Psychometrics 101: How to Design, Conduct, and Analyze Perceptual Experiments in Computer Graphics

Course organizer

James A. Ferwerda
Program of Computer Graphics
Cornell University

Presenters

James A. Ferwerda
Program of Computer Graphics
Cornell University
jaf@graphics.cornell.edu

Holly Rushmeier
IBM TJ Watson Research Center
hertjwr@us.ibm.com

Benjamin Watson
Dept. of Computer Science
Northwestern University
watsonb@cs.northwestern.edu

Summary statement

Psychometric methods from experimental psychology can be used to quantify the relationships between the properties of images and what people perceive. This course will provide an introduction to the use of psychometric methods in computer graphics, and will teach attendees how to design perceptual experiments that can be used to advance graphics research and applications. This course will be of interest to members of the graphics community who want to be able to interpret the results of perception psychology experiments and develop their own perceptual studies of computer graphics techniques.

Prerequisites

This course assumes a basic level understanding of issues in computer graphics and electronic imaging. Familiarity with freshman-level college math will be helpful. No specific knowledge of perception psychology or statistical methods is necessary.

Course syllabus

- **I. Welcome, Introductions, Schedule Review (Ferwerda, 10 mins)**

- **II. Motivation/Orientation (Rushmeier, 35 mins)**
  - **Why psychometrics?**
    - graphics are generated to be useful to people
    - we need to be able to determine what factors contribute to visual experience
    - we need to be able to assess what methods produce an effective visual experience
  - **Why don't we just use existing psychophysical results?**
    - graphics builds on psychophysical research (e.g. colorimetry)
    - goals of psychophysical research are different than graphics research
      - determining contrast sensitivity vs. designing a rendering method that uses a model of contrast sensitivity
  - **What are example problems to be addressed by psychometrics?**
    - realistic image synthesis
      - how accurate does the input need to be?
      - what input is needed?
      - how accurate does the light transfer need to be?
      - how are the results in physical units transformed to displays?
    - data visualization
      - how should data values be mapped to visual attributes?
      - how effective are different visual cues for conveying information about data
      - what are the interactions between these different cues?
      - how can we make sure that the images we create are faithful representations
    - virtual environments
      - what trade-offs are acceptable to maintain real time performance?
      - what spatial representations are adequate?
      - what are the perceptual differences between screen-based and immersive displays?
    - compression
      - what kinds of artifacts are visually acceptable in still images? In temporal sequences? In 3D geometric models?
    - animation
  - **How does psychometrics relate to physical measurement?**
    - human observers are responding to physical stimulus
    - depending on problem various physical measurements also needed
      - object shape/material properties; light energy from real scenes/displays
  - **What kind of results can we expect?**
    - more efficient graphics techniques -- computing only what is necessary
    - more effective graphics techniques -- choosing the right image to generate
  - **How do we make progress?**
    - adopt established experimental methods
    - build a literature of results relevant to graphics techniques

- **III. Psychophysical Methods (Ferwerda, 60 mins)**
  - **Introduction**
    - need for objective metrics of subjective visual experience
    - fundamental psychophysical metrics: **thresholds** and **scales**
    - history: Weber, Fechner
  - **Methods for measuring thresholds**
    - the method of adjustment
    - the method of limits
      - adaptive methods
    - the method of constant stimuli
  - **Threshold models**
    - psychometric functions
  - **Signal detection theory**
    - variation in threshold measurements
    - the signal detection problem
    - stimulus/response (SR) matricies
    - the decision criterion
    - measures of sensitivity and response bias
  - **Suprathreshold scaling methods**
    - types of psychophysical scales
      - nominal, ordinal, interval, ratio
    - indirect scaling methods
      - rating
      - pair comparison
      - ranking
      - category scaling
    - direct scaling methods
      - equisection/fractionation
      - magnitude production
      - magnitude estimation
  - **Scaling models**
    - Weber's law, Fechner's law
    - Steven's power law
  - **Multidimensional scaling (MDS)**
  - **Practicalities of running psychophysical experiments**
    - stimulus selection
    - display/interface issues
    - selecting subjects
    - experimental design
    - data analysis
  - **Summary**
  - **Resources**
    - books
    - papers/standards
    - software packages

- **Break**

- **IV. Experimental Design (Watson, 60 mins)**
  - **Why do an experiment?**
    - real world context: an example experiment from the graphics literature
  - **Global experimental issues**
    - internal/external validity
    - feasibility
    - non-null results
  - **Design components**
    - hypothesis
    - independent/dependent/control variables
    - test or task
  - **Threats to validity, feasibility and non-null results**
    - to internal validity
      - randomness, confounds, individual differences, carryover effects, reactivity, researcher bias
    - to external validity
      - unreliability, too much control, unrepresentative population
    - to feasibility
      - population too big, design too ambitious, task too hard
    - to non-null results
      - ceiling and floor effects, type 2 errors
  - **Standard experimental designs**
    - single factor designs
      - within/between subject designs
    - multi-factor designs
      - interactions
    - mixed designs
    - repeated measures designs
  - **Analysis of results**
    - analytical tools
      - descriptive statistics
      - inferential statistics
    - ethics of analysis
      - excluding participants
      - excluding results
  - **Practical questions**
    - how many participants?
    - getting approval (human subjects committees)
    - motivating participants
    - the hunt for significance
      - pilot studies

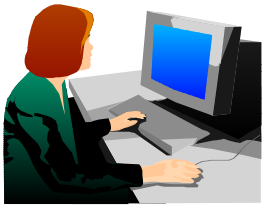
- **V. Case Studies of Psychometric Methods in Graphics (Rushmeier, 25 mins)**
  - realistic image synthesis
  - animation
  - data visualization
  - virtual environments

- **VI. Panel / Group Discussion (All, 20 mins)**
  - review of day's material
  - pointers to resources
  - open questions

- **VII. Supplementary Materials**
  - Rogowitz, B., and Rushmeier, H.E. (2001) Are image quality metrics adequate to evaluate the quality of geometric objects? Proceedings SPIE Vol 4299 Human Vision and Electronic Imaging VI,1-9.
  - Rushmeier, H.E., Rogowitz, B.E., Piatko, C. (2000) Perceptual issues in substituting texture for geometry. Proceedings SPIE Vol 3959 Human Vision and Electronic Imaging V, 372-383.
  - Meyer, G.W., Rushmeier, H.E., Cohen, M.F., Greenberg, D.P., and Torrance, K.E. (1986) An experimental evaluation of computer graphics imagery. ACM Transactions on Graphics, 5(1), 30-50.
  - Pellacini, F., Ferwerda, J.A., and Greenberg.D.P. (2000) Toward a psychophysically-based light reflection model for image synthesis. Proceedings SIGGRAPH '00, 55-64.
  - Watson, B.A., Friedman, A., McGaffey, A. (2001) Measuring and predicting visual fidelity. Proceedings SIGGRAPH 2001, 213-220.

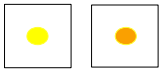
Motivation/Orientation

Why psychometrics?

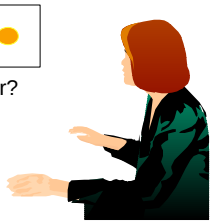


- graphics are generated to be useful to people
- we need to be able to determine what factors contribute to visual experience
- we need to be able to assess what methods produce an effective visual experience

Why not use existing psychophysical results?

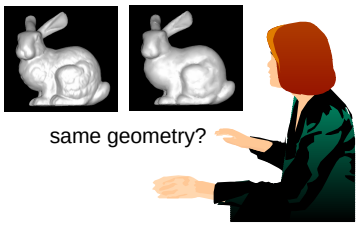


same color?



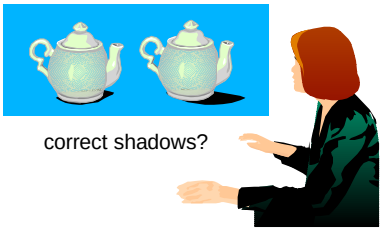
- graphics does use existing psychophysical research
- prime example is colorimetry, we routinely display colors that look the same as objects in the real world although their actual reflected spectra cannot be produced by the display device
- existing research is used in many different ways in graphics, ranging from "rules of thumb" to reliable numerical algorithms

We have new problems



- goals of psychophysical research are different than graphics research, e.g. determining contrast sensitivity (vision research) vs. designing a rendering method that uses a model of contrast sensitivity (computer graphics)
- we have issues that have never come up in vision research -- such as the spatial resolution required for faithful representation of a 3D surface

We have new capabilities



- New controlled experiments are possible using relatively new computer graphics techniques. Visual effects like shadows can be precisely controlled in synthetic imagery that can't be controlled physically, letting vision researchers explore questions such as how shadows affect our judgements of size and position. Graphics researchers can exploit these new insights in designing efficient rendering algorithms.
- Ref: M.S. Langer and H.H. Buelthoff, "Measuring Visual Shape using Computer Graphics Psychophysics" Rendering Techniques 2000, Springer-Verlag.

Problems that can be addressed by psychometrics

What should be displayed to achieve a particular goal?

Given limited resources what is most important to be computed/stored?

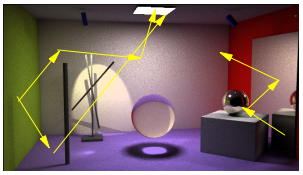
- What should be displayed to achieve a particular goal?
- - mapping visual attributes to non-visual data
- - identifying relevant input data required
- Given limited resources what is most important to be computed?
- - limited time
- - limited memory
- - limited display device
- - limited bandwidth

Example: Realistic Image Synthesis


What input ? How accurate?

- To simulate these photographs of two different teapots, how accurate do the geometric descriptions have to be to show the difference in shapes? How are the surface properties represented to show that one is light and diffuse and the other is dark and shiny? How accurately do we have to measure the input?
- Sample Refs:
- "Toward a psychophysically-based light reflection model for image synthesis" Fabio Pellacini, James A. Ferwerda Donald P. Greenberg, SIGGRAPH 2000.
- "Measuring and predicting visual fidelity", Benjamin Watson, Alinda Friedman, Aaron McGaffey, SIGGRAPH 2001

Example: Realistic Image Synthesis

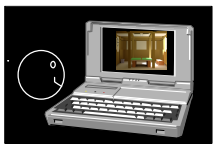
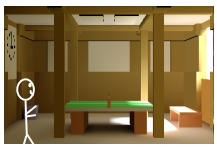


How accurate does the light transfer need to be?

- A realistic image can be generated by simulating the paths of millions of light rays. How many are enough to sample so that the result is indistinguishable from the real thing?
- Sample Refs:
- M.R. Bolin and G.W. Meyer. A Perceptually Based Adaptive Sampling Algorithm. In SIGGRAPH 98 Conference Proceedings, Annual Conference Series, pages 299–310, 1998.
- K. Myszkowski, P. Rokita, and T. Tawara. Perceptually-Informed Accelerated Rendering of High Quality Walkthrough Sequences. In Eurographics Rendering Workshop 1999, pages 5–18, 1999.
- M. Ramasubramanian, S.N. Pattanaik, and D.P. Greenberg. A Perceptually Based Physical Error Metric for Realistic Image Synthesis. In SIGGRAPH 99 Conference Proceedings, Annual Conference Series, pages 73–82, 1999.
- Y.L.H. Yee. Spatiotemporal Sensitivity and Visual Attention for Efficient Rendering of Dynamic Environments. M.Sc. thesis, Cornell University, 2000.

Example: Realistic Image Synthesis

Viewers should have same visual impression

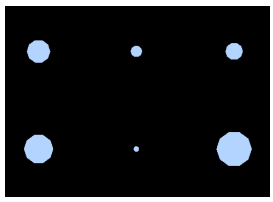
- We can calculate radiances that aren't displayable on our viewing device. What is the best way to map onto the limited device?
- Sample Refs:
- Ward-Larson, G., Rushmeier H., and Piatko, C. (1997) A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Trans. on Vis. and Comp. Graph., 3(4):291-306.
- Jack Tumblin , Jessica K. Hodgins , Brian K. Guenter, Two methods for display of high contrast images, ACM Transactions on Graphics (TOG), v.18 n.1, p.56-94, Jan. 1999
- Sumanta N. Pattanaik , James A. Ferwerda , Mark D. Fairchild , Donald P. Greenberg, A multiscale model of adaptation and spatial vision for realistic image display, SIGGRAPH p.287-298, July 19-24, 1998.

Example: Data Visualization

How do you make pictures of numbers?

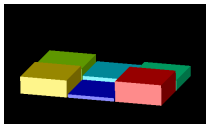
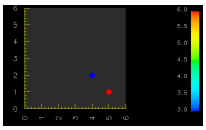
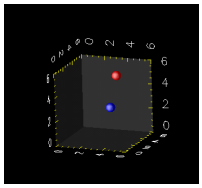
make a circle for each
with a size related to the number

4 2 3
5 1 6



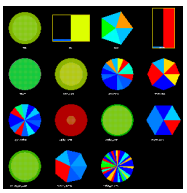
- How should data values be mapped to visual attributes? For some data this can be obvious, i.e. geographic data on a map. For other types of data, there are no obvious physical representations.

Example: Data Visualization

- Even for a small group of numbers, there are an infinite number of possible representations. Which if any are of any use? What does a person get out of looking at them?

Example: Data Visualization

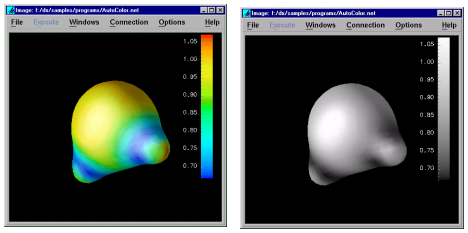


Visualizing 15
variables, numerical
and categorical

- It is possible to generate imagery of huge numbers of variables which may be scalar, vector, textual etc. How much can a person understand about such complex data from an image? Current graphics systems expand our choices about what we can display, and we need to keep up with understanding what is useful to display.
- Ref:
- Healey, C. G. and Enns, J. T. "Building Perceptual Textures to Visualize Multidimensional Datasets" In Proceedings IEEE Visualization '98 (Research Triangle Park, North Carolina, 1998), pp. 111-118.
-

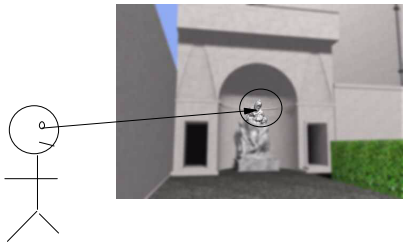
Example: Data Visualization

Example of specific study area: Color Maps



- Problem is not how do we reproduce a particular color, but what set of colors best represents a set of values.
- A Rule-based Tool for Assisting Colormap Selection, L. Bergman, B. Rogowitz and L. Treinish. Proceedings of the IEEE Computer Society Visualization '95 pp. 118-125, October 1995.
- The 'Which Blair Project:' A Quick Visual Method for Evaluating Perceptual Color Maps, Bernice E. Rogowitz and Alan D. Kalvin, Proceedings of IEEE Visualization 2001.
-

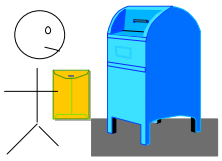
Example: Virtual Environments



- what trade-offs are acceptable to maintain real time performance? what spatial representations are adequate?
- what are the perceptual differences between screen-based and immersive displays? Can we take advantage of capabilities like eye tracking?
- Ref: B.A. Watson, N. Walker, L.F. Hodges & M. Reddy (1997). An evaluation of level of detail degradation in head-mounted display peripheries . Presence, 6, 6, 630-637.
-

Example: Virtual Environments

How are virtual environments modeled for effective training applications?



- How well does an environment need to be rendered to train for a task, i.e. such a putting an object (letter) in a specific place (mailbox)?

Example: Virtual Environments

What contributes to a sense of "presence"?

- What are the factors that make a person feel like they are actually in the environment that is presented virtually?
- Ref: M. Slater (1999)" Measuring Presence: A Response to the Witmer and Singer Questionnaire," Presence: Teleoperators and Virtual Environments, 8(5), 560-566.
-

Example: Compression/Transmission

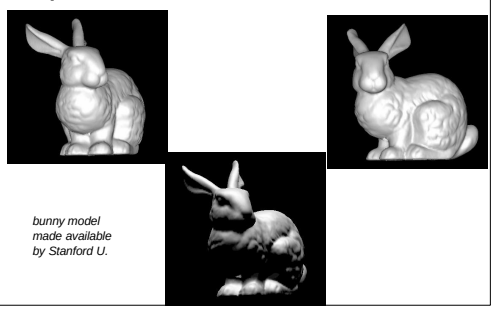
23 kB 11 kB 5 kB

- In image compression there has been great success in applying human vision characteristics to reducing data in a way that image file size correlates well with perceived image quality, as in these jpeg iimages shown here.

Example: Compression/Transmission

- We would like a similar scheme for geometry -- we can simplify geometry in many ways, but which techniques let us control perceived/quality versus file size?

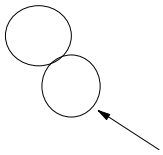
Example: Compression/Transmission



bunny model made available by Stanford U.

- The geometry problem is made more complex by the fact that there are an infinite number of views and lighting conditions.
- Ref.: B.E. Rogowitz and H.E. Rushmeier, "Are image quality metrics adequate to evaluate the quality of geometric objects?" Proceedings of SPIE Vol 4299 Human Vision and Electronic Imaging VI, 340-349, (2001)
-

Example: Animation

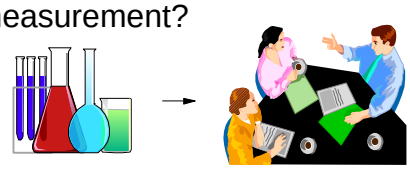


collisions

level of detail

- The fact that we perceive objects as moving when we flip through images, and how fast the images must be displayed is based on measurements of human perception.
- Computing the dynamics of animations can be hugely expensive -- how accurately do we need to compute to spheres colliding to "believe" they intersect? How accurately do we have to compute the fluttering of the jockey's jacket to be consistent with how fast the horse is running?
- Refs:Carlson D.A. and Hodgins J.K. – Simulation Levels of Detail for Real-Time Animation. Proc. of Graphics Interface '97. pp. 1-8.
- Collisions and Perception. O'Sullivan, C. Dingliana, J. ACM Transactions on Graphics. Vol. 20, No. 3. July 2001.
-

How does psychometrics relate to physical measurement?



Both have standard metrics and measurement methods

- Most people have school experience with physical sciences, and experiments involving measuring quantities such as time, length and temperature, but not experience in psychophysical experiments. As a result, we often rely on only anecdotal results rather than conducting experiments. The purpose of this course is to introduce to the standard metrics and methods used in psychophysics.

What kind of results can we expect?

More efficient graphics techniques: computing only what is necessary

What kind of results can we expect?

More effective graphics techniques: choosing the right image to generate

How do we make progress?

- adopt established experimental methods
- build a literature of results relevant to graphics techniques

- Besides mining the existing vision literature, new experiments are needed to gain insight into complex problems relevant to graphics applications. New experiments are also needed to examine the success or failure of new graphics techniques.

