Spatial Vision, Vol. 16, No. 5, pp. 429–457 (2003)

Also available online - www.vsppub.com

Psychophysical study of image orientation perception

JIEBO LUO*, DAVID CRANDALL, AMIT SINGHAL, MATTHEW BOUTELL and ROBERT T. GRAY

Electronic Imaging Products, R&D, Eastman Kodak Company, 343 State Street, Rochester, NY 14650-1816, USA

Received 21 December 2002; revised 4 February 2003; accepted 2 June 2003

Abstract—The experiment reported here investigates the perception of orientation of color photographic images. A collection of 1000 images (mix of professional photos and consumer snapshots) was used in this study. Each image was examined by at least five observers and shown at varying resolutions. At each resolution, observers were asked to indicate the image orientation, the level of confidence, and the cues they used to make the decision. The results show that for typical images, accuracy is close to 98% when using all available semantic cues from high-resolution images, and 84% when using only low-level vision features and coarse semantics from thumbnails. The accuracy by human observers suggests an upper bound for the performance of an automatic system. In addition, the use of a large, carefully chosen image set that spans the 'photo space' (in terms of occasions and subject matter) and extensive interaction with the human observers reveals cues used by humans at various image resolutions: sky and people are the most useful and reliable among a number of important semantic cues.

Keywords: Image orientation; human observer; semantic cues; low-level cues; photo space; image resolution.

1. MOTIVATIONS

The rapid growth of digital imaging has led to an increase in image-related tasks such as enhancement, manipulation, compression, understanding, organization, and retrieval. Knowledge of the correct image orientation can be of great importance for these tasks. Automatic image orientation can drastically reduce the human effort otherwise needed to orient the images for viewing (either on a computer monitor, a handheld device, or a TV) or for organizing an album. In addition, many automatic algorithms for object recognition, scene classification, and content-based image retrieval either require *a priori* knowledge of the correct image orientation, or can perform significantly better if image orientation is known. For example,

*To whom correspondence should be addressed. E-mail: jiebo.luo@kodak.com

face detection algorithms (Schneiderman and Kanade, 1998) usually assume the image is in an upright orientation. Otherwise, all four possible image orientations have to be examined, increasing the computation time and false positive detection rate. Most sky detection algorithms are designed to take advantage of the fact that sky often appears as a blue region at the top of an image, with the exception of the clear blue sky detection method by Luo and Etz (2002). Semantic features are becoming increasingly important for content-based image retrieval and annotation (Saber *et al.*, 1996; Naphade and Huang, 2000; Smeulders *et al.*, 2000). For classification of images into indoor-outdoor (Szummer and Picard, 1998), sunset, beach, field, fall foliage, mountain, and urban scenes (Vailaya *et al.*, 1998), images are assumed to be in the upright orientation so that scene layout of prototypical scenes can be learned through training.

1.1. Image orientation in computer vision literature

Automatic image orientation detection is a relatively new research area in computer vision. Most of the early work focused on documents, and success was largely due to the constrained nature of the problem (text cues). For natural images, the problem is considerably more challenging. Until recently (Vailaya *et al.*, 1999; Wang and Zhang, 2001), there had been little work on automatic image orientation detection for natural images. Humans appear to use scene context and semantic object recognition to identify the correct image orientation. However, it is difficult for a computer to perform the task in this way because current object recognition algorithms are extremely limited in their scope and robustness. Out of millions of possible objects that can appear in a natural scene, robust algorithms exist for only a handful of objects (e.g. face, sky). To date, scene classification is often approached by computing low-level features (e.g. color, texture, and edges) that are processed with a learning engine to directly infer high-level information about the image (Szummer and Picard, 1998; Vailaya *et al.*, 1999; Wang and Zhang, 2001). Recently, a new approach was proposed that combines low-level features with detectable semantic scene content in order to improve the accuracy of indoor-outdoor image classification (Luo and Savakis, 2001).

1.2. Object orientation in psychology literature

While a small portion of the psychology literature involves human perception of orientation of gratings (Dakin *et al.*, 1999; Mareschal *et al.*, 2001), and recognition of rotated letters and digits (Corballis *et al.*, 1978; Jolicoeur, 1992), most of the psychology literature involving orientation focuses on the interplay between orientation and recognition of objects, specifically the effect of in-plane rotation on the recognition of single objects represented by line drawings. Although this literature is vast (e.g. Corballis *et al.*, 1978; Braine *et al.*, 1981; Maki, 1986; Biederman, 1987; Tarr and Bulthoff, 1995 and 1998; Hamm and McMullen, 1998; DeCaro, 1998; Jolicoeur *et al.*, 1998; McKone and Grenfell, 1999; DeCaro and

Table 1.
Relationship between the prior literature and the present study

Percept	Task	
	Recognize object	Infer orientation
Letters/digits	Jolicoeur, Corballis	Corballis
Line drawings	Jolicoeur, Tarr, Maki, . . .	DeCaro
Full-cue color	Nicholson	Our work

Reeves, 2000; DeCaro and Reeves, 2002; Lawson and Jolicoeur, 2003), there is no general consensus in the means by which rotation affects object recognition.

The research consistently shows that humans show a greater response time when recognizing rotated images and that this time is reduced with practice, as in later stages (Jolicoeur, 1985). However, in the only study using full-cue color images (i.e. photographs containing not only shape but also color, texture and shading), Nicholson and Humphrey (2001) found this effect to be negligible.

In contrast to the great extent of literature on object recognition, few studies have involved subjects inferring 'object orientation'. Corballis *et al.*'s (1978) experiments involved rotated letters, while De Caro's (1998) involved rotated line drawings of objects. Both gave evidence that object recognition precedes orientation detection. We attempt to extend this work to unconstrained photographic images. Drawing an analogy with the surprise effects witnessed by Nicholson and Humphrey (2001) because of the additional cues available, we believe that our work is not a trivial extension of Corballis *et al*. or of De Caro, but is novel. We summarize the relationship between our work and the prior research in Table 1.

We are also interested in how human orientation of photographs is affected by the semantic content of the image. While this problem has not been addressed in the literature, the effects of semantic categorization and rotation upon recognition speed of objects in line drawings has been studied. Vannucci and Viggiano (2000) found that recognition speed of placed objects (e.g. desks) and animals, usually seen in one, fixed orientation, does depend on the orientation of the image. However, recognition speed of vegetables and unplaced objects (e.g. tools), which are routinely seen in all orientations, is independent of orientation.

Other research has focused on the effects of orientation on naming objects at various category levels, such as basic (e.g. dog), subordinate (e.g. poodle), and superordinate (e.g. animal). Hamm and McMullen (1998) distinguished object recognition at the basic level from that at the subordinate level, claiming that only recognition at the subordinate level is orientation-invariant. Lloyd-Jones and Luckhurst (2002) and Lawson and Jolicoeur (2003) rejected this due to the simplicity of the experimental task in Hamm and McMullen (1998), and demonstrated orientation effects even at the basic level.

This paper presents a psychophysical study on the perception of the orientation of color photographic images. We emphasize that we have a different motivation than

most psychophysical studies in general and those involving image orientation in particular. Our interest is primarily in 'what' humans do with the visual information presented to them, for the purpose of recognizing the correct image orientation, instead of 'how' humans process such information.

Specifically, the study is designed to answer a number of questions. First, some natural images are extremely difficult even for humans to orient correctly, or may not even have a 'correct' orientation. Assuming that humans have almost unlimited recognition power for the types of objects found in photographs compared to the current computer vision systems, this study provides an upper bound for the performance of an automatic system on a set of images reasonably representative of the 'photo space' in terms of occasions and subject matters. Second, discrepant detection rates based on purely low-level cues have been reported in the literature, ranging from exceptionally high (∼95%) in earlier work (Vailaya *et al.*, 1999) to more reasonable (∼78%) in recent work with a higher degree of sophistication (Wang and Zhang, 2001). The image databases used for the two studies are different and we suspect that the high accuracy numbers reported in the earlier work might be an artifact of the database used in that experiment. In other words, if most images fit into some prototypes, such as 'sky on top of grass', a low-level feature-based approach is expected to do well. This study allows us to put the reported results in the correct perspective. Finally, the use of a large, carefully chosen image set that spans the 'photo space' and extensive interaction with the human observers should reveal the various cues used by humans at various image resolutions. These can be used to design a robust orientation detection algorithm (Luo and Boutell, 2003).

In this study, images were shown at varying resolutions. On one hand, object recognition is expected to be much harder (and impossible for some images) at the lowest resolution and more likely as the resolution is increased. On the other hand, we believe that once the image resolution reaches a certain level, higher resolutions will not yield any additional benefits for a human to determine image orientation. At each resolution, observers were asked to indicate the image orientation, the level of confidence, and the cues they used to make the decision. Cues were selected from predefined low-level and semantic choices in a menu, or typed in if not in the list. Observers were also asked to make a general statement on whether they used the main subject, the background, the entire scene, or a unique object (e.g. labels on a cereal box) in making the decision. Observers might also comment about the scene or their decision process.

2. EXPERIMENT

As stated above, our motivation was to find out what humans can do with the visual information provided to them to determine image orientation without putting any limit on how they actually do that. Therefore, we designed each aspect of our experimental to enable, rather than control, our human observers.

2.1. Participants

Twenty-six observers (twenty males and six females) participated in the experiment. Most of them were imaging scientists and technicians. Twenty-four of the observers were aged 20–45 years and two were aged 45–60 years. All were regular users of computers and had normal or corrected-to-normal vision, although this was not as critical because our study allowed the observers to make adjustments for optimal viewing and therefore did not require high acuity. No observers were rejected due to outlier behaviors.

