McDougall et al. (1999) found that the Garcia et al. *concreteness* metric was not correlated with human judgments of concreteness but that it was strongly correlated ($r_s = .73, p < .01$) with their judgments of complexity.

Hochberg and Brooks (1960) developed a semiautomated measure of image complexity. They argued that relying solely on human judgments would mean that there would be no way of predicting how complex a novel image might be judged to be. Hochberg and Brooks's calculations demonstrated that it was possible to predict how viewers would "see" an image; the more interior angles, different angles, and lines in an image, the more likely it was that it would be perceived in three dimensions. The number of interior angles, the average number of different angles, and the average number of continuous lines can be combined to provide a measure of complexity.

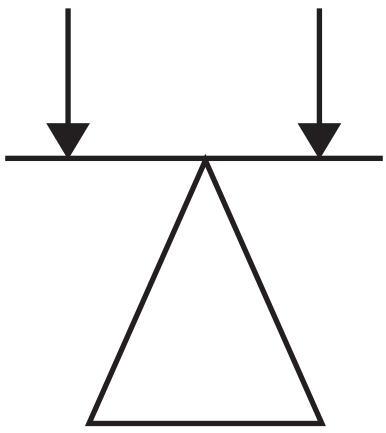


Figure 1. Complexity metric from Garcia, Badre, and Stasko (1994).

However, Attneave and Arnoult (1956) had argued that knowing how many dimensions are needed to explain a shape is not sufficient to judge its complexity, since some dimensions (e.g., reference axis or spaces) are more meaningful than others. In other words, the calculation of a metric based on increasing tridimensionality tells us very little about either the complexity of unfamiliar images or the learning processes that can influence the perception of form. Attneave and Arnoult wanted to understand the degree to which size, contrast, method, and familiarization influence the perception of form. They developed a system of calculations that could be used to generate nonsense shapes, the idea being that if testing using such a metric worked for images that had no meaningful relationship with real-world counterparts, it could be generalized to other stimuli.

Observers were given no advance information about the correct pattern but were required to make judgments regarding what the preceding element would look like. A simple image could be predicted from a limited amount of prior information. The observers made considerably more errors when predicating the structure of unpredictable shapes. Bartram (1973) applied the Attneave and Arnoult (1956) measure to generate nonsense shapes that differed from each other and had both the same mean association value and the same complexity value. People respond to familiar pictures more quickly than to novel pictures; however, with practice, even the speed of response to nonsense shapes can be improved, demonstrating that there is a familiarity component in the perception of complexity.

There are several practical reasons why researchers do not find this particular metric particularly attractive. The degree of detailed measurement involved in the identification, calculation, and documentation of primitive image components is time consuming. Moreover, should research teams change, it could be difficult to replicate the results of such meticulous measurement, because of the constant removal and addition of individual differences.

Symmetry and Higher Order Regularities

There are other geometrical components that interact with how complexity is judged by humans. Good symmetry is as likely to reduce perceived complexity as much as vertexes, objects, and holes are to increase it. For example, Attneave and Arnoult (1956) found that the most important stimulus property for predicting perceived complexity was the number of *turns* in a shape (i.e., changes in direction or corners; accounting for 78% of the variance), a finding that has been replicated by others. Chipman (1977) found that turns were important in the measurement of complexity and that turns interacted with judgments of symmetry. An image with horizontal symmetry would reduce perceived complexity at a rate equal to a 50% reduction in the number of turns, further indicating that humans are not particularly good at judging the "physical" complexity of a 2-D shape. Chipman also found that the amount of contour was a determinant of pattern complexity; however, turns were still the most significant contributor to perceived complexity. For example, an image with a fixed number of turns but a large perimeter area would be perceived as less complex than an image with a smaller perimeter area but more turns. In other words, small

jagged objects would be perceived as more complex than larger, less jagged objects. There has been some attempt to apply this metric to the measurement of architectural complexity; however, symmetry was reported as having a much smaller effect (Stamps, 2000), with a complexity trade-off between symmetry and turns of 25%. In that study, vertexes (extreme points) were the most important predictor of perceived complexity ($f = 53.9, p < .001, \hat{w}^2 = 42.8\%$).

Measuring Complexity: Why Should Automated Measurement Be Possible?

Some of the most successful theories of image processing (Biederman, 1987; Marr, 1982; Treisman & Gelade, 1980) consider measurable characteristics, such as the degree of detail within an image, as fundamental. Following this line of argument, Forsythe, Sheehy, and Sawey (2003b) have pursued one implication—namely, that a computer-based system capable of processing visual primitives might offer a valid measure of complexity for all 2-D stimuli.

For example, it is often argued that adding additional elements (e.g., primitive components, objects, or shading) to an object will increase its concreteness (Horton, 1994; Nielsen, 1993). Depicting both an object and an operation usually involves the inclusion of more elements to clearly communicate the intended meaning. Additional attentional processes are involved in building a cognitive representation, and as such, there is the need to integrate an increased number of constituent elements. Simple image properties are extracted from an image in parallel, and these properties are then combined to form objects of a particular shape, color, and size (see, e.g., Treisman, 1986; Treisman & Gelade, 1980; Treisman & Souther, 1985).

Forsythe et al. (2003b) tested the proposition that the number of discrete objects within a picture and the number of holes within those objects can be used as a measure of visual complexity. It was considered that a picture rated as complex would contain more elements than an abstract picture and that these elements would themselves be complex in nature, having more local detail (i.e., holes). The analysis was based on the connectedness between pixels, so that discrete objects were detected only when there were breaks between pixels. The holes within those objects were counted by calculating the Euler number of the image.

Although correlations between objects counts and subjective complexity were high, there was little evidence that they were psychologically plausible. For example, Figure 2 was rated by observers as containing two objects; the computer metric also rated the object as having two objects. We can see, however, that one object is the bug and the spray can and the second object is the droplet of spray (shown as a rogue gray pixel). Local detail (holes) presented similar anomalies; the sensitivity of the system enabled the detection of spaces between pixels that were not visible to the human eye.

Zhang and Lu (2004) identified several characteristics that an effective shape representation and description technique should have. The system should be robust and should be able to determine shapes in much the same way as a human observer; it should be stable, and there should be clarity about the ways in which measurement, identification, and description are attained. Although some of the measures reported by Forsythe et al. (2003b) failed on these criteria, automated measures based on psychophysical evidence proved more stable.

Figure 2. Debug (Forsythe, Sheehy, & Sawey, 2003b).

These measures were informed by arguments that changes in intensity, such as course and fine lines, are critical in providing information about a stimulus. The brain registers variations in an image as changes in intensity, and it is these coarse and fine changes that provide detail and local information about a stimulus (Beck, Graham, & Sutter, 1991; Harwerth & Levi, 1978; Sutter, Beck, & Graham, 1989; Vassilev & Mitov, 1976). Coarse scales are thought to be treated by the brain as low-frequency components obtained from local information. This difference in processing speed would seem to be a function of image complexity: When an object is of a detailed nature, its global attributes are processed much more quickly than its local ones (Hoeger, 1997; Parker, Lishman, & Hughes, 1997).