Psychophysical methods

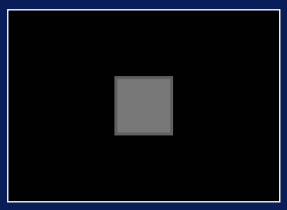
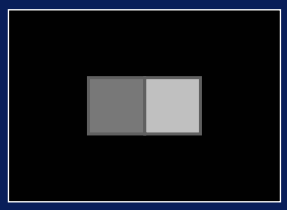
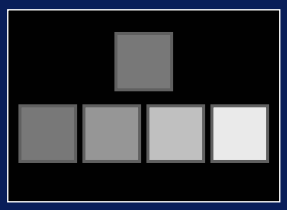
Jim Ferwerda
Program of Computer Graphics
Cornell University

Psychophysics

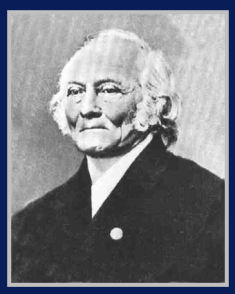
- Physical properties of objects (length, weight, intensity)
- Perceptual impressions on observer (size, heaviness, brightness)
- Measured directly
- Inferred from observer's responses

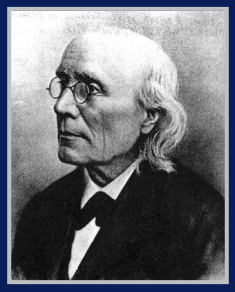
Goal: tools to quantify the relationships between physical stimulation and perceptual sensation

Psychophysical issues: thresholds and scales

detection	discrimination	scaling
		
how bright?	how much brighter?	twice as bright?
absolute threshold	difference threshold (JND)	suprathreshold appearance

History



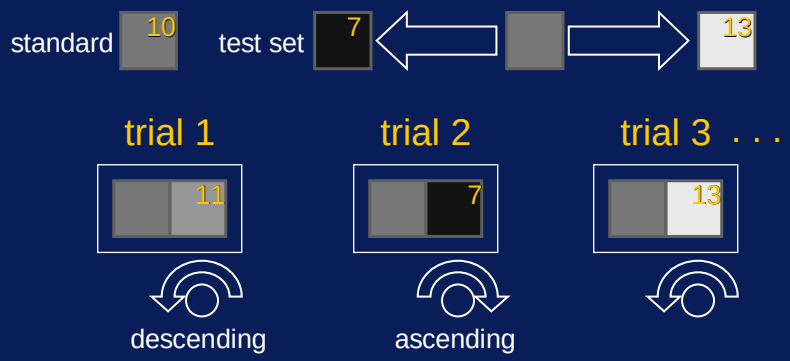


- Ernst Heinrich Weber (1795-1878)
  - experiments with weights (1830's)
  - just noticeable differences (JNDs) are proportional to stimulus magnitude
    - $\Delta I = k\ I$ (Weber's law)
- Gustav Teodore Fechner (1801-1887)
  - how to measure sensations?
  - need a zero and a unit
  - zero = absolute threshold
  - unit = difference threshold (JND)
  - "Elements of Psychophysics" (1860)
    - psychophysical methods
    - Fechner's Law ($S = k \log I$)

Psychophysical methods for measuring thresholds

- method of adjustment
- method of limits
- method of constant stimuli

Method of adjustment



start intensity	11	7	13
final intensity	10.19	10.41	9.52

- adjust intensity until target is just visibly different than the standard

Method of adjustment: analysis

number of trials 20

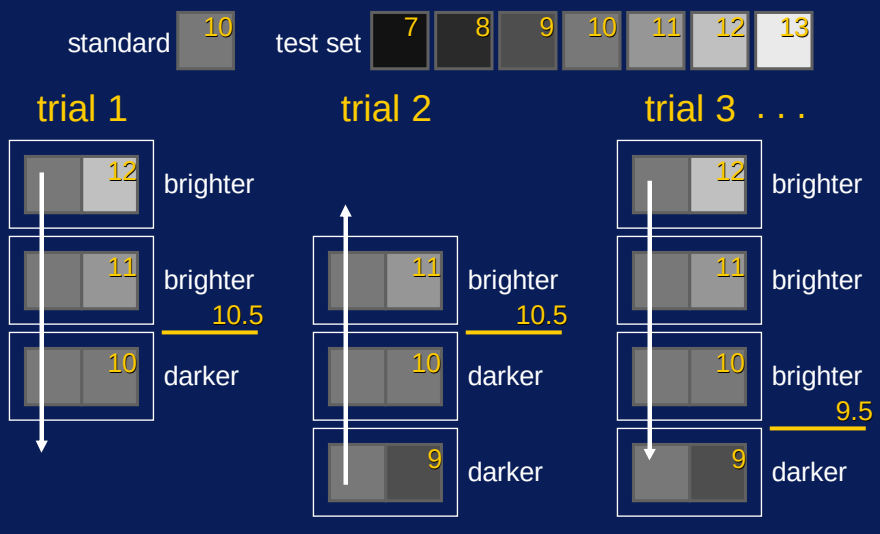
trial	1	2	3	4	5	6	7	8	9	10
series	d	a	d	a	d	a	d	a	d	a
start intensity	11.00	7.00	13.00	9.00	12.50	8.50	11.50	7.50	12.00	8.00
final intensity d	10.19		9.52		11.15		11.79		12.86	
final intensity a		10.41		7.08		8.15		12.22		9.29

series	mean	stdev
d	9.89	1.76
a	10.08	1.81

grand	mean	stdev
	9.99	1.74

- point of subjective equality (PSE) = grand mean = 9.99
- just noticeable difference (JND) = 0.67449 * stdev = 1.17
- upper threshold (UL) = PSE + JND = 11.16
- lower threshold (LL) = PSE - JND = 8.81
- interval of uncertainty (IU)= UL - LL = 2.35

Method of limits



- is the target brighter or darker than the standard?

Method of limits: analysis

number of trials: 20

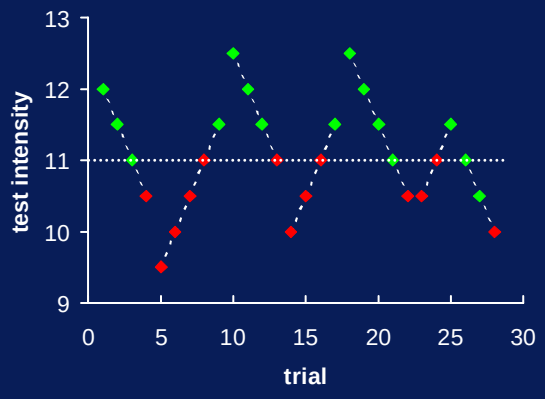
trial		1	2	3	4	5	6	7	8	9	10
series		d	a	d	a	d	a	d	a	d	a
test intensity	13								B		
	12	B		B		B			D		
	11	B	B	B		D		B			B
	10	D	D	B				B		B	D
	9		D	D				B		B	D
	8				B		B	D		D	D
	7				D		D				
crosspts.	d	10.5		9.5		11.5		8.5		8.5	
	a		10.5		7.5		7.5		12.5		10.5

series	mean	stdev
d	10.20	1.77
a	9.80	1.77

grand	mean	stdev
	10.00	1.73

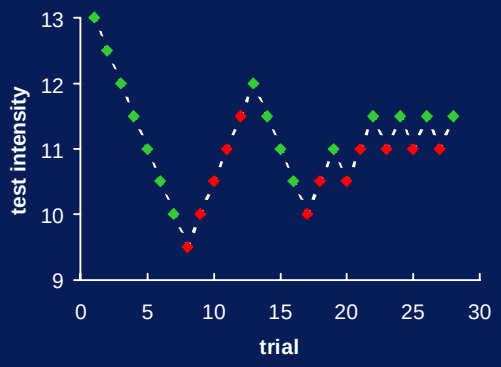
- point of subjective equality (PSE) = grand mean = 10.00
- just noticeable difference (JND) = 0.67449 * stdev = 1.17
- upper threshold (UL) = PSE + JND = 11.17
- lower threshold (LL) = PSE - JND = 8.83
- interval of uncertainty (IU)= UL - LL = 2.34

Variations on the method of limits



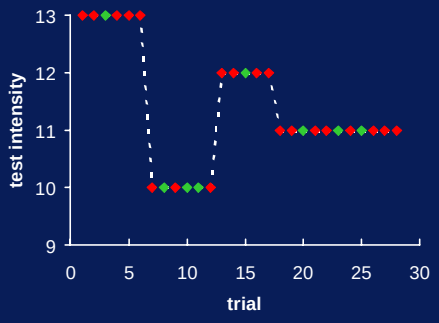
- staircase methods
- PEST
- QUEST

Staircase methods



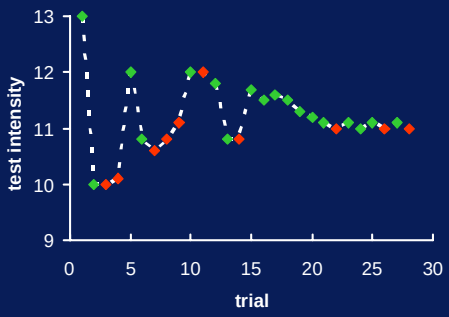
- next series starts where last ends
- increased efficiency over method of limits

PEST



- Parameter Estimation by Sequential Tracking
- test intensity chosen to achieve target performance level
- step size reduced on reversals

QUEST



- QUEST
- test intensity chosen by fit to psychometric function
- maximum information gain per trial

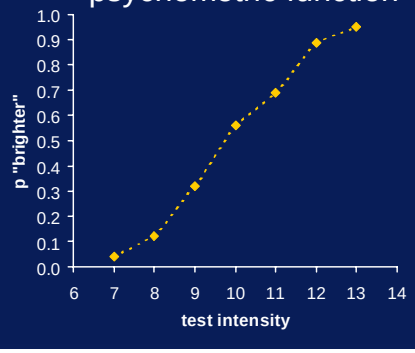
Method of constant stimuli

- is the target brighter or darker than the standard?

Method of constant stimuli: analysis 1

	20 trials/stimulus			
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
standard → 10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

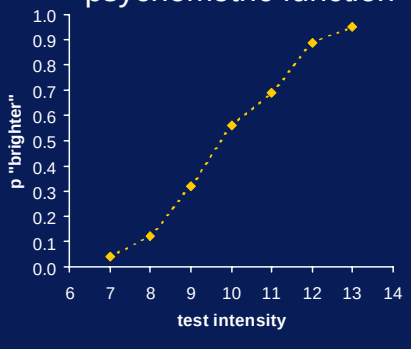
psychometric function



Method of constant stimuli: analysis 2

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function



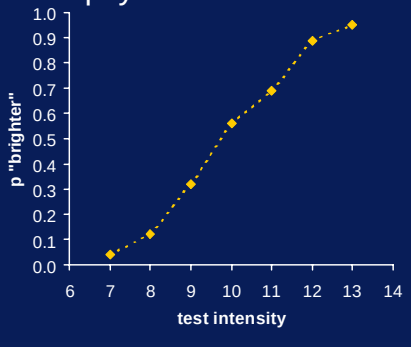
psychometric models

- cumulative normal
  - (probit analysis)
- logistic
- Weibull

Method of constant stimuli: analysis 3

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function



psychometric models

- cumulative normal

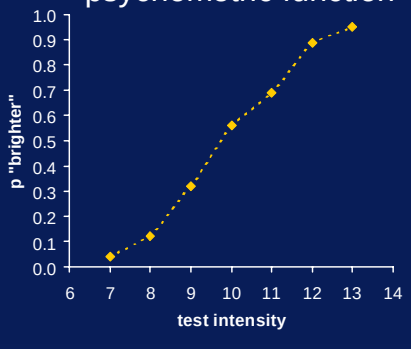
$$p_i = \int_{-\infty}^{z_i} e^{\left(z_i^2/2\right)} dz$$

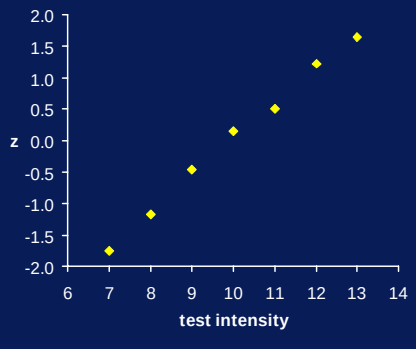
$$z = \left(i - \mu_I\right) / \sigma_I$$

Method of constant stimuli: analysis 4

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function

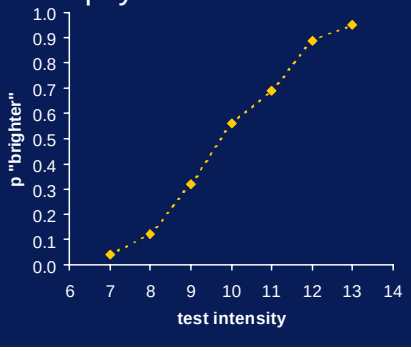


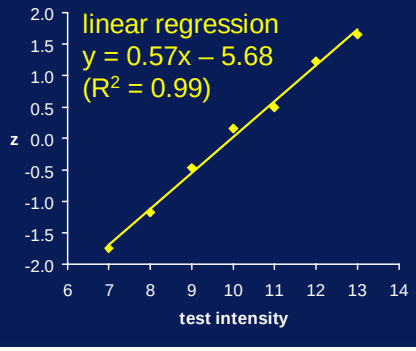


Method of constant stimuli: analysis 5

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function

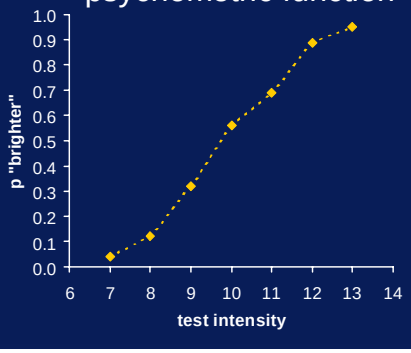


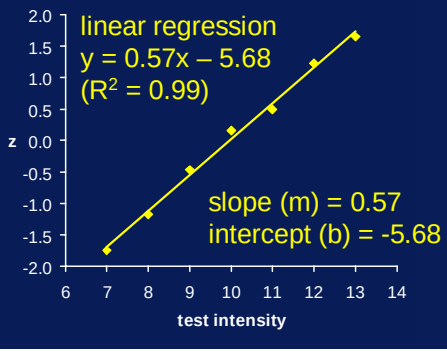


Method of constant stimuli: analysis 6

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function

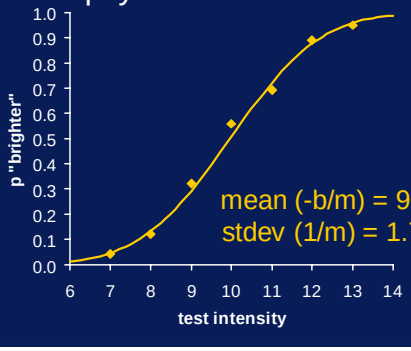


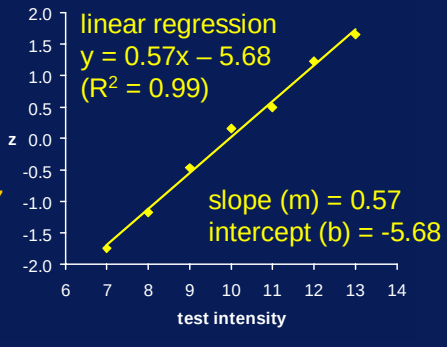


Method of constant stimuli: analysis 7

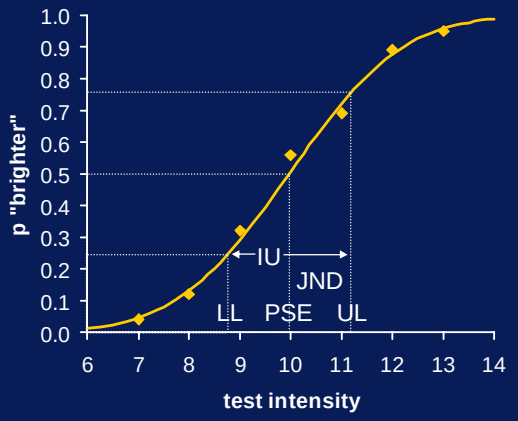
20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function





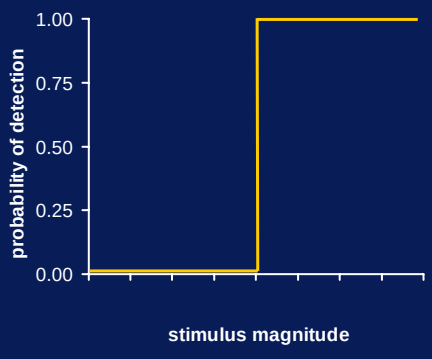
Method of constant stimuli: analysis 8



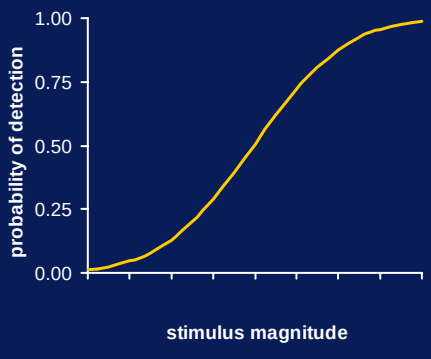
- point of subj. equality (PSE) = mean (p(.50)) = 9.97
- just noticeable diff. (JND) = 0.67449 * stdev = 1.18
- upper threshold (UL) = PSE + JND = p(.75) = 11.17
- lower threshold (LL) = PSE - JND = p(.25) = 8.83
- interval of uncertainty (IU)= UL - LL = p(.75) - p(.25) = 2.34

Variation in threshold measures

theoretical



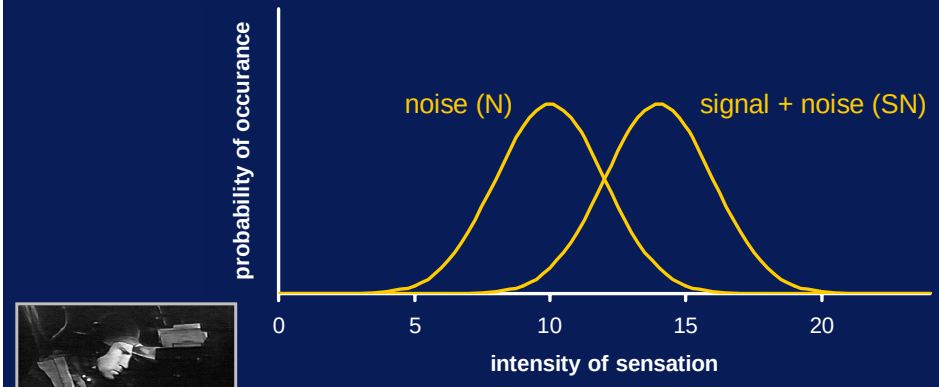
actual

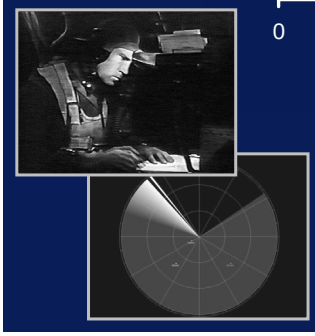


- no true cutoff
- threshold value determined by convention

The signal detection problem

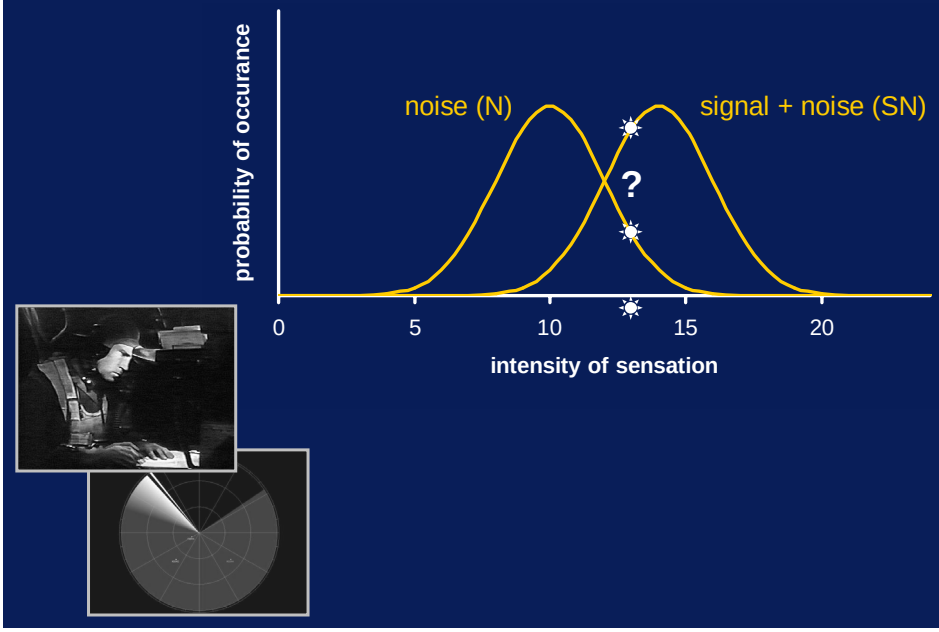
The signal detection problem





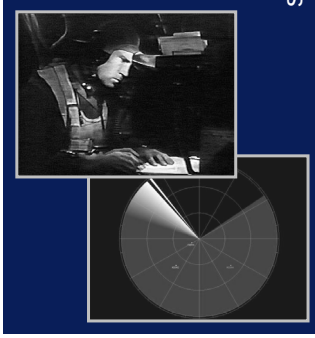
- sources of noise
  - external - photon emission, light scattering, ...
  - internal - neural transduction/transmission, adaptation, ...

