Theories and Results in Object Recognition

Thomas Breuel

goal today

Major controversy in computer vision and psychology:

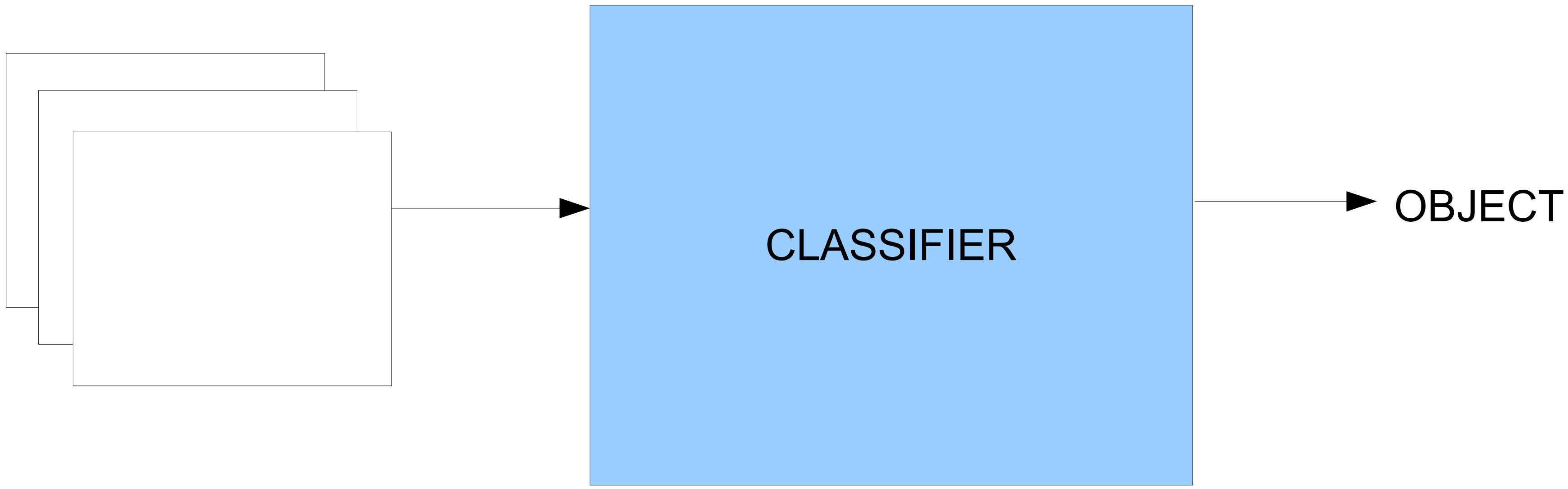
- How does visual object recognition work?
- How does it represent objects?

Today:

- what is the controversy about?
- how did people go about resolving it?
- has it been resolved?

so far...

- **"pattern recognition/machine learning view"**



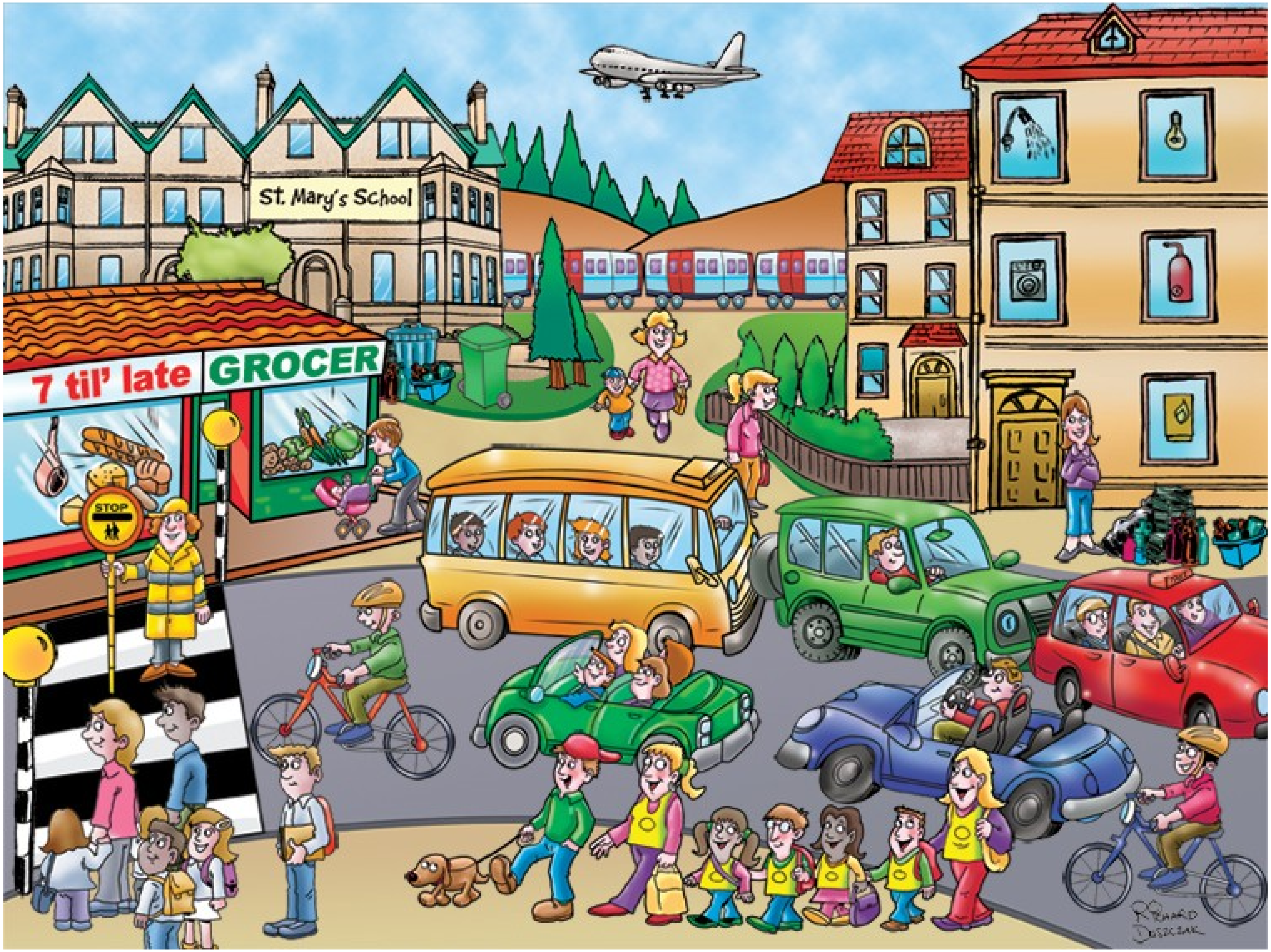
think about the task

- 3D objects, transformations, pose
- lighting
- clutter
- occlusions
- categorical levels
- context

object constancy

Objects can be recognized (i.e., their identity remains constant) across a wide range of viewing conditions.

海垦果园
放心水果 放心吃
HSF
海垦果园
放心水果 放心吃
琼A·V9627


St. Mary's School
7 til' late
GROCER
STOP
RICHARD DUSZCZAK

what are we recognizing?

- **individual level**
  - "my horse 'Alfie'"
- **subordinate level**
  - "Friesian horse"
- **base level**
  - "horse"
- **superordinate level**
  - "quadruped", "animal"

how do we account for this?

Three main theories in computer vision and psychology:

- **3D model-based recognition**
- **parts / components-based models**
- **view-based / appearance-based recognition**

Some evidence for each of them.

general questions

- what is explicitly represented?
- what algorithms are needed/used?
- how are new objects learned?
- what objects are distinguishable / indistinguishable?
- how does the method compare to human performance?
- what engineering problems does the approach solve?

3D MODEL-BASED RECOGNITION

3D model-based recognition

- three tasks:

- model matching
- model acquisition
- model generalization

3D model matching

- **input**
  - collection of points/lines in 3D
  - collection of points/lines in 2D
  - error bound
- **output**
  - a 3D transformation *T* such that the following score is maximized:

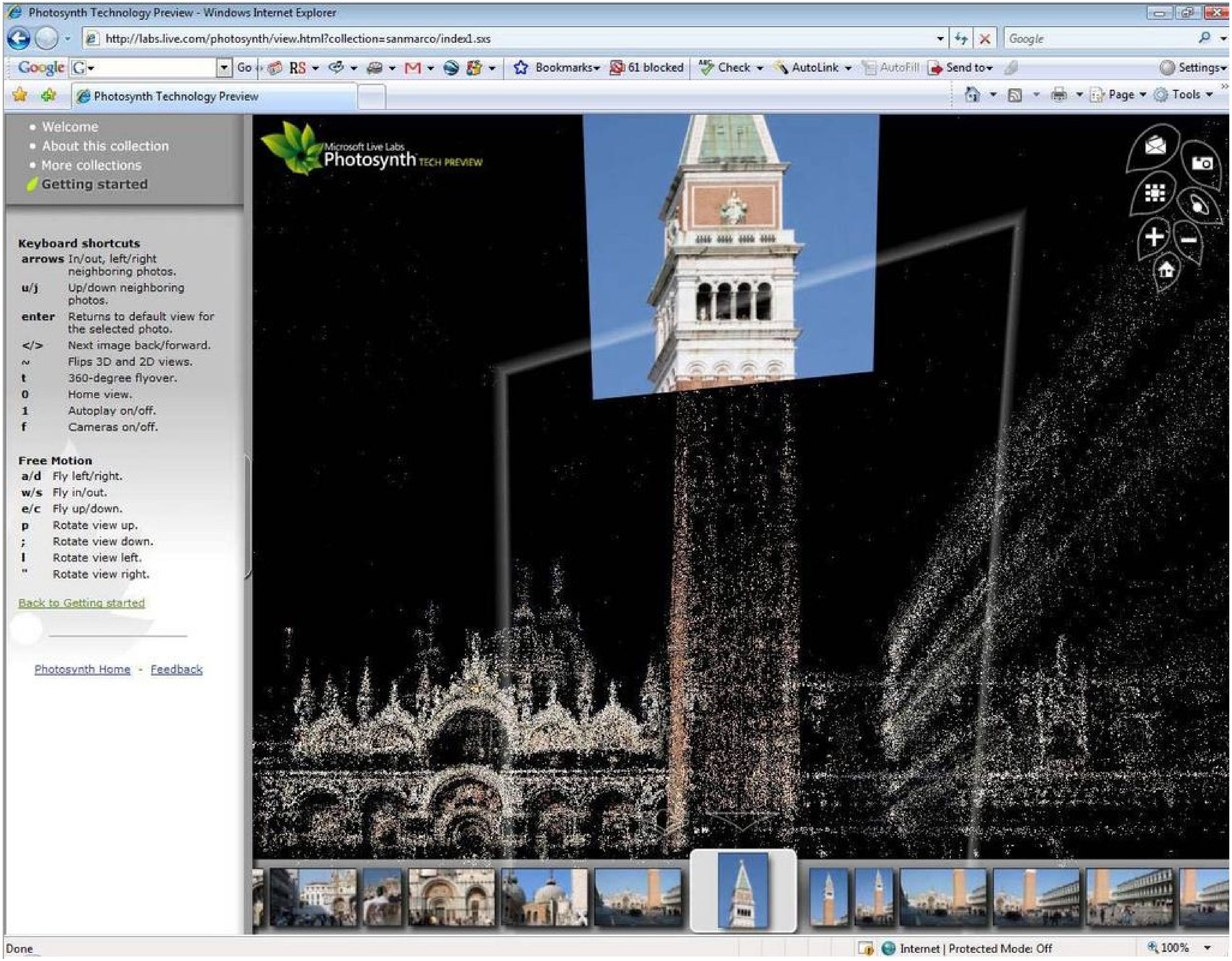
$$S = \sum_{j=1}^{M} min_{i=1\ldots N} \; \varphi(b_i - T m_j)$$

$$\varphi(x) = max(0, 1 - \alpha x^2)$$

(This can be justified as a maximum likelihood or MAP estimator.)