Forsythe et al. (2003b) showed that these basic perceptual components (i.e., edges) are important in the measurement of complexity. Two edge detection techniques were tested: the Canny edge detection algorithm and perimeter detection. Both techniques measure the changes in intensity that occur at the edges of an image element. The Canny technique is particularly useful in the detection of fine lines or gray shading. It works by using two thresholds to detect strong and weak edges and includes the weak edges in the output only if they are connected to strong edges. This means that truly weak edges will be detected in the analysis, but noise—such as shadow or shading—will be ignored. Perimeter detection measures (outlined in detail in the Method section of Experiment 1) more rapid changes in image intensity and performs well for images with sharp changes in contrast, such as line drawings. The extent to which an image is measured as having edges correlated highly with subjective judgments of image complexity. For example, the perimeter detection metric correlated ($r_s = .64, p < .001$) with a random set ($n = 68$) of the McDougall et al. icons and symbols and also correlated ($r_s = .66, p < .001$) with measures in Garcia et al. (1994). This perimeter metric has reasonably good predictive validity when applied to other pictorial images (Forsythe, Sheehy, & Sawey, 2003a). For example, it produces complexity scores that approximate human judgments when icons

are systematically manipulated; a simple contrast (black–white) inversion of the entire picture produces complexity scores that approximate users' judgments of complexity ($n = 239, r_s = .46, p < .001$). It is thought that these measures are effective because the algorithm underpinning perimeter detection takes into consideration the extent to which a pictorial image has edges; edges combine to form small shapes and add detail, and these are perhaps what make an image to be perceived as complex.

Perimeter measures have been described by Zhang and Lu (2004) as contour-based, global measures of shape. Perimeter measures do not divide a shape into parts; rather, the whole shape contour is used to describe the shape. This makes this type of measure very straightforward for users to implement and, as such, it tends to be a popular method of image measurement. An alternative automated measure of picture complexity that is also very straightforward to implement is based on the size of the compressed image file (Bates et al., 2003; Donderi, 2006a, 2006b; Vitevitch, Armbrüster, & Chu, 2004). Bates et al. used JPEG compression on black-and-white line drawings, and Vitevitch et al. applied JPEG compression to the standardization of visual stimuli consisting of words. JPEG compression is often a *lossy* type of data compression. This type of compression does not allow the exact reconstruction of an original image, and although the image tends to be "good enough," the process of removing small details and fine edges makes it particularly unsuitable for line drawings and textual or pictorial graphics (Taubman & Marcellin, 2001). Furthermore, the system also adds additional information, known as compression artifacts, that were not contained in the original image. Figure 3 shows the original Bitmap image and the resulting artifacts following JPEG compression (enhanced for visibility). Compression file size is also influenced by a number of factors other than image complexity (e.g., luminance and chrominance). These problems suggest a lack of clarity and stability (Zhang & Lu, 2004) in the application of JPEG as a measure of complexity.

Where the application of JPEG compression techniques is perhaps more justified is in the measurement of images such as electronic charts and radar screens (Donderi, 2006a, 2006b; Donderi & McFadden, 2003). In these highly detailed and colorized environments, JPEG compression file sizes can correlate highly with subjective measures of image complexity (between 25% and 85% of the variance).

Donderi (2006a, 2006b) has revisited information theory (Shannon & Weaver, 1949) as a possible framework that could explain the success of compression size as a determinant of complexity. Information theory treats a message as a series of components to be communicated, and this framework was adopted both by Attneave (1959) and by Hochberg and McAlister (1953) to explain the information content of visual images. The message components in a visual image are small image elements, such as angles and lines. As the number of different elements increases, so does the unpredictability of the message. This predictability improves when other image components can be used to determine meaning—that is, symmetry. Donderi (2006a, 2006b) argued that when a picture is compressed, the string of numbers that represent the organization of that picture is a measure of its information content. When the image contains few elements or is more homogenous in design, there are few message alternatives, and, as such, the file string contains mostly numbers to be repeated. A more complex picture will have more image elements, and these elements will be less predictable. The file string will be longer and will contain an increasing number of alternatives.

Figure 3. Line nonsense shapes showing compression artifacts.

All computational measures have some way to go before they will be able to account for and measure all the processes involved in the perception and cognition of images. Nevertheless, the current move toward the development of a quick approximation of human judgments of complexity suggests that further exploration of these types of automated measures is warranted. Computer-based measures also offer the potential to remove the familiarity effect in judgments of complexity present in all of the major studies of picture norms, thereby avoiding the need to conduct supplementary normalization studies.

Four image measurement techniques (Perimeter, Canny, JPEG, and GIF) were applied here to four sets of published data: R&P (2005), Bonin et al. (2003), and the classic set of adult ratings (S&V, 1980) and ratings for the S&V pictures collected from children (Cycowicz et al., 1997). The first experiment tested two propositions: first, that the automated metrics are a good approximation of how humans judge complexity in a picture, and second, that it accounts for the familiarity/complexity bias because, as was argued by Alario and Ferrand (1999), visually complex pictures *tend* to be unfamiliar and more novel.

EXPERIMENT 1

Method

The S&V (1980; including ratings for children, Cycowicz et al., 1997) and R&P (2005) picture sets ($n = 260$ pictures per set) and the data in Bonin et al. (2003; $n = 290$ pictures) were analyzed using four measures: perimeter, Canny, JPEG, and GIF.

Edge detection: Perimeter and Canny. MATLAB (MathWorks, 2001) is an integrated commercial package with powerful mathematical algorithms and visualization utilities for the acquisition, analysis of, and exploration of data. When preparing picture sets for processing, MATLAB treats a binary (black-and-white) image as an array of 1s and 0s. On white paper, black normally (but not always) represents the foreground, and white represents the background. MATLAB, on the other hand, considers white to be an *on* pixel, giving it the value of 1, and black to be an *off* pixel, giving it the value of 0. Thus, before any analysis was carried out, the representation of all the pictures in MATLAB was reversed, with 1s becoming 0s and vice versa.

The perimeter detection metric examines the changes in intensity occurring at the edges of an image. Edges are located with two criteria that are used to examine areas in the pictorial image where there is a rapid change in image intensity. Either a change in intensity must be larger than a predetermined threshold (edge detection provides a number of estimators that can used to specify sensitivity), or an edge will be detected where the intensity derivative has a zero crossing. Zero crossings are considered to occur at the places where negative and positive pixels are adjacent. For a pixel to be considered an edge pixel, it must be activated (on), and it must be connected to at least one nonactivated (off) pixel. This is a simplified version of more general detectors, such as Canny, which calculate the gradient of intensity values for close-by pixels in color or grayscale images. A limitation of the perimeter measure is that thicker lines are awarded higher scores than are thinner lines, because this measure rates a thick line as having two edges, rather than one. The selection of a four-connected neighborhood (rather than an eight-connected neighborhood) compensates for this problem to some degree, since it produces a finer image (see Forsythe et al., 2003b, for a fuller treatment).

Edges that are blurred or difficult to detect may, however, be included superfluously in a MATLAB perimeter detection calculation. To allow for these considerations, the Canny perimeter detection calculation was included. The advantage of the Canny method is that it works by using two thresholds to detect strong and weak edges and includes the weak edges in the output only if they are connected to strong edges. This means that truly weak edges will be detected in the analysis but noise—such as shadow or shading—will be ignored.

Compression: JPEG and GIF. Lossy compression using JPEG is contrasted here with a *lossless* compression, a technique that permits a reconstruction of the exact original image from the compressed data. GIF compression works better on pictures with limited colorization (<245) and performs particularly well on sharp transitions, such as diagrams or text (or in this study of line drawings). GIF compression can reduce a file size only to about half of its original size. To control for this difficulty, JPEG compression was also calculated to a 50% compression size.

Results

Table 1 shows the means, standard deviations, kurtosis, skew, and minimum and maximum automated counts for the automated measures for each of the three picture sets. For the S&V (1980) and R&P (2005) picture sets, the distribution for several of the measures was skewed at more than twice the standard error; similarly, Table 1 shows evidence of kurtosis in the distribution. A $\log_{10}$ transformation was considered in order to correct the distribution of these measures, but this made virtually no difference to the subsequent nonparametric correlational analysis. Similarly, scores were standardized into a 5-point scale to permit direct comparisons with published ratings. This adjustment tended to inflate the positive results slightly, but this increase caused slightly reduced variance within the data.