The signal detection problem

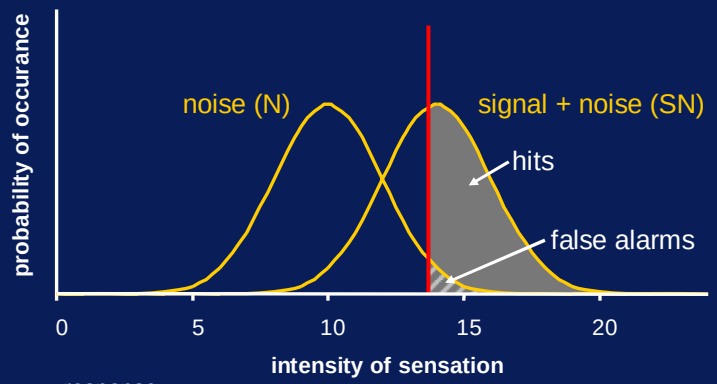


stimulus/response (SR) matrix

stimulus	response: yes	response: no
signal+ noise	**hits** p(yes|SN)	**misses** p(no|SN)
noise	**false alarms** p(yes|N)	**correct rejects** p(no|N)



Decision criterion



stimulus	response: yes	response: no
signal+ noise	**hits** p(yes|SN)	**misses** p(no|SN)
noise	**false alarms** p(yes|N)	**correct rejects** p(no|N)

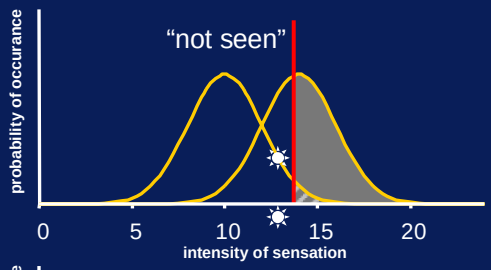
Effects of criterion on detection

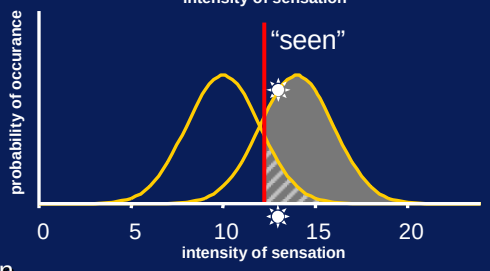
strict criterion

stimulus	response: yes	response: no
signal+ noise	**hits 75%**	**misses 25%**
noise	**false alarms 15%**	**correct rejects 85%**

lax criterion

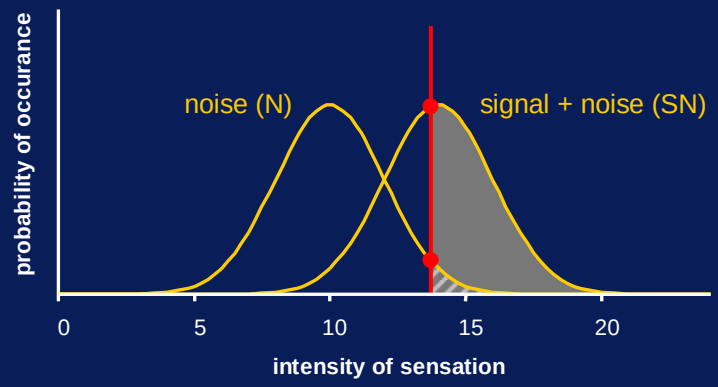
stimulus	response: yes	response: no
signal+ noise	**hits 60%**	**misses 40%**
noise	**false alarms 5%**	**correct rejects 95%**





- factors affecting the criterion
  - payoff, expectation, attention, learning

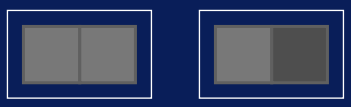
Measuring sensitivity and response bias



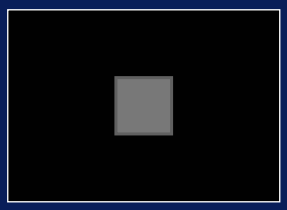
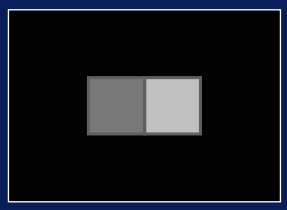
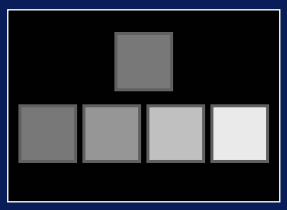
$$d' = \frac{M_{SN} - M_N}{\sigma_N} \qquad \beta = \frac{\text{value of SN dist. at criterion}}{\text{value of N dist. at criterion}}$$

Contributions of SDT

- no true threshold cutoffs
- detection/discrimination processes probabilistic
- measured thresholds affected by sensory and psychological factors
- effects can be teased apart with SDT methods
- Two alternative forced choice method (2AFC)



Psychophysical issues: thresholds and scales

detection	discrimination	scaling
		
how bright?	how much brighter?	twice as bright?
absolute threshold	difference threshold (JND)	suprathreshold appearance

Types of scales

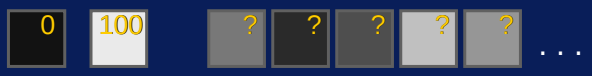
- nominal – teams
- ordinal – 1st, 2nd, 3rd place
- interval – relative times
- ratio – absolute times

Scaling methods

- indirect
  - rating
  - pair comparison
  - ranking
  - category scaling
- direct
  - equisection / fractionation
  - magnitude production
  - magnitude estimation

Rating

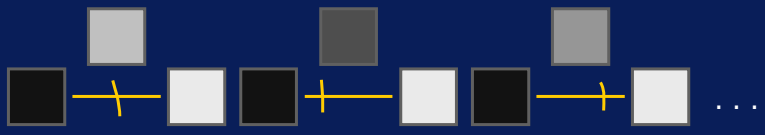
numerical

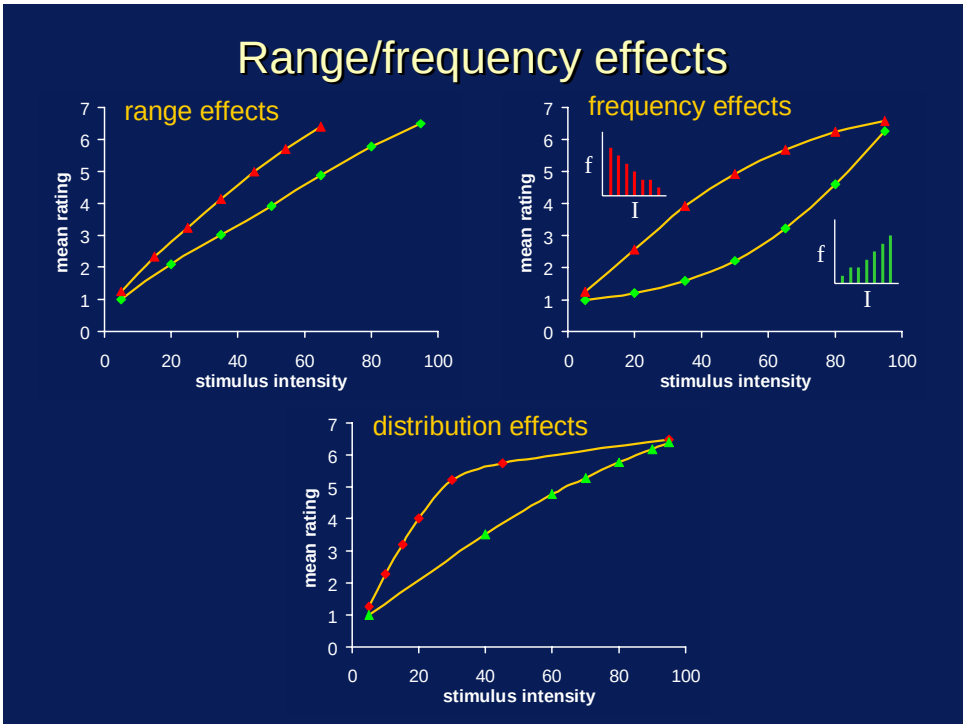


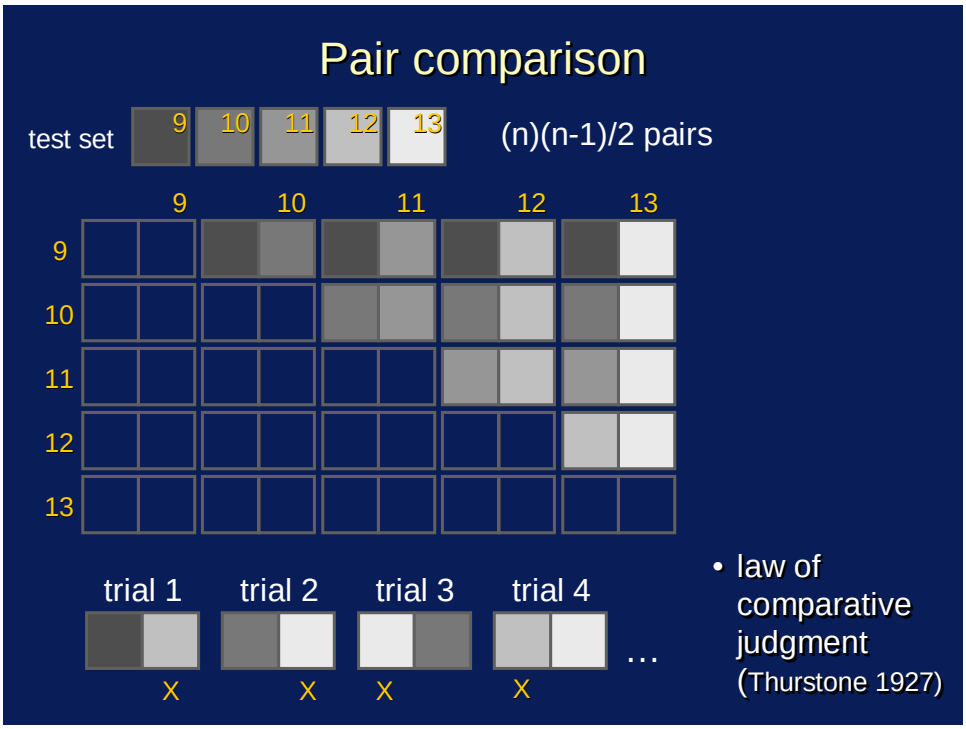
adjectival

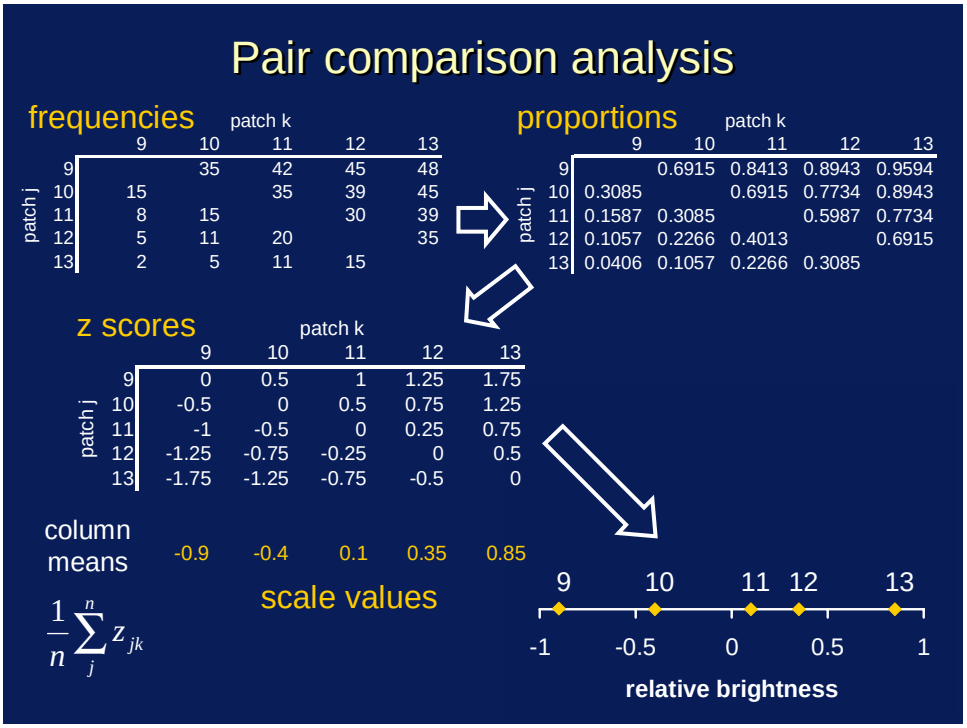
- very bright
- bright
- moderate
- dark
- very dark

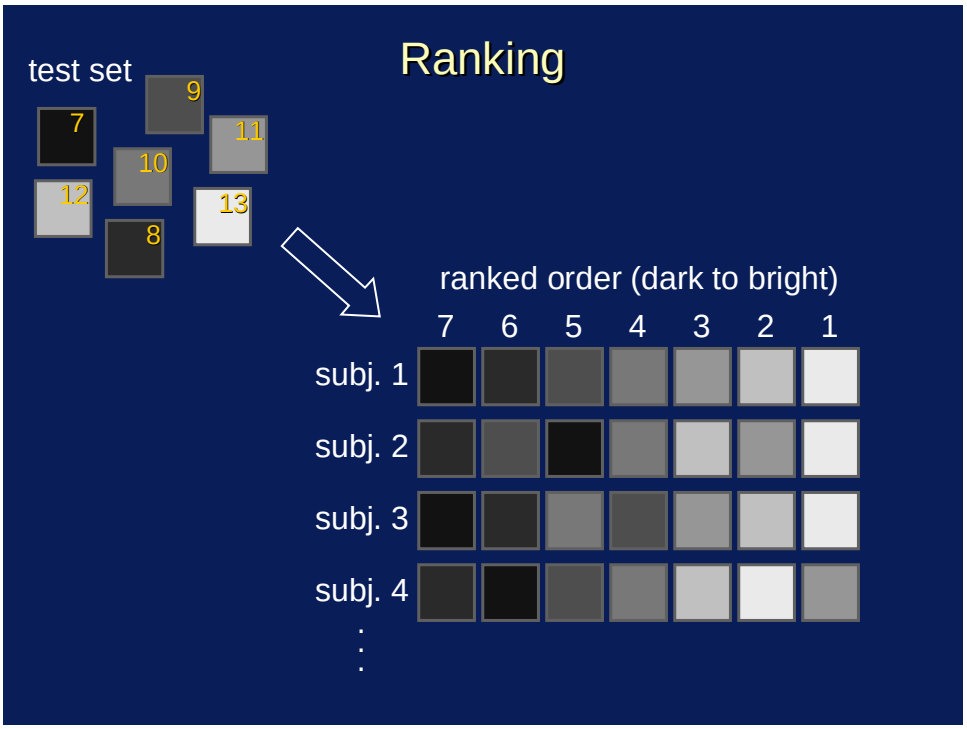
graphical




Range/frequency effects
range effects
mean rating
stimulus intensity
frequency effects
f
I
distribution effects


Pair comparison
test set
9
10
11
12
13
(n)(n-1)/2 pairs
trial 1
trial 2
trial 3
trial 4
X
…
• law of comparative judgment (Thurstone 1927)


Pair comparison analysis
frequencies
patch k
patch j
9 10 11 12 13
9 35 42 45 48
10 15 35 39 45
11 8 15 30 39
12 5 11 20 35
13 2 5 11 15
proportions
patch k
patch j
9 10 11 12 13
9 0.6915 0.8413 0.8943 0.9594
10 0.3085 0.6915 0.7734 0.8943
11 0.1587 0.3085 0.5987 0.7734
12 0.1057 0.2266 0.4013 0.6915
13 0.0406 0.1057 0.2266 0.3085
z scores
patch k
patch j
9 10 11 12 13
9 0 0.5 1 1.25 1.75
10 -0.5 0 0.5 0.75 1.25
11 -1 -0.5 0 0.25 0.75
12 -1.25 -0.75 -0.25 0 0.5
13 -1.75 -1.25 -0.75 -0.5 0
column means
-0.9 -0.4 0.1 0.35 0.85
$\frac{1}{n}\sum_{j}^{n} z_{jk}$
scale values
9 10 11 12 13
-1 -0.5 0 0.5 1
relative brightness


Ranking
test set
9
7
11
10
12
13
8
ranked order (dark to bright)
7 6 5 4 3 2 1
subj. 1
subj. 2
subj. 3
subj. 4
:

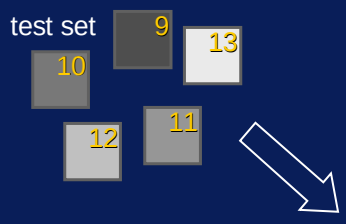
Ranking analysis

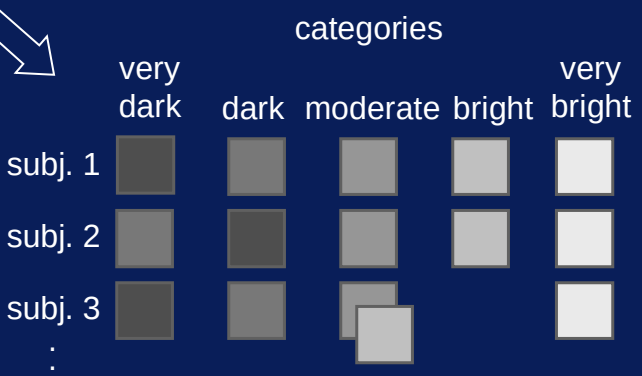
judged rank	patch intensity 7	8	9	10	11	12	13
7th	78	8	10	0	0	0	0
6th	11	81	10	2	0	0	0
5th	0	1	65	33	1	0	0
4th	10	1	10	52	10	17	0
3rd	0	0	2	11	75	12	0
2nd	0	8	0	2	12	68	16
1st	1	1	3	0	2	3	84
mean rank (M_r)	6.53	5.68	5.04	4.22	2.96	2.43	1.16
	7	6	5	4	3	2	1
proportions $p = (N_r - M_r)/(N_r - 1)$	0.078	0.220	0.327	0.463	0.673	0.762	0.973
z scores z(p)	-1.416	-0.772	-0.449	-0.092	0.449	0.712	1.932

ordinal scale

interval scale

Category scaling





- law of categorical judgment (Thurstone 1927)