3D model building

- **input**
  - collection of images, each represented as 2D points
- **output**
  - a 3D model representing the object


Photosynth Technology Preview - Windows Internet Explorer
http://labs.live.com/photosynth/view.html?collection=sanmarco/index1.sxs
Google
Go
RS
Bookmarks
61 blocked
Check
AutoLink
AutoFill
Send to
Settings
Photosynth Technology Preview
Page
Tools
Welcome
About this collection
More collections
Getting started
Microsoft Live Labs
Photosynth TECH PREVIEW
Back to Getting started
Photosynth Home - Feedback
Done
Internet | Protected Mode: Off
100%

Keyboard shortcuts

arrows	In/out, left/right neighboring photos.
u/j	Up/down neighboring photos.
enter	Returns to default view for the selected photo.
</>	Next image back/forward.
~	Flips 3D and 2D views.
t	360-degree flyover.
0	Home view.
1	Autoplay on/off.
f	Cameras on/off.

Free Motion

a/d	Fly left/right.
w/s	Fly in/out.
e/c	Fly up/down.
p	Rotate view up.
;	Rotate view down.
l	Rotate view left.
"	Rotate view right.

3D model-based recognition

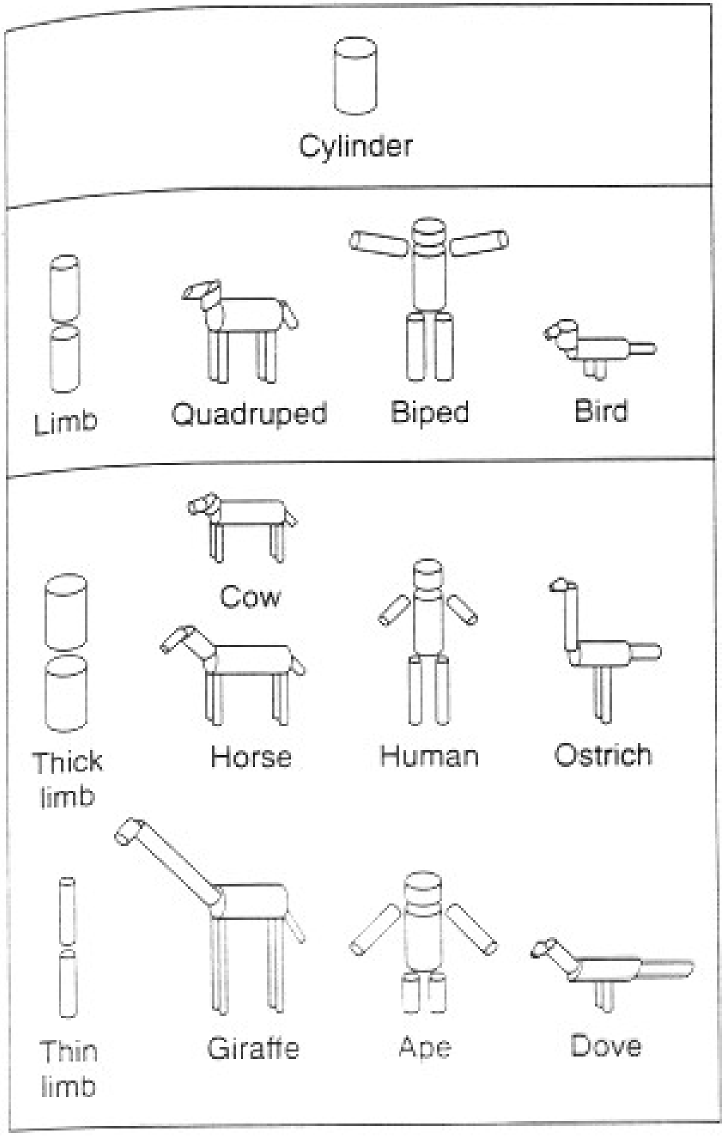
- combinatorial explosion during matching (correspondence problems)
- error handling is difficult
- 3D shape variation does not seem to correspond well to objects

PARTS BASED RECOGNITION

components-based recognition

- objects are composed of simple components / parts / ...
- the identity + connectivity of parts determines what an object is

generalized cylinders



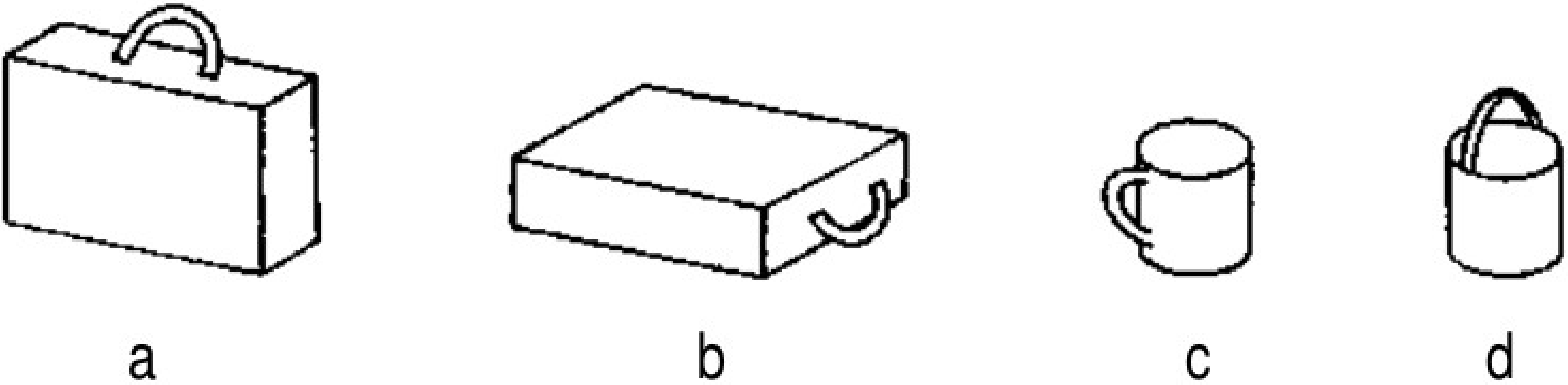
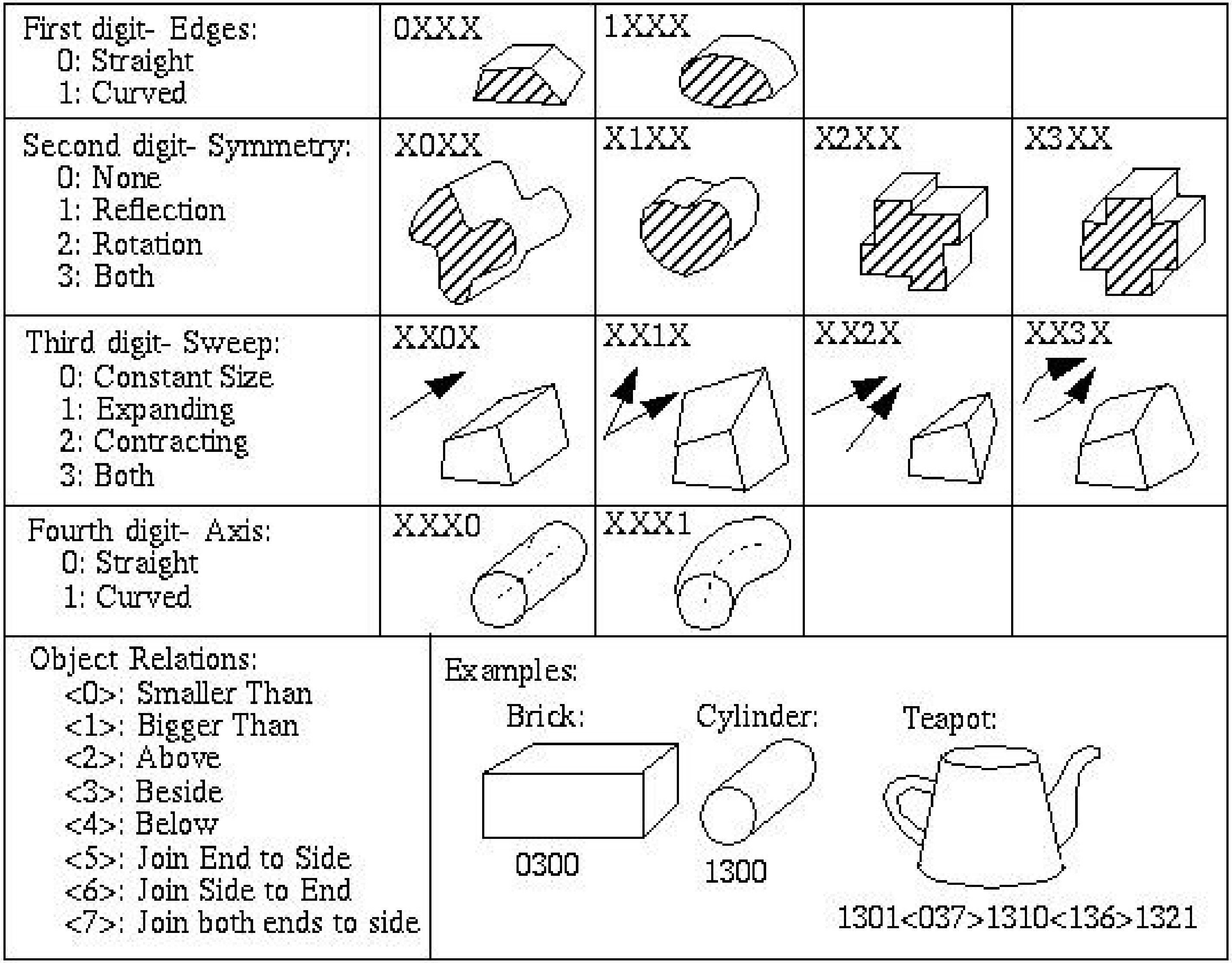


figure 3. Different arrangements of the same components can produce different objects.

geons



First digit- Edges: 0: Straight 1: Curved	0XXX	1XXX		
Second digit- Symmetry: 0: None 1: Reflection 2: Rotation 3: Both	X0XX	X1XX	X2XX	X3XX
Third digit- Sweep: 0: Constant Size 1: Expanding 2: Contracting 3: Both	XX0X	XX1X	XX2X	XX3X
Fourth digit- Axis: 0: Straight 1: Curved	XXX0	XXX1		
Object Relations: <0>: Smaller Than <1>: Bigger Than <2>: Above <3>: Beside <4>: Below <5>: Join End to Side <6>: Join Side to End <7>: Join both ends to side	Examples: Brick: 0300 Cylinder: 1300 Teapot: 1301<037>1310<136>1321			

parts-based recognition

- how do we implement this?
- how are new objects learned?
- what objects are distinguishable / indistinguishable?

implementation

- **possible recognition of parts**
  - specialized algorithms (e.g., generalized cylinders)
  - 3D modeling and matching
  - machine learning and pattern recognition
- **recognition of connectivity**
  - note: more complex than adjacency

VIEW-BASED APPROACH

view-based approach

- **no explicit 3D models**
- **strictly view-based models**
  - objects are represented as collections of views
  - matches only happen between input images and views
- **weakly view-based models**
  - collections of views are pre-processed in some way into a model
  - no attempt is made to reconstruct 3D information (but the learning algorithms may do so implicitly)
  - unknown images are matched against the model

view-based approach

- **largely retains pattern recognition approach**
- **approaches to variation**
  - 3D rotation, pose → manifold learning
  - lighting → feature extraction, learning
  - clutter, occlusion → pre-segmentation, hidden vars., noise
  - instance / object / category → statistical modeling
  - context → priors, Bayesian methods

view interpolation models

- a 2D view is a linear transformation of points in 3D followed by a projection
- as a result, the coordinates of points in 2D views consisting of *n* points form a low-dimensional linear subspace of R^{2n}

RBF-based models

- the input image is compared separately to different known views of each object
- if it is similar enough to a stored view, it "votes" for the corresponding image; the more similar it is, the more it contributes to the voting
- very similar to SVMs, kernel-based methods, and neural networks

QUESTIONS

general questions

- what is explicitly represented?
- what algorithms are needed/used?
- how are new objects learned?
- what objects are distinguishable / indistinguishable?
- how does the method compare to human performance?
- what engineering problems does the approach solve?