The reduction of scores was larger for the Bonin et al. (2003) picture set. Although this picture set presents a normal distribution for human judgments of complexity, the automated measures indicate a distribution that is significantly leptokurtic. Skew and kurtosis are commonly caused by sampling bias, nonnormal distribution of the characteristics of the items being measured, or the sensitivity of the measurement tool. Analysis using the automated measures raised two issues with this set. First, several pictures in the set were calculated as being extremely complex; these pictures attracted scores more than two standard deviations above the mean (thus forming the tail). Although humans also identified these images as highly complex, the range of scores available (1–5) did not adequately reflect the range of scores awarded by the automated measure (293–10,866); moreover, there were insufficient pictures available in this range of scores to correct the distribution

Table 1
Descriptive Statistics for Automated Measures of Complexity

	Complexity		Skew		Kurtosis			
Data Set	*M*	*SD*	*M*	*SE*	*M*	*SE*	Min.	Max.
Snodgrass & Vanderwart (1980)								
Perimeter	2,583.34	1,197.72	0.76	0.153	0.128	0.305	461	8,107
GIF	1,468.16	452.01	0.66	0.153	−0.065	0.305	646	3,172
JPEG	7,042.45	2,311.64	0.59	0.153	0.101	0.305	2,091	13,633
Canny	2,240.87	1,030.85	0.79	0.153	0.151	0.405	338	5,954
Bonin, Peereman, Malardier, Méot, & Chalard (2003)								
Perimeter	2,343.35	1,599.522	1.754	0.141	4.589	0.281	293	10,866
GIF	4,245.88	2,143.839	1.971	0.141	5.384	0.281	1,708	16,821
JPEG	6,838.12	2,425.858	0.963	0.141	0.915	0.281	3,211	17,140
Canny	2,315.03	1,437.954	1.534	0.141	4.122	0.281	264	10,448
Rossion & Pourtois (2005)								
Line drawings								
Perimeter	1,970.08	894.10	0.70	0.153	−0.01	0.305	381	6,144
GIF	3,099.93	902.49	0.53	0.153	−0.16	0.305	999	7,285
JPEG	2,689.36	571.26	0.39	0.153	0.00	0.305	1,560	4,690
Canny	2,188.29	931.03	0.54	0.153	−0.26	0.305	600	6,242
Gray-shaded drawings								
Perimeter	649.89	350.77	1.09	0.153	1.23	0.305	64	3,553
GIF	4,721.55	1,388.44	0.19	0.153	−0.49	0.305	1,698	33,056
JPEG	2,472.28	490.91	0.68	0.153	0.80	0.305	1,533	9,519
Canny	784.38	264.48	0.19	0.153	−0.11	0.305	210	5,827
Colorized drawings								
Perimeter	653.41	347.68	1.51	0.153	4.46	0.305	125	4,955
GIF	4,757.18	1,387.85	0.17	0.153	−0.60	0.305	1,811	29,895
JPEG	2,647.59	524.28	0.75	0.153	1.00	0.305	1,689	11,470
Canny	766.56	263.83	0.19	0.153	−0.21	0.305	207	5,671

($n = 7$). This problem can be resolved to some degree by removing these pictures as outliers, but this only reduces the leptokurtic distribution; it does not dissolve it.

Relative to the other picture sets reported here, Bonin et al.'s (2003) is larger ($n = 299$). The additional pictures are not distributed within the expected range of complexity (as determined by the automated measures). The picture set contains more pictures rated very simple than pictures toward the midpoint of the distribution. In the first quintile, this amounted to more than double the number of simple pictures contained in the R&P (2005) and S&V (1980) picture sets. This possibly explains why Bonin et al. reported one of the smallest correlations between complexity and familiarity ($r = -.22$, $p < .01$). A larger number of pictures varying in VC at all levels would perhaps present correlations closer to other published ratings.

Log_{10} transformation and histogram equalization made no discernable difference to the subsequent correlations; therefore, correlations with the Bonin et al. (2003) means are reported here. Given the difference in distributions, one would expect smaller correlations with the automated measures.

Outliers were removed from all the picture sets (Bonin et al. [2003], $n = 7$; S&V [1980], $n = 7$; R&P [2005], $n = 5$), reducing the picture set sample sizes to 252 (R&P and S&V) and 292 (Bonin et al.). Given the large data sets and the four different measures of JPEG, GIF, Canny, and perimeter, a Bonferroni adjustment was set at .0002. A caveat to this adjustment is that it greatly increases the likelihood of Type II error.

Validity of the perimeter detection metric. One would predict that for line drawings, techniques such as perimeter detection and GIF compression will show stronger correlations with human judgments of complexity than techniques such as JPEG. The relationship should be strongest in data sets that show less evidence of a familiarity bias (i.e., Cycowicz et al., 1997; S&V, 1980) and a strong but reduced relationship with the data sets containing a different range of complex to simple pictures (Bonin et al., 2003). For the S&V data set, correlations between the different automated measures are strong and broadly comparable. There is only a slight advantage in using GIF or perimeter detection over JPEG, and similar results can be observed (Table 2) for the data relating to children's judgments (Cycowicz et al., 1997). The data published by Bonin et al. also correlates well with the automated measures. The largest correlate is with perimeter detection, although the correlations are smaller because of the larger number of simple pictures (Table 3).

The stronger familiarity/complexity bias found in the R&P (2005) data set accounts for the smaller correlations (Table 4). Despite this problem, some slight variations between different methods of image measurement are detectable: JPEG compression, for example, is the strongest correlate of complexity in the colorized picture set ($r_s = .53$, $p < .0002$) and in the grayscale set ($r_s = .60$, $p < .0002$). For line drawing, GIF compression shows some advantage ($r_s = .65$, $p < .0002$). These differences are small, but they are predictable from the recommendations regarding the correct application of different image compression techniques.

Tables 2–4 also show the correlations between subjective complexity and familiarity. This pattern is less apparent in the automated measures. There are only two significant but small correlations between familiarity and JPEG or GIF compression techniques for the S&V (1980) data set, ($r_s = -.24$ and $r_s = -.26$, respectively). Although comparable in size, similar correlations between familiarity and the automated measures did not reach statistical significance (as determined by the Bonferroni adjustment) for the R&P (2005) data set. No relationship between automated complexity and familiarity was detected in the Bonin et al. (2003) picture set.

Judgments of complexity as a function of familiarity. The following analyses explore the argument that visually complex pictures *tend* to be unfamiliar (Alario & Ferrand, 1999). If we consider perimeter detection as an unbiased measure of VC, it is possible to investigate the extent to which human judgments of familiarity vary across stimuli of varying complexity. Very complex pictures should be less familiar than more simple stimuli.

Human judgments of complexity fall within the range of 1–5, whereas perimeter detection and other automated measures have a much larger range (see Table 1). To permit direct comparisons, all the scores were standardized on a 5-point scale (reflecting the human judgments obtained using 5-point rating scales). The standardization was calculated, through histogram equalization, into five intervals (or quintiles). Although standardizing the scores was unnecessary for the previous correlational analysis, the adjustment was necessary here to examine differences between perimeter detection and human judgments at the high and low ends of the rating scales.

It was predicted that relative to pictures falling in the middle of the 5-point scale (Quintiles 2–4), very complex pictures (falling in the fifth quintile) would attract significantly lower familiarity scores, whereas the simplest pictures (falling in the first quintile) would be rated as more familiar.