Category scaling analysis 1

frequencies — categories

patch intensity	V. Dk.	Dark	Mod.	Bright	V. Brt.
9	100	38	49	11	2
10	84	27	47	23	19
11	13	32	110	39	6
12	62	14	32	23	69
13	4	9	49	58	80

cumulative frequencies — categories

patch intensity	V. Dk.	Dark	Mod.	Bright	V. Brt.
9	100	138	187	198	200
10	84	111	158	181	200
11	13	45	155	194	200
12	62	76	108	131	200
13	4	13	62	120	200

cumulative proportions — category boundaries

patch intensity	VD/D	D/M	M/B	B/VB
9	0.50	0.69	0.94	0.99
10	0.42	0.56	0.79	0.91
11	0.07	0.23	0.78	0.97
12	0.31	0.38	0.54	0.66
13	0.02	0.07	0.31	0.60

z scores — category boundaries

patch intensity	VD/D	D/M	M/B	B/VB
9	0.00	0.50	1.51	2.33
10	-0.20	0.14	0.81	1.31
11	-1.51	-0.76	0.76	1.88
12	-0.50	-0.31	0.10	0.40
13	-2.05	-1.51	-0.50	0.25

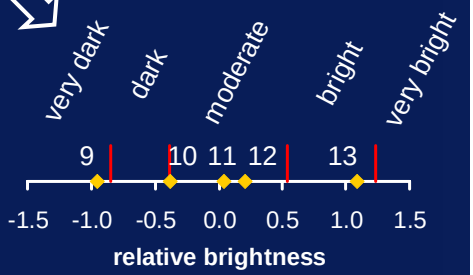
Category scaling analysis 2

z scores — category boundaries

patch intensity	VD/D	D/M	M/B	B/VB	gnd. mean – row mean
9	0.00	0.50	1.51	2.33	-0.95
10	-0.20	0.14	0.81	1.31	-0.38
11	-1.51	-0.76	0.76	1.88	0.04
12	-0.50	-0.31	0.10	0.40	0.21
13	-2.05	-1.51	-0.50	0.25	1.08
col. means	-0.85	-0.39	0.54	1.23	

scale values

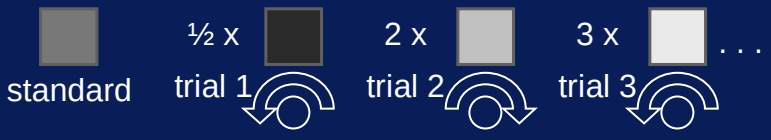
category boundaries



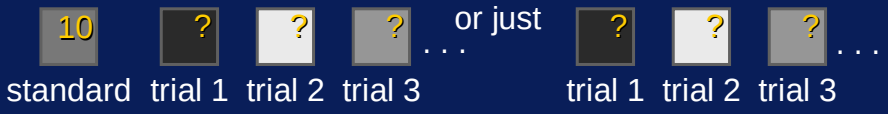
Ratio scaling methods

- equisection

- magnitude production



- magnitude estimation



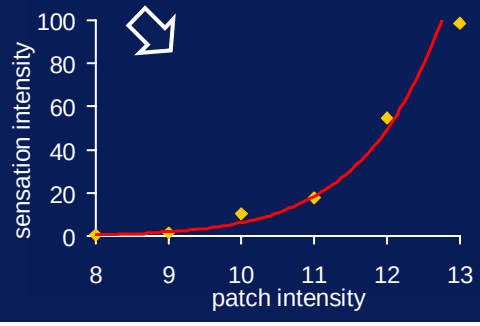
Magnitude estimation analysis

subject	patch intensity 8	9	10	11	12	13
A	1.1	3	35	565	1651	788
B	0.5	1	9	2	12	39
C	0.3	6	7	14	38	166
D	0.5	1	7	8	6	19
E	0.2	0.5	2	8	49	50
F	0.4	1	28	18	13	75
G	0.1	1	9	5	60	49
H	1.2	3	14	110	434	499
geometric means	0.4	1.5	10.0	17.4	54.5	98.6

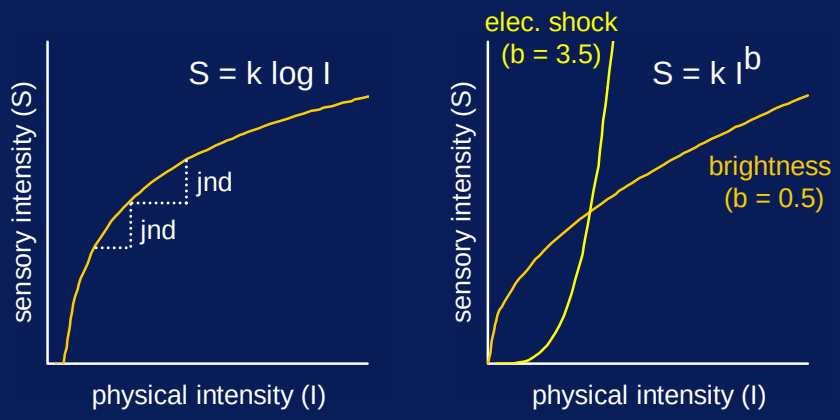
scale values

$$S_C = 10^{\frac{1}{R}\sum \log_{10} X_r}$$

power function



Scaling models



- Fechner's law
  - $\Delta I = k\ I$ (Weber's law)
  - $S = k \log I$
- Steven's power law
  - $S = k\ I^b$
  - different powers for different modalities

Multidimensional scaling

	Atl	Chi	Den	Hou	LA	Mia	NYC	SF	Sea	DC
Atlanta	0									
Chicago	587	0								
Denver	1212	920	0							
Houston	701	940	879	0						
LA	1936	1745	831	1374	0					
Miami	604	1188	1726	968	2339	0				
NYC	748	713	1631	1420	2451	1092	0			
SF	2139	1858	949	1645	347	2594	2571	0		
Seattle	2182	1737	1021	1891	959	2734	2406	678	0	
DC	543	597	1494	1220	2300	923	205	2442	2329	0

- distances -> locations

Practicalities of running psychophysical experiments

- stimulus selection
- display/interface issues
- selecting subjects
  - population, human subjects committees
- experimental design
  - randomization, control, counterbalancing
  - # of subjects, # trials, # repetitions
- data analysis
  - tests of significance/fit, confidence intervals

Summary

- psychophysics: quantify relationships between physical stimulation and perceptual sensation
- psychophysical issues: thresholds and scales
- thresholds are a product of sensory and psychological factors
- four types of scales: nominal, ordinal, interval, ratio
- multidimensional scaling

Resources -books

- Fechner, G.T. (1966) *Elements of Psychophysics*. Holt, Rinehart & Winston.
- Gescheider, G.A. (1997) *Psychophysics: The Fundamentals*, 3rd Edition. Erlbaum.
- Guilford, J.P. (1954) *Psychometric methods*. Mcgraw-Hill.
- Torgerson, W.S. (1960) *Theory and Methods of Scaling*. Wiley.
- Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*. Wiley.
- Engeldrum, P.G. (2000) *Psychometric scaling: A Toolkit for Imaging Systems Development*. Imcotek Press.

Resources - papers/standards

- Use of computers and cathode-ray-tube displays in visual psychophysics - special issues of the journal Spatial Vision 10(4) and 11(1) - http://www.lrz-muenchen.de/~Hans_Strasburger/contents.html
- ASTM (American Society for Testing and Materials), Standard Guide for Conducting Visual Experiments, E1808-96
- ASTM (American Society for Testing and Materials), Standard Guide for Selection, Evaluation, and Training of Observers, E1499-94
- CIE Technical Committee 1-34 Testing Color-Appearance Models: Guidelines for Coordinated Research - Alessi, P.J. (1994) Color Research and Applications,19, 48-58.

Resources - software

- Psychophysics Toolbox
  - Matlab-based - http://psychtoolbox.org/
- Psychophysica/Cinematica
  - Mathematica/Quicktime-based - http://vision.arc.nasa.gov/mathematica/psychophysica/
- Strasburger's review of psychophysics software -
  - http://www.lrzmuenchen.de/~Hans_Strasburger/psy_soft.html

Updates/Errata

- http://www.graphics.cornell.edu/~jaf

Acknowledgements

- Cornell Program of Computer Graphics
- NSF ASC-8920219, IIS-0113310

Psychometrics 101: Designing an Experiment

Benjamin Watson

Dept. Computer Science
Northwestern University
watson@northwestern.edu

What's coming

Design goals

Design elements

Threats

Design types

Analysis

Practice

What's coming: *examples*

Example given for many points

- *Titled with* "Example"
- *Presented in white*

One main running example

- *Watson, Friedman & McGaffey*
- *SIGGRAPH 2001*
- *Included in notes*

Why experiment?

Because…

- *You have a pet theory*
- *Literature doesn't support it*
- *You want to support your theory*

Why not?

- *It's a lot of work*
- *You may not get answer you want*

Running example: *why exp't?*

Because…

Had many visual fidelity theories

E.g. in model simplification:

Algorithm affects perceived quality

Algorithms show mettle at small output

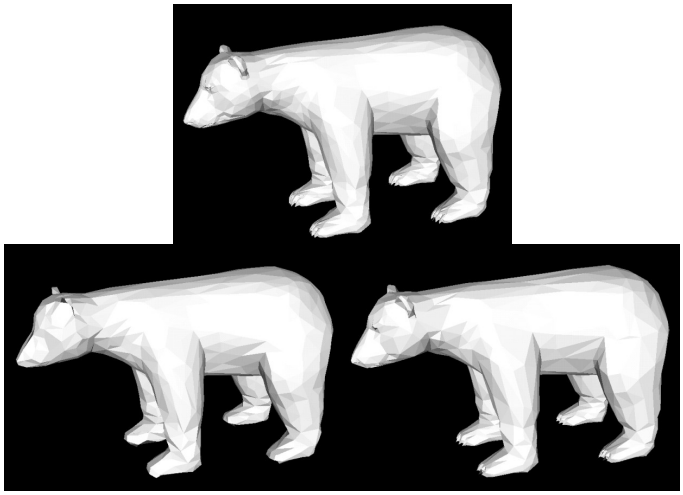
Why not?

Most of two years' work (not full time)

Example: *why experiment?*

original model

50% simplified



Algorithm A

Algorithm B

Design goals: *causality*

Experiments probe causes & effects:

Condition A causes effect B

Causality is easier to publish!

Sure, no causality can be interesting

But may also indicate experimental flaws

Design goals: *feasibility*

Can you perform experiment without:

Unreasonable time

Unreasonable cost

Trouble finding participants

Exhausting participants

Harming participants

Design goals: *validity*

Internal validity:

Is there really causality here?

External validity:

Is this true or important outside the lab?

Without validity, you wasted your time

Design elements: *task/process*

Phenomenon catching your interest

May be human (task) or natural (process)

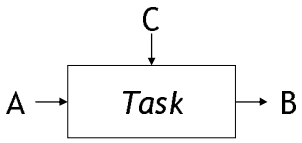
Must be measurable

For internal validity

Must be meaningful

For external validity

Design elements: *task/process*



Experiment probes black box *Task* that:

Operates on or with inputs A and C

Produces output B

Design elements: *variables*

Independent variable (input A)

Manipulated experimental condition

Set to many discrete "levels" (A1,A2...)

Dependent variable (output B)

Observed experimental condition (measure)

Control variable (input C)

Fixed (constant) experimental condition

Design elements: *hypothesis*

Verifiable claim about task or process:

If A1 causes B1, then:

When A1 happens, B1 happens

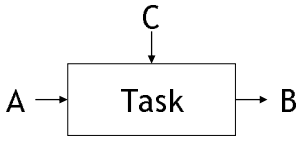
When A1 doesn't, B1 doesn't

Null hypothesis

Competing possibility: A1 doesn't cause B1

"Null results" support null hypothesis

Design elements: *hypothesis*



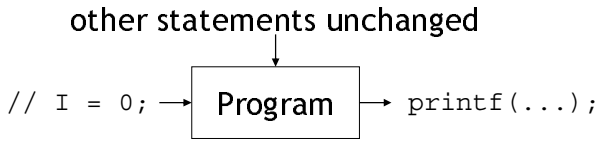
Claim about black box *Task*:

If A is set to A1 then B == B1

Else B == B2

(While C is held constant)

New example: *debugging*



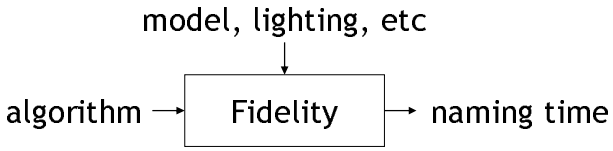
Process: *computer program*

Variables:

- Independent: *if statement executed*
- Controls: *other statements*
- Dependent: *output statements*

Hypothesis: *stmt execution causing bug*

Running example: *elements*



Process: *visual fidelity judgements*

Variables:

- Independent: *simplification method (QSlim,Clust)*
- Controls: *models, lighting, viewpoint...*
- Dependent: *time to verbalize name*

Hypothesis: *QSlim lowers naming time*

Progress

Design goals

Design elements

Threats

Design types

Analysis

Practice

Threats to: *internal validity*

Chance - data is atypical
Solution: more data or participants!

Confounds - uncontrolled variables
Solution: hold variables constant

Individual differences -
Confound independent level & participants
Solutions: matched; random assignment

Threats to: *internal validity*

Carryover - previous level affects current
Solutions: random; counterbalanced order

Reactivity - participant acts unnaturally
Solution: deception or "blind study"

Researcher bias
Solutions: treat all same; "double blind"

Threats to: *external validity*

Unrepresentative participants
Solutions: choose well; get more

Artificial setting
Solution: minimize control; vary settings

Unrealistic independent variables
Solutions: choose levels well; use more

Threats to: *feasibility*

Too many, too few participants
It may take you too long to get/find them

Too many variables, levels
Each variable requires more time

Task too challenging, boring
Causing fatigue, reducing retention

Threats to: *causality*

Ceiling/floor effects -
Measure never leaves max/min of range
Solution: adjust task difficulty
Solution: adjust measure resolution

Type 2 error -
Stats incorrectly accept null hypothesis
Solution: more data/participants ("power")

Threats: *conflicting goals*

Control
- *More: internal validity*
- *Less: external validity*

Participants
- *More: internal/external validity, causality*
- *Less: feasibility*

No ideal solution: it's "design"

Running example: *threats*

Internal validity: individual differences
- *All saw both QSlim and Clust*

Internal validity: carryover
- *Half saw QSlim first, half Clust*

Internal validity: reactivity
- *Naming time is subsecond and subconscious*

Running example: *threats*

External validity: artificial setting

Used 36 everyday stimuli

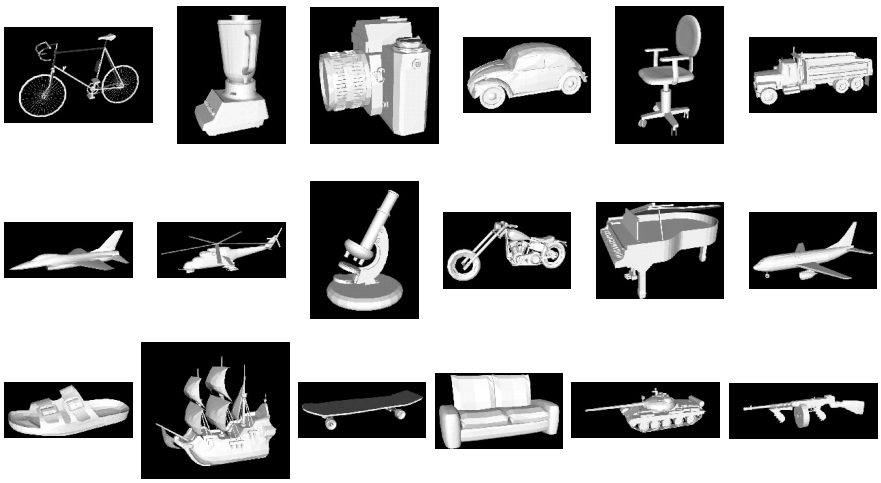
Feasibility: challenging task

Together, all naming took roughly 5 mins

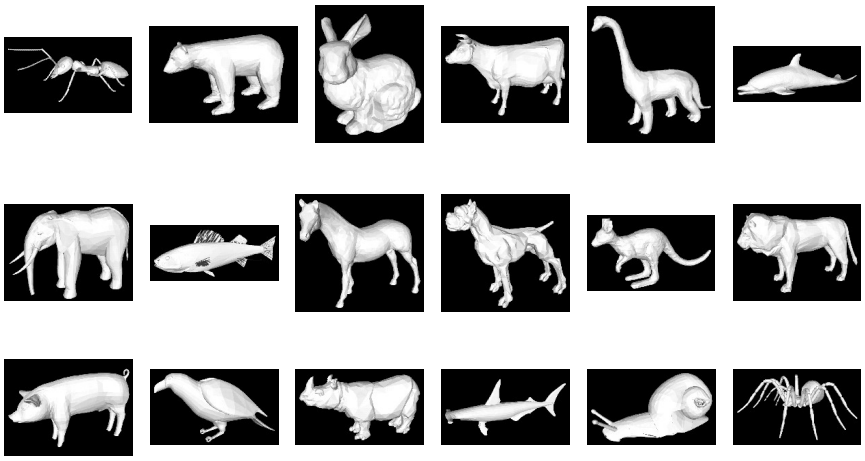
Causality: ceiling effect

Stimuli were easily named

Running example: *threats*



Running example: *threats*



Design types: *single factor*

I.e. single independent variable

Nice simple designs

May have arbitrary number of measures

With computers, measures often painless

More measures increase information gain

Design types: *within subjects*

Each participant sees every factor level

- *No problems with individual differences*
- *Generally requires fewer participants*

Major concern: carryover effects

- *Can randomize order levels experienced*
- *Or, can counterbalance order*

Design types: *within subjects*

Complete counterbalancing

- *Use all orderings, e.g. 2 levels => 2 orders*
- *But n levels => n! orders*

Partial counterbalancing

- *Complete often not feasible*
- *Common scheme: Latin square*

Design types: *within subjects*

Latin Square: *Each level occurs equal times in same experimental portion*

Levels A,B,C,D:

Participant	Order 1st	2nd	3rd	4th
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Design types: *between subjs*

Each participants sees one factor level

No problems with carryover effects

Generally requires more participants

Major concern: individual differences

Can randomize assignment to level

Or, can match level groups by some criteria

What if participant doesn't finish?

Design types: *multi-factor*

I.e. multiple independent variables

More complex, but quite common

Two-way interactions

One variable's effect depends on other

Three-way interactions

2-way interaction depends on 3rd variable

Etc...