2.2. Stimuli and equipment

Image selection for any study is a non-trivial task. First, we need a sufficient number of images to draw statistically significant conclusions. Second, we need to have a representative set of images in terms of scene content because certain types of scenes are easy to recognize (e.g. outdoor, sky over an open field) and not much can be learned from them. However, if most of the images are difficult (e.g. flowers) or do not have a preferred orientation (e.g. texture patterns) the study would be skewed as well. Third, because each observer is asked to determine image orientation at multiple resolution levels for multiple images, the amount of labor limits the number of images that can be shown to each observer.

We used a total of 1000 images in order to have reasonably good coverage of the 'photo space'. It is also desirable to have a balance between professional stock photos and amateur consumer photos. External research has concentrated on stock photos, e.g. the Corel collection, which were taken with attention to image composition and exposure. Such photos are more likely to fall into the prototypes for which a learning engine can be effective. However, the validity of results based on such data can be questionable when applied to general digital imaging applications. Therefore, we decided to use 500 images from the Corel collection and another 500 from a consumer database called JBJL. The Corel collection has over 100 000 images in various categories such as sunset, coast, field, city, animal, people, textures and patterns, etc. JBJL has 1870 images organized according to the four seasons. Examples of the images can be found in Appendix B.

Table 2 shows a detailed breakdown of the images used in this study. Conscientious effort was made to select a mix of easy and difficult images to cover the most likely picture occasions and locations, including indoor and outdoor pictures and pictures with and without people. Extremely challenging images (e.g. fireworks, underwater, specifically collected textures and patterns, etc.) were avoided. A few texture patterns came naturally from the random sampling of pictures in the specified categories.

The photographic images were presented in the typical 3 : 2 aspect ratio of photographs (uncropped — see Note 1) on test displays comprised of various calibrated and uncalibrated computer monitors attached to various PC and Sun workstations.

Table 2.
Image data

Corel (500)	Easy (150)	Fields (30)
		Mountains (30)
		Alaska/California Coast (30)
		National Parks (30)
		Dawn/Dusk (30)
	Hard (150)	New York/Los Angeles (30)
		Interior (30)
		Christmas (30)
		Japan/Rural Africa (30)
		People/Indigenous People (30)
	Other (200)	Caribbean
		Autumn
		Automobile
		Barns/Farms
		Ski
		Garden
		Fishing
		Animals (horses/dogs/cats)
		Lakes/Rivers
		Winter
		City Signs
JBJL (500)		Spring
		Summer
		Fall
		Winter

Consistent with our motivation, we did not control viewing conditions, such as viewing distance, screen brightness, color gamut, and room lighting.

2.3. GUI design

A tool with a graphical user interface (GUI) was used to conduct the experiment. The tool presented the images to the observer in a sequence of increasing resolution. The observer was instructed to take a 'best' guess early on and not to wait until he was 100% sure. The GUI provided menus and text boxes that were used by observers to indicate orientation decisions, confidences, cues, and comments. The GUI allowed the user to quit and restart at any time. The remainder of this section describes the GUI in more detail.

2.4. Procedure

Because we were interested in what observers can do with visual information, the procedure allowed them to manipulate the picture displayed on the screen in a fashion virtually like holding a photograph in their hands to help them optimally

utilize the information embedded in the pixels presented to them. The environment was typical room lighting by normal in-the-ceiling fluorescent bulbs in an office without windows, although a few observers may have dimmed the lights somewhat to reduce glare. All the images were pre-rendered to suit the color gamut of a typical computer monitor (with gamma between 2.0 and 2.5). They were able to adjust the brightness of an image on the monitor to see the details in the shadows and highlights (we noted that brightness adjustment is not possible when holding an actual photograph, but only 1.5% of the observations in this study were made when the default brightness setting was changed). They were able to rotate an image freely without having to keep track of the rotation. They were also allowed to adjust the viewing distance freely in front of the monitor (typically between 8 to 20 inches) or even take the eyeglasses off. In addition, observers were able to see zoomed/enlarged versions of an image. However, to streamline the workflow, observers could not zoom an image unless there was a need for it, which was indicated naturally by the confidence level. In other words, if the confidence was lower than the maximum, the observer was presented with a higher resolution version of the image (unless the maximum resolution was reached). Several menus and text fields were provided for observers to record their decisions, including:

Cue #1. This menu was used for recording the use of low-level cues. The pull-down menu included none, color, texture, lines, and other. This was expected to be useful at lower resolution levels.

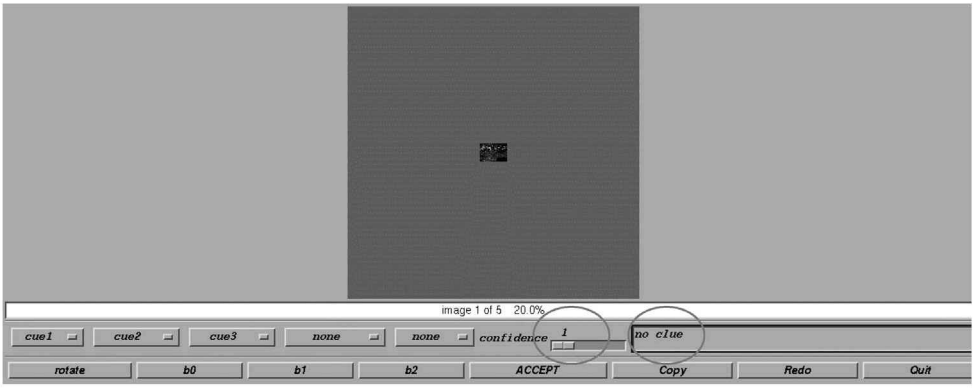
Cue #2. This menu was used for recording the use of semantic-level cues. The pull-down menu included none, face, people, animal, car, sky, cloud, grass, tree, flower, snow, water, road, ground, window, ceiling, furniture, building, bridge, mountain, text, and other. This was expected to be useful at higher resolution levels.

Cue #3. This menu could be used to specify the use of additional cues, either low-level or semantic-level. The pull-down menu included none, color, texture, lines, face, people, animal, car, sky, cloud, grass, tree, flower, snow, water, road, ground, window, ceiling, furniture, building, bridge, mountain, text, and other.

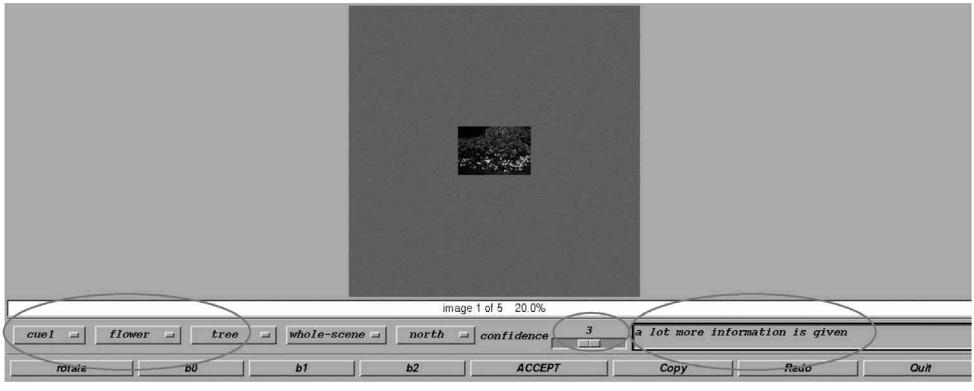
Cue Type. We were interested in knowing whether an observer uses cues from primarily the main subject, the background, or the whole scene. Further, image orientation can sometimes be determined solely from a unique object. The pull-down menu featured these four choices.

Orientation. Image orientation was defined as which side (north, east, west, south, unknown, don't care) of the image was upright relative to the currently displayed image.

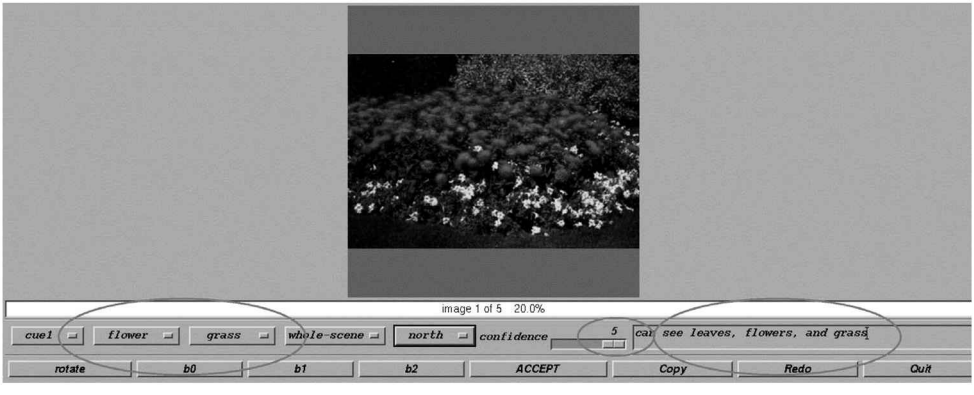
Confidence. Confidence (1: no clue — 5: absolutely sure) was considered very important because we could gauge how the observer felt about the task and their decisions. It was also a control signal for streamlining the workflow.



(a) Screen of the First Resolution — after user input



(b) Screen of the Third Resolution — after user input



(c) Screen of the Fifth (final) Resolution — after user input

Figure 1. Sample screenshots of test application at various stages of user interaction.