Human complexity with human familiarity. A one-way ANOVA was performed on human judgments of familiarity, with perimeter quintiles as a factor. Although there is a trend for familiar objects to be judged as less complex (Figure 4), there is very little supporting statistical evidence. For the S&V (1980) picture set, a significant effect was

Table 2
Spearman Correlations: Snodgrass and Vanderwart (1980) and Cycowicz, Friedman, and Rothstein (1997)

	Complexity	Familiarity	Perimeter	Canny	JPEG
		Snodgrass & Vanderwart (1980)			
Familiarity	−.46*	1.00			
Perimeter	.73*	−.19	1.00		
Canny	.68*	−.16	.99*	1.00	
JPEG	.72*	−.24*	.96*	.96*	1.00
GIF	.75*	−.26*	.94*	.93*	.97*
		Cycowicz et al. (1997)			
Familiarity	−.21*	1.00			
Perimeter	.65*	−.06			
Canny	.61*	−.05			
JPEG	.65*	−.07			
GIF	.65*	−.10			

Note—For Cycowicz et al., perimeter, Canny, and JPEG, see Snodgrass & Vanderwart. *$p < .0002$.

Table 3
Spearman Correlations: Bonin, Peereman, Malardier, Méot, and Chalard (2003)

	Complexity	Familiarity	Perimeter	Canny	JPEG
Familiarity	−.23*	1.00			
Perimeter	.45*	−.13	1.00		
Canny	.39*	−.10	.93*	1.00	
JPEG	.41*	−.12	.92*	.95*	1.00
GIF	.39*	−.11	.81*	.85*	.84*

*$p < .0002$.

found for adult ratings of familiarity [$F(4,255) = 3.05$, $p < .01$; $M^2 = 4.39$, $\eta^2 = .05$], but post hoc comparisons (Tukey HSD) determined this to occur only between Quintiles 1 and 4 ($p < .05$). No difference was found in children's ratings for familiarity (Cycowicz et al., 1997) as a function of VC. Bonin et al. (2003) presented a comparable trend, but again, this pattern was not statistically significant.

Figure 4 can be used to determine the extent to which this effect is a function of unfamiliar pictures being more complex. Although there is a trend toward complex images being less familiar, the statistical evidence is not clear.

R&P (2005) presented stronger evidence that complexity may be related to familiarity [$F(4,255) = 24.98$, $p < .01$; $M^2 = 36.33$, $\eta^2 = .28$], with Quintiles 1 and 5 being significantly different from all complexity quintiles (Tukey HSD). One reason why it is possible to detect this trend in the R&P data set is that each quintile contains a similar number of stimuli (between 50 and 55 pictures per quintile). Other data sets contain much greater variation across the perimeter quintiles, with Quintile 5 being most problematic (S&V, $n = 7$; Bonin et al., $n = 9$). This makes statistical analysis of this quintile set problematic. Moreover, given that R&P adopted the unusual procedure of asking their participants to score for both VC *and* familiarity, it is conceivable that the ratings of perceived familiarity used in this analysis could be confounded by judgments of VC.

Discussion

Automated measures of complexity. Four image-processing measures were used to examine the relationship between subjective visual complexity and the ability of a computer to closely approximate those judgments. Several metrics were applied to four published sets of standardized norms for pictures. The perimeter and Canny measures correlated moderately well with human judgments of complexity for all four data sets: S&V (1980), $r_s = .73$, $p < .0002$ (Perimeter) and $r_s = .68$, $p < .0002$ (Canny); Cycowicz et al. (1997), $r_s = .65$ and $r_s = .61$, $p < .0002$, respectively; Bonin et al. (2003), $r_s = .45$, $p < .0002$, and $r_s = .39$, $p < .0002$; and R&P (2005), $r = .57$, $p < .0002$, and $r = .60$, $p < .0002$ (line drawings).

Compression techniques also performed comparably well as measures of complexity (cf. Tables 2–4): for example, S&V (1980), $r_s = .72$, $p < .0002$ (JPEG), $r_s = .75$, $p < .0002$ (GIF); Cycowicz et al. (1997), $r_s = .65$, $p < .0002$ (JPEG and GIF); and Bonin et al. (2003), $r_s = .41$, $p < .0002$, and $r_s = .31$, $p < .0002$, respectively. Results are broadly comparable for the R&P (2005) data set. The correlations are slightly smaller, but the larger familiarity bias in the original data would account for the reduction in coefficient size.

Despite concerns that compression (particularly JPEG) is not suitable for the treatment of simple images, researchers have applied such measures in the standardization of such pictures (Bates et al., 2003; Vitevitch et al., 2004). The correlations reported here suggest that complexity could be reasonably approximated through a compression metric, but there are some subtle differences in the results, depending on which automated metric is applied. With these limitations in mind, compression can be used to make quick approximations of human judgments of VC.

Children's judgments of picture complexity. Children's judgments of complexity are broadly similar to those of adults, but the effect is smaller for children than for adults. This may partly be due to the nonstandard instructions used by Cycowicz et al. (1997), and it may also be explained by the fact that children's semantic networks are at an earlier stage of development (Wright & Wanley, 2003) and, as such, their judgments of complexity may be less influenced by their familiarity with a picture.

Unfamiliar pictures being more complex. Evidence that complex pictures tend to be perceived as less familiar is equivocal. Data sets that present a systematic relationship between complexity and familiarity are known to be flawed (e.g., R&P, 2005), and a fundamental assumption of the analysis reported here is that although human VC can be confounded by familiarity, familiarity is not confounded by VC. An analysis of less biased data sets (Bonin et al., 2003; Cycowicz et al., 1997; S&V, 1980) presents weaker evidence in support of the complexity–familiarity relationship, but very large differences in the number of shapes falling across the range from simple to complex images (S&V had just seven images falling into the simplest category; Bonin et al. had nine images) makes interpretation of the statistical analysis problematic. Further research is required, based on equivalent numbers of images falling across the range of visual complexity; such an analysis will help determine whether complex images are actually less familiar.

Table 4
Spearman Correlations: Rossion and Pourtois (2005)

	Complexity	Familiarity	Perimeter	Canny	JPEG
Line drawings					
Familiarity	−.50*	1.00			
Perimeter	.57*	−.22	1.00		
Canny	.60*	−.20	.94*	1.00	
JPEG	.59*	−.27	.90*	.89*	1.00
GIF	.65*	−.30	.92*	.93*	.93*
Gray drawings					
Familiarity	−.41*	1.00			
Perimeter	.52*	−.28	1.00		
Canny	.46*	−.08	.70*	1.00	
JPEG	.60*	−.27	.79*	.81*	1.00
GIF	.45*	−.14	.67*	.86*	.80*
Colorized drawings					
Familiarity	−.50*	1.00			
Perimeter	.47*	−.18	1.00		
Canny	.43*	−.07	.71*	1.00	
JPEG	.53*	−.26	.78*	.81*	1.00
GIF	.40*	−.11	.56*	.85*	.77*

*$p < .0002$.