BULTHOFF AND EDELMAN PAPER

Psychophysical support for a two-dimensional view interpolation theory of object recognition

(regularization networks/computer graphics psychophysics/generalization)

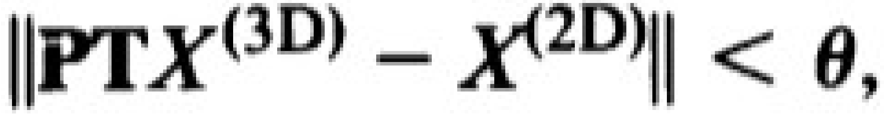
HEINRICH H. BÜLTHOFF*† AND SHIMON EDELMAN‡

*Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912; and ‡Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel

Communicated by Richard Held, September 19, 1991

ABSTRACT **Does the human brain represent objects for recognition by storing a series of two-dimensional snapshots, or are the object models, in some sense, three-dimensional analogs of the objects they represent? One way to address this question is to explore the ability of the human visual system to generalize recognition from familiar to unfamiliar views of three-dimensional objects. Three recently proposed theories of object recognition—viewpoint normalization or alignment of three-dimensional models [Ullman, S. (1989) *Cognition* 32, 193–254], linear combination of two-dimensional views [Ullman, S. & Basri, R. (1990) *Recognition by Linear Combinations of Models* (Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge), A. I. Memo No. 1152], and view approximation [Poggio, T. & Edelman, S. (1990) *Nature (London)* 343, 263–266]—predict different patterns of generalization to unfamiliar views. We have exploited the conflicting predictions to test the three theories directly in a psychophysical experiment involving computer-generated three-dimensional objects. Our results suggest that the human visual system is better described as recognizing these objects by two-dimensional view interpolation than by alignment or other methods that rely on object-centered three-dimensional models.**

3D Model



$$\|\mathbf{PT}X^{(3\mathrm{D})} - X^{(2\mathrm{D})}\| < \theta,$$

2D View Interpolation

$$\|\sum_i \alpha_i X_i^{(2D)} - X^{(2D)}\| < \theta,$$

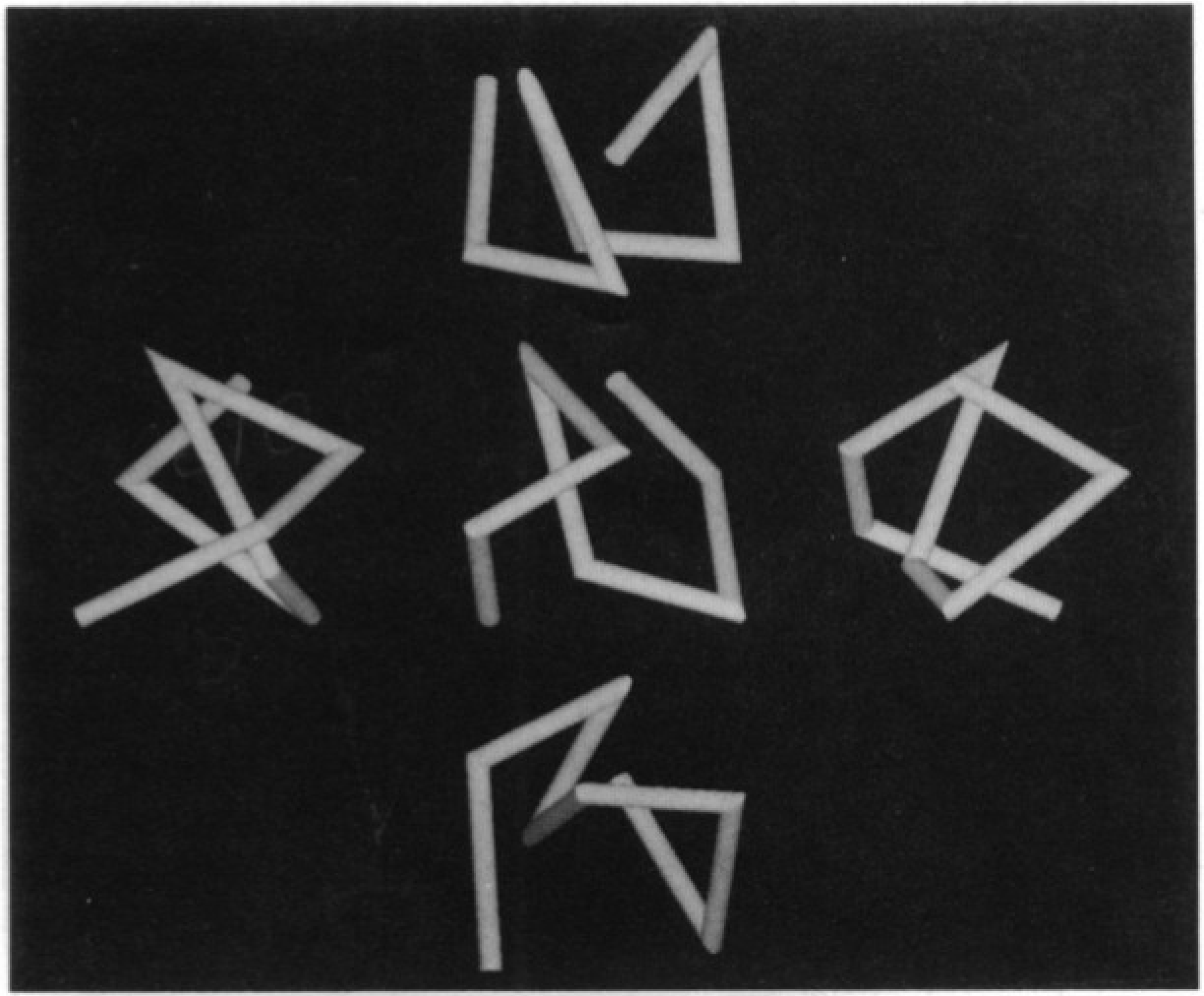
RBF Networks

$$|\sum_{k} c_k G(\|X^{(2\mathrm{D})} - X_k^{(2\mathrm{D})}\|) - 1| < \theta,$$

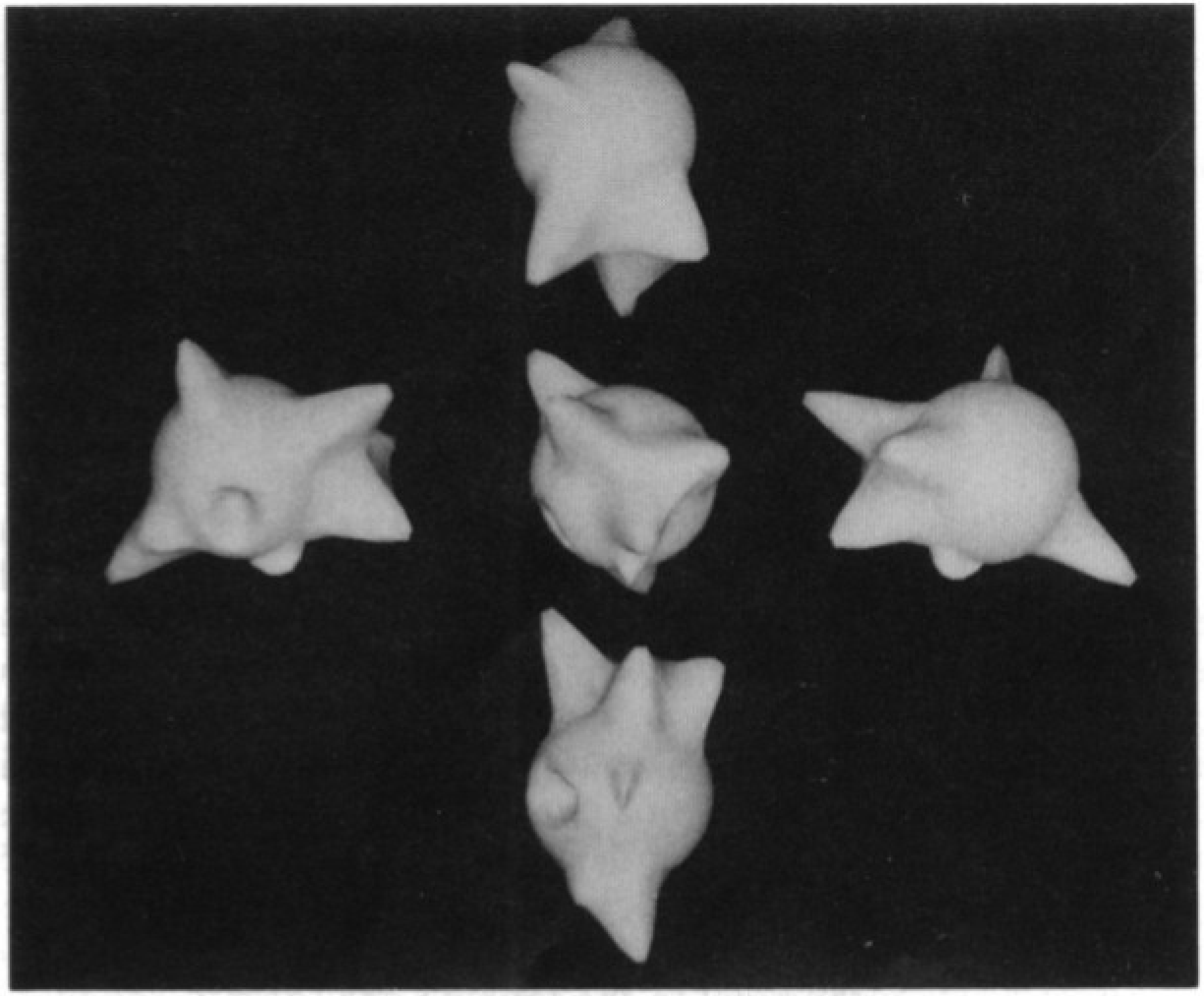
experimental idea

- show objects in a way that there is enough information for the brain to reconstruct 3D structure
- test how well human subjects recognize objects from novel views

stimulus 1



stimulus 2



stimulus generation

- stimuli are computer generated
- large numbers of stimuli can be generated automatically
- lighting etc. can be controlled precisely
- easy to control statistics of objects