A typical observer workflow is as follows. First, the GUI presents the observer with an image at the lowest resolution (24×36), as shown in Fig. 1a. The observer examines the image, selects confidence, orientation, and cue choices from the drop-down menus, optionally enters a comment, and presses the 'accept' button. If the observer chooses a confidence value less than the maximum (5), the GUI clears the previous selections and presents the second resolution level (64×96). The second resolution is a simple interpolation of the lowest resolution, i.e. no new information is provided. We are interested in seeing if this version of the image can be of value. The observer again selects cues from the pull-down menus and presses the 'accept' button. If the confidence level is less than 5, the third resolution is then presented (128×192, Fig. 1b). This image contains a significant amount of new information, and we expect that most observers can determine the orientation of most images at a high confidence ($\geqslant 3$) at this resolution. The GUI continues to display larger images (256×384, 512×768) in sequence until either the fifth (maximum) resolution level is reached or the user indicates a confidence value of 5.

Note that it was possible that the confidence level did not reach the highest level for some challenging images even at the fifth resolution level. This information was recorded but an observer was not prompted to see a higher resolution. All of the inputs and interactions by each user on each image were stored in a log file.

The set of 1000 images was randomly partitioned into 5 non-overlapping sets of 200 images. Each observer was presented with only one of these sets so the task was more manageable. We later found that, although randomly partitioned, some sets were indeed more challenging than others because all the observers on such sets turned in lower scores. The session was always preceded by a block of 10–20 practice trials using images not used elsewhere in the study.

3. ANALYSIS

3.1. Observer confidence

Intuitively, confidence should increase monotonically with increasing image resolution. The average confidence level at each image resolution was calculated across all observers and all images. Because observers may stop at a lower resolution level for a particular image once their confidence level reaches the maximum value, we assigned the maximum confidence value to the remaining higher resolution levels, if any, even though they were never shown to the observer; we refer to the resulting distribution as the 'corrected' one (*vs.* the raw, 'uncorrected' distribution obtained if one does not perform this assignment). Our assumption is that if an observer gave a confidence of 5, he or she was absolutely sure of the orientation, and would make the same decision if presented with a higher resolution image. In the following analysis, only the corrected values will be presented unless the raw values are also informative and explicitly stated. The distributions of aver-

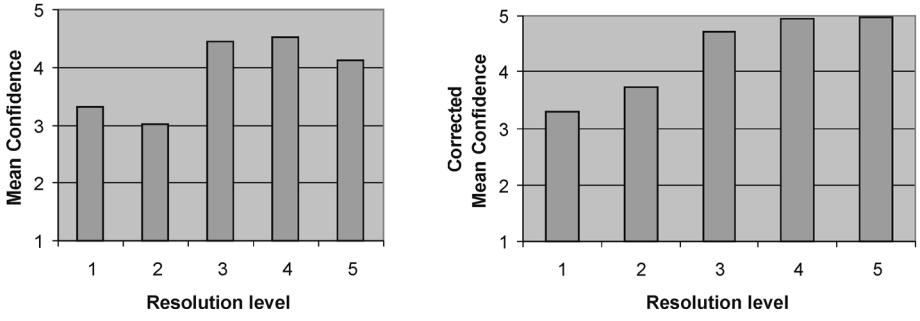


Figure 2. Average confidence levels (vertical axis) at each resolution level (horizontal axis): (a) uncorrected distribution, (b) corrected distribution (see text for description).

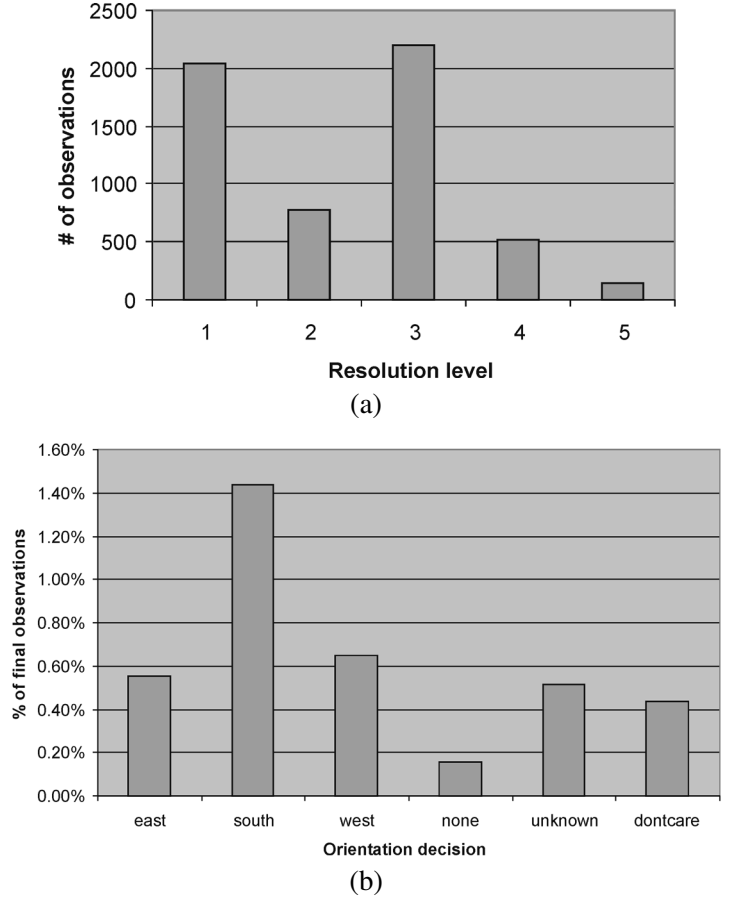


Figure 3. (a) Frequency of different resolution levels as the final resolution, and (b) frequency of incorrect orientation at the final resolution.

age confidence levels at each resolution level are shown in Fig. 2. Some observations are:

– The average confidence reaches above 3 (out of 5) even at resolution level 1.
– Resolution level 2 does have benefit (even though it is a simple interpolation of level 1).
– Resolution level 3 represents the largest jump in confidence.
– Confidence approaches the maximum level at resolution level 4.
– Resolution level 5 adds little benefit.

The uncorrected distribution is also informative. In Fig. 2a, the fact that the average confidence value for resolution level 5 is below those of resolution levels 3 and 4 suggests that the maximum resolution version was requested only for those images with high difficulty. In fact, we can see from the distributions of the final resolution levels in Fig. 3a that

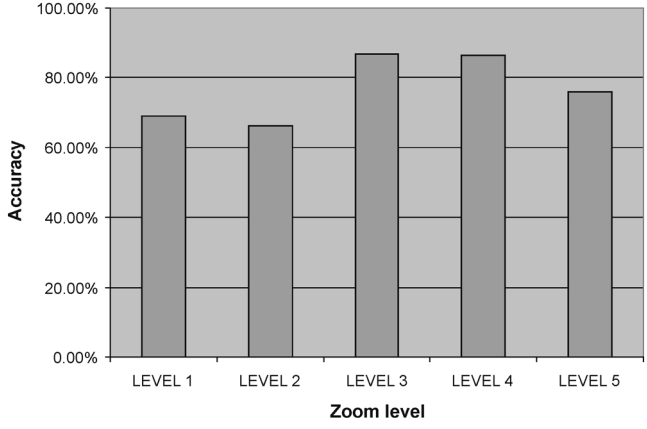
– Most (94%) of the time, the orientation task was completed at resolution levels 1–3.
– A thumbnail (resolution levels 1 and 2) is adequate 50% of the time.
– Resolution level 5 was only requested 1% of the time.

3.2. Orientation accuracy

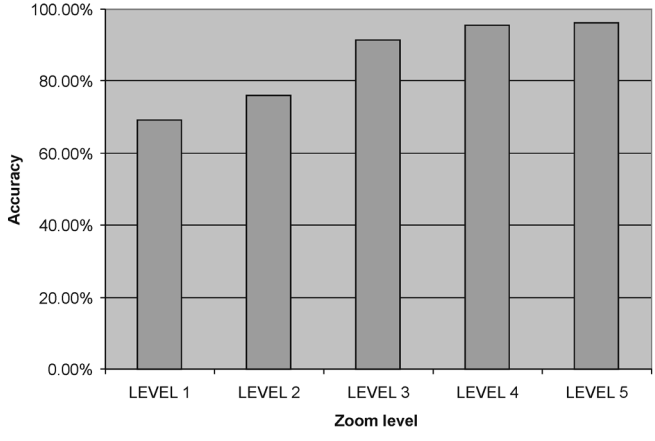
Mean accuracy at the final resolution level was 96.2%. Figure 4 shows orientation accuracy by zoom level, using both uncorrected and corrected values because both are interesting. We note that 69.2% of observations at the first zoom level (24×36) were correct. The (corrected) accuracy at the second zoom level was 76.2%, a significant increase over zoom level 1. Note once again the first two zoom levels share the same number of pixels and hence same amount of information; zoom level two is simply a larger, interpolated version of zoom level one. It is interesting that the larger image size caused a significant increase in accuracy, even though no actual additional information is contained in the larger image.

By the third resolution, the (corrected) accuracy is 91.6%, and 95.7% by zoom level 4. The final resolution level helped very little, increasing (corrected) accuracy to 96.2%. When resolution level 5 was actually needed, its uncorrected accuracy was less than 80%. This is because only the most difficult images required viewing at resolution level 5.