Design types: *multi-factor*

Factorial designs

All combinations of variables, levels tested

Each combination called a "treatment"

Three variables with i, j and k levels:

An "i x j x k factorial design"

With i x j x k treatments

Running example: *types*

Additional factors to algorithm:
- *Degree of simplification (0%, 50%, 80%)*
- *Object type (animals,artifacts)*

A 2x3x2 factorial within subject design
- *Counterbalanced*

"Mettle" hypothesis: 2-way interaction
- *QSlim advantage greater at 80%*

Progress

Design goals

Design elements

Threats

Design types

Analysis

Practice

Analysis: *descriptive stats*

Central tendency
- *Mean or median*
- *Median less sensitive to outliers*

Dispersion
- *Variance: avg'd squared diffs from mean*
- *Standard deviation σ: sqrt(variance)*

Relationship
- *correlations (no causality!)*

Analysis: *inferential stats*

Testing for causal relationships
- p *is probability chance explains results*
- *Standard: causality at p < .05, 1 in 20*
- *Results are then called "significant"*

Single factor, two level designs
- *T-test (special version for within subj)*

Analysis: *inferential stats*

More complex designs

- *Analysis of variance (ANOVA)*
- *Special ANOVA for within subjects*
- *One* p *for each factor, interaction*

Post hoc tests in ANOVAs

- *Zoom in on level differences*
- *Common in analyzing interactions*

Analysis: *ethics & standards*

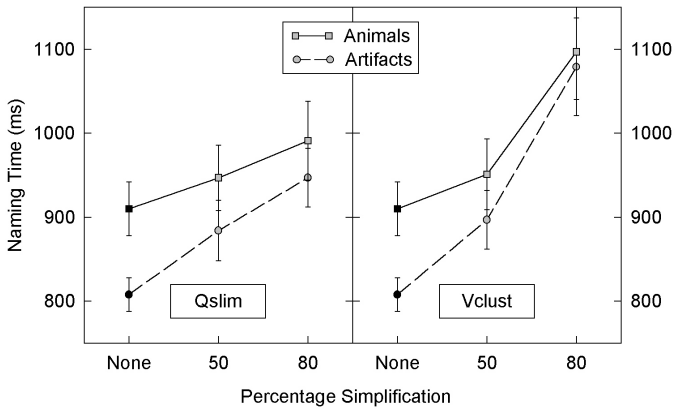
Discarding data or participants

- *Real danger of experimenter bias!*
- *Justifiable if a few samples distort result*
- *Should be a small (few %) data portion*
- *Ideally 3 standard deviations from mean*

In presentation

- *Show standard error w/ means: σ/sqrt(n)*

Running example: *analysis*



Significance:

Algorithm
Simp %
Type
Alg x Simp %

Practice: *pilot studies*

Prelim studies with few participants

For assessing threats, e.g.

Causality: ceiling/floor effects
Validity: confounds
Feasiblity: too many variables

For working out tool/procedure bugs

Practice: *how many subjects?*

This is a black art, however:

More are always better
- *Subject to feasibility concerns*

Pilot gives a clue about needed power
- *Estimate σ & aim for small std error*

Permissible to add subjects
- *Counterbalancing often fixes increments*

Practice: *board approval*

The bible: APA ethics guidelines
- *www.apa.org/ethics/code.html*

In consent form, tell participants
- *As much as possible (deceive sparingly)*
- *They can stop when they like*
- *How confidential their data is*
- *If they will experience discomfort*

Sources:

D. Elmes, B. Kantowitz & Henry Roediger (1992). *Research Methods in Psychology (4th ed)*. West Publishing: St. Paul, MN.

A. Graziano & M. Raulin (1989). *Research Methods: A Process of Inquiry*. Harper Collins: NY, NY.

Case Studies

- Our purpose here is to highlight a sampling of the work that has appeared in the computer graphics literature that uses psychometrics to gain additional insight into human perception, and/or examines the success of computer graphics techniques.

Case Study: Realistic Image Synthesis

"An experimental evaluation of computer graphics imagery"
Gary Meyer, Holly Rushmeier, Michael Cohen, Donald Greenberg, and Kenneth Torrance, ACM Transactions on Graphics, Volume 5 , Issue 1 (January 1986).

Case Study: Realistic Image Synthesis

Graphics Problem:
Synthetic images look artificial

Possible Solution:
Use radiosity to calculate diffuse interreflections

Psychophysics:
Does radiosity produce more realistic images?

Meyer et al.

- One reason that synthetic images can look artificial is that they take short cuts in represent the light transfer that accounts for the formation of the image. Radiosity is a method for more accurately calculating light transfer. Radiosity can be computationally intensive, and the question arose whether the extra effort was really worthwhile considering the quality of the image produced.

Case Study: Realistic Image Synthesis

From the vision literature:

- Previous work comparing view of real scene and a picture suggested limiting view.
- Previous work in color metamerism to convert spectral radiosity results to RGB for calibrated display

Meyer et al.

Case Study: Realistic Image Synthesis

The Experiment:

Meyer et al.

- Comparing a displayed image to a photograph requires simulating the entire film chain. A more direct comparison is to view the actual real scene and the displayed image. The view that would reveal which is which was restricted using view cameras. One advantage of the view cameras was that the subject could see both images at the same time.
- The light coming from the scene was measured to compare to numerical values to ensure that the radiosity simulation was physically accurate. The goal of the experiment was to see if this physical accuracy had an impact on the visual accuracy.

Case Study: Realistic Image Synthesis

Insights

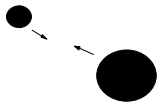
- When asked which is the real model and which is the synthetic image 9 out of 20 picked the wrong one.
- For this scene, it was possible to generate a realistic image from principles of light transfer and color metamerism.

Meyer et al.

- As a group, subjects did no better than they would have with random guessing in choosing which image was synthetic and which was a view of the real scene.
- The experiment proved that at least for certain classes of scenes radiosity was capable of generating realistic images, with realism in this case defined as indistinguishable from a view of the physical scene being simulated.

Case Study: Animation

"Collisions and Perception"
Carol O'Sullivan and John Dingliana
ACM Transactions on Graphics
Volume 20 , Issue 3 (July 2001)



Case Study: Animation

Graphics Problem:
Collisions expensive to compute

Possible Solution:
Degradable collision calculations

Psychophysics:
When is a collision plausible?

O'Sullivan and Dingliana

- Collisions are expensive to compute, especially when there are large numbers of potentially interacting objects. Taking the time to accurately compute the collisions can bring a real time application to a complete halt.
- Previously O'Sullivan and Dingliana considered degradable collisions, a method for computing collisions at different levels of accuracy.
- If you can adjust the level of accuracy of a collision, how much can you get away with without the observer feeling that the results are implausible?
- Issue is not "is this a **faithful simulation**" but "is this **plausible**"

Case Study: Animation

From Vision Literature

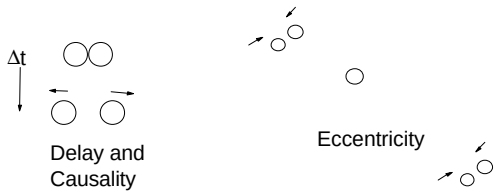
- causality
- eccentricity
- distractors
- velocity

O'Sullivan and Dingliana

- Rather than run through hundreds of cases with arbitrary variations, important elements in perception of collisions are identified from the vision literature.

Case Study: Animation

The Experiments



Delay and Causality

Eccentricity

O'Sullivan and Dingliana

- Psychophysical experiments conducted to examine each effect.
- Perception of causality, whether the collision is causing the objects to move, depends on the delay in the motion. Experiment verified that the more the delay the less plausible the collision.
- In the proposed degradable collision method, less accurate collisions leave gaps between the two colliding objects. The experiment verified that detectability of these gaps decreases as the collision is further away from the point where the observer is directing attention.

Case Study: Animation

Distractors

Velocity

O'Sullivan and Dingliana

- In the proposed degradable collision method, less accurate collisions leave gaps between the two colliding objects. The experiment verified that detectability of these gaps decreases as the collisions as there are more objects "distracting" the observer.
- In the proposed degradable collision method, less accurate collisions also affect the angular momentum imparted on collision. In a surprising result, for higher velocities highly accurate collisions actually rated less plausible than medium accuracy.

Case Study: Animation

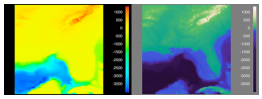
Insights:

- Time delays are important
- Eccentricity and distractors can be exploited
- Simulation accuracy != plausibility

O'Sullivan and Dingliana

Case Study: Data Visualization

"The Which Blair Project:
A Quick Visual Method for Evaluating
Perceptual Color Maps"
Bernice Rogowitz and Alan Kalvin
IEEE Visualization, Oct. 2001



Case Study: Data Visualization

Graphics Problem:
Mapping colors to data values

Possible Solution:
Perceptual color maps

Psychophysics:
Can a simple test identify a good map?

Rogowitz and Kalvin

- One way to represent a scalar data value visually is to map a range of color to a range of values. To be effective, the relative changes perceived colors should be the same as the relative changes in the data,e.g. if a value of 0.5 is significantly low compared to 0.7 they should not be mapped to imperceptibly different shades of yellow. Perceptual color maps has been developed to address this issue. Unfortunately, the color displayed by a value 0-255 depends on the particular display be used. Without tedious calibration, it is impossible to know what map should be used on a particular display to produce a perceptually uniform mapping.

Case Study: Data Visualization

From Vision Literature

- perception of luminance versus hue, saturation
- sensitivity of luminance variation on faces

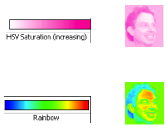
Rogowitz and Kalvin

- The goal of this work is to find an easy test for evaluating the best color map without calibration. The work starts with the insights that monotonic increase in luminance is critical in a useful color map, and that people are very sensitive to luminance variations on human faces.

Case Study: Data Visualization

Experiment:

different color maps applied to Blair image

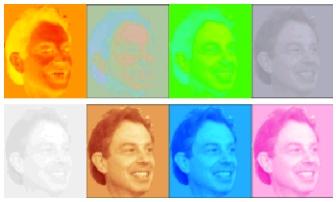


Rogowitz and Kalvin

- In the experiment various color maps were used on a grey scale image of a well-known face.

Case Study: Data Visualization

Experiment:



Rogowitz and Kalvin

- Users judged whether the picture was very bad (-2), somewhat bad (-1), neutral (0), somewhat good (1) or very good(2).

Case Study: Data Visualization

Experiment:

- scales with monotonic increase rated positive
- higher luminance contrast, higher rating

Rogowitz and Kalvin

- The experimental results showed that images generated with the perceptual color maps received positive ratings.

Case Study: Data Visualization

Insight:

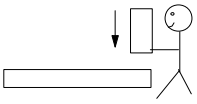
Applying various sets of color maps to a face image can aid in selection of perceptual color map on uncalibrated display.

Rogowitz and Kalvin

- The results of this experiment indicated that applying color maps on a face may be an effective method for choosing a color map from a set of color maps that is best for a particular uncalibrated display. This may help when people share data among remote sites, to make sure that the same features are discernible in a visualization, even when the data is being viewed on vastly different display devices.

Case Study: Virtual Environments

"Visual Cues for Imminent Object Contact"
Helen Hu, Amy Gooch, William Thompson, Brian Smits, John Rieser, Peter Shirley
IEEE Visualization, Oct. 2000.



Case Study: Virtual Environments

Graphics Problem:
Judging distances in V.E.'s is difficult

Possible Solution:
Add visual cues

Psychophysics:
What visual cues are worth computing to improve judgements?

Hu et al.

- A sense of imminent contact is needed for virtual environments to have a sense of naturalness -- i.e. its unnatural to not know you are about to hit something until you feel it. Visual cues can help, but some of them can be very time-consuming to compute, so which are really helpful?

Case Study: Virtual Environments

From Vision Literature

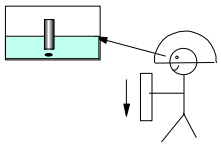
- binocular stereo
- shadows
- interreflections

Hu et al.

- The study begins with a review of relevant vision literature on depth perception. Stereo is important and has been studied a lot. There are a few studies on shadows, and very little on interreflections and their impact on depth perception.

Case Study: Virtual Environments

The Experiment



Hu et al.

- In the experiment the user moved a real cylinder, constrained to vertical motion, while viewing a virtual cylinder which was being moved above a horizontal virtual surface. The cylinder was to be lowered as close as possible to the horizontal surface without hitting it.

Case Study: Virtual Environments

The Experiment

All eight combinations of stereo/no stereo, shadows/no shadows, interreflections/no interreflections tested.

Hu et al.

Case Study: Virtual Environments

The Experiment

- All results better for cases including stereo than for cases which did not include stereo.
- Overall all subjects used all three cues.
- Benefit of shadows and interreflections varied between subjects.

Hu et al.

Case Study: Virtual Environments

Insights

- Stereo important
- Shadows help
- Benefit of interreflections unclear
- Data for more configurations needed

Hu et al.

- A single experiment of this type can indicate useful effects, but it will take much more additional testing for different configurations and tasks before the value of various cues can before quantitative results can be obtained.

Summary

Just a small sampling of work using perceptual studies in computer graphics.

No single experiment completely solves a problem, but each gives new insights.

- The original papers for these examples should be consulted for details on the experimental methods and design. Reading about the experiments in these papers in the context of the principles presented in this course is a good way to start using the information presented today.

Are Image Quality Metrics Adequate to Evaluate the Quality of Geometric Objects?

Bernice E. Rogowitz and Holly E. Rushmeier

IBM TJ Watson Research Center, P.O. Box 704, Yorktown Heights, NY USA

ABSTRACT

Geometric objects are often represented by many millions of triangles or polygons, which limits the ease with which they can be transmitted and displayed electronically. This has led to the development of many algorithms for simplifying geometric models, and to the recognition that metrics are required to evaluate their success. The goal is to create computer graphic renderings of the object that do not appear to be degraded to a human observer. The perceptual evaluation of simplified objects is a new topic. One approach has been to use image-based metrics to predict the perceived degradation of simplified 3-D models [1] Since that 2-D images of 3-D objects can have significantly different perceived quality, depending on the direction of the illumination, [2] 2-D measures of image quality may not adequately capture the perceived quality of 3-D objects. To address this question, we conducted experiments in which we explicitly compared the perceived quality of animated 3-D objects and their corresponding 2-D still image projections. Our results suggest that 2-D judgements do not provide a good predictor of 3-D image quality, and identify a need to develop "object quality metrics."

Keywords: computer graphics, geometric simplification, perception

1. INTRODUCTION

Three dimensional computer graphics was once used primarily in specialized computer-aided design (CAD) systems, and off-line rendering applications such as feature film production. Increasingly 3-D graphics is being used in widespread interactive, networked applications such as computer games and e-commerce. The representation and display of 3-D objects can require substantial computer resources. A critical issue in designing effective interactive systems is to find the minimum representation of a 3-D object that does not compromise the visual quality of the object when it is rendered in two dimensions. In this paper we address how to evaluate the quality of the representation of a 3-D object. To do so, we compare the degree to which a simplified object appears degraded relative to an "original". In one condition, the observers judge the perceived quality of the simplified object by comparing static images of the object; in the other condition, observers compare animated sequences of the original and simplified objects rotating through an angle of 90 degrees.

1.1. Geometric Representation and Simplification

A wide variety of numerical forms can be used to represent 3-D objects. For interactive applications however, the most widely used representation is triangle meshes since they can be displayed rapidly by computer graphics cards commonly available on personal computers. In the past, 3-D objects for games and other interactive applications were carefully designed to use small numbers of triangles. More recently, objects are modeled by sampling continuous representations designed in CAD systems, or by capturing physical objects using 3-D scanning systems. Typically CAD or scanned objects are over sampled. Simplification algorithms have been used to reduce the number of triangles. A review of simplification algorithms can be found in an article by Cignoni et al..[3] In general, geometric metrics, such as maximum distance from the original unsimplified surface, are used to drive algorithms that reduce the number of triangles. Simplified models are produced by trial and error using different values of the geometric metric until the model judged to be of adequate visual quality with the least number of triangles is obtained.

Recent algorithms have been designed based on the realization that ultimately the perceived quality of the object is the critical issue. Driving the simplification by the resulting 2D display of the object, image metrics, rather than geometric metrics, have been proposed. In particular, Lindstrom and Turk [4] have developed a simplification method driven by minimizing the root-mean-squared difference in images generated from a large number of views of the simplified object relative to the original object. The algorithm simplifies a shape for a particular surface color variation and reflectance. In all the views used in their algorithm, light coincident with the viewer is used.

Further author information: (Send correspondence to B.E.R.) E-mail:{rogowtz,hertjwr}@us.ibm.com

1.2. Perceptual Evaluations of Geometric Representations

Although many simplification algorithms take into account geometric measures that are related to perception, only two studies have appeared that document psychophysical experiments to evaluate object quality.

Watson et al.[1] used naming times to study the quality of object representations. Observers were presented with images of objects that had been simplified to various extents, and the time it took for observers to name the objects was recorded. Generally naming times correlated poorly with both geometric measures and image-based measures. In one condition however the authors found a correlation between naming time and the output of a perceptually-based image quality metric. This suggests a role for image-based metrics for evaluating geometric object quality.

Rushmeier et al.[2] studied the effectiveness of replacing geometric detail with texture maps as a method of simplification. Observers were asked to rate the quality of match between high resolution models and various simplifications of the model. Consistent results were obtained across observers. For the simple objects used, it was found that the effectiveness of textures to replace geometric detail depended on the spatial detail of the object .

In both of these studies, the authors assumed that the quality of still images can be used to assess the quality of an object representation. Since one essential feature of an interactive application is that objects are observed in motion, we wonder whether the perceived quality of 2D projections correlates with the perceived quality of the rotating object. It may be that the requirements for the representation of an object in motion may be reduced because the observer is able to detect less detail in frames that pass quickly. On the other hand, motion may make some artifacts in simplified objects more apparent if they result in unnatural jumps between frames. If the observer's judgements are the same for animations and for the still images that the animations are composed of, asking observers to rate comparisons of animations may be a more efficient method of examining object quality in future experiments.

2. DESIGN OF EXPERIMENT

In general, 3-D objects have varying color and surface finish. To narrow the scope of this experiment, we considered only gray objects with a uniform surface finish. We used objects from the Georgia Tech Large Model database, that were originally obtained using 3-D scanning by researchers at Stanford University. In addition to being publically available, these models are of interest because they are commonly used in comparisons of geometric simplification algorithms by the computer graphics community.

2.1. Preparation of Stimuli

The two objects used were the models "bunny" **b** and "happy Buddha" **h** in the database. The models are shown in Figs. 1, 2 and 3. The highest resolution model for the bunny was composed of 69,451 triangles, and for the Buddha 143,206 triangles. Two simplified versions of each model were generated using the *Simplify* module in the OpenDX open source visualization system. The *Simplify* module uses Gueziec's simplification method[5] that specifies a bound on the distance of each vertex in the simplified model as a percentage of the diagonal of a rectangle bounding box (BB) for the original model. For the bunny an error bound of 0.4 per cent of the BB diagonal produced a simplified model of 6467 triangles, and a bound of 1.0 per cent produced a very simplified model of 1679 triangles. The three versions of the bunny model viewed and lit from the front are shown in Fig. 1. For the happy Buddha bounds of 0.25 and 1.40 percent of the BB diagonal produced simplified and very simplified models of 27,168 and 6389 triangles respectively. The three happy Buddha models are shown in Fig. 2.

For each model and level of simplification two sequences of 15 images were produced, one with a light collocated with the view position, as in Figs. 1 and 2, and one with the light directly above the object. In each sequence the object rotated in 5 degree increments from a front to a side view. Examples of side views of the models lit from above are shown in Fig. 3. The image sequences were assembled into animations to be played back at a rate of 15 frames/second, giving a smooth rotation from front to side that could be run in a forwards/backwards loop.