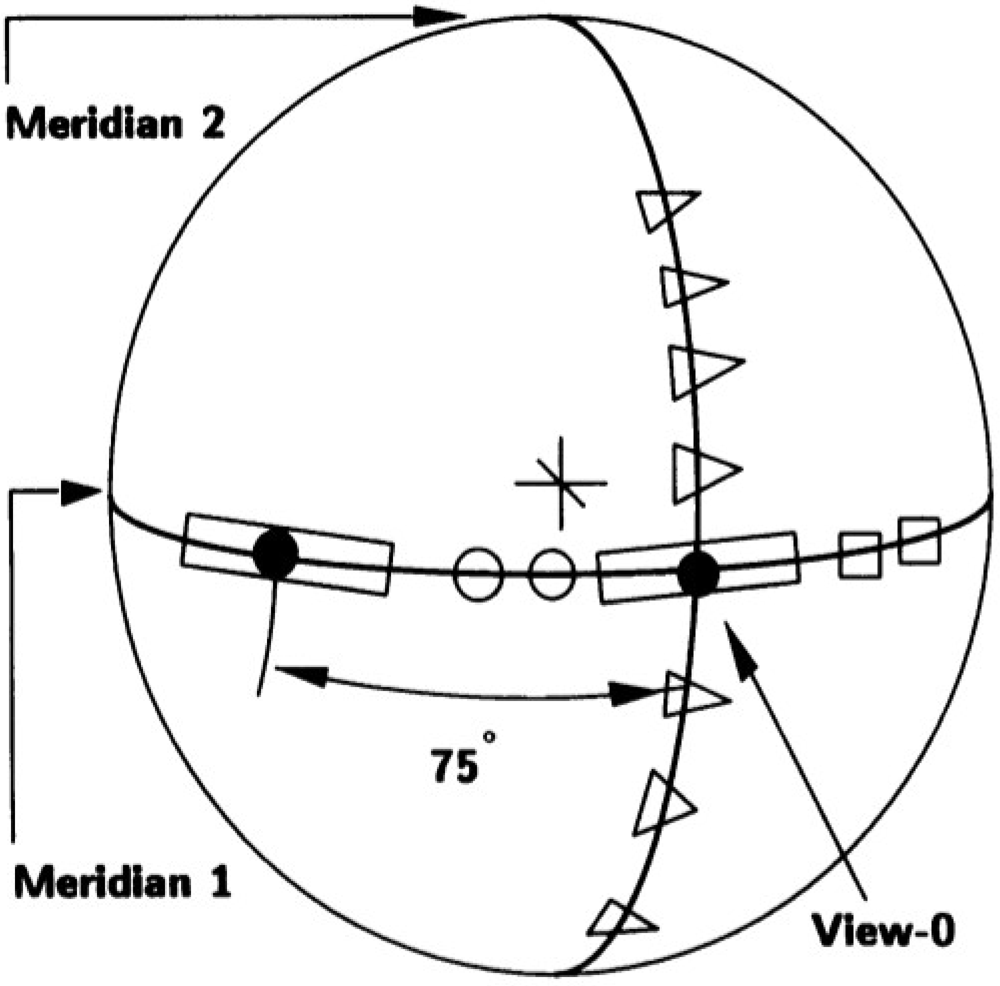
training phase

- show target object with slight motion in frontal view and side view (two training views)
- motion provides (provably) enough information to reconstruct 3D structure
- test recognition from different views

recognition tests

- **2AFC – two alternative forced-choice task**
  - provide human subject with two alternatives
  - force a choice between them
- **is this the target object or not?**
- **generally**
  - miss rates of about 30%
  - error rates of about 30%

test views



VIEWING SPHERE

(centered at the object)

▭ = TRAIN (view sequences)

○ = interpolation ⎤ same
□ = extrapolation ⎦ meridian ⎤ TEST
◺ = ortho meridian ⎦ (static views)

predictions based on each hypothesis

- **3D recognition**
  - all three conditions should be recognized equally well
- **2D view interpolation**
  - inter and extra conditions should be equal, but ortho should be worse
- **2D RBF interpolation**
  - inter should be best, followed by extra, followed by ortho

What do you predict?

results

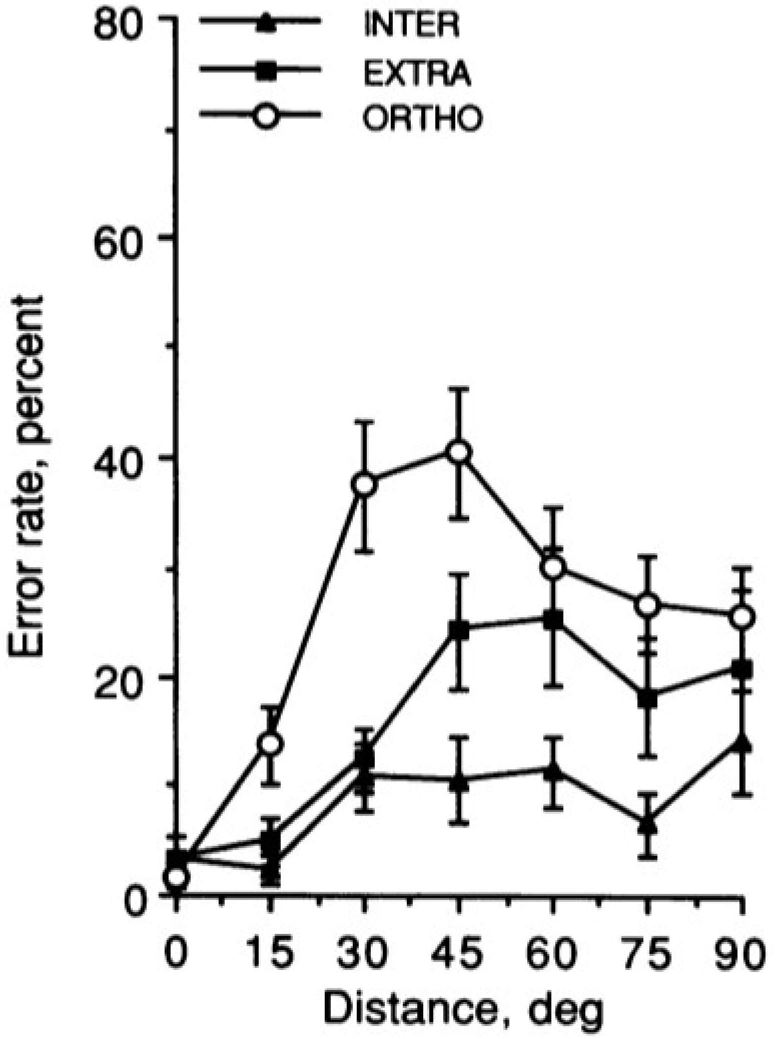


FIG. 3. Unbalanced objects. Error rate (type I errors only) vs. great-circle distance (D) from the reference view-0 (four subjects; error bars denote ± SEM). A three-way (subject × condition × D) general linear model analysis showed highly significant effects of condition [$F(2, 524) = 23.84$, $P < 0.0001$] and D [$F(6, 524) = 6.75$, $P < 0.0001$]. The mean error rates in the INTER, EXTRA, and ORTHO conditions were 9.4%, 17.8%, and 26.9%, respectively. Subjects tended to perform slightly worse on one of the training views (INTER condition, 75°) than on the other (0°), possibly because this view always appeared as the second one in the training phase. A repeated experiment involving the same subjects and stimuli yielded shorter and more uniform response times but an identical pattern of error rates. deg, Degrees.

additional result

- strong preference in terms of horizontal generalization over vertical generalization
- (i.e., turn the experiment by 90 degrees and performance gets worse)

discussion

- results disagree strongly with predictions of 3D model or view interpolation theories
- results are consistent with RBF-like view-based models

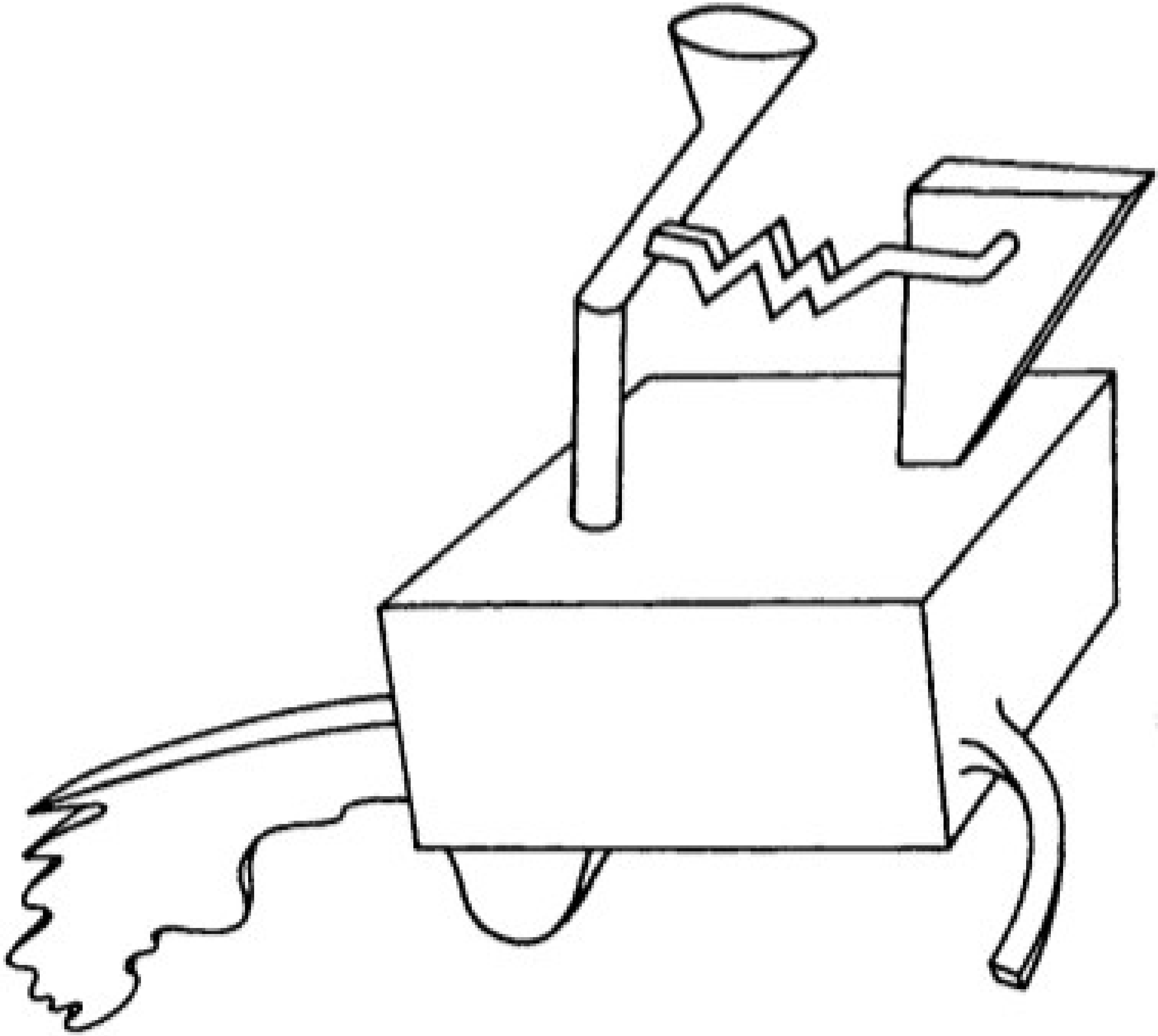
What do you think?

GEON THEORY

Recognizing Depth-Rotated Objects: Evidence and Conditions for Three-Dimensional Viewpoint Invariance

Irving Biederman and Peter C. Gerhardstein

Five experiments on the effects of changes of depth orientation on (a) priming the naming of briefly flashed familiar objects, (b) matching individual sample volumes (geons), and (c) classifying unfamiliar objects (that could readily be decomposed into an arrangement of distinctive geons) all revealed immediate (i.e., not requiring practice) depth invariance. The results can be understood in terms of 3 conditions derived from a model of object recognition (I. Biederman, 1987; J. E. Hummel & I. Biederman, 1992) that have to be satisfied for immediate depth invariance: (a) that the stimuli be capable of activating viewpoint-invariant (e.g., geon) structural descriptions (GSDs), (b) that the GSDs be distinctive (different) for each stimulus, and (c) that the same GSD be activated in original and tested views. The stimuli used in several recent experiments documenting extraordinary viewpoint dependence violated these conditions.