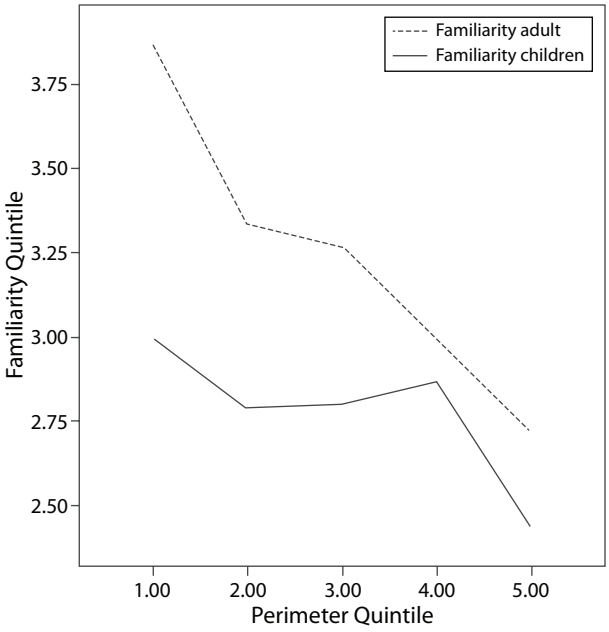


Figure 4. Familiarity judgments of adults (Snodgrass & Vanderwart, 1980) and children (Cycowicz, Friedman, & Rothstein, 1997) as a function of objective visual complexity (perimeter detection).

EXPERIMENT 2

Explaining the Complexity–Familiarity Bias

Attneave (1954, 1971) and Hochberg (1968) suspected that humans are not particularly good at making objective judgments of image complexity, and the data reported here support their suspicion. The strong negative correlations between VC and familiarity (cf. Tables 2–4) add to evidence that human observers cannot process the structure of an image independently of its familiarity. When observers are asked to consider the complexity of an image, they also process task-irrelevant information, such as its familiarity and meaningfulness (Boucart & Humphreys, 1992; Carmichael, Hogan, & Walter, 1932).

A stronger test of the familiarity bias would be to examine the responses to unfamiliar nonsense shapes. For these unfamiliar stimuli, there should be a higher correlation between perceived complexity and an objective metric, such as a compression technique. Moreover, training on a subset of these images should show that as familiarity increases, the biasing influence of familiarity on perceived complexity becomes stronger. Specifically, if there is a tendency for human raters to inflate the complexity ratings of familiar objects, raters who are familiar with a subset of nonsense shapes will rate those shapes as being less complex than will naive raters.

Method

The stimuli consisted of 100 nonsense shapes collected from several Internet data bases and used in discrimination studies—for example, Gauthier, James, and Curby (2003); Shatzman and McQueen (2006). Twenty of these shapes were geometric nonsense shapes. Geometric nonsense shapes have more regularity in their structure and are easier to learn and recall, and this warrants limiting their number (Appendixes A and B). Twenty geometric nonsense shapes and 80 random nonsense shapes were presented to 76 participants for rating. The participants were randomly placed in one of three groups.

Group 1 (n = 23). A subset of 22 nonsense shapes was selected from the larger corpus of 100 images. All the images were of random design, since it was considered that geometric shapes, being more regular, could be easier to learn. The selection of 22 shapes was based on their automated complexity scores, using perimeter detection. The shapes represented a relatively normal complexity distribution range.

The participants in Group 1 were asked to familiarize themselves with these shapes over 7 days. The participants did not receive specific instructions as to how they should become familiar with the shapes, but only that they should not spend any more than 5–10 min per day learning the shapes. Follow-up interviews confirmed that the participants had limited their time to only 5–6 min per day. The participants reported using strategies to memorize the shapes. For example, some tried to make links between the random shape and something in the real world, whereas others gave names to the shapes.

Group 2 (n = 32). The participants received training on 22 shapes that would not reappear later in testing.

After 7 days, Group 1 and Group 2 participants were presented with the entire corpus of 100 shapes. They were instructed to use a 10-point scale to indicate how complex they perceived the shape to be. Complexity was defined as the *amount of detail or intricacy* (S&V, 1980). A score of 1 was an *extremely complex shape*; a score of 10 was an *extremely simple shape*. The 10-point rating scale was used in this instance to permit greater differentiation between the shapes.

Group 3 (n = 21). There is a human tendency, termed *pareidolia*, to recognize shapes, see patterns, and establish order in otherwise vague and random stimuli—for example, seeing faces in clouds or the "man in the moon." Similarly, gestalt grouping processes can put order and stability into a random shape. Group 3 participants considered the extent to which shapes resembled something. They were asked to rate the shapes for "how like something" they were. A score of 1 indicated a shape that was "extremely like something"; a score of 10 represented a shape that was *extremely like nothing*.

Results

Perceiving complexity in familiar shapes. The means, standard deviations, kurtosis, and the skew for ratings for the trained and naive groups are shown in Table 5. Group 1 tended to rate the 22 images on which they were trained as less complex than did Group 2, the untrained naive group (as per inverted scoring, larger mean scores equate with less complexity). A two-way ANOVA with

Table 5
Summary Statistics per Group Over All Image Types

		Complexity		Skew		Kurtosis	
Image Type	Group	*M*	*SD*	*M*	*SE*	*M*	*SE*
Familiar (n = 22)	Trained	5.21	1.23	0.37	0.49	−0.89	0.95
	Naive	4.27	1.24	0.48	0.49	−0.38	0.94
Unfamiliar (n = 78)	Trained	4.86	1.77	0.04	0.27	−0.99	0.54
	Naive	4.99	1.70	0.05	0.27	−0.72	0.54

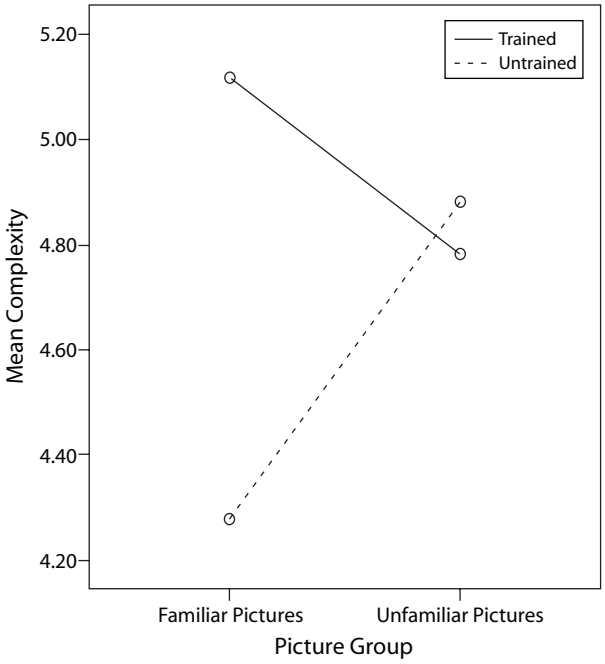


Figure 5. Interaction between training and familiarity on judgments of complexity. Larger scores equate with less complexity.

group and familiarity as factors detected a significant group × familiarity interaction [$F(1,52) = 17.53$, $p < .05$; $M^2 = 5.74$; see Figure 5].

Measuring complexity in unfamiliar shapes. Compression scores (GIF and JPEG) were obtained using the methods outlined in Experiment 1. The extent to which compression techniques and the perimeter detection measure are able to predict human judgments of perceived complexity was determined through correlational analysis (Table 6). Perimeter scores were standardized onto a 10-point scale using histogram normalization. This adjustment made no difference to the correlation coefficient. As such, the raw scores retaining all of the variance were used in the following analysis.

Bonferroni adjustment was computed on the basis of the 100 stimuli, with separate analyses for 80 nonsense shapes and 20 geometric nonsense shapes. This placed the significance level at .0006 for the nonsense shapes and .0002 for the geometric shapes.

Correlations between the automated measures and human judgments from the untrained group were compared (scores inverted for ease of understanding). Perimeter correlated moderately well ($r_s = .64$, $p < .0006$) with human judgments of complexity and with JPEG to a lesser degree ($r_s = .50$, $p < .0006$). The correlation between human judgments and GIF compression was not significant using the Bonferroni-adjusted criterion. There was no relationship between human judgments of complexity and the tendency for random pictures to be considered to "look like something" (Group 3 pareidolia scores), suggesting that the images were truly random. For the geometric shapes, the strongest correlation was between human judgments and the perimeter measure ($r_s = .76$, $p < .0002$).