Figure 3b shows the frequency of each of the incorrect orientation decisions at the final resolution level. The orientations have been translated such that 'north' is always the correct answer. We note that while the frequency of east and west misorientations are approximately the same, south misorientations occur twice as frequently. This suggests that when humans make orientation errors, they are more likely to misorient by 180° ('upside-down') than by either of the 90° possibilities. This actually occurred with both 'landscape' and 'portrait' pictures, indicating that these errors could not have been avoided simply by using square pictures (see Note 1).



(a)



(b)

Figure 4. Accuracy by zoom level, (a) uncorrected and (b) corrected.

3.3. Observer confidence vs. orientation accuracy

Next, we look at the relationship between confidence (what observers think they know) and accuracy (what the truth is). Because we have translated the relative orientations of the displayed image back to absolute orientations, a translated 'north' decision is always correct. Overall for all observations, observers were indeed more confident when they made the right decision (4.18) or declared the orientation as 'don't care' (4.18), and less confident when they were wrong (2.77) or declared the orientation as 'unknown' (2.84). As shown in Table 3, both confidence and accuracy increase as zoom level increases. There is extremely strong correlation between confidence and accuracy (a linear trend line has $R^2 = 0.9996$!). Note that this is also true for the 'fake' zoom level where no new information is provided. While such exceptionally strong correlation may be somewhat accidental, high correlation also exists when the data are broken down by sub-categories of scenes (as shown

in Fig. 6), further indicating that human observers are well aware of their level of accuracy across stimuli.

One challenge in computer vision and pattern recognition is for an algorithm itself to produce a confidence measure that is highly correlated with the difficulty the algorithm has on classifying a particular sample. This is extremely useful; for example, the difficult images for an automatic algorithm can be prompted for real-time human intervention or set aside for later human inspection. Alternatively, easy images can be processed through a fast algorithm while difficult cases can be presented to a more expensive and more accurate algorithm to maximize overall throughput of a fully automatic system. Unfortunately, in addition to classification accuracy, computer vision also often lags behind humans in the arena of measuring self-confidence.

3.4. Accuracy across observers

At low resolutions, we found that orientation accuracy varied widely from observer to observer. At the first resolution level, observer accuracy ranged from 55% to 91%, with a median of 75% and a standard deviation of 16.2%. This large variation still existed even after we discarded an obvious outlier (12.9%). At the second resolution level, the accuracies ranged from 74% to 94%, with a median of 82% and a standard deviation of 13.3%. As resolution increases, the range of scores across different observers continues to decrease. There are several explanations for this result. Some observers may be better at using low-level (e.g. color) features to determine image orientation than others. It is also possible that some observers were simply more tenacious than others at low resolution levels (i.e. some observers gave up quickly while others carefully examined low-resolution images to develop a reasonable guess). Finally, environmental factors (e.g. small monitor size or poor contrast settings) may have placed some observers at a disadvantage although the monitor size was between 17″ and 21″ and the screen resolution was approximately 768×1024.

Recall that zoom level 2 is simply a larger, super-sampled version of zoom level 1 (i.e. the effective pixel resolution is the same) and hence contains no new information. Note that this zoom level cannot be achieved by moving closer to the monitor because the latter does not provide more pixels (albeit the same amount of real information). Interestingly, the accuracy data show that some observers found zoom level 2 to be very helpful, while others found it not helpful at all. About 28% of observers showed almost no increase (i.e. <2 percentage points) in corrected

Table 3.
Improved accuracy and increased confidence due to increase in resolution

	Level 1	Level 2	Level 3	Level 4	Level 5
Confidence	3.30	3.72	4.71	4.93	4.97
Accuracy	69.1%	76.2%	91.6%	95.7%	96.2%

accuracy between zoom levels 1 and 2. For about 38% of observers, accuracy increased between 2 and 10 percentage points in level 2, and for the remaining 33% of observers, corrected accuracy increased 10 percentage points or more. At the final resolution level, observer accuracies ranged from 87% to 100%, with a median of 97% and a standard deviation of 3.1%.

3.5. Orientation cues

The study was also intended to give an idea of how observers determine image orientation. To capture this, observers were asked to specify which (primary) cue(s)

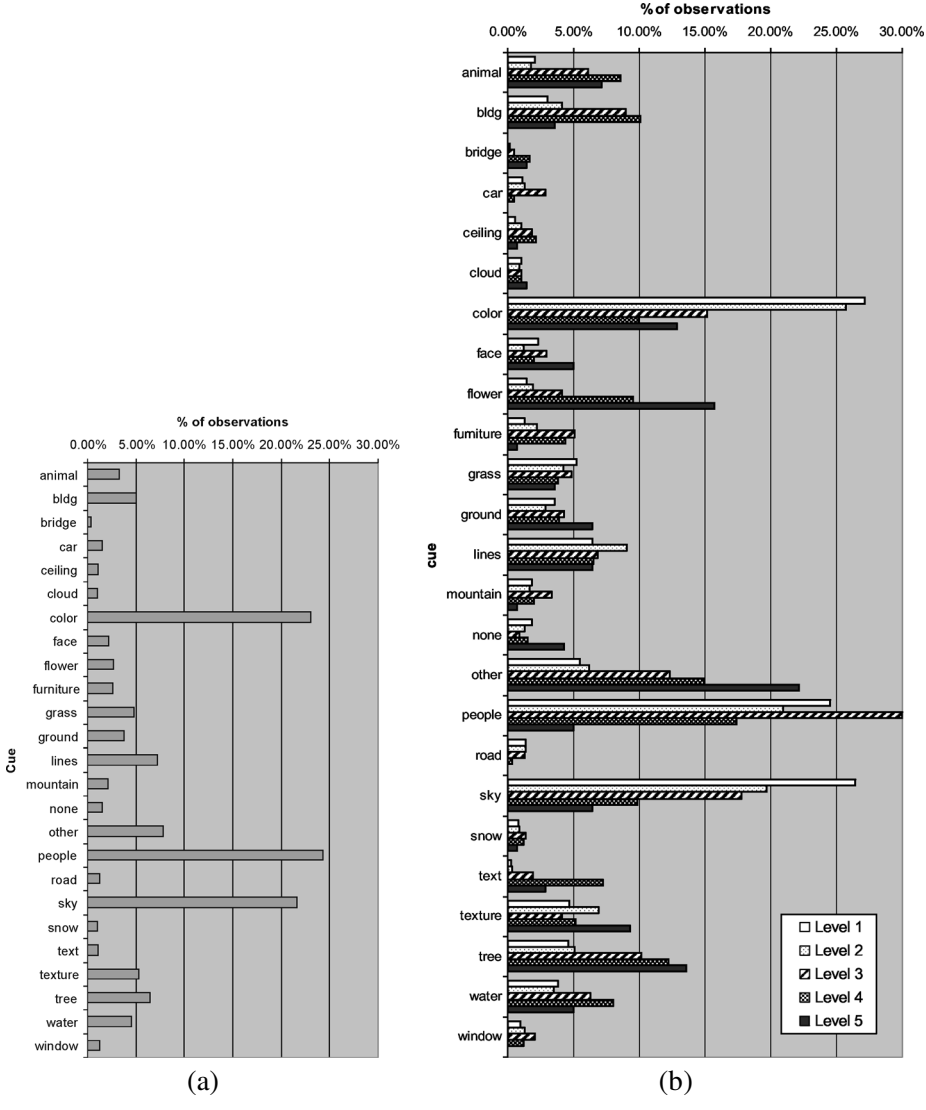


Figure 5. Summary of cues used by observers, (a) overall and (b) by resolution level.

they used in determining orientation for each image at each orientation. Observers could choose from a list of pre-defined cues in a pull-down menu, or enter an arbitrary cue into a free-form 'comments' field.

Figure 5a shows a histogram of the major cues mentioned. We see that people, color, and sky were, by far, the most common cues; mentioned in 24.3%, 23.1%, and 21.6% of observations, respectively. Other common cues included lines, trees, texture, buildings, grass, and water. Figure 5b shows a histogram of the cues mentioned by resolution level. As one might expect, color is mentioned frequently at the first and second zoom levels, but is used much less frequently at higher resolution levels. This suggests that at higher levels, other cues become more important or easier to recognize. We note, however, that even at zoom level 5, color was still mentioned in over 10% of the actual observations. Use of the two other low-level cues, texture and lines, increased slightly as resolution increased. This makes sense since texture features may not be apparent at low resolutions.

Like color, sky was mentioned frequently at the first zoom level (26.4% of observations) but its use diminished with higher resolutions (6.4% of observations at zoom level 5). This indicates that humans perform sky detection at low resolution levels, but shift their focus to other cues as resolution increases. Another less obvious but important fact is that the images that required viewing at higher resolutions are less likely to contain sky. Grass, on the other hand, was mentioned infrequently ($<5\%$ of observations) at all zoom levels. This is somewhat surprising, because grass has been identified as an important low-level semantic cue for automatic orientation determination by computers. However, it is possible that observers ignored grass because other more prominent cues (e.g. sky) were also available.

Figure 5b shows that people are a very important cue in orientation determination. In fact, people (including faces) were mentioned in 24% of observations at zoom level 1, suggesting that a 24×36 image contains sufficient information for humans to recognize people (or, at least, to think they recognize people). This underscores the importance of people detection in image orientation determination. In fact, several observers noted that if they thought they recognized a person in an image, they would give that cue priority over all other cues. In some cases, they noted that what they thought was a person in a low resolution image turned out to be something else entirely when viewed at higher resolutions. The people cue was mentioned most often (30% of observations) at zoom level 3, and decreased as resolution further increased. This suggests that zoom level 3 is sufficient for most observers to recognize people in most images. However, we must be cautious in making the same statement for automatic face/people detection algorithms.