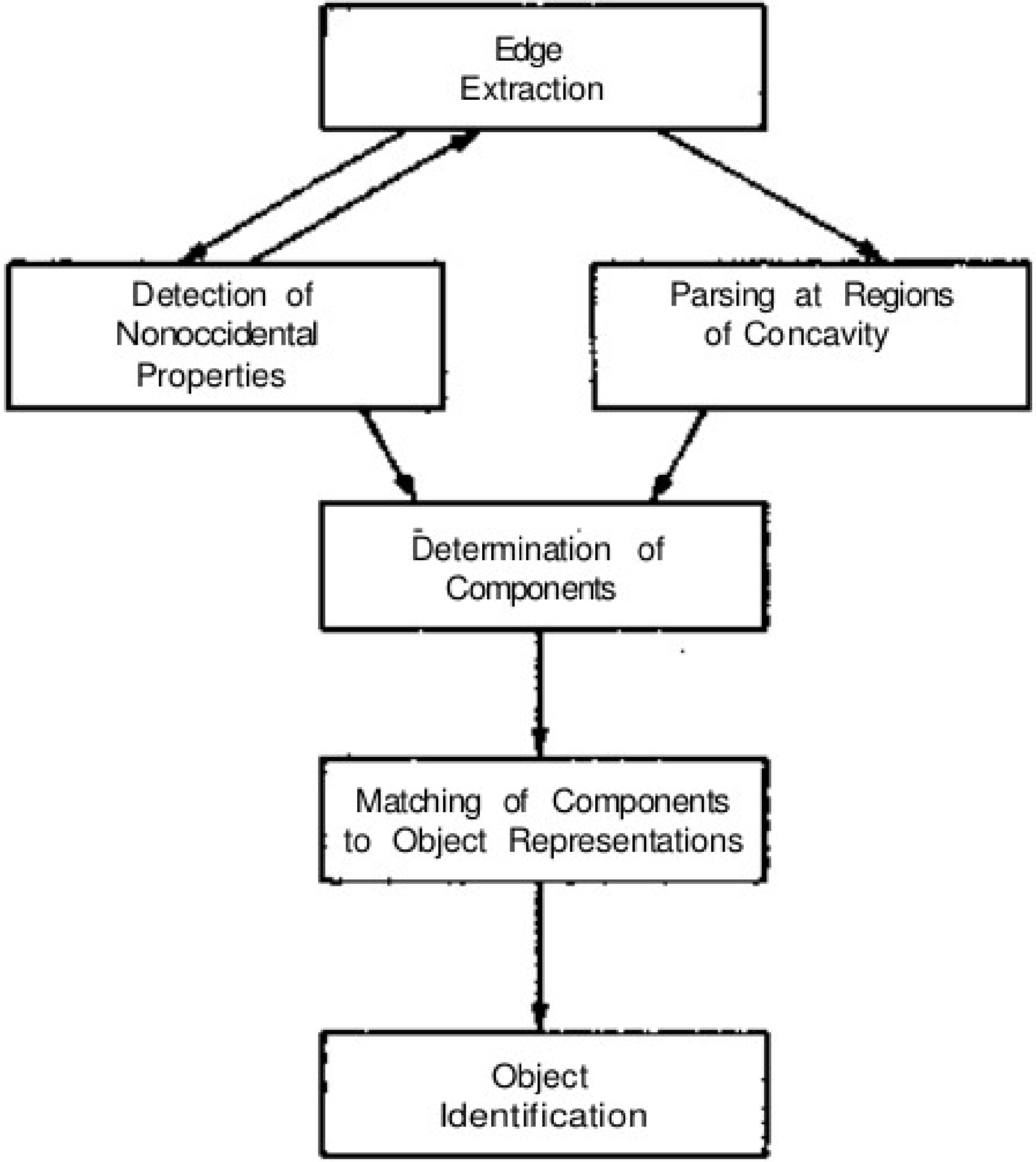

Edge Extraction
Detection of Nonoccidental Properties
Parsing at Regions of Concavity
Determination of Components
Matching of Components to Object Representations
Object Identification

Figure 2. Presumed processing stages in object recognition.

view-based model

- **you store a bunch of key views in memory: *view-specific templates***
- **the 3D object “is” the collection of templates**
- **you match a novel view to its closest stored view, taking time proportional to the angle of rotation**

evidence for view-based models

- recognition speed is slower for novel viewpoints
- human recognition of unfamiliar objects from novel viewpoints is poor
- there are consistent relationships between training poses and recognition of novel views

problem

that model doesn't work for familiar object classes

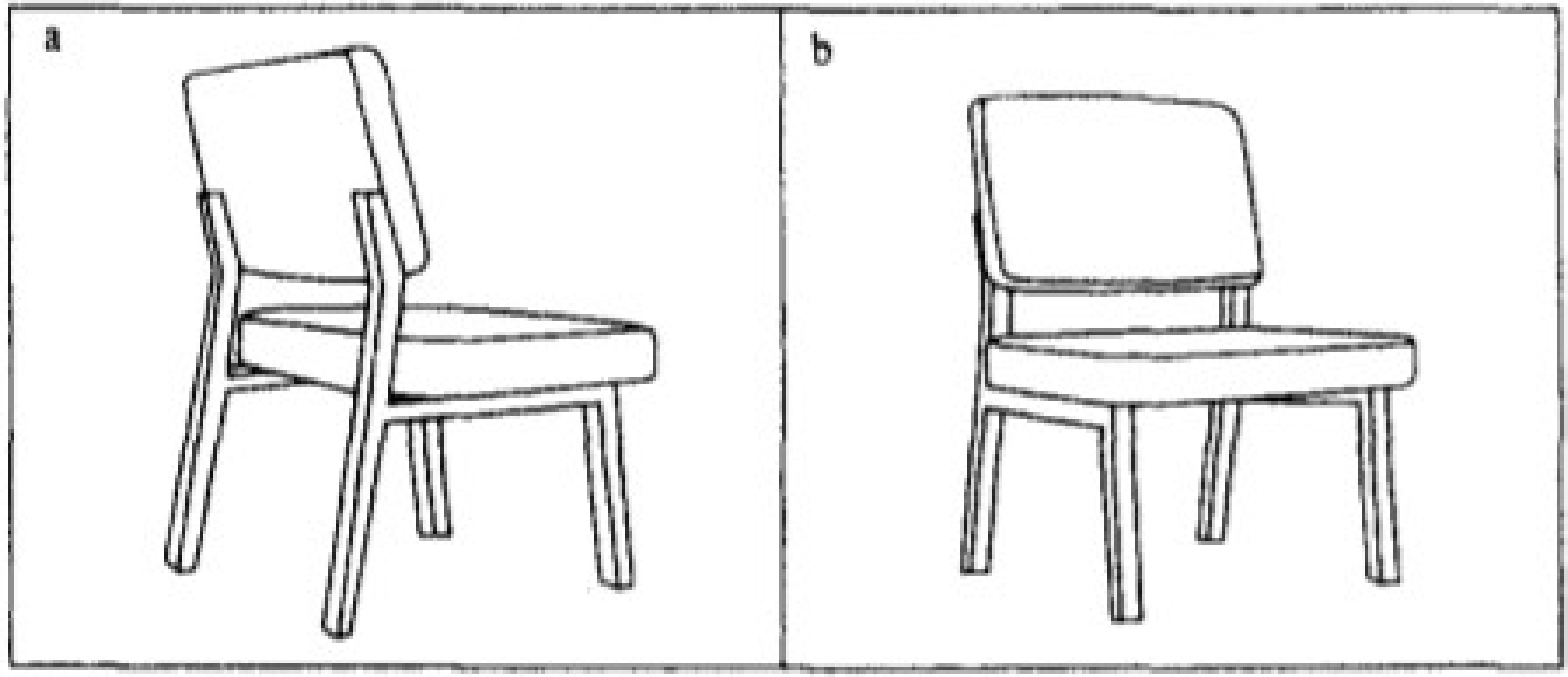


Figure 1. Two views of a chair. (The pose depicted in Panel a is a 90° clockwise rotation of the pose depicted in Panel b.)

- You can easily generalize from the view on the left to the view on the right.

- Although you have seen other chairs from other views, you haven't seen this exact geometry from other views.

parts-based recognition

- for familiar objects...
- humans recognize primitive parts
- assemble those primitive parts into a structural description
- use that structure to recognize the object in novel views

Biederman's paper

- Observed results in Bulthoff study are the result of selecting unnatural stimuli.
- Object recognition is frequently viewpoint invariant.
- Viewpoint invariance results from recognition of "geon" parts.

conditions for viewpoint invariance

- **paper starts off by examining different conditions for viewpoint invariance**

condition 1

- readily identifiable invariant parts
- prediction: required to show large degrees of viewpoint invariance
- explains absence of viewpoint invariance for crumpled paper, bent paperclips, lumps of clay: no parts, no GSD

condition 2

- distinctive structural descriptions
- prediction: objects with the same GSD will be easily confused across viewpoints
- explains absence of viewpoint invariance in experiments with multiple objects having the same GSD

condition 3

- identical structural descriptions in different viewpoints

- unusual viewpoints sometimes lead to different structural descriptions for the same 3D object
- these cause problems for viewpoint invariance in humans

Biederman's Experiments

- choose stimuli that have recognizable, distinctive parts
- choose views across which parts do not change