Discussion

Familiarity bias. Training users even for a short time (1 week) on a set of nonsense shapes introduced a familiarity bias into ratings of subjective complexity. In practical terms, this means that asking observers to rate an image only for "detail and intricacy" or "complexity" is not possible, because they cannot prevent their familiarity with the content of the image from biasing their judgment. Computer-based measures could eliminate this confound, because they are unaffected by the familiarity of an image. For example, the influence of familiarity on perceived complexity sometimes led to extreme divergence in the complexity ratings obtained from humans and those produced by an objective metric. In Experiment 1, human judgments for the highly familiar picture *sun* (score = 4.9) attracted a very low VC rating (score = 1.20). The perimeter detection metric identified *sun* as considerably more complex than was judged by human observers. The difference in rank orders between the two measures of complexity are considerable: Humans put *sun* 7th out of 260,

Table 6
Spearman Correlations Between Nonsense Shapes and Subjective Complexity

	Complexity (Group 2)	Pareidolia (Group 3)	Perimeter	Canny	GIF
		Random			
Human complexity (Group 2)	1.00				
Pareidolia (Group 3)	.02	1.00			
Perimeter	.64*	.12	1.00		
Canny	.50*	.01	.76*	1.00	
GIF	.35	.18	.33*	.46*	1.00
JPEG	.50*	.05	.85*	.88*	.44*
		Geometric			
Human complexity (Group 2)	1.00				
Pareidolia (Group 3)	.17	1.00			
Perimeter	.76*	.14	1.00		
Canny	.55	.08	.79*	1.00	
GIF	.23	.03	.04	.05	1.00
JPEG	.66	.00	.89*	.80*	.12

*Random shapes, $p < .0006$; geometric shapes, $p < .0002$.

whereas perimeter detection placed *sun* 131st out of 260. Conversely, the complexity of the somewhat less familiar picture *flute* was ranked 232nd out of 260 by humans, but only 13th out of 260 by the automated measure.

Choosing between metrics. The Bonferroni adjustment was developed to aid decision making, and not to assess evidence in the data. This makes its application controversial, and there seems to be little consensus among statisticians regarding its use (Perneger, 1998). It is also important to consider that just because a finding achieves significance against a Bonferroni-adjusted criterion does not mean that it is "more significant" (Cohen, 1990, 1994). That being said, its application here supports the logic that when repeated decisions are made over many trials, error rates will be reduced (Neyman & Pearson, 1928).

The perimeter measure correlated moderately strongly with subjective complexity on both sets of pictures (random shapes, $r_s = .64, p < .0006$; geometric shapes, $r_s = .76, p < .0002$). Canny correlated moderately well with perceived complexity on the random shapes ($r_s = .50$, $p < .0006$), and JPEG also presented moderately strong correlations; however, GIF compression failed to correlate strongly with perceived complexity on any of the picture sets.

The Canny edge detection measure is better suited to the detection of fine or blurred edges in a picture; this perhaps explains why the correlations tended to be lower. GIF compression, however, is intended for use with pictures with sharp transitions and should have produced a stronger association with human judgments. One explanation for the reduction in correlations with both the JPEG and the GIF compression measures is that the randomness of these objects precluded a reduction in the number of bits required to store the object. Most of these objects would be rated as complex by a compression system because there would have been fewer commonly occurring sequences of pixels that could be replaced with shorter codes. Perimeter, however, simply measures the existing object; it does not remove or add information. Given that this was a large corpus of stimuli ($n = 100$), it seems that perimeter is a more robust and stable measure of complexity (Zhang & Lu, 2004). There is also clarity in relation to the ways in which the image is measured: Coding rules are explicit, and there are no issues in relation to the addition of artifacts or the removal of pieces of information. Although the perimeter measure is unable to capture an object in the way a human does, it can determine an edge in much the same way as a human can. These edges combine to form small shapes and add detail, and these factors perhaps contribute to human judgments of complexity.

GENERAL DISCUSSION

Several studies have reported a significant inverse correlation between human judgments of familiarity and complexity (Alario & Ferrand, 1999; Bonin et al., 2003; Cycowicz et al., 1997; S&V, 1980), and others have failed to document its existence (R&P, 2005). None have considered the important implication that the reported norms for picture complexity are systematically flawed by the presence of an underlying familiarity interference effect. A valid measure of image complexity would be one in which ratings of complexity are unaffected by judgments of familiarity.

There have been several attempts to develop valid and reliable measures of image complexity. Attneave (1954) and Hochberg and Brooks (1960) acknowledged that shape is a multidimensional variable that varies with the complexity of an image and that relying solely on human judgments means that there is no way of predicting how complex an image might be judged to be. Forsythe et al. (2003b) argued that an objective measure of visual primitives (i.e., edges) may provide a valid index of complexity for all 2-D stimuli. Their perimeter measure is based on an approach originally articulated by Attneave (1954, 1971) and has overcome the practical difficulties that thwarted its earlier adoption. Basic perceptual components (i.e., edges) are important in the measurement of complexity because edges combine to form small shapes and add detail. These are perhaps what lead an image to be perceived as complex (Beck et al., 1991; Harwerth & Levi, 1978; Sutter et al., 1989; Vassilev & Mitov, 1976). The perimeter detection metric locates edges by examining sudden changes in intensity that occur at image boundaries and then counts the number of such changes.

These results point to a measure of complexity for 2-D stimuli that is more valid because it is unaffected by judgments of familiarity. It is suggested that when researchers are seeking to select images on the basis of their complexity, they should treat the perimeter criterion as superior to human judgments.

Although correlations between nonsense shapes and the perimeter detection measure were larger in Experiment 2, the earlier analysis (Experiment 1) demonstrated that GIF (lossless) and JPEG (lossy) compression measures are able to approximate human judgments of complexity, particularly in colorized or grayscale pictures. Information theory (Shannon & Weaver, 1949) is a useful framework through which to further evaluate the effectiveness of compression. Difficulties are, however, likely to arise concerning the importance of the data (or information) that are "thrown away" and the addition of erroneous artifacts. JPEG operates in such a way that the file size is reduced and disk space and transmission time are reduced; it was never intended as a measure of visual complexity. The extent to which humans and compression techniques manage information in a similar way requires further exploration. Any shape representation and description technique should have clarity, be stable, and determine shapes in much the same way as a human observer (Zhang & Lu, 2004). Notwithstanding problems relating to image artifacts and the exactness of the compression scores awarded, availability and usability may make compression an efficient choice for complexity measurement in colorized and grayscale pictures.

Further Examination of the Effect of Familiarity on Complexity

Automated metrics provide a measure of complexity that is unaffected by the familiarity of the content of a picture. Their success in predicting how human observers will judge the complexity of an image is mitigated by the fact that image familiarity is not taken into account.

As has been shown by Bartram (1973), even if researchers perfect a metric that will generate 2-D complex and simple shapes that represent no meaningful stimuli, observers will overcome any complexity effects—such as a reduction in response latency—as familiarity increases.

These predictable changes have been demonstrated throughout the literature and seem to be related largely to redundant information content. For example, when Garner (1970) asked participants to retrace from memory geometric shapes that had gaps in their structure, the spaces would be omitted. Garner suggested that these gaps were not needed to recover the shape information; this detail was, effectively, redundant information. Similarly, when relevant detail is exaggerated in a picture, it becomes even larger on recall (Donderi, 1973). As was argued by Fodor and Pylyshyn (1981), "What you see when you see a thing depends on what the thing you see is. But what you see the thing as depends on what you know about what you are seeing" (p. 189).