At zoom level 5, flowers were the most commonly mentioned cue (16% of observations). It is expected that flowers, or at least their internal structures, cannot be identified easily at lower resolutions because of their usually small size in a scene. Also, some scenes in our image set are close-ups of flowers, which are usually symmetric, and lack many obvious orientation cues, e.g. stems, pedals, and ground.

Because such images are difficult, observers tend to wait until higher resolutions to make orientation decisions. Both of these factors may explain the high occurrence of flowers at zoom level 5. It is noteworthy that flowers, even at zoom level 5, are not necessarily reliable cues for image orientation.

Observers were also free to enter their own cues into a free-form 'comments' field (see Appendix A). In general, the written-in cues were mentioned much less frequently than the predefined cues. Only 'sun/moon', 'lamps', and 'shadows' were mentioned more often than the least-frequent predefined cue ('bridges').

3.6. Accuracy and confidence by cue

Figures 6a and b show the confidence and accuracy of observations by the cues that were mentioned. The data in these figures represent all observations at all

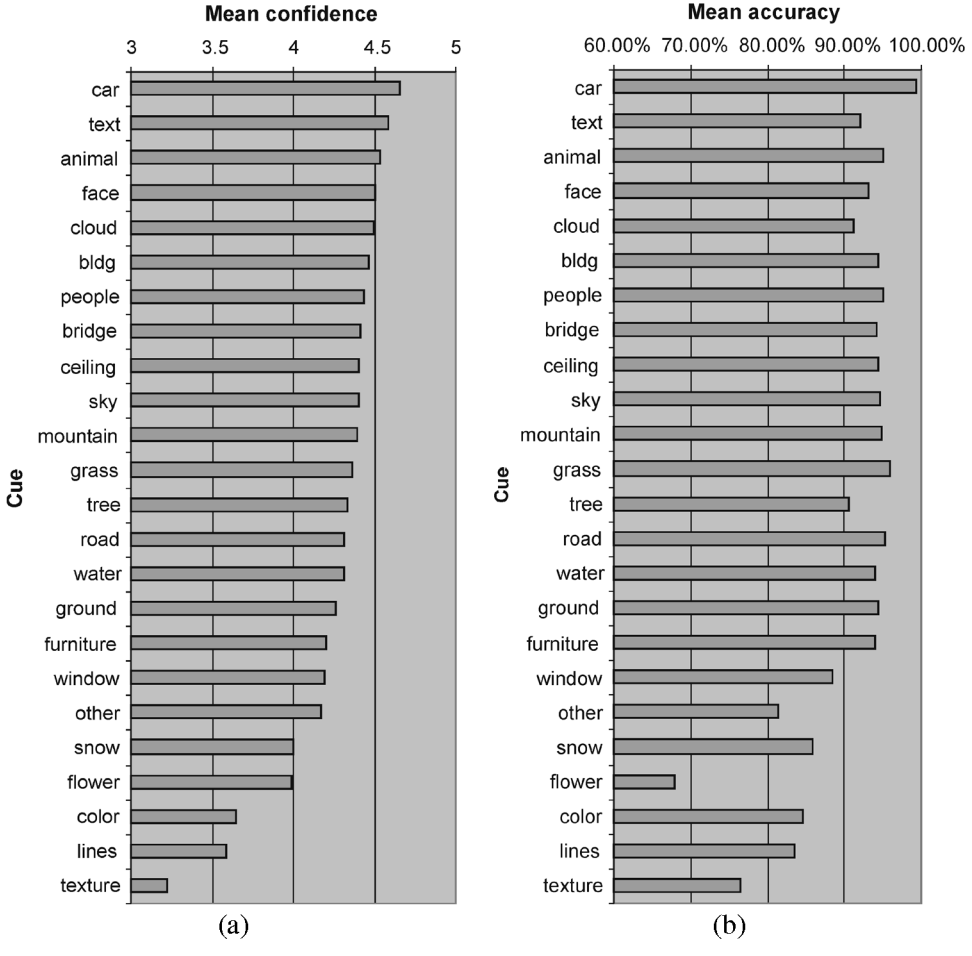


Figure 6. Mean confidence and accuracy by cues mentioned, for all observations at all zoom levels.

zoom levels. Note that the cues have been sorted in order of decreasing confidence. It is observed from Fig. 6a that the three low-level cues — color, lines, and texture — have the lowest mean confidences. This suggests that observers are more confident of their decisions when semantic cues are available. Figure 6b suggests that orientation decisions are also more accurate when semantic cues are available. Except for 'flower', the three low-level cues exhibited the lowest accuracies.

'Car' was the best cue in terms of confidence and accuracy, with an average confidence of 4.7 and an accuracy of 99.4%. Animals, buildings, and people were also high in both confidence and accuracy. Interestingly, even though observers were very confident of observations involving text (mean confidence = 4.6), the accuracy of text was among the worst of the semantic cues (92%). This suggests either that observers think that text is a more reliable cue than it actually is, or that observers think they can recognize text better than they actually can. Some of the text appearing in the image sets was written in languages and scripts unfamiliar to many of the observers. The 'clouds' cue was also high in confidence (4.5) but low in accuracy (91.1%).

The most accurate cues were 'car', 'grass', 'road', 'animal', and 'people', each of which showed an accuracy above 95%. The least accurate semantic cues were 'flower', 'snow', 'window', 'tree', 'cloud', and 'text'. Flowers were mentioned in some of the most difficult images, such as close-ups of flowerbeds. The similar color of snow and clouds, which often point to exactly opposite image orientations, could be confused at low resolutions, resulting in incorrect orientation decisions. Trees can be a misleading cue because of their fractal nature; i.e. branches of trees can be confused as whole trees, confusing orientation decisions.

In similar analysis, we examined which cues were associated with incorrect classifications at the final zoom level. Figure 7a presents a plot of the fraction of incorrect final observations in which each cue was mentioned. 'Color', 'sky', and 'flower' were the biggest culprits, responsible for 27%, 24.6%, and 20.9% of incorrect orientations, respectively. 'People' and 'trees' were also mentioned in more than 10% of incorrect final observations. Note that we need to separate concept failure (e.g. a person lying on the beach) from recognition failure (e.g. water or reflection of sky mistaken as sky).

Figure 7b shows the reliability of each cue, expressed as the percentage of times that the cue was mentioned but the orientation was incorrect. 'Flowers' and 'buildings' were the least reliable cues. The orientation was incorrect 13.1% of the times that 'flower' was mentioned as a cue and 11.9% of the times that 'building' was mentioned. Most other cues showed a 1–3% rate of incorrect orientation.

Figures 8a and 8b present similar graphs broken down by resolution level. Some caution should be taken when trying to draw conclusions from the higher zoom levels, as the sample size is quite small (e.g. at zoom level 5, only 31 observations were incorrect) and the difficulty is greater.

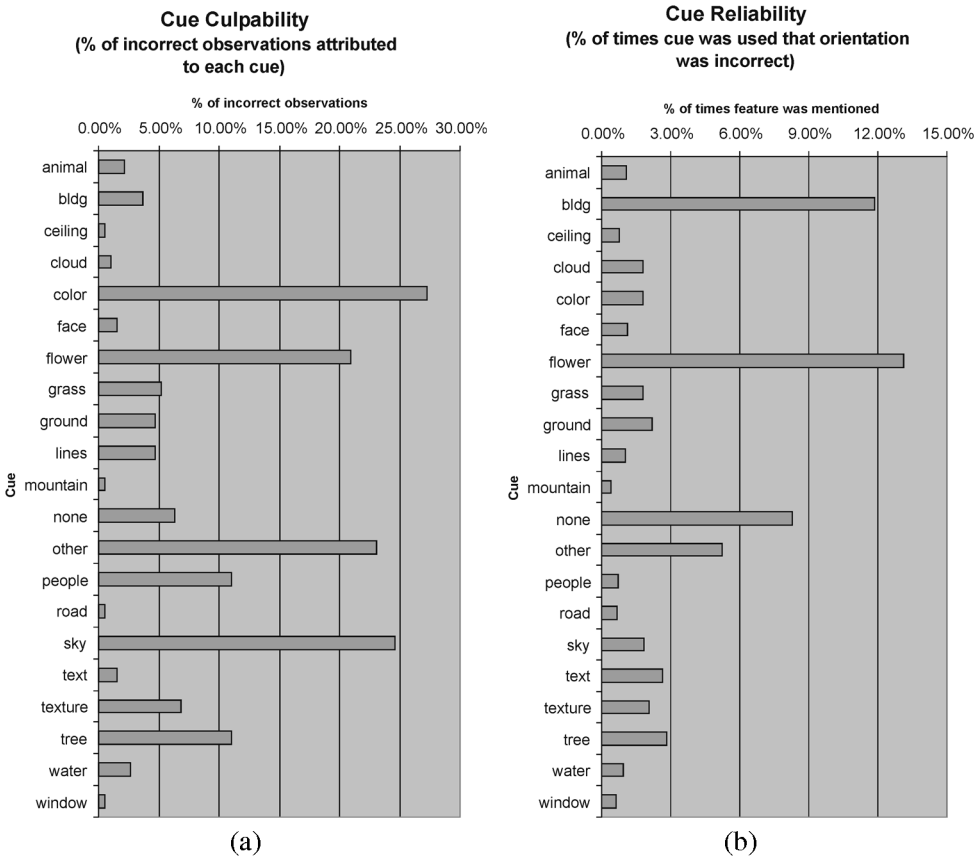


Figure 7. Analysis of misleading cues at the final zoom level.