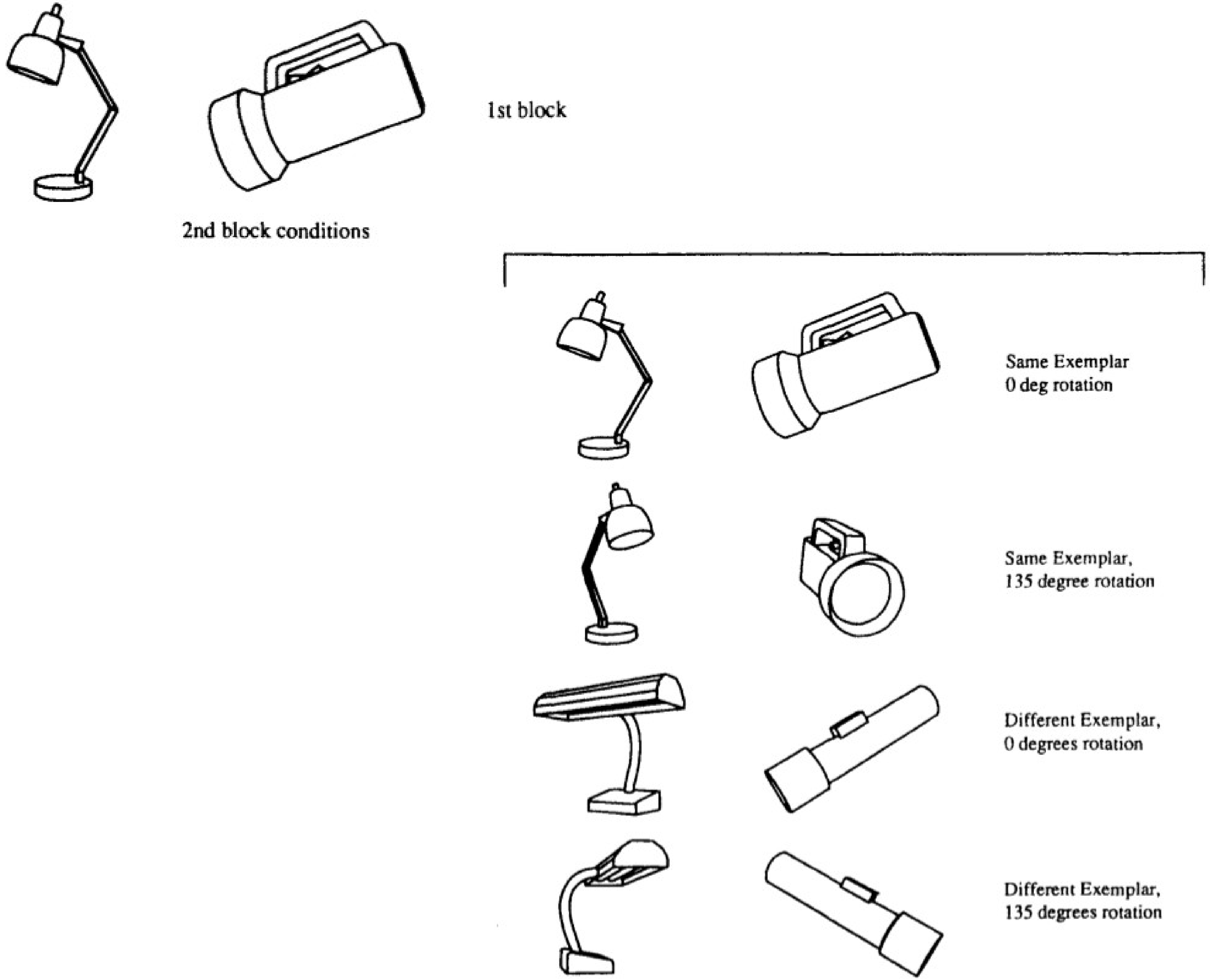
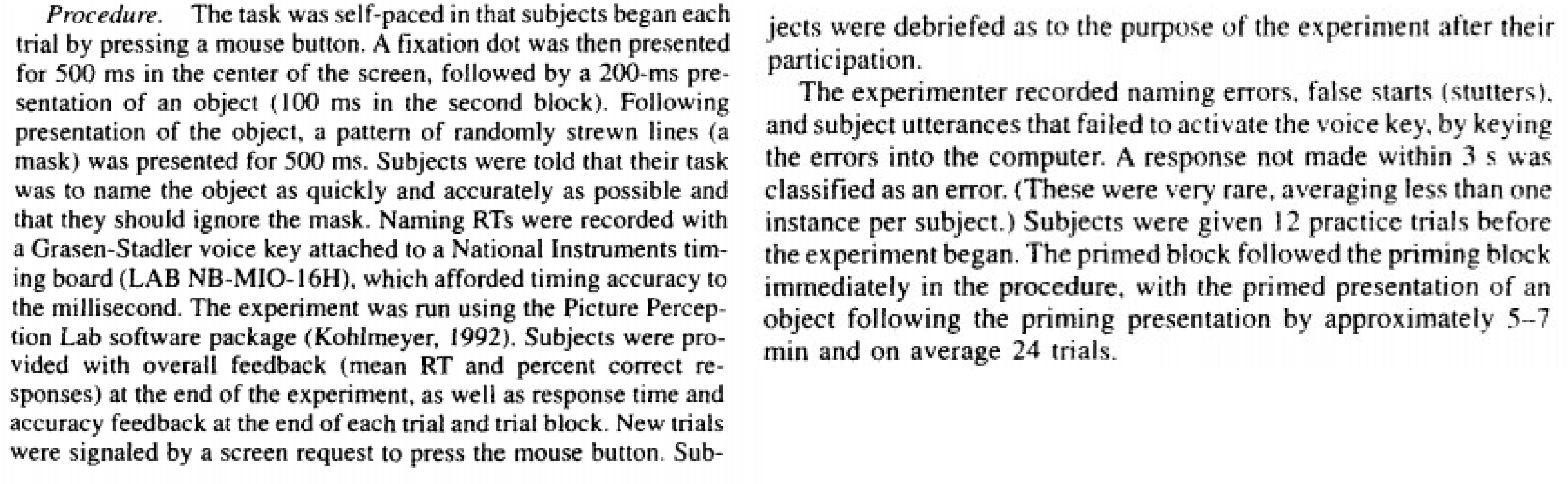


Figure 5. Two sample stimuli illustrating the same and different exemplar conditions, and the 0° and 135° rotation angles of Experiment 1 on name priming of familiar objects.

experimental task

- **task**
  - 500 ms fixation point
  - 100-200 ms presentation of object
  - 500 ms mask
  - name the object quickly and accurately (automatic detection)
- **primed and unprimed conditions**
  - subjects either saw or didn't see the target object beforehand
- **reduce parts changes between views**

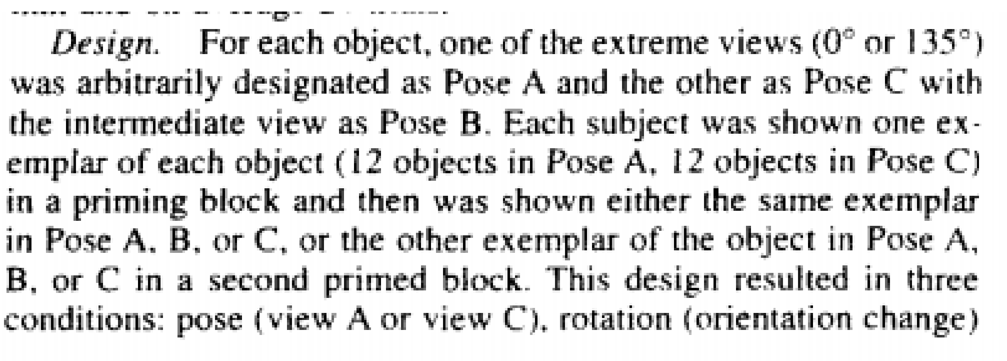
procedure

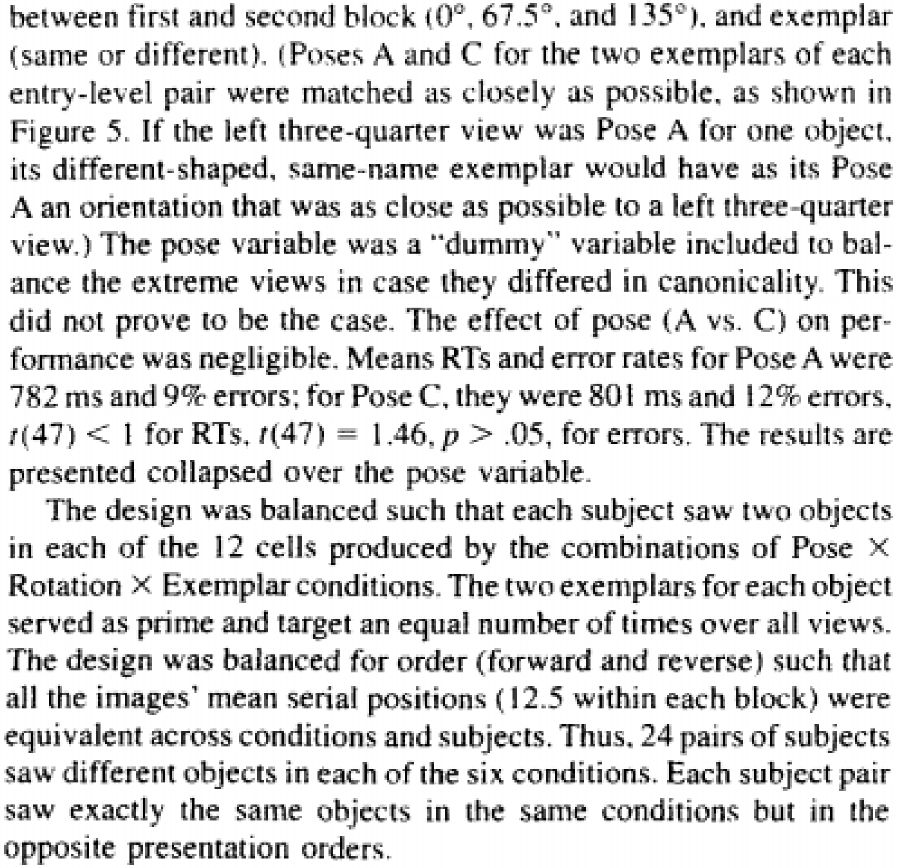


Procedure. The task was self-paced in that subjects began each trial by pressing a mouse button. A fixation dot was then presented for 500 ms in the center of the screen, followed by a 200-ms presentation of an object (100 ms in the second block). Following presentation of the object, a pattern of randomly strewn lines (a mask) was presented for 500 ms. Subjects were told that their task was to name the object as quickly and accurately as possible and that they should ignore the mask. Naming RTs were recorded with a Grasen-Stadler voice key attached to a National Instruments timing board (LAB NB-MIO-16H), which afforded timing accuracy to the millisecond. The experiment was run using the Picture Perception Lab software package (Kohlmeyer, 1992). Subjects were provided with overall feedback (mean RT and percent correct responses) at the end of the experiment, as well as response time and accuracy feedback at the end of each trial and trial block. New trials were signaled by a screen request to press the mouse button. Subjects were debriefed as to the purpose of the experiment after their participation.

The experimenter recorded naming errors, false starts (stutters), and subject utterances that failed to activate the voice key, by keying the errors into the computer. A response not made within 3 s was classified as an error. (These were very rare, averaging less than one instance per subject.) Subjects were given 12 practice trials before the experiment began. The primed block followed the priming block immediately in the procedure, with the primed presentation of an object following the priming presentation by approximately 5–7 min and on average 24 trials.

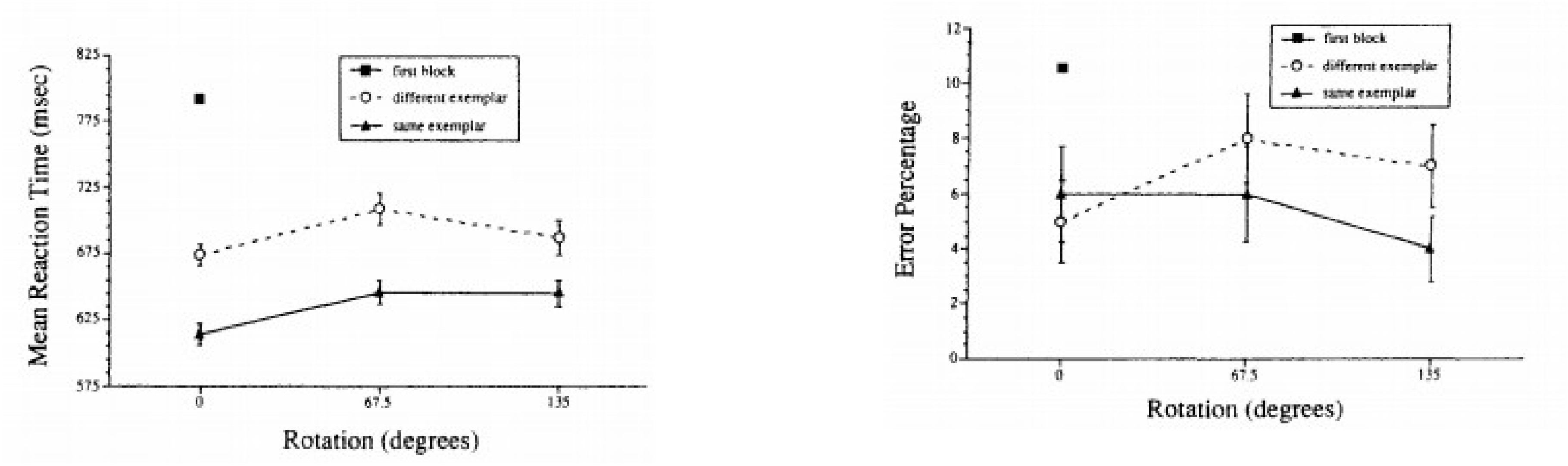
design





Design. For each object, one of the extreme views (0° or 135°) was arbitrarily designated as Pose A and the other as Pose C with the intermediate view as Pose B. Each subject was shown one exemplar of each object (12 objects in Pose A, 12 objects in Pose C) in a priming block and then was shown either the same exemplar in Pose A, B, or C, or the other exemplar of the object in Pose A, B, or C in a second primed block. This design resulted in three conditions: pose (view A or view C), rotation (orientation change) between first and second block (0°, 67.5°, and 135°), and exemplar (same or different). (Poses A and C for the two exemplars of each entry-level pair were matched as closely as possible, as shown in Figure 5. If the left three-quarter view was Pose A for one object, its different-shaped, same-name exemplar would have as its Pose A an orientation that was as close as possible to a left three-quarter view.) The pose variable was a "dummy" variable included to balance the extreme views in case they differed in canonicality. This did not prove to be the case. The effect of pose (A vs. C) on performance was negligible. Means RTs and error rates for Pose A were 782 ms and 9% errors; for Pose C, they were 801 ms and 12% errors, $t(47) < 1$ for RTs, $t(47) = 1.46$, $p > .05$, for errors. The results are presented collapsed over the pose variable.