2.2. Procedure

There were 8 basic conditions: two objects (**b**unny and **h**appy Buddha), each at two levels of simplification (**s**implified and **v**ery simplified), with each level of simplification viewed under two lighting conditions (lit from **a**bove or from the **f**ront). We will refer to these conditions by specifying object-simplification-lighting, e.g. **bva** refers to the **b**unny model, **v**ery simplified lit from **a**bove. The experiment consisted of two parts. In the first part the observers were asked to rate the images of the simplified and very simplified models relative to images of the full resolution model under corresponding view and lighting conditions. The images were presented pairwise (full and simplified resolution in each pair) in random order (model types, simplification

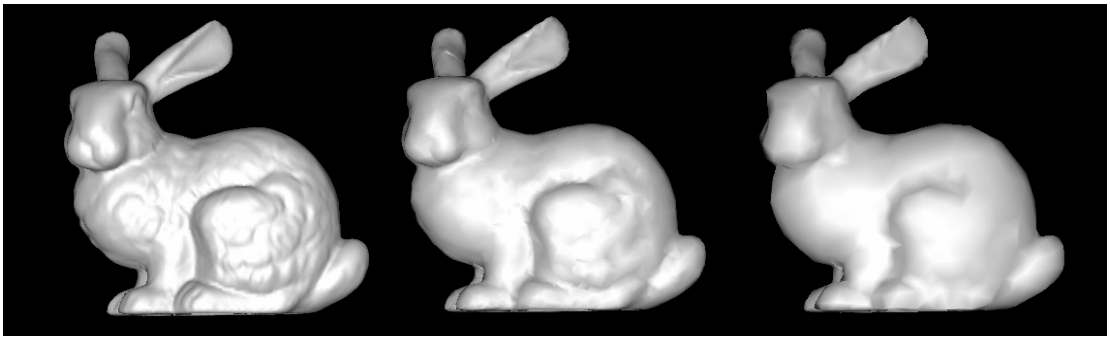


Figure 1. Three versions of the bunny **b** model were used. On the left is the full resolution model, in the center the simplified model, and on the right the very simplified model. All three versions are shown here viewed and lit from the front.

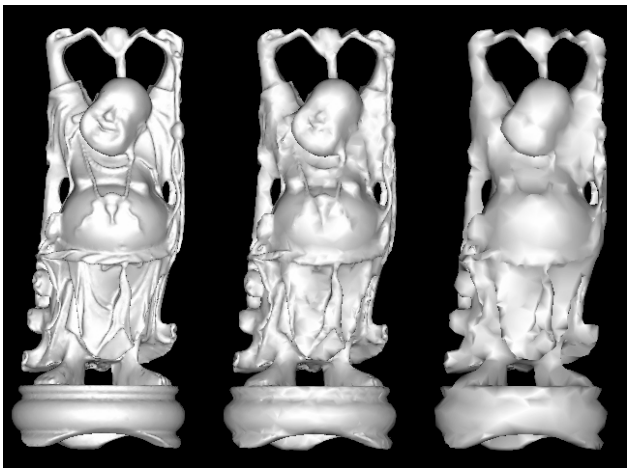


Figure 2. Three versions of the happy Buddha **h** model were used. On the left is the full resolution model, in the center the simplified model, and on the right the very simplified model. All three versions are shown here viewed and lit from the front.

Figure 3. Sequences of each model were generated rotating from a front to a side view. Two sequences were constructed of each of the three models. They were either lit from the front and viewed from the front, as in Figs. 1 and 2, or lit from the top and viewed from the front as shown here.

levels and light conditions all mixed) to each observer in an html form. The observer rankings were indicated on an integer scale of 0 (worst) to 10 (perfect) using radio buttons under each test stimulus.

For each of the 8 basic conditions images were generated for 15 different view positions, for a total of 120 image comparisons to be made by each observer. The observers were free to view all of the images to be compared before assigning scores, to determine their own calibration for "worst" comparison. Ten observers with normal or corrected to normal vision participated in the experiment. All were professionals at the Watson Research Center, but naive to the purpose of the experiment.

In the second part of the experiment the observers were presented animations of the objects. They were asked to judge the quality of each animated simplified object relative to the full resolution animated object under the same lighting condition. The animations for each comparison were embedded in a Lotus Freelance presentation file to allow the observers to view the animations for as long as they desired without being able to stop the animations and examine individual frames. The animations for each comparison were viewed in succession, rather than side by side. As in the image comparisons, the observers were free to review the animations as many times as they wanted, and to go back a review previous pairs of animations before recording their final scores.

3. RESULTS

Ten observers judged the degree to which a simplified geometric object matched the perceived image quality of the original object and assigned a rank to the perceived quality. The higher the rank, the greater the perceived similarity between the original and simplified object. This judgment was made for two different geometric objects, under two different illumination conditions, for a sequence of still images, and for an animated set of sequences.

3.1. Analyzing Data from Rating Experiments

The data from these experiments are rank judgments. Observers use numerical values on a scale from 0 to 10, and these judgments provide a measure of perceived quality. These judgments are ordinal. Consider three still images rated "2," "5" and "8". These values are ordered, the images increase in perceived image quality, but the distances between these judgments do not necessarily represent perceptual distances. That is, although the number assigned to the highest quality image is four times the number assigned to the lowest quality image, its perceived image quality is not necessarily four times as great. Since these are ordinal, not nominal data, we cannot simply use mean rating scores to summarize the data, and must instead treat the data as ranks. We can compute, instead, statistical summaries appropriate to ordinal data, such as the proportion of scores at or above a certain value, rank-order correlations, etc. We call this out explicitly since it is a common practice to simply compute means and standard errors for rating data. Since these methods assume that the data are interval, using them on ordinal data can lead to biases in data interpretation.[6]

3.2. The Perceived Quality of Simplified Objects

Figure 4a shows results for one observer from one of the test patterns, **hsa** (**h**appy Buddha, **s**implified, lit from **a**bove). The graph shows the rating score for each of the 15 viewing positions, plus, at the value indicated by **A** the score representing the degree of degradation of that target, relative to its original, when animated. Figure 4b shows the results for all 10 observers. Observations for two of the observers have been connected by lines to aid in visualizing the data. These data reveal the large variability in the observers' rating responses. For this target, rating scores ranged from 2 to 8. Whatever effect there might be of viewing position, it is small relative to the variability in the data.

Since we found no systematic effect of viewing position, we combined the rating scores for the 15 different viewing positions. Figure 5 shows a set of histograms representing quality scores under each of the eight conditions. Each histogram represents rating scores for ten observers at 15 different viewing positions (150 scores). The first column shows histograms; the second column shows cumulative histograms. The histogram representing the data **hsa** in Fig. 4 are in the fifth row. Images of the test objects are provided to aid in interpretation. For ease of viewing, the histograms have been ordered by mean rank score. Ordering them by the percentage of scores at or above the rank of 6 produces the same ordering.

Several results emerge from this plot. First, for all targets, independent of lighting, the more simplified the object, the greater the perceived degradation. Second, the simpler model, the bunny, was less sensitive to degradation than the more complex model, the happy Buddha. When compared with its original, the bunny was systematically rated as being less degraded, independent of lighting angle. Third, in three of the four cases, the object lit from above was rated less positively than the same object, at the same simplification, lit from the front. In the cumulative histograms, the distribution of scores shifts to the right with higher scores when the object is lit from the front.

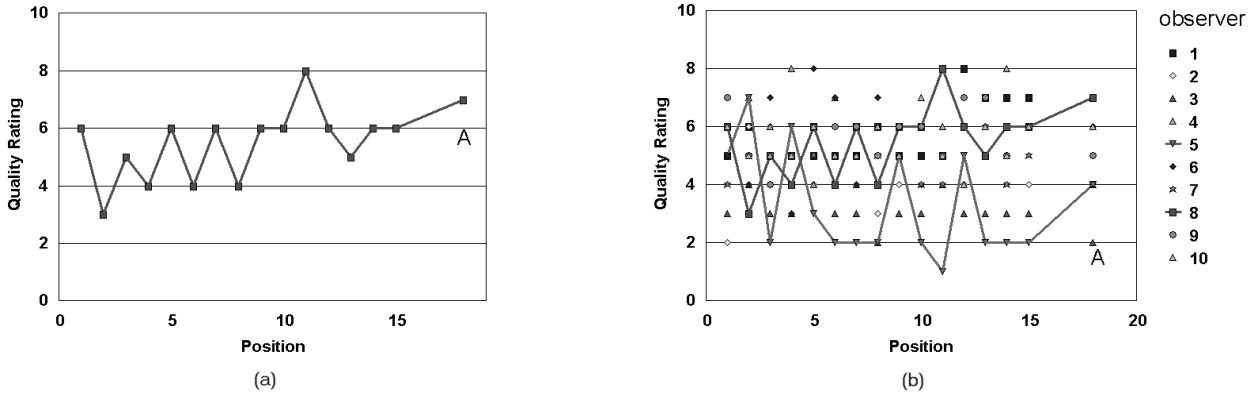


Figure 4. Example results for the target **hsa** (**h**appy Buddha, **s**implified, lit from **a**bove). Fig. (a) shows the rating results for a single observer at each of 15 position of the object, and the score assigned to the quality of **hsa** when animated relative to the (animated) original, **A**. Fig. (b) shows the results for all observers for **hsa**.

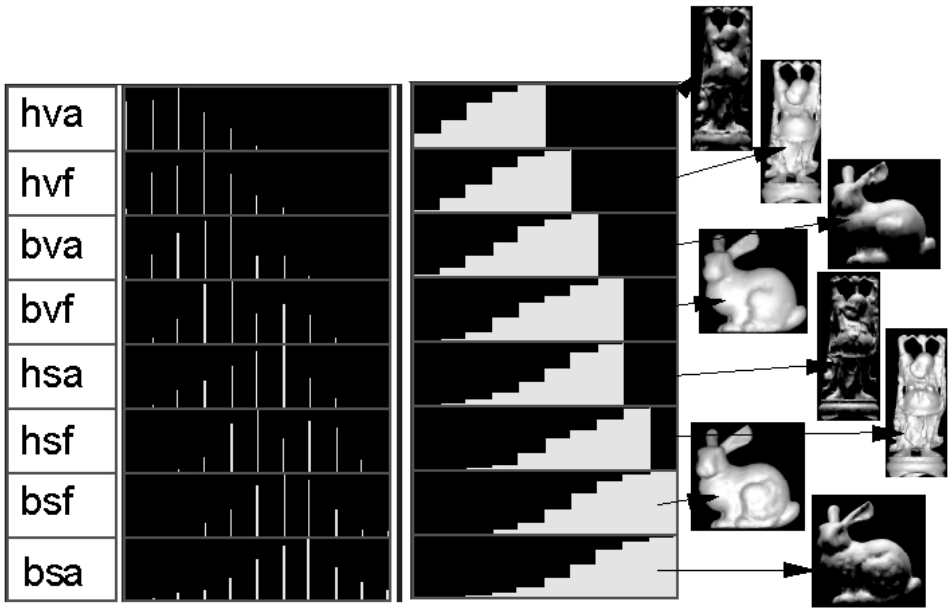


Figure 5. Histogram results for all eight conditions in the still image experiment. The first column on the left shows histograms, the second column shows cumulative histograms, and images representing each condition are shown on the far right.

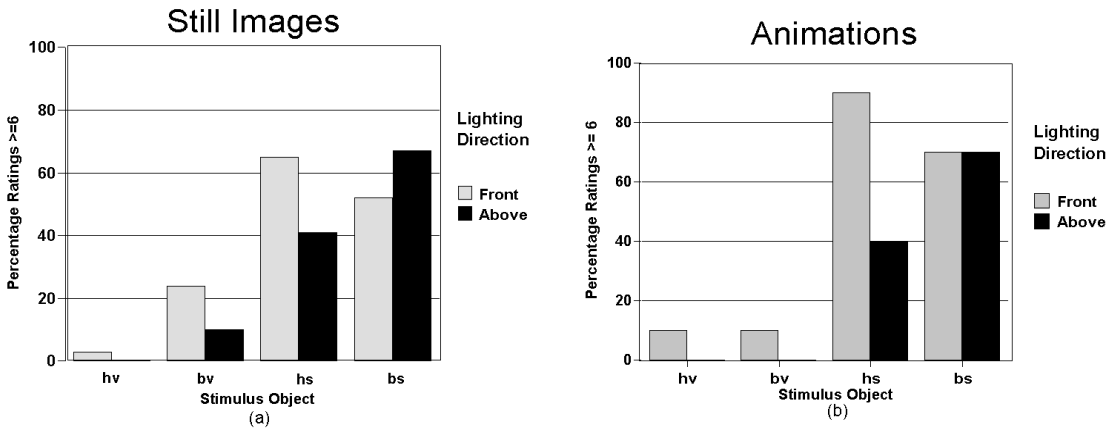


Figure 6. Charts showing the percentage of trials with high image quality (% scores $>= 6$) for the 8 conditions. Fig. (a) shows the results for the still images, Fig.(b) shows the results for the animations.

3.3. Effects of Lighting Direction on Perceived Quality

Figure 6 summarizes the effects of lighting on the perceived degradation of simplified objects. As a dependent measure, we compute the proportion of total trials where the simplified object is rated as having good quality relative to the original (% scores $>= 6$). The light bars indicate those conditions in which the object is lit from above; the dark bars indicate those conditions in which the object is lit from the front. The chart to the left shows data when judging 2-D images of the objects and provides a summary of the data in figure 5. Each bar represents the percentage of trials across observers (10) and positions (15), $n = 150$, rated $>= 6$. The chart to the right shows data when judging animations of the 3-D objects. Each bar represents the percentage of trials across observers ($n = 10$) rated $>= 6$. In Fig. 6b we see that the quality judgments made for animations of 3-D objects are in some ways similar to those made for their 2-D projections. For example, the very simplified models receive systematically lower scores than their less simplified counterparts. The greater robustness of the bunny model relative to the happy Buddha however is not replicated, and most strikingly the effects of lighting are quite different for the better quality (less simplified) models. In particular the superiority of lighting from the front is much more pronounced for the simplified happy Buddha model **hs**. Lighting from above produces comparable results, but under animation lighting from the front produces a considerably higher proportion (80 %) positive scores. That is , the visual effects produced by simplification are reduced under lighting from the front when that object is animated. Under animation, the perceived quality of the object when lit from the front, is increased.

3.4. Geometric Metrics and Perceptual Quality

Figure 7 explores the degree to which a standard computer graphics metric for measuring object simplification captures the perceived quality of these objects. Here rated quality is plotted as a function of a standard measure of object degradation. The lower the maximum distance as a percentage of the bounding box, the less distortion. If this measure captures perceived quality, then the quality rating scores should decrease linearly with this measure, and there should be no difference between objects lit from above and those lit from the front.

In this figure, we consider the results of the two experiments separately. The results are remarkably similar. Whether the judgments are made on static 2-D images or animated 3-D objects, perceived quality decreases monotonically and linearly for models lit from the front. For front lighting r^2 is -0.98 for still images and -0.95 for animated sequences. However perceived quality is not monotonic for models lit from above. The large difference between the two lighting conditions, suggests that a purely geometric description will not be adequate to describe these results. As with the still images, the perceived quality of 3-D animated geometric objects depends critically on the incident angle of the lighting.

3.5. Object and Image Quality

In our experiments, we measured the degree to which animated objects and their 2-D image projections were degraded perceptually by geometric simplification. In order to use an image-based quality metric to describe the perceived degradation of a

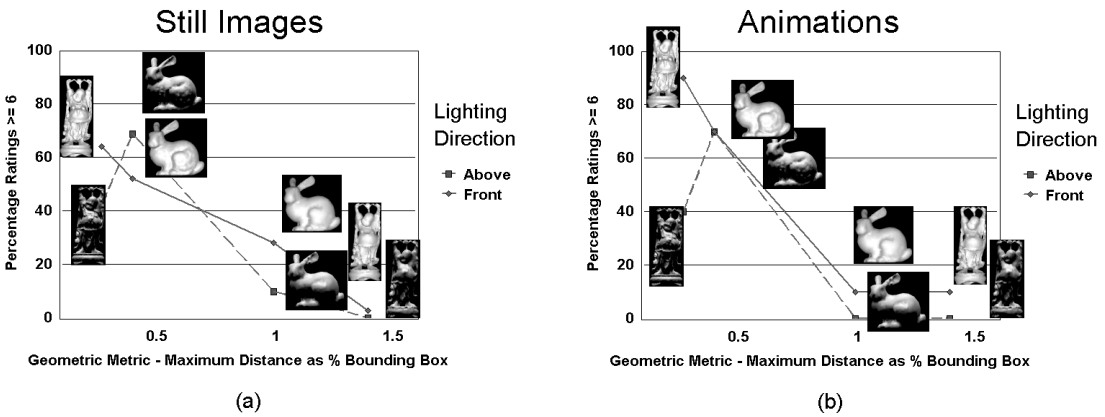


Figure 7. Perceptual ratings versus a standard geometric metric for simplification, the maximum distance of the simplified geometry from the original as a percentage of bounding box diagonal. Fig. (a) shows the ratings for the still images, (b) shows the ratings for the animations. Although perceived quality is linear with this metric with lighting from the front, it is not even monotonic with lighting from above. Furthermore, this metric does not account for differences between lighting conditions.

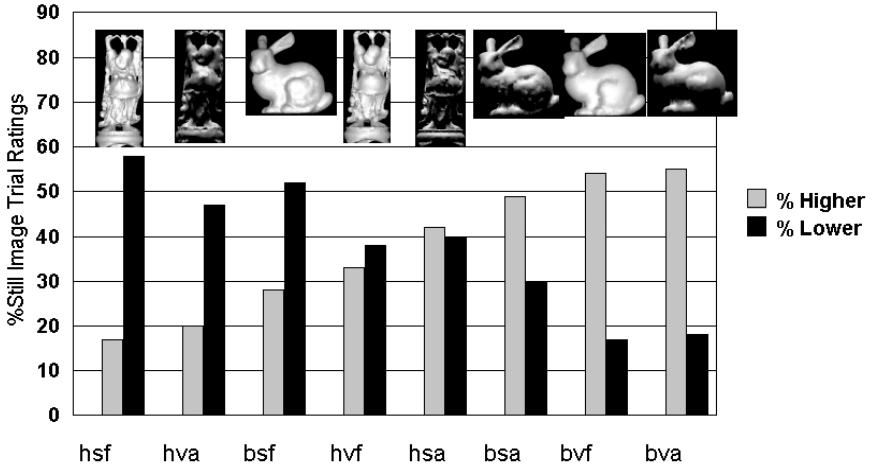


Figure 8. Comparisons of the ratings of the still and animated cases. The proportion of trials in which the still image is given a higher score than the animated object is shown by the light bars, and the proportion of trials in which the still image is given a lower score is shown by the dark bars.

3-D object, we would need to show that the degree of perceived degradation of the still image is comparable to the perceived degradation of the animated object. Figure 8a explores this hypothesis.

Instead of counting the percentage of "good scores," we count the proportion of trials when the still image was given a higher score than the animated object and the proportion of trials when the still image was given a lower score. If the degree to which geometric simplification degrades perceived quality is the same for images and animations, then we expect the proportion of lower and higher scores to be the equal. For example, if an observer assigns a rank of 6 to an animated object, then the expected value for the 2-D projections should also be 6, with an equal proportion of scores above and below this expected value.

In Fig. 8, the light bars show the proportion of the still images trials rated higher than the animated object; the dark bars show the proportion of trials where the still images were rated lower in quality than the animated object. In this chart, we see that for three of the test stimuli, the still images are consistently rated lower than the animated object (dark bars smaller); for three of the stimuli, the still images are consistently rated higher in quality than the animated object. For two of the stimuli, the perceived degradation due to the geometric simplification is the same, whether the observer is rating the images or the animation.

It is clear that the degree of perceived degradation of the still images does not adequately predict the perceived degradation in the equivalent animated images.