3.7. Analysis of semantic cues

Our study has confirmed that semantic cues, when available, are very important for orientation determination by humans. When no semantic cues were mentioned at the final zoom level, mean observer accuracy was 77.5%, significantly lower than the 96% overall accuracy. In fact, of the correct orientations at the final zoom level, 98.4% mentioned at least one semantic cue. Humans seem to strongly prefer using semantic cues. Unfortunately, semantic cue detection is often the weakest link in a computer vision system.

'Sky', 'grass', and 'people' were identified as important cues for the image orientation problem. Indeed, the majority (69.9%) of correct final observations used at least one of these three cues. When only semantic cues were mentioned in a correct final observation, only 21.5% did not include sky, grass, or people. In other words, perfect recognition of sky, grass, and people can carry about 80% of the images (though humans are not perfect at recognizing such objects, in particular people, at low resolutions). In addition, when recognizing these cues, humans have

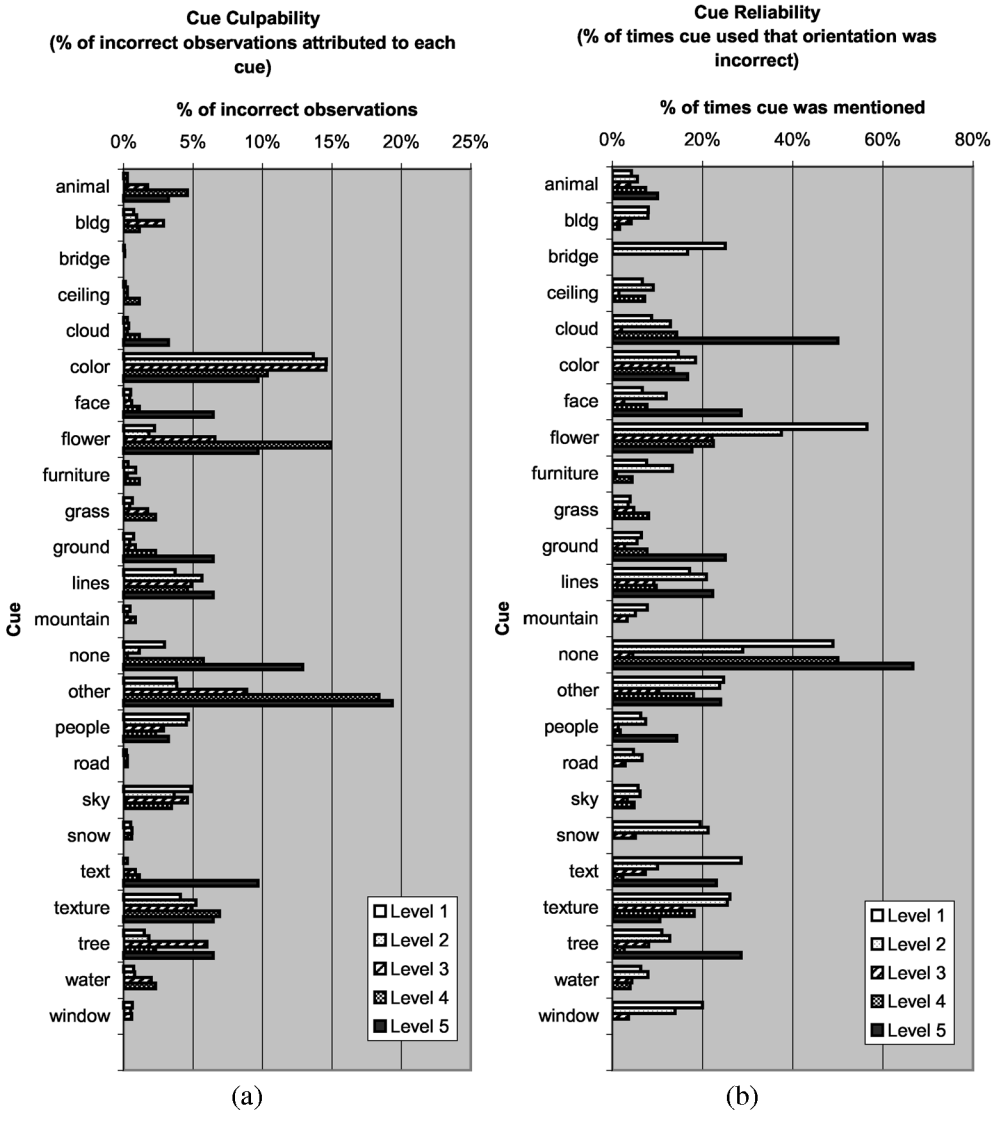


Figure 8. Analysis of misleading cues, by zoom level.

remarkable ability in dealing with occlusion (e.g. seeing sky through tree branches), and color or geometric variations.

It is also of interest to see how accuracy increases with zoom level for each subcategory of scenes that contain these semantic cues. It is important from the point of human vision to know whether different categories of scenes show different patterns in terms of psychophysical curves, and what might determine those categories. In Fig. 9, the curves correspond to four of the most interesting categories, namely scenes containing sky, grass, people, and flowers. Although we did not explicitly use

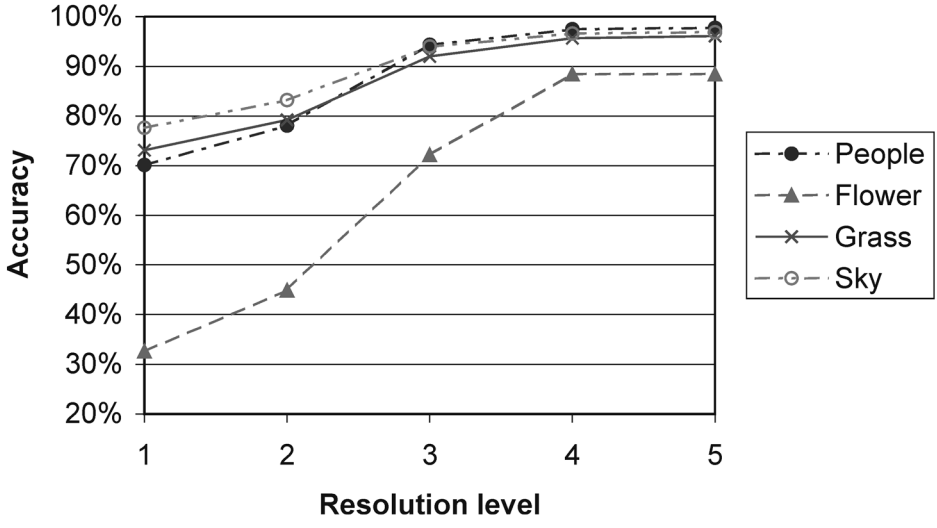


Figure 9. Psychophysical curves showing how accuracy increases with zoom for major scene subcategories.

the size of the semantic features in selecting images for the categories, it is assumed that these features are significant in size because the observers mentioned them in the top three primary cues (for at least one zoom level). The psychophysical curves for large or asymmetric features (e.g. sky, grass, people) have modest slopes while that for flower clearly shows more dramatic effect with increased image resolution. It is also interesting to note that sky is a reliable feature even at the lowest resolution and its accuracy was influenced the least by image resolution.

3.8. Analysis of difficult images

Of the 1000 image set, 11.6% of the images were incorrectly oriented by at least one observer at the final zoom level. 1.5% of the images were more often incorrectly oriented than correctly oriented at the final zoom level. Of those 15 images, all but one are from the Corel image collection. Four of the images are close-ups of flowers, three are drawings (e.g. graffiti), two are reflections of landscapes onto water, two are sandy landscapes (beaches or deserts), and two have extremely subtle features (e.g. hidden or tiny birds) indicating the true image orientation. These images are shown in Fig. 10.

3.9. Difficulty of professional vs. consumer images

There has been some debate on whether consumer photos or professional photos are harder to orient correctly. On one hand, professional photos tend to have better composition, better exposure control, and better overall image quality. On the other hand, professional photographers tend to take artistic pictures (e.g. close-ups of