The design was balanced such that each subject saw two objects in each of the 12 cells produced by the combinations of Pose × Rotation × Exemplar conditions. The two exemplars for each object served as prime and target an equal number of times over all views. The design was balanced for order (forward and reverse) such that all the images' mean serial positions (12.5 within each block) were equivalent across conditions and subjects. Thus, 24 pairs of subjects saw different objects in each of the six conditions. Each subject pair saw exactly the same objects in the same conditions but in the opposite presentation orders.



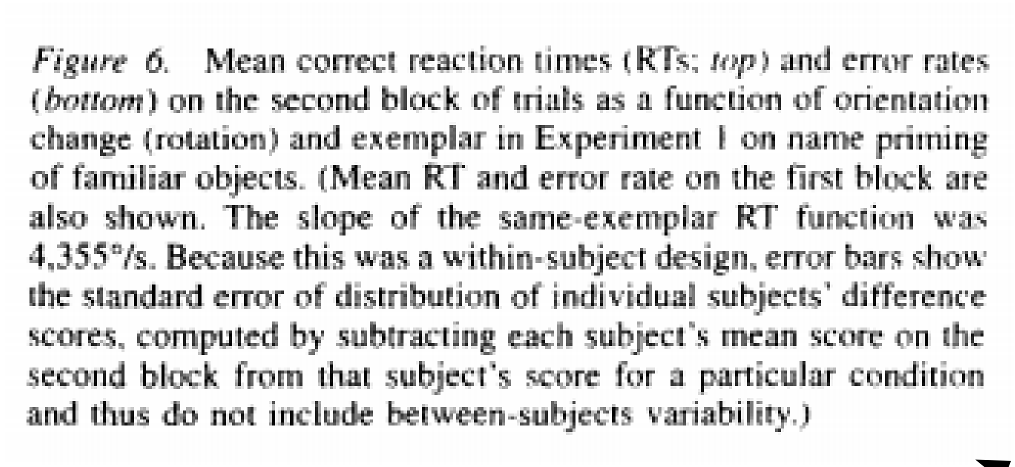
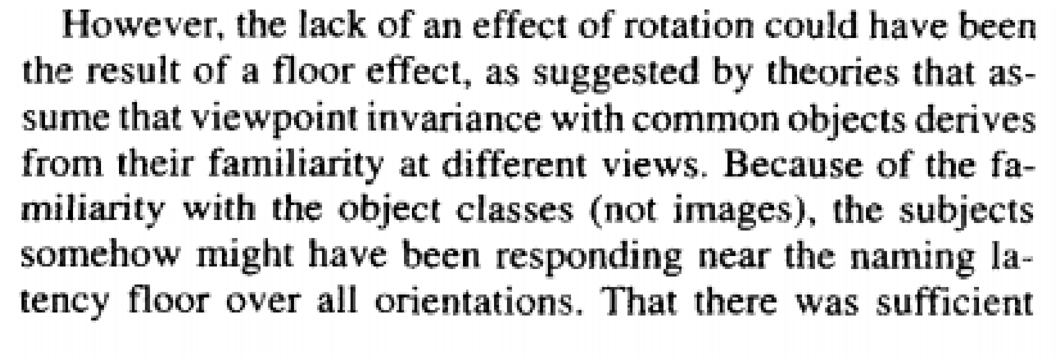


Figure 6. Mean correct reaction times (RTs; *top*) and error rates (*bottom*) on the second block of trials as a function of orientation change (rotation) and exemplar in Experiment 1 on name priming of familiar objects. (Mean RT and error rate on the first block are also shown. The slope of the same-exemplar RT function was 4,355°/s. Because this was a within-subject design, error bars show the standard error of distribution of individual subjects' difference scores, computed by subtracting each subject's mean score on the second block from that subject's score for a particular condition and thus do not include between-subjects variability.)

"These results indicate that visual priming of naming latency for familiar objects was relatively insensitive to changes in depth orientation occurring between priming and primed images."

note good figure caption

possible objection



However, the lack of an effect of rotation could have been the result of a floor effect, as suggested by theories that assume that viewpoint invariance with common objects derives from their familiarity at different views. Because of the familiarity with the object classes (not images), the subjects somehow might have been responding near the naming latency floor over all orientations. That there was sufficient

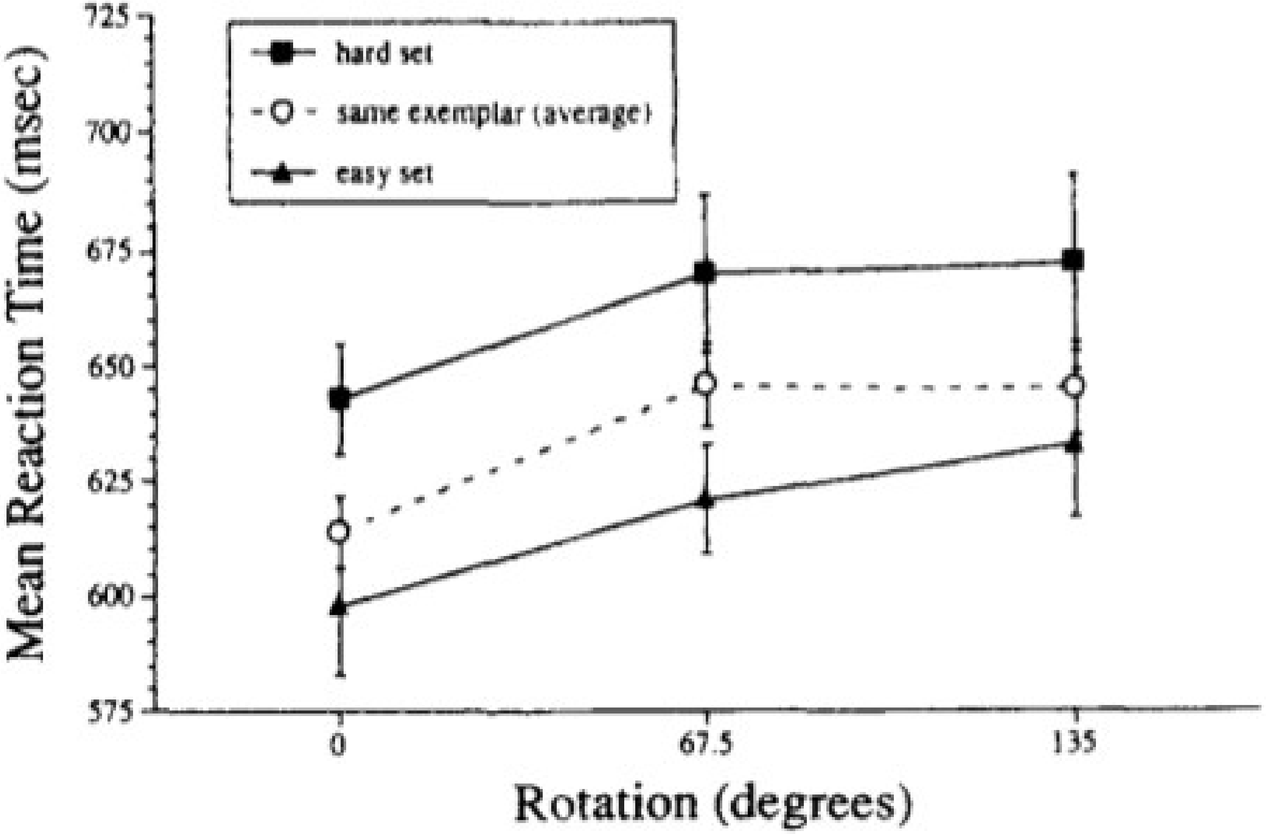
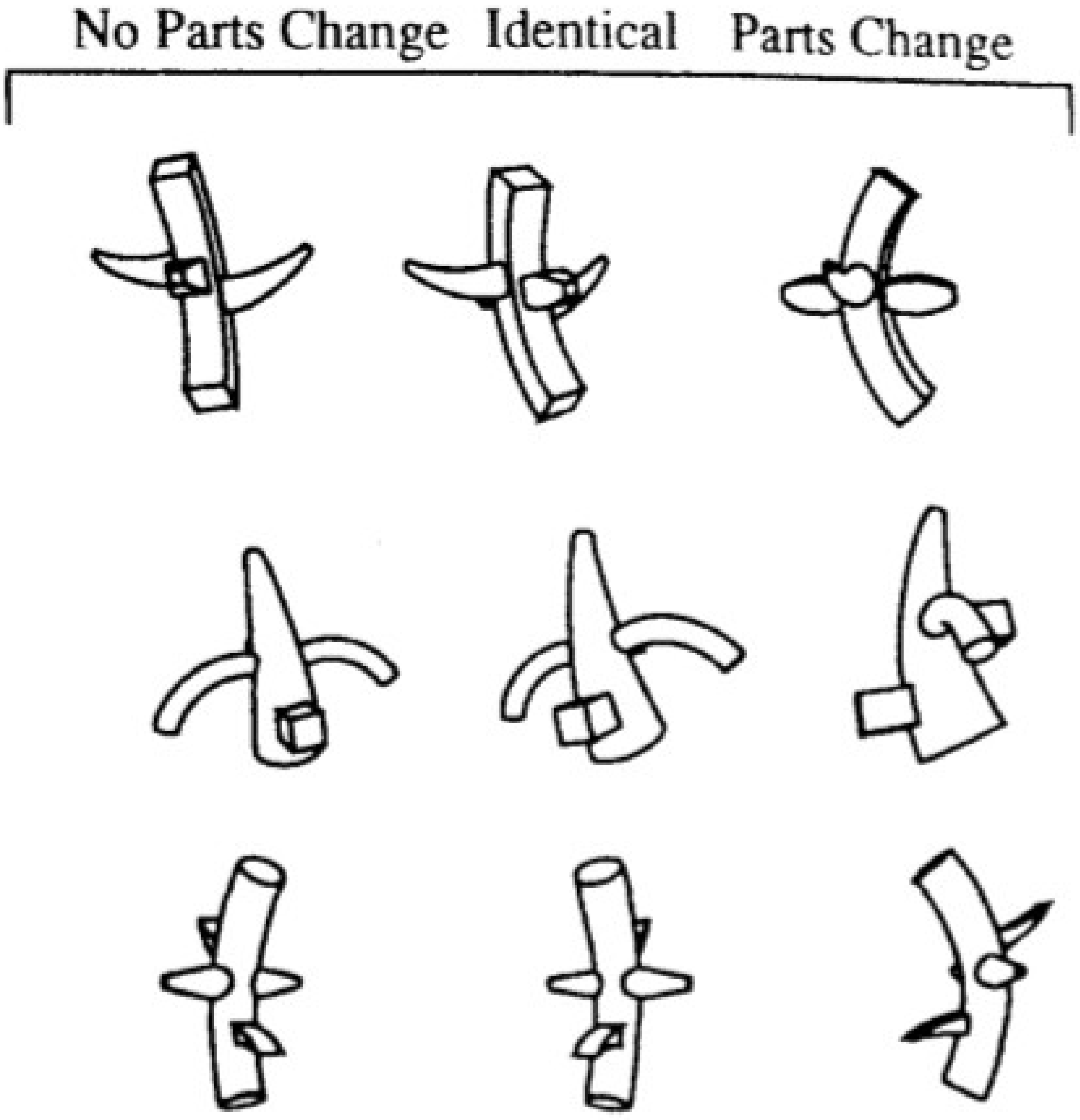