Familiarity can help reduce the amount of information required to communicate a message, because small redundancies can be overlooked. When something is less likely, it will require more pieces of information to determine its meaning, and the overlooking of small pieces of information becomes less desirable. This possibly explains the drop in correction coefficients for the nonsense objects for the GIF and JPEG compression techniques. Compression techniques operate to remove as much information as possible, whereas small visual elements become highly valuable to the human viewer because they enable discrimination between subtle degrees of complexity.

Conclusions

Human judgments of complexity are influenced by familiarity; since this reflects the reality of the user, familiarity with a complex object is something that researchers should take into consideration before testing for complexity effects. The Forsythe et al. (2003b) perimeter detection metric is a measurement of complexity based on the extent to which a picture has edges. This measure correlated moderately well with human judgments of complexity for four standardized sets of picture ratings (Bonin et al., 2003; Cycowicz et al., 1997; R&P, 2005; S&V, 1980). Published ratings of complexity correlate significantly with judgments of familiarity, whereas the perimeter detection metric does not. Humans are influenced proportionately by the degree to which they have previously encountered a picture, with familiar nonsense shapes receiving judgments of complexity that are lower than those indicated by the objective metric. The Forsythe et al. (2003b) perimeter detection metric is a measure of the level of complexity for 2-D stimuli that is unaffected by task-irrelevant information, such as the meaningfulness or familiarity of a picture; the metric is a pure measure of geometric primitives (edges). Compression techniques also present a good approximation of subjective image complexity; however, the ways in which the techniques operate on particular sets of pictures require further exploration. Although there is a trend for visually complex pictures to be rated as less familiar, further research is required to determine whether visually complex pictures are actually more unfamiliar.

AUTHOR NOTE

We thank Don Donderi and the two reviewers, Sine McDougall and Patrick Bonin, for their comments on an earlier version of the manuscript. Correspondence concerning this article should be addressed to A. Forsythe, School of Psychology, John Moores University, Room 451, Webster Street, L3 2ET, England (e-mail: a.m.forsythe@ljmu.ac.uk).

REFERENCES

ALARIO, F.-X., & FERRAND, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, **31**, 531-552.

ATTNEAVE, F. (1954). Some informational aspects of visual perception. *Psychological Review*, **61**, 183-193.

ATTNEAVE, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results*. New York: Holt, Rinehart & Winston.

ATTNEAVE, F. (1971). Multistability in perception. *Scientific American*, **225**(6), 62-71.

ATTNEAVE, F., & ARNOULT, M. D. (1956). The quantitative study of shape and pattern perception. *Psychological Bulletin*, **53**, 452-471.

BARRY, C., MORRISON, C. M., & ELLIS, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *Quarterly Journal of Experimental Psychology*, **50A**, 560-585.

BARTRAM, D. J. (1973). The effects of familiarity and practice on naming pictures of objects. *Memory & Cognition*, **1**, 101-105.

BATES, E., D'AMICO, S., JACOBSEN, T., SZÉKELY, A., ANDONAVA, E., DEVESCOVI, A., ET AL. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, **10**, 344-380.

BECK, H., GRAHAM, N., & SUTTER, A. (1991). Lightness differences and the perceived segregation of regions and populations. *Perception & Psychophysics*, **49**, 257-269.

BIEDERMAN, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, **94**, 115-147.

BONIN, P., PEEREMAN, R., MALARDIER, N., MÉOT, A., & CHALARD, M. (2003). A new set of 299 pictures for psycholinguisitic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, **35**, 158-167.

BOUCART, M., & HUMPHREYS, G. W. (1992). Global shape cannot be attended without object identification. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 785-806.

CARMICHAEL, L., HOGAN, H. P., & WALTER, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived forms. *Journal of Experimental Psychology*, **15**, 73-86.

CHIPMAN, S. F. (1977). Complexity in visual structure. *Journal of Experimental Psychology: General*, **106**, 269-301.

COHEN, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.

COHEN, J. (1994). The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

CYCOWICZ, Y. M., FRIEDMAN, D., & ROTHSTEIN, M. (1997). Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology*, **65**, 171-237.

DELL'ACQUA, R., LOTTO, L., & JOB, R. (2000). Naming times and standardized norms for the Italian PD/DPSS set of 266 pictures: Direct comparisons with American, English, French, and Spanish published databases. *Behavior Research Methods, Instruments, & Computers*, **32**, 588-615.

DONDERI, D. (1973). Changes in visual recall memory following discrimination learning. *Canadian Journal of Psychology*, **27**, 210-219.

DONDERI, D. (2006a). An information theory analysis of visual complexity and dissimilarity. *Perception*, **35**, 823-835.

DONDERI, D. (2006b). Visual complexity: A review. *Psychological Bulletin*, **132**, 73-97.

DONDERI, D. C., & MCFADDEN, S. (2003). A single marine overlay is more efficient than separate chart and radar displays. *Displays*, **24**, 147-155.

FODOR, J., & PYLYSHYN, Z. (1981). How direct is visual perception? Some reflections on Gibsons "ecological approach." *Cognition*, **9**, 139-196.

FORSYTHE, A., SHEEHY, N., & SAWEY, M. (2003a). The automated measurement of pictorial image complexity: A feasibility study. In D. Harris, V. Duffy, M. Smith, & C. Shephanidis (Eds.), *Human-centred computing: Cognitive, social and ergonomic aspects* (Vol. 3, pp. 205-209). Mahwah, NJ: Erlbaum.

FORSYTHE, A., SHEEHY, N., & SAWEY, M. (2003b). Measuring icon complexity: An automated analysis. *Behavior Research Methods, Instruments, & Computer*s, **35**, 334-342.

GARCIA, M., BADRE, A. N., & STASKO, J. T. (1994). Development and validation of pictorial images varying in their abstractness. *Interacting With Computers*, **6**, 191-211.

GARNER, W. R. (1970). Good patterns have few alternatives. *American Scientist*, **58**, 34-42.

GAUTHIER, I., JAMES, T. W., & CURBY, K. M. (2003). The influence on conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, **20**, 507-523.

GEISELMAN, R. E., LANDEE, B. M., & CHRISTEN, F. G. (1982). Perceptual discriminability as a basis for selecting graphic symbols. *Human Factors*, **24**, 329-337.

HARWERTH, R. S., & LEVI, D. M. (1978). Reaction time as a measure of suprathreshold grating detection. *Vision Research*, **18**, 1579-1586.

HOCHBERG, J. E. (1968). *Perception* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

HOCHBERG, J. E., & BROOKS, V. (1960). The psychophysics of form: Reversible perspective drawings of spatial objects. *American Journal of Psychology*, **73**, 337-354.

HOCHBERG, J. E., & MCALISTER, E. (1953). A quantitative approach to figural "goodness." *Journal of Experimental Psychology*, **46**, 361-364.

HOEGER, R. (1997). Speed of processing and stimulus complexity in low-frequency and high-frequency channels. *Perception*, **26**, 1039-1045.

HORTON, W. (1994). *The icon book: Visual symbols for computer systems and documentation*. New York: Wiley.

JOHNSON, C. J., PAIVIO, A., & CLARK, J. M. (1996). Cognitive components of picture naming. *Psychological Bulletin*, **120**, 113-139.

MARR, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

MATHWORKS (2001). *Image Processing Toolbox user's guide.* Boca Raton, FL: CRC Press.