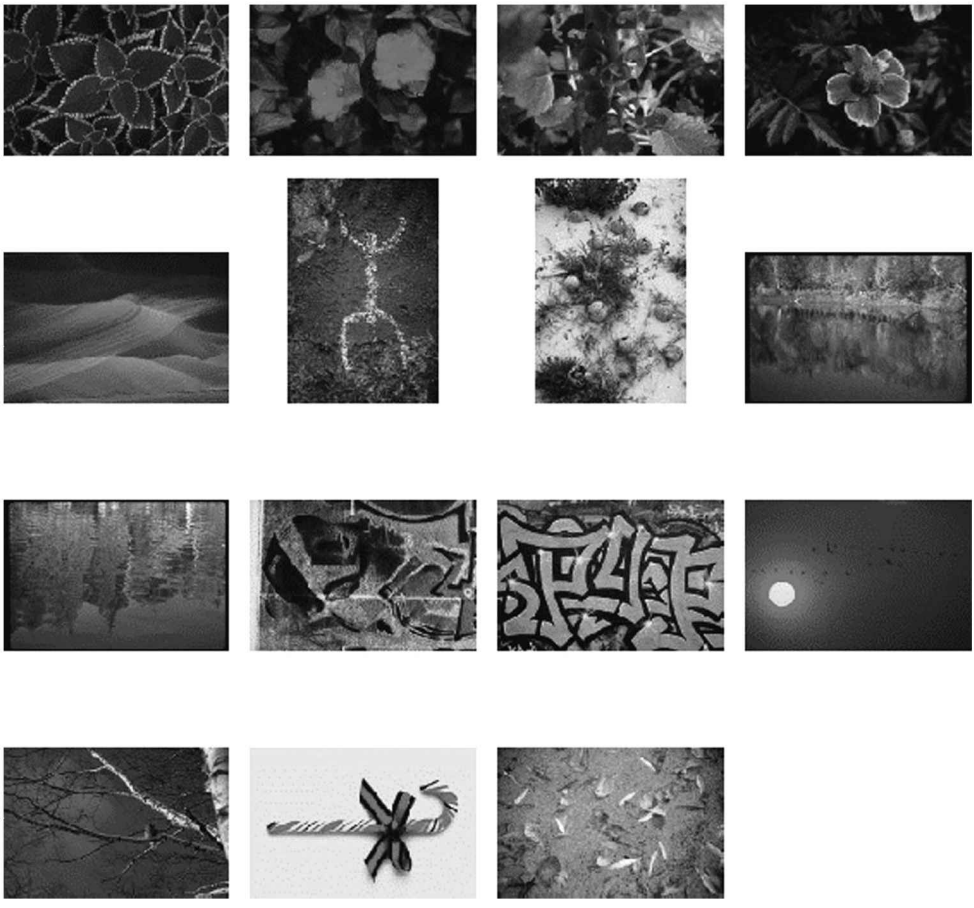


Figure 10. The 15 most difficult scenes.

flowers or some interesting patterns in nature) more often than consumers, and these can be difficult to orient correctly.

To resolve this debate, we analyzed the observations on the 500 JBJL consumer images separately. The accuracy at the final zoom level was 98.1% on the JBJL images, compared to 96.2% on the full image set. Figures 11a and b show the differences in accuracy by resolution level. The accuracy on JBJL was 2–6% better at all resolution levels. On the other hand, there were higher percentages of incorrect decisions at the final zoom level on the whole set than on JBJL, as shown in Fig. 11c. In addition, there were more images judged as 'don't care' or 'unknown'. Our experiment suggests that for humans, orienting professional photos is more difficult than orienting consumer photos because of higher occurrences of peculiar subjects and composition. Interestingly, Vailaya *et al*. (1999) reported that an automatic single-frame orientation algorithm based on low-level features achieved higher accuracy on the Corel images than on the personal images probably similar

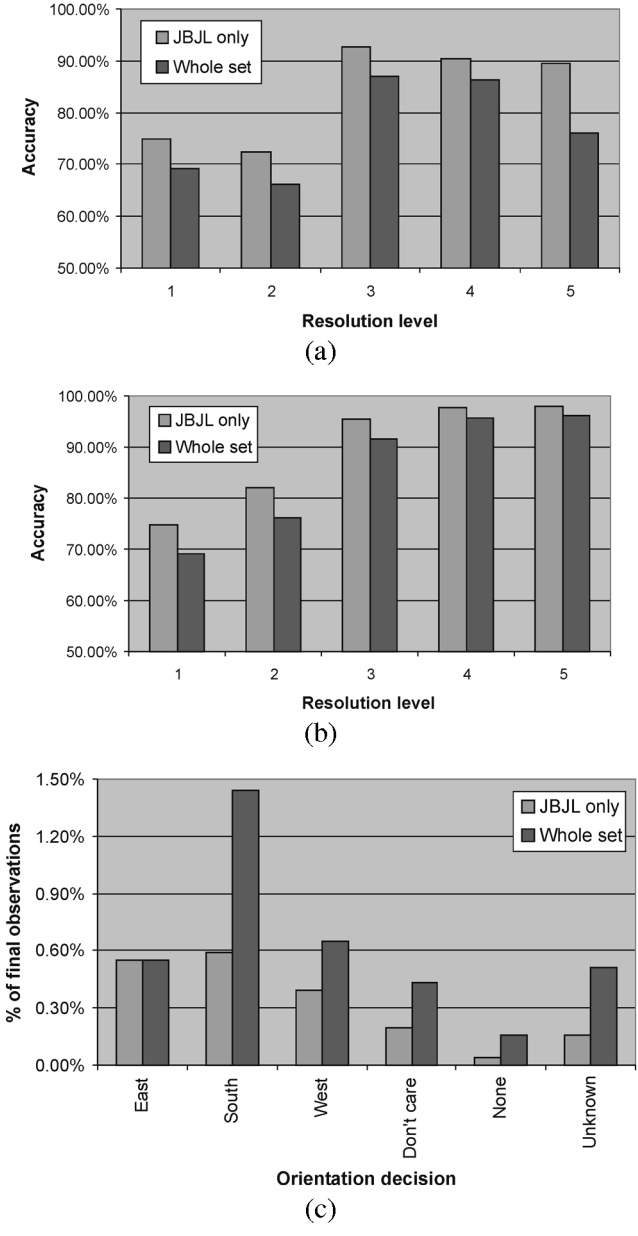


Figure 11. Comparison of orientation accuracy on only JBJL images and the whole image set (JBJL + Corel): (a) uncorrected accuracy by zoom level, (b) corrected accuracy by zoom level, (c) histogram of incorrected decisions at the final zoom level.

to the JBJL images. Therefore, it appears that consumer images are easier than professional images for humans to orient but harder for computer vision systems, which are severely limited in their ability to recognize semantic cues.

3.10. Observer comments

Participants made a variety of interesting comments. Many mentioned that their orientation mistakes at low resolutions were due to incorrect object recognition. Several observers reported that they thought they saw a person in a low-resolution image and made an orientation decision accordingly, but at higher resolutions realized that there was no person. This underscores that people are a very important cue, and that the brain tries hard to locate people in images. One observer noted that his confidence in orientation was actually a measure of confidence in object recognition, and assuming that the object recognition was correct, the confidence of correct orientation was 100%. There was also a strong tendency to place blue colors on top and green colors on the bottom. Several observers noted that they were tricked by grass patches near the actual top of the image and blue objects near the bottom. Several people mentioned that text was not helpful because it was in an unfamiliar script. In several cases, observers said that they could narrow the orientation down to north/south or east/west, but could not choose the specific orientation. Some comments indicate that observers could sometimes not explain an orientation decision, instead saying that it was a 'hunch' or that it 'just feels right'.

For some observations, the comments reveal that very specific and unique objects were used. For example, a number of observers mentioned that a Santa Claus doll was easy to identify at low resolutions because of its unique clothing. The skyline of a specific city, Seattle, was identified in one comment. The faucet of a sink was used in one case. The unique shape of a baseball field was used in another. Some observers reported using subtle cues caused by gravity, like the curve of a plant stem and the texture features of falling water. At least one observer used the fact that red is above green in traffic signals. Use of these very specific cues signals a problem for automatic orientation detection algorithms. While humans recognize thousands of objects and use them to make complex inferences about orientation, robust detection algorithms exist only for a handful of objects. This is a substantial handicap for automatic orientation algorithms and will likely prevent them from surpassing or rivaling human orientation accuracy.

4. DISCUSSION

We gained a number of insights from this psychophysical study.

4.1. Image resolution

Our study found that observer accuracy increases steadily with increasing resolution until what is referred to as 'Base/4' (i.e. 256×384), at which point accuracy was 95.7%. Increasing to the next resolution level, 512×768, increased accuracy by less than a percentage point. A conclusion is that Base/4 is an adequate resolution for accurate orientation by human observers, and therefore, is probably a reasonably

adequate resolution for automatic algorithms as well, especially considering the limitations of such algorithms in recognizing semantic objects.

4.2. Upper bounds on accuracy

It is safe to assume that an automatic orientation algorithm will not surpass the accuracy of an average human observer on an unconstrained set of images, given that inferring orientation is a task that humans are trained to do well. Humans are able to recognize thousands of objects and use high-level reasoning to deduce orientation. An automatic algorithm cannot rival that level of sophistication in the foreseeable future. Human performance, therefore, represents an upper bound on the accuracy that an algorithm can attain. We conclude that an upper bound for accuracy of an algorithm using all available semantic cues is about 96%. If only coarse semantics from thumbnails are used, the upper bound is about 84%. Of course, these bounds depend on the nature of the image set. An algorithm could achieve vastly different detection rates on different image sets, even 100% detection on a conveniently chosen dataset.

4.3. Relative frequencies of incorrect answers

Our study found that observers are twice as likely to misorient images by 180° than by either of the 90° possibilities. This suggests that observers use cues that can distinguish between 'north and south' or 'east and west', but are unable to distinguish between the remaining possibilities. Such cues could include horizon lines.

4.4. Orientation confidence

It was found that accuracy and confidence of observations were highly correlated ($R^2 = 0.9996$), indicating that humans are very good at judging the quality of their decisions. This would be a very desirable characteristic of an automatic algorithm. When the confidence of the algorithm is low, the input image could be flagged and judged by a human, thus improving the overall accuracy of the system.