4. CONCLUSIONS

In these experiments, we explored the perceived quality of two different representations of 3-D objects, in order to better understand how to characterize and measure the effects of geometric simplification. To do so, we selected a simple and a complex geometric model, Bunny and Happy Buddah, and created two simplified versions of each. We used these stimuli to study the effect of geometric simplification on perceived quality by having observers rate the quality of these stimuli relative to their unsimplified originals. Since, in previous experiments, we had observed striking differences in the quality of 2-D images of 3-D objects depending on the direction of the lighting, we varied lighting explicitly, making all measurements with lighting from above and with lighting from the front, aligned with the observers viewing direction. Although geometric measures of model simplification are based on the 3-D geometry of those models, the two existing experiments aimed at developing a more perceptual model were based on the evaluation of 2-D projections of those objects. In particular, it had been suggested that the quality of 3-D objects could be predicted based on the quality of static 2-D projections. To explore this hypothesis, we conducted our experiments under two different conditions. In the first condition, the observers rated the quality of 2-D static images, 15 projections of the 3-D object, at 5-degree angles along a quarter-rotation from side-to-front view. In the second condition, the observers rated an animated sequence of these 15 images, showing the object rock back and forth from side-to-front-to side.

When compared with the unsimplified original, the more the object was simplified, the lower the rating scores, for both animated and static presentations. Perhaps the most striking result of these experiments is the remarkable effect of lighting on perceived quality. In 7 of the 8 conditions tested, judgments of perceived quality depended on the direction of the lighting. Furthermore, the degree to which lighting direction affected the quality judgments depended on whether static or animated images were being judged. For the two more complex objects, the simplified Buddah (**hs**) and the simplified Bunny (**bs**), the perceived quality of the rendering increased significantly when the object was animated, if that object was lit from the front. That is the degree of geometric simplification was much less noticeable in animation for simplified objects lit from the front. This suggests that certain simplified geometric objects might be best lit from the front, rotating. Since the human visual system is less sensitive to high spatial frequencies when an object is moving, it may be that the rotation reduces the detectability of high spatial frequency artifacts introduced by the simplification process. This result may not be observed when the object is lit from the front, since the contrast of these artifacts is so high that their effect cannot be effectively attenuated. It would be interesting to explore this idea further by explicitly varying the spatial and temporal frequency composition of the stimuli.

Since one goal of this work is to develop a metric for characterizing the perceived quality of geometrically simplified graphical objects, we examined the perceptual rating scores as a function of a standard metric for characterizing the geometric effect of simplification, the maximum distance of the distortion as a percentage of the bounding box. We found that for objects lit from the front, perceived quality decreased linearly with this measure of geometric difference. Rated quality did not decrease monotonically for these same objects when lit from above. According to this metric, the simplified happy Buddah (**hs**), with $\%BoundingBox = 0.27$ should have much higher perceived quality than the simplified bunny (**bs**), with $\%Boundingbox = 0.4$. Instead, we found for both static and animated viewing conditions, the opposite was true. The **hs** stimulus had consistently lower scores, indicating much lower perceived quality.

Perhaps the most important observation to make regarding the geometric metric is that it does not predict any difference between the two lighting conditions. Since the geometry of the object is unchanged when the lighting direction is changed, the metric sees these as identical. Perceptually, however, lighting direction is a critical factor in judging the amount of distortion produced by the simplification.

An important criterion for selecting a metric is that address the difference between the perceived quality of still and animated images. This difference is not a simple shift in sensitivity that can be accounted for with a normalization factor. In some cases, the static images were consistently rated higher in quality than the animated sequence; in some cases the static images were consistently rated lower than the animated sequence. The pattern of these results is not straightforward, and more work needs to be done. The clear conclusion, however, is that even if we had a metric that completely characterized the perceived quality of static 2-D projections of 3-D objects, this metric would not predict the quality of 3-D animated sequences of those same images.

REFERENCES

1. B. Watson, A. Friedman, and A. McGaffey, "Using naming time to evaluate quality predictors for model simplification," in *Proceedings of ACM SIG CHI Conference*, pp. 113–120, 2000.
2. H. Rushmeier, B. Rogowitz, and C. Piatko, "Perceptual issues in substituting texture for geometry," in *Proc. SPIE Human Vision and Electronic V*, vol. 3959, pp. 372–383, 2000.
3. P. Cignoni, C. Montani, and R. Scopigno, "A comparison of mesh simplification algorithms," *Computers and Graphics* **15**(1), pp. 37–54, 1998.
4. P. Lindstrom and G. Turk, "Image-driven simplification," *ACM Transactions on Graphics* **19**, pp. 204–241, 2000.
5. A. Gueziec, "Locally toleranced surface simplification," *IEEE Transactions on Visualization and Computer Graphics* **5**, pp. 168–169, 1999.
6. P. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, MA, 2001.

Perceptual Issues in Substituting Texture for Geometry

Holly Rushmeier[a], Bernice E. Rogowitz[a], Christine Piatko[b]

[a]IBM TJ Watson Research Center, P.O. Box 704, Yorktown, NY, USA
[b]Johns Hopkins Applied Physics Laboratory, 1100 Johns Hopkins Rd., Laurel MD, USA

ABSTRACT

An important goal in interactive computer graphics is to allow the user to interact dynamically with three-dimensional objects. The computing resources required to represent, transmit and display a three dimensional object depends on the number of polygons used to represent it. Many geometric simplification algorithms have been developed to represent the geometry with as few polygons as possible, without substantially changing the appearance of the rendered object. A popular method for achieving geometric simplification is to replace fine scale geometric detail with texture images mapped onto the simplified geometry. However the effectiveness of replacing geometry with texture has not been explored experimentally.

In this paper we describe a visual experiment in which we examine the perceived quality of various representations of textured, geometric objects, viewed under direct and oblique illumination. We used a pair of simple large scale objects with different fine-scale geometric detail. For each object we generated many representations, varying the resources allocated to geometry and texture. The experimental results show that while replacing geometry with texture can be very effective, in some cases the addition of texture does not improve perceived quality, and can sometimes reduce the perceived quality.

Keywords: computer graphics, geometric simplification, texture, perception

1. INTRODUCTION

Graphics systems can make intensive use of available computational resources. Given the complexity and detail of geometric models available, trade-offs must be made in graphics rendering to balance between interactivity and perceptual quality. Advances in networking have also focused attention on the use of rendering approximations in order to enable the efficient use of bandwidth-constrained resources for distributed graphics applications. Different choices of rendering approximations will affect perception in different ways. Thus, it is desirable to develop perceptual measures to understand the impact of texture and geometry approximations. In this paper we study the implications of substituting texture for geometry.

Many geometric simplification algorithms have been developed to provide interactivity and reduce bandwidth requirements. The goal of these algorithms is to achieve a perceptually acceptable representation with minimal resource requirements. Ultimately, both the acceptable perceptual quality and resource requirements are dependent on the particular computing environment. The acceptable perceptual quality would be quite different for an e-commerce application where a customer is examining an object to make a decision whether to buy it, versus a game environment in which the player may be willing to accept a level of artificiality. The resource limitations would be quite different for an application in which transmission over the Internet is the bottleneck, versus an application where interactive frame rate on a specific workstation is required. Before we can address these issues, however, we need to enhance our understanding of the interplay of graphics resource allocation and perceptual quality.

Very little work has been done to explore the perceptual effects of different simplification schemes. For simple sprite representations, Horovitz and Lengyel [1] considered trade-offs in the perceptual and computational costs. Watson et al. [2] consider naming time as a predictor of the perceptual quality for various levels of geometric simplification. They considered only triangle reduction – not methods which replace geometry with texture. They found that both geometric and image metrics correlated poorly with the quality indicated by naming time.

The goal of our work is to provide a general framework for evaluating the perceptual effects of geometric simplification, and geometry-texture trade-offs. We do not specify a task or a particular network/workstation setting. Instead, we create simple, well-controlled stimuli, explicitly varying the geometry, texture and illumination. We introduce a simple scaling procedure as a method for exploring fundamental questions in the area of geometry/texture allocation. Does texture replacement always result in better perceptual quality for a given resource allocation? Are certain types of geometry more suitable for texture

Further author information: (Send correspondence to H.R.)
H.R.: holly@watson.ibm.com, B.R.: rogowtz@us.ibm.com, C.P.: Christine.Piatko@jhuapl.edu

Figure 1. In computer graphics, a solid object (leftmost image) is often represented as a dense mesh of triangles (second image from left.) To facilitate interactive display performance, objects are often simplified (third image from left), to reduce the number of triangles (rightmost image), while attempting to preserve the appearance of the object.

replacement? Are there different rules for different classes of objects, different viewing conditions? What is the impact of lighting?

To evaluate the degree to which adding texture could compensate for simplifications in geometry, we measured the perceived fidelity of two objects, a smooth sphere and a crinkly sphere, lit from the front or obliquely from the side. In each experiment, twelve test stimuli were created using three levels of geometric resource and four levels of texture resource (including a no-texture condition). In these experiments, psychophysical data were obtained from eight observers who rated the degree to which each of the test stimuli compared with a comparison stimulus with full geometry and full texture resource.

2. GEOMETRIC SIMPLIFICATION AND TEXTURE MAPPING

A wide range of representations can be used to model the geometry of an object – including tensor spline surfaces and quadric patches. For interactive display, typically all representations are converted to a set of polygons approximating the surface. For curved or complicated objects, thousands of polygons may be used. Usually networks of triangles are used, such as the example shown in Fig. 1 since triangles are guaranteed to be non-self-intersecting, facilitating hidden surface calculations. The time to display a surface depends directly on the number of triangles. Interactive display rates require that the number of triangles be limited. To maintain interactive rates, many algorithms have been developed to reduce the number of triangles in a given model while maintaining the visual appearance.

2.1. Simplification

Many different simplification methods have been proposed. Here we present just a brief overview. A comprehensive discussion can be found in an article by Cignoni et al.[3] Virtually all of the methods are associated with parameters that affect appearance, but none are based on perceptual data. One general class of methods creates simplified geometries by a series of small incremental simplifications. Popular incremental approaches are vertex removal and edge collapse. In vertex removal methods such as the one proposed by Schroeder et al.[4] one point in the mesh is removed, and new triangles are defined to fill the resulting hole. In edge collapse methods such as the one developed by Gueziec,[5] one edge is removed, and the two vertices at the ends of the edge are merged into one.

In any type of incremental method, a choice needs to be made to determine the next best vertex or edge to remove. The choice is generally made on the basis of a geometric metric. Common metrics include selecting the change that results in the smallest change in surface area, or selecting the change that results in a surface that is the closest to the original vertices.

Other methods work globally, such as clustering or volumetric methods.[6] These impose a maximum length scale on the simplified model by coalescing all vertices within a prescribed distance to one another into a single vertex. These methods have the advantage that they allow the object topology to change.

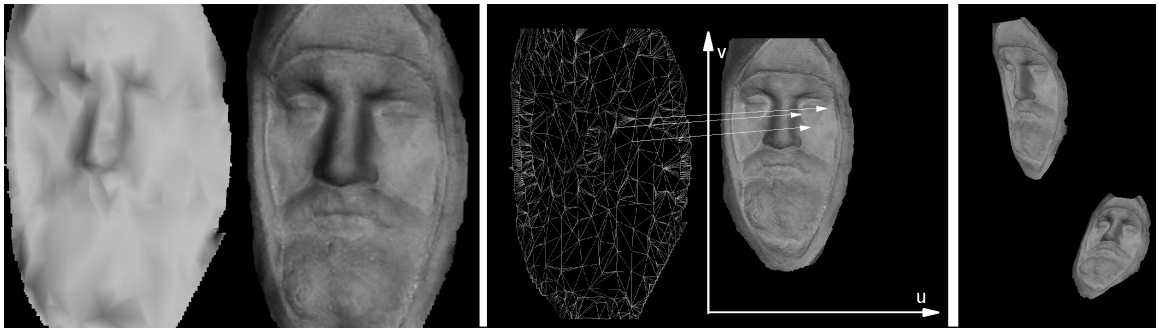


Figure 2. A simplified geometry can be given a detailed appearance using a texture map, as shown in the before and after images on the far left. At each vertex (u, v) indices are stored , which are locations in an image, as shown in the center diagram. Details inside the triangle are "painted" on the triangle using the image pixels within the corresponding triangular area on the texture image. A texture-mapped object can be viewed from any direction (right-most image).

Recently more attention has been paid to visual impression rather than purely geometric measures. On one extreme some efforts seek to find the simplest symbolic representation of an object – for example representing a hand with a skeleton of five line segments. At the other extreme, some efforts attempt to produce a pixel by pixel identical image of the object by selecting the correct level from a hierarchical description of the object.[7] The object is rendered to the image plane progressively, at each step refining the representation until further refinement does not influence the image. This approach can account for both the particular view and lighting conditions. However, it does require the observer to wait while the object is progressively rendering on the screen.

2.2. Replacing geometry with texture

The most successful efforts to date to maintain visual appearance with a relatively small number of triangles make use of texture mapping. Texture maps provide the illusion of detail. Even though texture is a flat image, when they are attached to the geometry new views can be obtained. Texture maps that contain color and fine scale detail associated with lower resolution geometry have been used for decades in computer graphics, [8] [9] .

In texture mapping, each vertex in a model is associated with a (u, v) coordinate pair, where u and v each vary from zero to one. As diagrammed in Fig. 2, the (u, v) coordinates correspond to a location in an image. The detailed value for any point in the geometry is found by interpolating the (u, v) coordinates of the vertices and looking up the corresponding location in the texture image. Storing detailed data in image maps is efficient because images do not require the explicit storage of positional or connectivity data. Images to represent detail may simply store colors. Bump maps are a type of texture image in which each pixel value represents a small height deviation of the detailed surface from the underlying surface. Normals maps store a vector quantity at each pixel representing surface normal.

There have been many successful methods for simplifying large sections of complex scenes by generating a texture map for a single polygon, or small number of polygons, [10] [11] . A major feature of the Talisman graphics architecture proposed by Torborg and Kajiya[12] was the representation of three dimensional objects as small polygons, or "sprites" that could be translated and warped. Soucy et al.[13] moved beyond mapping textures to single polygon by developing a systematic method for deriving color texture maps to represent color detail on a simplified version of a dense triangle mesh.

Display hardware for texture mapping colors is available at relatively low cost on personal computers, and is routinely used in computer games. A difficulty with using a color image for a texture map is lighting the object. There are two options: compute the lighting based on the simplified geometry, or use a color image that includes the effect of light. The problem with the first option is that the lighting variations make the simplified geometry visible, the problem with the second option is that the lighting cannot be changed.

Display hardware that performs dynamic lighting changes using bump or normals maps is rapidly becoming more widely available. The hardware interpolates the detailed value of the surface normal from maps and performs the dot product with the light direction in real time. The calculation performed in hardware to compute a detailed lit texture is diagrammed in Fig. 3. Cohen et al.[14] recognized the importance of dynamically relighting details and developed a method for computing both normals and color maps for a simplified version of a complex mesh.

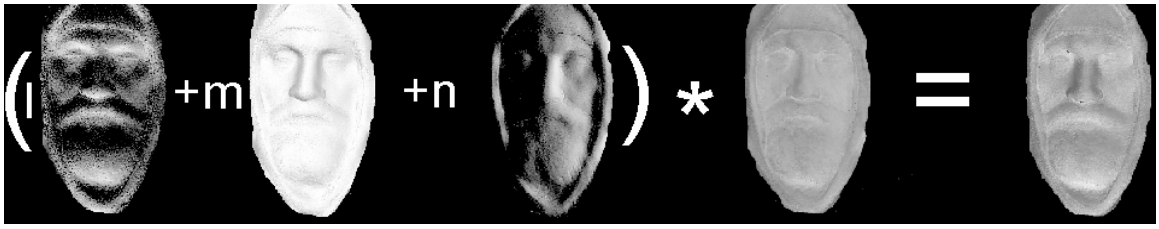


Figure 3. Texture maps can be dynamically relit by taking the dot product of light source direction $[l, m, n]$ (as shown in first three images from left), multiply by the image storing surface color (or simply albedo) and producing a new image (far right).

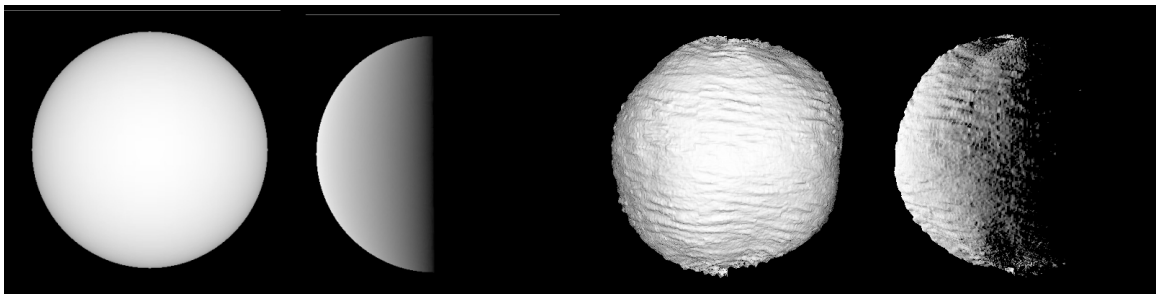


Figure 4. The objects used in our experiments were a sphere lit from the lef (leftmost image), a sphere lit from the right (image second from left), a crinkled sphere lit from the front (image thrid from left) and a crinkled sphere lit from the left (rightmost image.)

3. EXPERIMENTAL DESIGN

We examined perceptual trade-offs in replacing geometry with texture for objects of uniform color and surface finish. We examined abstract shapes, to avoid issues of semantic interpretation. We constructed our experimental stimuli in order to have object representations that allowed us to control the geometric and texture simplifications separately. Suitable numerical object representations were not readily available, since existing collections of test objects did not include the two-dimensional parameterization required for the texture maps. A few texture mapped objects were available on the Internet, but the textures were in the form of one texture per triangle, and so were not suitable for filtering to varying resolutions. For our experiment, test objects were constructed using IBM Visualization Data Explorer (DX). DX is a flexible visual programming system, and is available without cost as open source from http://www.opendx.org/

The stimuli used were a sphere and a sphere with a crinkled surface, each represented by varying levels of geometry and texture, viewed under direct or oblique lighting. We used a psychophysical scaling procedure to measure the perceived fidelity of each representation of each object, relative to a "perfect" reference representation.

3.1. Preparation of Stimuli

The two test objects are shown in Fig. 4. The objects were defined starting with a 512 x 512 grid of points, and warping the grid into a sphere with unit radius. Normals were computed at each grid vertex, a dot product was performed with lighting direction, and the results were stored in a 512 x 512 image to be used as a texture map. Texture (u, v) coordinates were stored at each vertex, with a one-to-one correspondence between object vertices and texture image pixels for the full representation.

The geometry was simplified using the DX *Simplify* module which is an implementation of Gueziec's geometric simplification.[5] Gueziec's method produces a simplification that guarantees that the resulting surface is within a specified distance of all the original vertices. *Simplify* maintains the (u, v) indices at the vertices remaining after each simplification, so all version of the sphere could be texture mapped with all versions of the texture image. Lower resolution versions of the lit normals maps were computed with box filtering to reduce aliasing artifacts.

The distribution of crinkles on the second object was obtained by painting intensity variations on a 512x512 image. This grey scale image was imported into DX, and the grey levels were interpreted as radial perturbations ranging between 0 and 0.1

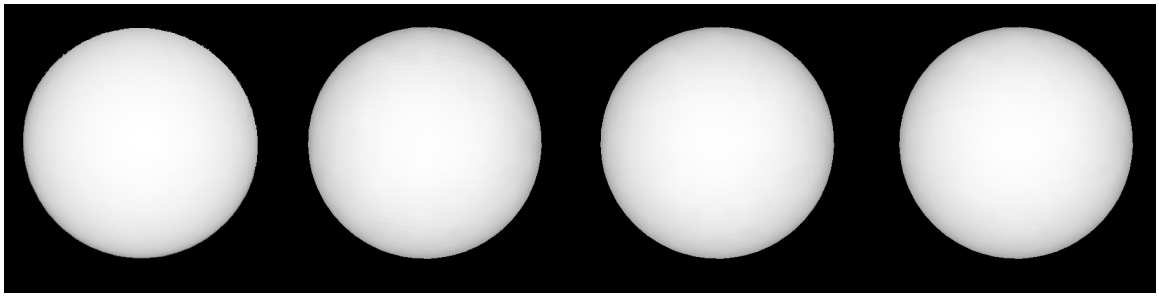


Figure 5. The medium geometry sphere lit from the front, with geometry only (far left), with simple texture (second from left), with medium texture (second from right) and with full texture (right).

on the unit sphere. After the full crinkled geometry was constructed, the normals maps and simplified geometries and images were computed as for the case of the sphere.