MCDOUGALL, S. J. P., CURRY, M. B., & DE BRUIJN, O. (1999). Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers*, **31**, 487-519.

MCDOUGALL, S. J. P., DE BRUIJN, O., & CURRY, M. B. (2000). Exploring the effects of icon characteristics on user performance: The role of icon concreteness, complexity, and distinctiveness. *Journal of Experimental Psychology: Applied*, **6**, 291-306.

NEYMAN, J., & PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20A**, 175-240, 263-297.

NIELSEN, J. (1993). Noncommand user interfaces. *Communications of the ACM*, **36**, 83-99.

PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, **76**(1, Part 2), 1-25.

PARKER, D. M., LISHMAN, J. R., & HUGHES, J. (1997). Integration of spatial information in human vision is temporally anisotropic: Evidence from a spatiotemporal discrimination task. *Perception*, **26**, 1169-1180.

PERNEGER, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal*, **216**, 1236-1238.

PIND, J., JONSDOTTIR, H., TRGGVADOTTIR, H. B., & JONSSON, F. (2000). Icelandic norms for the Snodgrass and Vanderwart (1980) pictures: Name and image agreement, familiarity, and age of acquisition. *Scandinavian Journal of Psychology*, **41**, 41-48.

PROCTOR, R. W., & VU, K.-P. L. (1999). Index of norms and ratings published in the Psychonomic Society journals. *Behavior Research Methods, Instruments, & Computers*, **31**, 659-667.

ROSSION, B., & POURTOIS, G. (2005). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface information in basic-level object recognition. *Perception*, **33**, 217-236.

SHANNON, C. E., & WEAVER, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

SHATZMAN, K. B., & MCQUEEN, J. M. (2006). Prosodic knowledge affects the recognition of newly acquired words. *Psychological Science*, **17**, 372-377.

SNODGRASS, J. G., & VANDERWART, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 174-215.

STAMPS, A. E., III (2000). *Psychology and the aesthetics of the built environment*. Boston: Kluwer.

SUTTER, A., BECK, J., & GRAHAM, N. (1989). Contrast and spatial variables in texture segregation: Testing a simple spatial-frequency channels model. *Perception & Psychophysics*, **46**, 312-332.

TAUBMAN, D., & MARCELLIN, M. (2001). *JPEG2000: Image compression fundamentals, standards and practice.* London: Kluwer.

TREISMAN, A. (1986). Features and objects in visual processing. *Scientific American*, **255**(5), 114-125.

TREISMAN, A., & GELADE, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, **12**, 97-136.

TREISMAN, A., & SOUTHER, J. (1985). Search asymmetry: A diagnostic preattentive processing of separable features. *Journal of Experimental Psychology: General*, **114**, 285-310.

VASSILEV, A., & MITOV, D. (1976). Perception time and spatial frequency. *Vision Research*, **16**, 89-92.

VITEVITCH, M. S., ARMBRÜSTER, J., & CHU, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 514-529.

VITKOVITCH, M., & TYRRELL, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology*, **48A**, 822-848.

WRIGHT, B. C., & WANLEY, A. (2003). Adults' versus children's performance on the Stroop task: Interference and facilitation. *British Journal of Psychology*, **94**, 475-485.

ZHANG, D., & LU, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, **37**, 1-19.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, www.psychonomic.org/archive.

To access these files, search the archive for this article using the journal name (*Behavior Research Methods*), the first author's name (Forsythe), and the publication year (2008).

FILE: Forsythe-BRM-2008.zip

DESCRIPTION: The compressed archive file contains 6 files:

SVC_picture_ratings.xls, Rank of picture ratings taken from Snodgrass and Vanderwart (1980), and Cycowicz et al. (1997);

RP_picture_ratings.xls, Rank of picture ratings taken from Rossion and Pourtois (2005);

EBonin_picture_ratings.xls, Rank of picture ratings taken from Bonin, Peerman, Malardier, Méot, & Chalard, (2003);

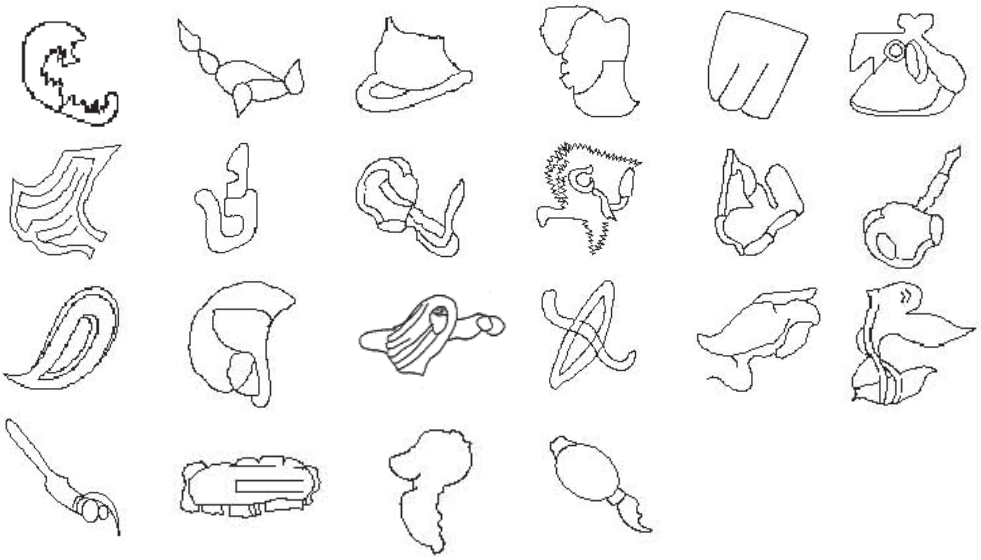
SVC_picture_ratings.txt, a tab-delimited text version of the first Excel file;

RP_picture_ratings.txt, a tab-delimited text version of the second Excel file;

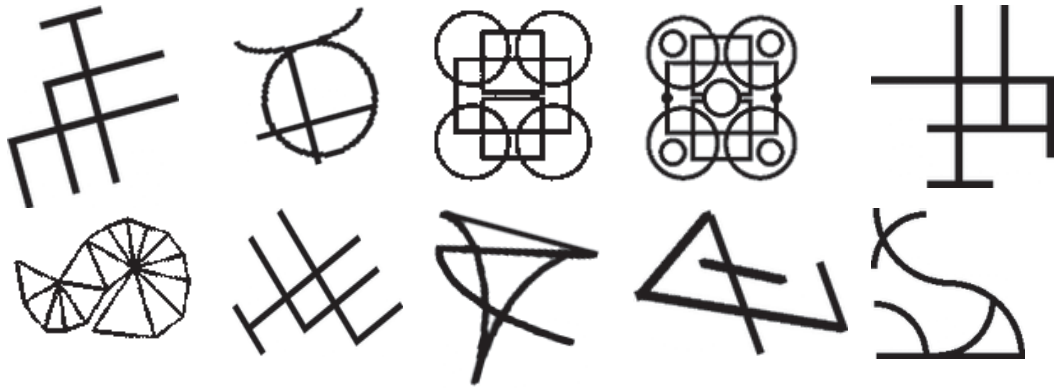
EBonin_picture_ratings.txt, a tab-delimited text version of the third Excel file.

AUTHOR'S E-MAIL ADDRESS: a.m.forsythe@ljmu.ac.uk.

APPENDIX A
Examples of Nonsense Shapes (Shatzman & McQueen, 2006)



APPENDIX B
Examples of Geometric Nonsense Shapes (Gauthier, James, & Curby, 2003)



(Manuscript received October 16, 2006;
revision accepted for publication March 28, 2007.)