4.5. Semantic cues

Semantic cues are very important for image orientation. In our study, only 1.6% of images were oriented correctly without semantic cues. Some cues stood out as being very important, such as 'sky' (used in 31% of the correct observations), and 'people' (36.7% of the correct observations). Other important semantic cues include 'cloud', 'water', 'grass', and 'trees'. In fact, a combination of 'sky', 'grass', or 'people' were used in over 70% of the correct observations. These objects are all fairly well defined. We are in the process of developing more robust automatic algorithms for detecting these types of objects (Luo and Boutell, 2003). Other cues mentioned by observers are not as well defined, making detection of them by an

automatic algorithm more difficult. Such cues include categories of objects, such as animals (all species), buildings (all types and styles), ground (dirt, carpet, tiles, etc.), furniture (all types), and vehicles (all types and models). Among them, it is possible and beneficial to develop automatic detectors for sub-categories of objects: the most promising include skyscrapers (Iqbal and Aggarwal, 2002), passenger cars (Schneiderman and Kanade, 2000), paved road (Campbell *et al.*, 1997) and sand (Naphade and Huang, 2000). We do note that many of the published semantic object detectors actually use location cues (therefore explicitly assuming the correct image orientation). The least accurate semantic cues were flowers and snow. Text was found to be a low-payoff cue, because it occurs infrequently in typical photographic images, and the variety of languages and scripts makes it difficult to use.

In conclusion, we have conducted a psychophysical study of perception of orientation of full-cue (color, texture, shape, and shading) photographic images (see Note 2). Using a large set of images representative of the photographic space and extensive interaction by a large group of observers, we were able to obtain valuable information for development of automatic single-frame orientation detection algorithms, including realistic accuracy goals and beneficial types of semantic cues.

Acknowledgements

The authors wish to acknowledge the individuals who contributed their time and knowledge towards this study, a large undertaking. Their efforts have been invaluable in obtaining the important data to facilitate the analysis of the problem and gain insight into how future algorithm development should be conducted.

NOTES

1. We were aware of the potential bias a rectangular photograph may pose and considered using squarely-cropped images. However, in order to stay true to the real use scenarios, we included images in both 'landscape' and 'portrait' orientations and emphasized this aspect in the practice session with each observer.
2. Adding B&W pictures would be an interesting experiment but would require the same level of effort to be repeated.

REFERENCES

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding, *Psychol. Rev.* **84**, 115–147.

Braine, L., Relyea, L. and Davidman, L. (1981). On how adults identify the orientation of a shape, *Perception and Psychophysics* **29**, 138–144.

Campbell, N. W., Thomas, B. T. and Troscianko, T. (1997). Automatic segmentation and classification of outdoor images using neural networks, *Intern. J. Neural Systems* **8** (1), 137–144.

Corballis, M. C., Zbrodoff, N. I., Shetzer, L. L. and Butler, P. B. (1978). Decisions about identity and orientation of rotated letters and digits, *Memory and Cognition* **6**, 98–107.

Dakin, S., Williams, C. and Hess, R. (1999). The interaction of first- and second-order cues to orientation, *Vision Research* **39**, 2867–2884.

DeCaro, S. (1998). On the perception of objects and their orientations, *Spatial Vision* **11**, 385–399.

DeCaro, S. and Reeves, A. (2000). Rotating objects to determine orientation, not identity: Evidence from a backward-masking/dual-task procedure, *Perception and Psychophysics* **62**, 1356–1366.

DeCaro, S. and Reeves, A. (2002). The use of word-picture verification to study entry-level object recognition: Further support for view-invariant mechanisms, *Memory and Cognition* **30**, 811–821.

Hamm, J. and McMullen, P. (1998). Effects of orientation on the identification of rotated objects depend on the level of identity, *J. Exp. Psychol.: Human Perception and Performance* **24**, 413–426.

Iqbal, Q. and Aggarwal, J. K. (2002). Retrieval by classification of images containing large manmade objects using perceptual grouping, *Pattern Recognition* **35**, 1463–1479.

Jolicoeur, P. (1985). The time to name disoriented natural objects, *Memory and Cognition* **13**, 289–303.

Jolicoeur, P. (1992). Orientation congruency effects in visual search, *Canad. J. Psychol.* **46**, 280–305.

Jolicoeur, P., Corballis, M. and Lawson, R. (1998). The influence of perceived rotary motion on the recognition of rotated objects, *Psychonomic Bull. Rev.* **5**, 140–146.

Lawson, R. and Jolicoeur, P. (2003). Recognition thresholds for plane-rotated pictures of familiar objects, *Acta Psychologica* **112**, 17–41.

Lloyd-Jones, T. J. and Luckhurst, L. (2002). Effects of plane rotation, task, and complexity on recognition of familiar and chimeric objects, *Memory & Cognition* **30**, 499–510.

Luo, J. and Boutell, M. (2003). An integrated approach to image orientation detection based on low-level and semantic cues, in: *Proc. IEEE Intern. Conf. on Computer Vision* (submitted).

Luo, J. and Etz, S. (2002). A physical model-based approach to detecting sky in photographic images, *IEEE Trans. Image Processing* **11**, 201–212.

Luo, J. and Savakis, A. (2001). Indoor vs. outdoor classification of consumer photographs using low-level and semantic features, in: *Proc. IEEE Intern. Conf. Image Processing*, Thessaloniki, Greece, pp. 745–748.

Maki, R. (1986). Naming and locating the tops of rotated pictures, *Canad. J. Psychol.* **40**, 368–387.

Mareschal, I., Sceniak, M. and Shapley, R. (2001). Contextual influences on orientation discrimination: Binding local and global cues, *Vision Research* **41**, 1915–1930.

McKone, E. and Grenfell, T. (1999). Orientation invariance in naming rotated objects: Individual differences and repetition priming, *Perception and Psychophysics* **61**, 1590–1603.

Naphade, M. and Huang, T. S. (2000). A probabilistic framework for semantic indexing and retrieval in video, in: *Proc. IEEE Intern. Conf. on Multimedia and Expo (ICME)*, New York, NY, pp. 745–748.

Nicholson, K. and Humphrey, G. (2001). Surface cues reduce the latency to name rotated images of objects, *Perception* **30**, 1057–1081.

Saber, E., Tekalp, A. M., Eschbach, R. and Knox, K. (1996). Automatic image annotation using adaptive color classification, in: *CVGIP: Graphical Models Image Processing*, Vol. 58, pp. 115–126.

Schneiderman, H. and Kanade, T. (1998). Probabilistic modeling of local appearance and spatial relationships of object recognition, in: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 45–51.

Schneiderman, H. and Kanade, T. (2000). A statistical approach to 3D object detection applied to faces and cars, in: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Head, SC, pp. 746–751.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000). Content-based image retrieval: The end of the early years, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22** (12), 1349–1380.

Szummer, M. and Picard, R. (1998). Indoor-outdoor image classification, *Proc. IEEE Intern. Workshop on Content-Based Access Image Video Database*, Santa Barbara, CA, pp. 42–51.

Tarr, M. and Bulthoff, H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993), *J. Exp. Psychol.: Human Perception and Performance* **21**, 1494–1505.

Tarr, M. and Bulthoff, H. (1998). Image-based object recognition in man, monkey, and machine, *Cognition* **67**, 1–20.

Vailaya, A., Jain, A. and Zhang, H.-J. (1998). On image classification: City images vs. landscapes, *Pattern Recognition* **31**, 1921–1936.

Vailaya, A., Jain, A. and Zhang, H.-J. (1999). Automatic image orientation detection, in: *Proc. IEEE Intern. Conf. on Image Processing*, Kobe, Japan, pp. 600–604.

Vannucci, M. and Viggiano, M. P. (2000). Category effects on the processing of plane-rotated objects, *Perception* **29**, 287–302.

Wang, Y. and Zhang, H. (2001). Content-based image orientation detection with support vector machines, in: *Proc. IEEE Workshop on Content-Based Access Image Video Libraries (CBAIVL2001)*, Kauai, Hawaii, pp. 17–23.

APPENDIX A. LIST OF WRITE-IN CUES MENTIONED BY OBSERVERS

Cue	# of obs	% of obs
balloon	12	0.09%
bicycle	4	0.03%
bird nest	1	0.01%
boat	35	0.27%
bottle	5	0.04%
bow and arrow	1	0.01%
bubbles	2	0.02%
candy cane	7	0.05%
clothing	1	0.01%
cross	2	0.02%
curtain	1	0.01%
dishes	3	0.02%
door	27	0.21%
fence	13	0.10%
fireplace	14	0.11%
flag	2	0.02%
food	2	0.02%
heart shape	5	0.04%
highlights from sun	1	0.01%
horizon	3	0.02%
lamps	43	0.33%
lighting	26	0.20%
moon	3	0.02%
ornaments	7	0.05%
plants	33	0.25%

Cue	# of obs	% of obs
post/pole	8	0.06%
railroad tracks	3	0.02%
rocks	4	0.03%
shading	5	0.04%
shadows	41	0.32%
ski gear	2	0.02%
ski lift	3	0.02%
slide	1	0.01%
smoke	3	0.02%
stairs	22	0.17%
statue	20	0.15%
stockings	4	0.03%
stop light	1	0.01%
sun	20	0.15%
sunset	22	0.17%
swing	2	0.02%
symmetry	2	0.02%
tractor	3	0.02%
train	3	0.02%
umbrella	3	0.02%
vanishing point	3	0.02%
vase	2	0.02%
walls	20	0.15%
waterfall	9	0.07%
wreath	7	0.05%

APPENDIX B1. EXAMPLES OF THE CONSUMER PHOTOS USED IN THIS STUDY

APPENDIX B2. EXAMPLES OF THE PROFESSIONAL PHOTOS USED IN THIS STUDY