Figure 7. Mean correct naming reaction times for the hard and easy members of each exemplar pair, as determined by a post hoc ranking by mean reaction time for each pair of same-name exemplars, collapsed across object and plotted as a function of orientation change [rotation] from Block 1 in Experiment 1. (Because this was a within-subject design, error bars show the standard error of distribution of individual subjects' difference scores, computed by subtracting each subject's mean score on the second block from that subject's score for a particular condition, and thus do not include between-subjects variability.)

no floor effect - “hard and easy groups vary in the same way across conditions”

recognition with and without parts changes



NB: inspired as geon-variant of another experiment prediction?

results

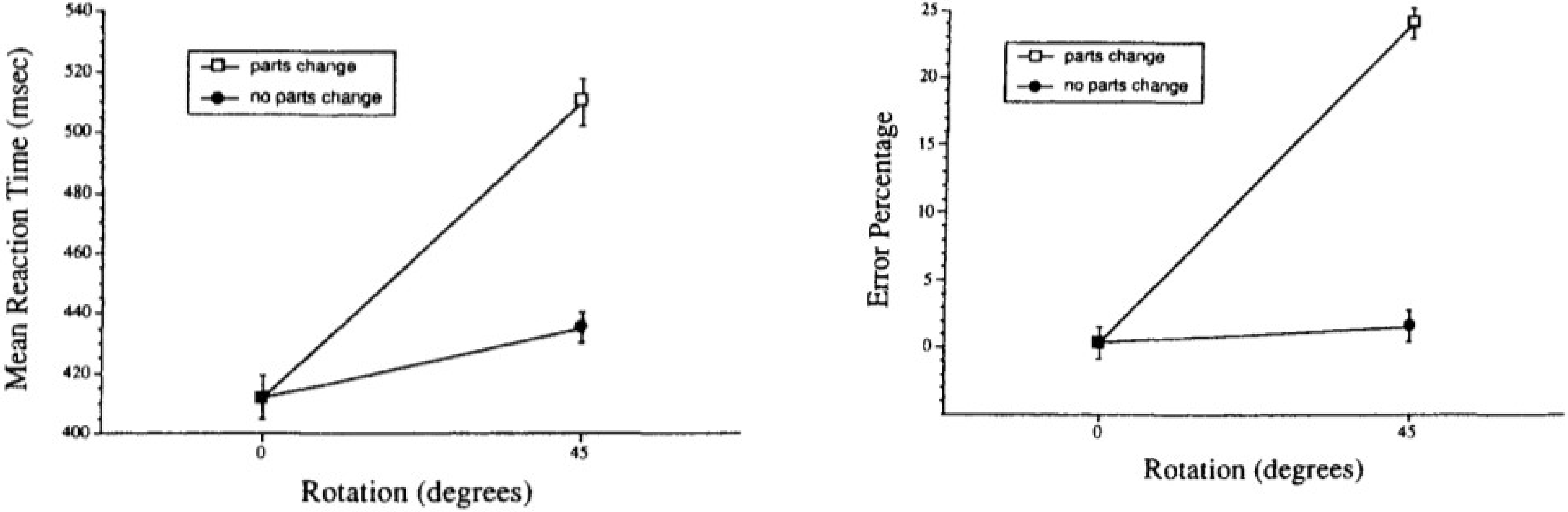
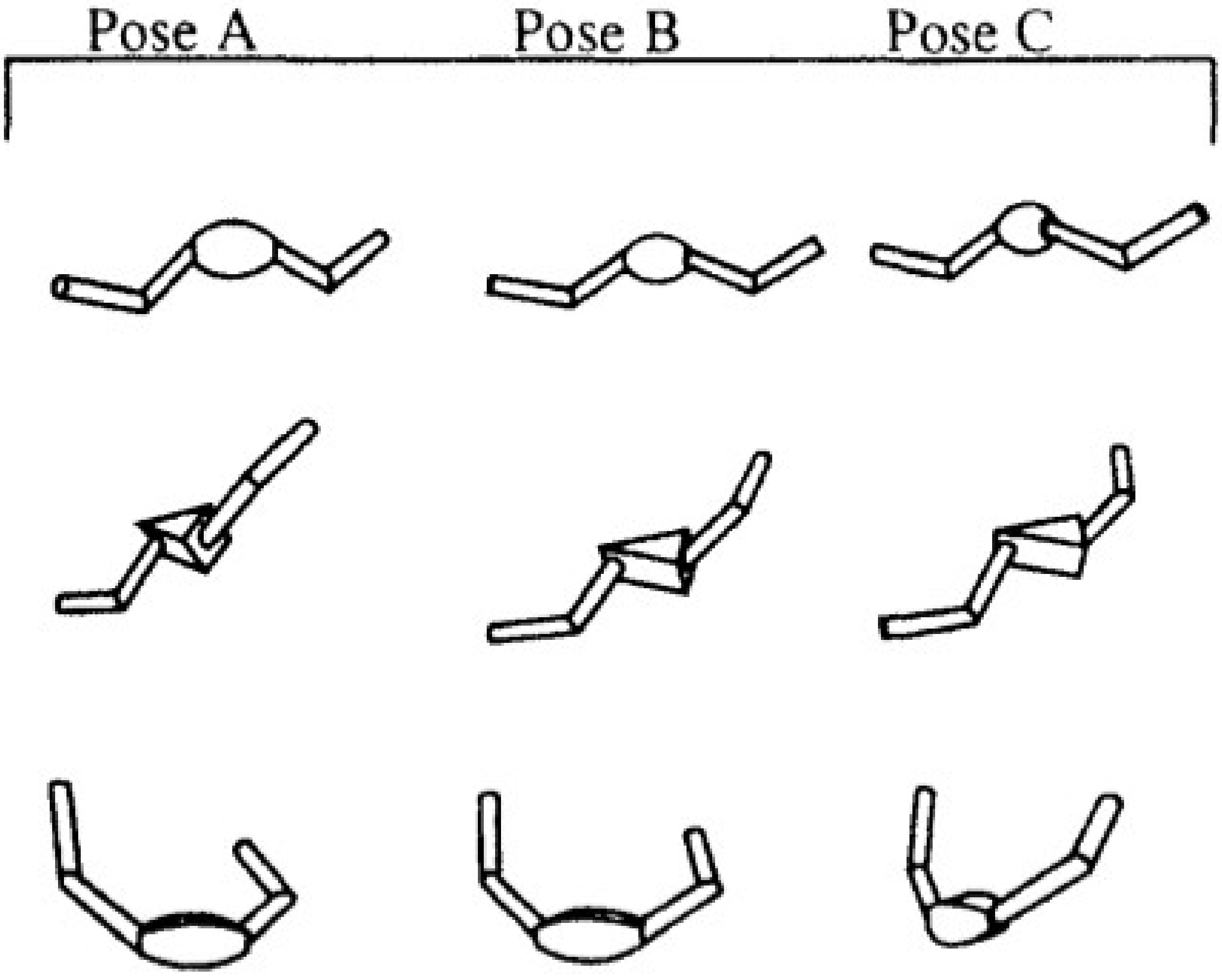


Figure 11. Mean correct reaction times (*top*) and error rates (*bottom*) for same–different judgments of unfamiliar objects in Experiment 3 as a function of the degree of angular change between the first and second exposures on a trial and indication of whether that rotation produced a part change. (The speed for the no-parts-change reaction time function was 1,875°/s. For the parts-change function, the speed was 459°/s. Because this was a within-subject design, error bars show the standard error of distribution of individual subjects' difference scores, computed by subtracting each subject's mean score from that subject's score for a particular condition, and thus do not include between-subjects variability.)

consistent with geon model

adding a distinctive geon to viewpoint-dependent objects



prediction?

results on paperclips with geons

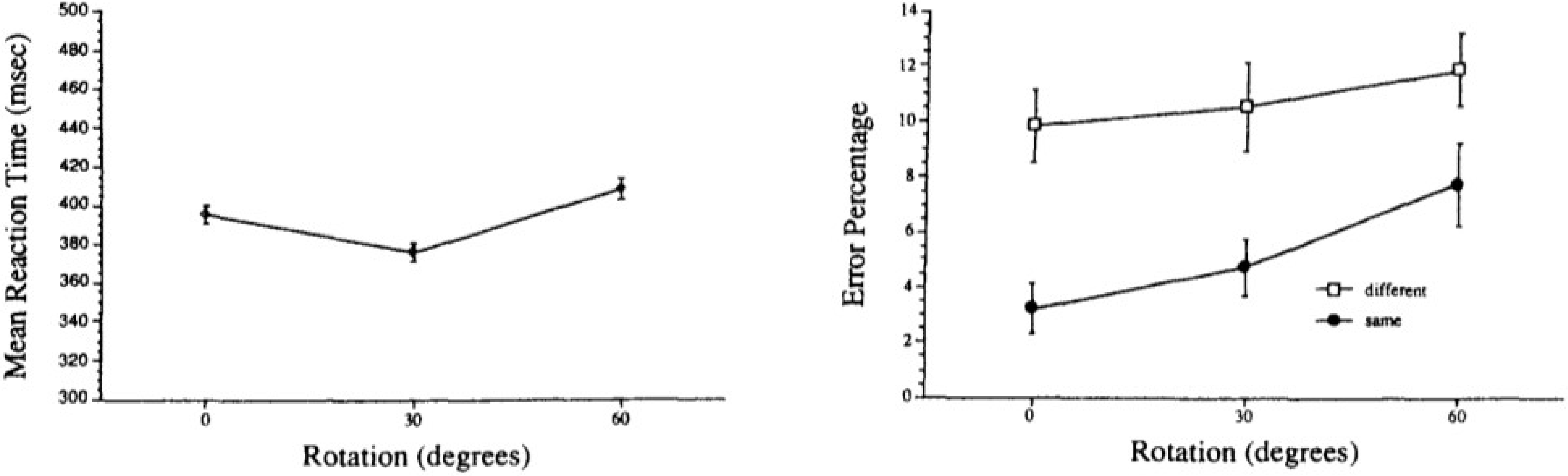


Figure 15. Mean correct reaction times (RTs; *top*) and error rates (*bottom*) in Experiment 5 for same–different judgments of two nonsense objects (the charm bracelets) as a function of rotation angle relative to the studied target orientation. (The RT rotation rate was 5,000°/s. As this was a go–no-go task, only "yes" RTs were obtained. Because this was a within-subject design, error bars show the standard error of distribution of individual subjects' difference scores, computed by subtracting each subject's mean score from that subject's score for a particular condition, and thus do not include between-subjects variability.)

Biederman's conclusions

- 3D invariance is common and based on recognizing parts
- Bulthoff's results are the consequence of picking very unusual, unfamiliar kinds of stimuli

Discussion?