For each of the objects we generated two levels of simplification by experimenting with the distance error parameter in Gueziec's method to produce simplified objects that spanned a range of visual quality. For each object then we have representations we will refer to as full, medium and simple. We measure the resource required for each geometry as the compressed ascii file size of the DX representation. Clearly, this size would vary depending on the particular format used, and the particular geometry viewer used. The levels of geometric simplification were selected based on the complexity of the model. For the smooth sphere, the geometric resource was 4.4 Mb for the "full" geometry, .093 Mb (a reduction by a factor of 47) for the "medium" geometry and .047Mb (a further reduction by a factor of 2) for the "simple" geometry. For the crinkled sphere, the geometric resource was 14.4 Mb for the "full" geometry, 6.39 Mb (a reduction of a factor 2.3) for the "medium" geometry and .24 Mb (an additional reduction by a factor of 26.6) for the "small" geometry.

The same four levels of texture were selected for all four stimuli: the original 512x512 image as the "full" case, a 256x256 (factor of 2 reduction in resolution) image "medium" case, and 64x64 (additional factor of 4 reduction in resolution) as the "simple" case. The resource required is the uncompressed tif file for each of these images, since the texture memory required is the expanded image size. In terms of memory the "full" image is .787 Mb, the "medium" image is .197 Mb (a factor of 4 reduction), the "small image is .0122 Mb (an factor of 16 reduction) and 0 Mb for no texture. We use this size measure to illustrate the general framework for evaluation; the correct measures to use would depend on the specific task and computing environment.

We consider the effect of two lighting conditions, illustrated in Fig. 4. The first condition is with light parallel to the view direction. This is typical of the "headlight" lighting used in most computer graphics systems used to view individual objects. The second lighting condition was a light perpendicular to the view direction that causes a relatively abrupt, attached, shadow. Such shading could occur when an object is used as part of a virtual environment.

For each object and lighting condition, images were computed in which the object height corresponded to approximately 370 pixels on an image with black background. This size was chosen so that a few of the images could be displayed simultaneously on a 1280 x 1024 monitor. Twelve images were computed for each object/lighting combination, covering all combinations of simple/medium/full geometry and no/simple/medium/full texture. Samples of the various combinations are showin in Figs. 5 and 6. For the geometry only case, the image was formed by taking the dot product of vertex normal and light source direction at each vertex, and then using Gouraud shading for the smooth surface display. Note that for the full geometry case, the result of this technique is pixel-by-pixel identical to mapping the full texture on the geometry.

3.2. Procedure

Eight observers participated in the experiment. They were told that they would be asked to score images of various representations of an object relative to a "perfect" object (i.e. the image of the full geometry) using a scale of 0 to 100. A score of 100 indicated a perfect match. The observers were shown a variety of the comparisons they were to make from the four tests sets, and told to try to assign 0 to the worst match or matches. The observers were then asked to scale each of the twelve images in each of the 4 test sets, with the 4 test sets presented in a random order. The observers viewed a full-size, "perfect" image in one corner of the display screen, and a matrix of thumbnails of the objects to be rated. For each evaluation, the observer clicked on the thumbnail to expand it to full size. The observers were free to score each of the 12 objects in any order they liked, and were given no time limit. The time to complete all 48 comparisons was typically half an hour.

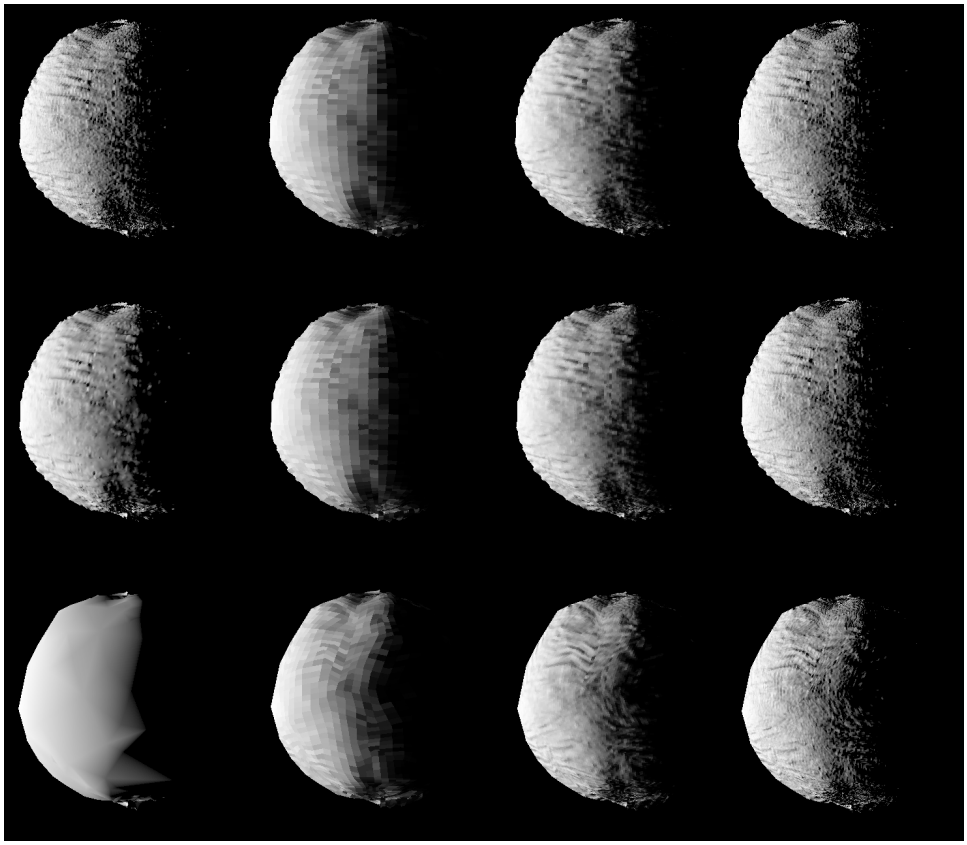


Figure 6. All of the representaions of the side lit crinkled sphere presented to viewers: geometry varies from full (top row) to simple (bottom row), texture varies from none (left column), to simple (second column from left) to full (right column).

4. RESULTS

The results of the observer scores are summarized in Figs. 7 to 9.

4.1. Validity of Results

Figure 7 shows the distribution of scores assigned by the eight observers in this experiment, across all four stimulus conditions. There was a high degree of intra-subject concordance. All observers used the whole range of scores from 0 to 100, and the responses are evenly distributed over that range. The number to the right of each distribution gives the value of r^2 for the correlation of that observer's scores with the mean scores for all observers. When each test set is examined individually, r^2 is systematically higher.

4.2. Perceptual Scaling Results

Figure 8 shows the mean rating scores for the four test stimuli, the smooth sphere viewed from the front (Sphere-Front), the smooth sphere viewed from the side (Sphere-Side), the crinkly sphere viewed from the front (Crinkle-Front) and the crinkly sphere viewed from the side (Crinkle-Side). For each stimulus the observer compared the full geometry original with 12 alternate representations varying in texture and geometry. The four texture levels are shown along the x axis, with the no-texture condition at the extreme left. The three geometry conditions are shown along the y axis.

Sphere-Front. The plot in the top left quadrant of Fig. 8 contains results for the smooth regular sphere, lit front on. Data for the control stimulus (full geometry and full texture) is shown by the top right column in the set. Since this is identical to the comparison stimulus, it should be judged as equal (100). For this "sphere-front" stimulus, there was a clear effect of geometry. Increasing the geometry by a factor of 2 (from .047 to .093) caused the quality ranking to roughly double. An additional factor of 47 increase in the geometric resource, however, did not produce a further increase in perceived fidelity. Said the other way

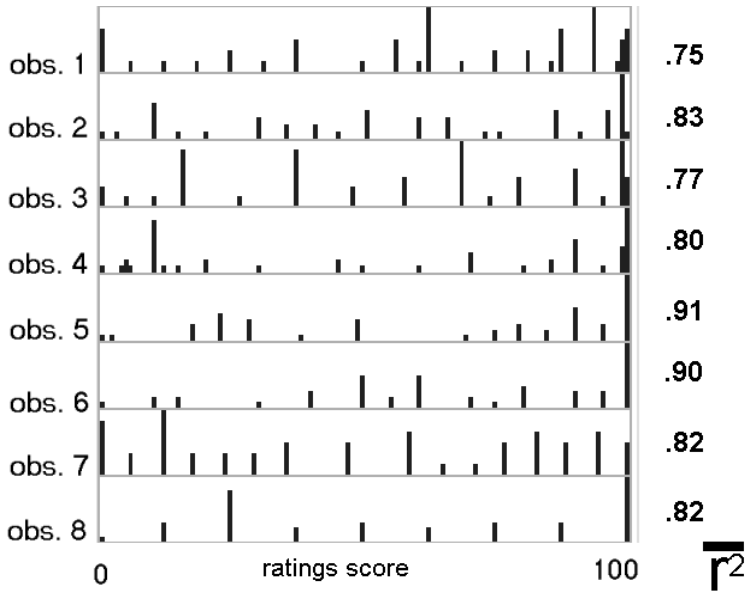


Figure 7. The distribution of scores assigned by each observer across the 4 test sets. The correlation coefficient r^2 shows the degree of correlation between each observer's score and the mean score for all observers. Each observer used the full range of scores. The individual scores correlated well with the mean scores.

around, this smooth shape, viewed front on, is impervious to geometric distortion. Reducing the geometric resource by a factor of 47 had no significant effect on the ratings. It was only perceived to be of reduced quality when the geometry was decreased by an additional factor of 2.

On this smooth sphere, variations in the texture resource had no effect on perceived image quality. The left-most set of columns shows the rating data for the three levels of geometric simplification with no texture added. For all three levels of geometric simplification, there was no change in the rating score with increases in the quality of the texture. For this low spatial-frequency texture, illuminated from the front, the rating score was driven entirely by the underlying geometry and was independent of texture.

Sphere-Side. When the sphere was viewed under oblique illumination (bottom left panel), the mean rating score dropped systematically for each decrease in geometry. This result occurred in the no-texture condition, and at all three levels of .

Under this illumination, adding texture had a significant effect on the perceived quality. Low resolution texture (.0122 Mb) degraded the perceived quality for all levels of geometric simplification. This may be due to the fact that under these lighting conditions, the pixel structure of the undersampled texture is very visible, producing an image which is distinctly different from the smoothly-shaded comparison stimulus. For the most simplified geometry (.047 Mb), additional texture did little to improve the perceived quality of the smooth sphere. For the less simplified models, using a less simplified texture significantly improved perceived quality, but with diminishing returns. The first factor of 16 (from .0122 to .197 Mb), produced a big improvement, but the next factor of four produced no additional effect. This may be because the factor of 16 jump in texture resolution was sufficient to eliminate most of the pixelation noise.

Can the degradation in perceived quality produced by the decrease in geometry be compensated for by adding texture? For the middle level (.093 Mb) geometry, the perceived quality of the geometry-plus-texture stimuli was always less than the perceived quality of geometry alone. Trying to compensate for reduced quality by adding texture would be a waste of resource.

Crinkle-Front and Crinkle-Side. The second model is the sphere with a highly textured surface, viewed under direct illumination (top right quadrant) and under oblique illumination (bottom right quadrant). Since the data for the two cases are quite similar, we will discuss them together. Looking first at the left-most columns, we see that reducing the geometry significantly reduced the perceived fidelity. For most simplifed geometry (.24 Mb), every increase in texture resource, even if that texture was highly subsampled, produced an increase in perceived fidelity. This can be seen clearly by looking at the bottom row of Fig. 6. The figure at the bottom left is the low-resolution geometric object without texture. The three images to its right

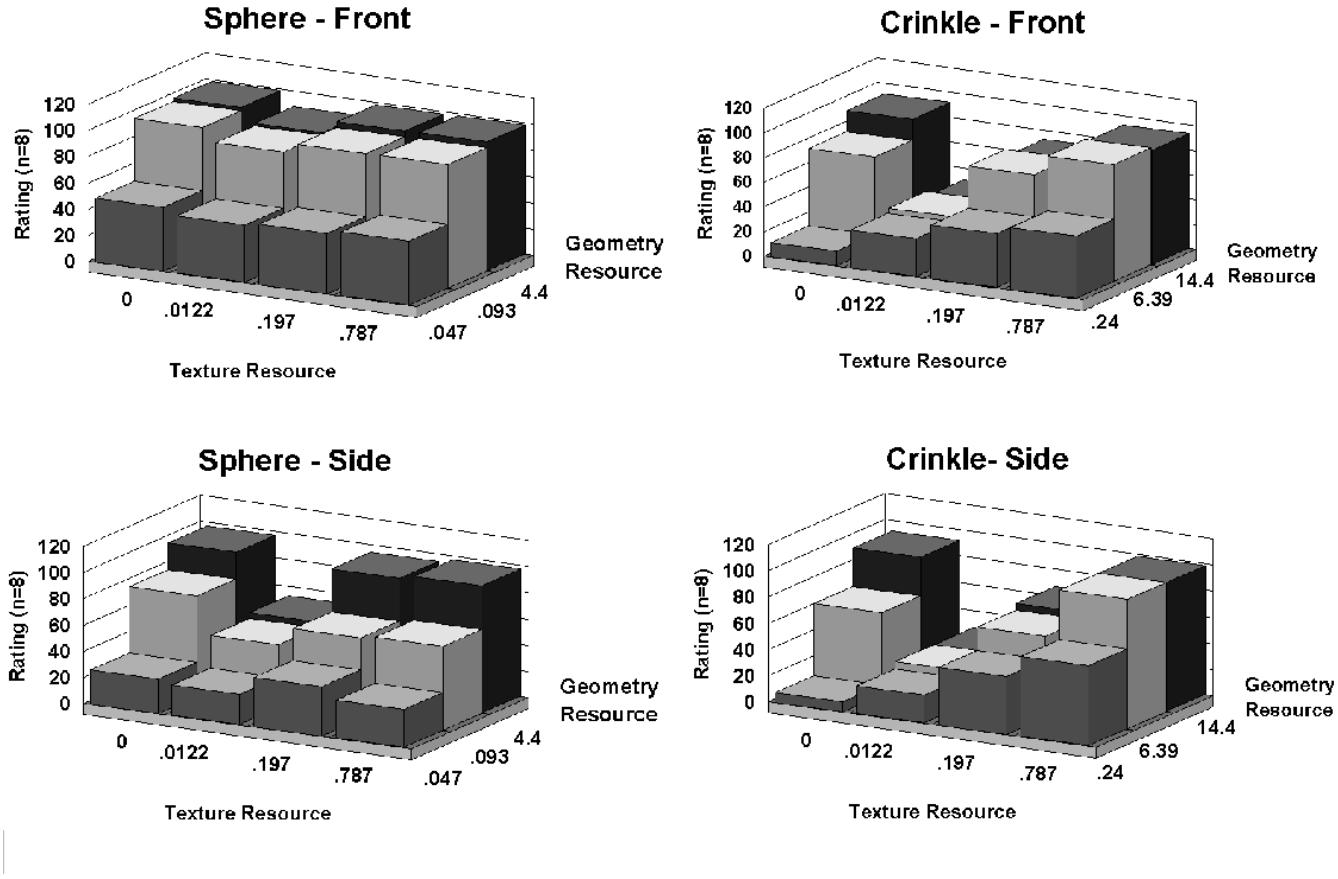


Figure 8. The mean scores for each of the four test sets are shown as bar charts. The bar in the back left of each chart corresponds to the full geometry. A comparison of the front- and side-lit sphere demonstrates that illumination can have a significant impact on perceived quality. A comparison of the sphere and crinkled sphere demonstrates that the effect of adding texture is different for these two objects. In particular, a comparison of the front row of bars (simplest geometry) shows that increasing texture does not improve perceived quality for the sphere, but does improve perceived quality for the crinkled sphere.

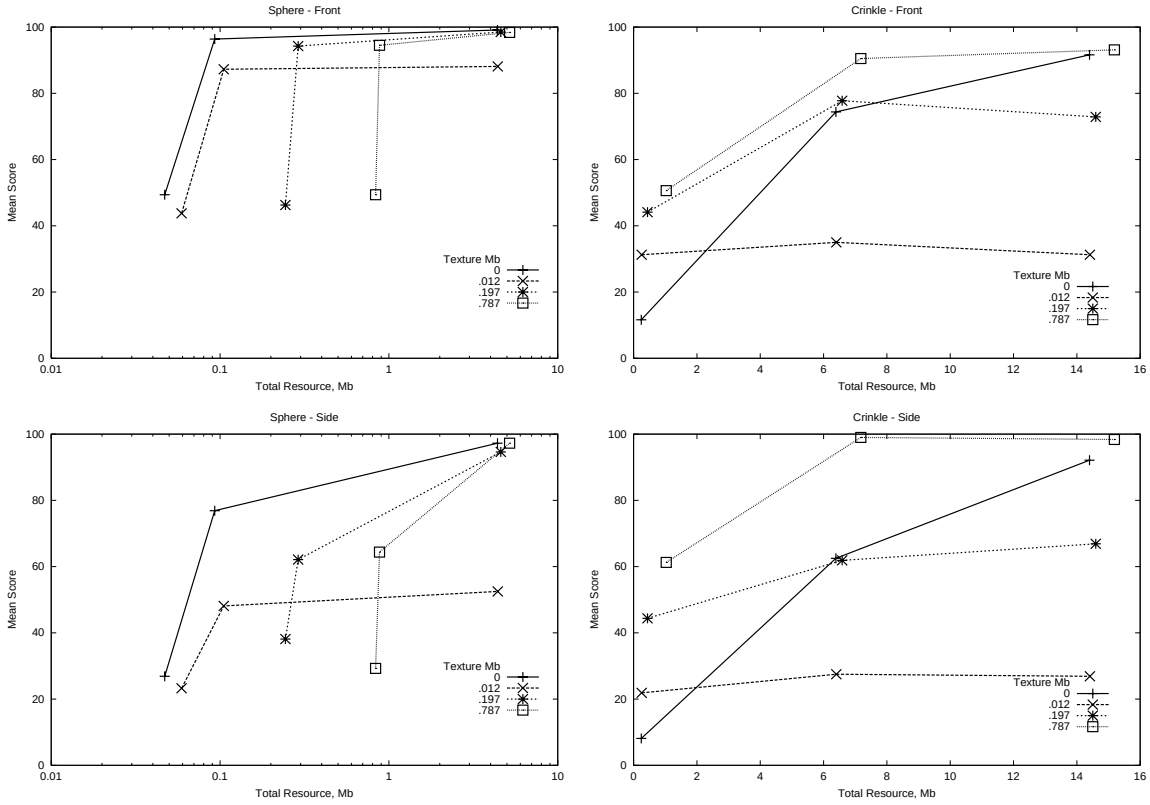


Figure 9. The mean scores for each of the four test sets are shown as line charts with mean score versus total memory resource. The 0 Mb texture case corresponds to geometry alone. For the sphere, the geometry alone always gives the best perceived quality at each resource level. For the crinkled sphere the full texture (.787 Mb) always gives the best perceived quality at each resource level.

show the effect of adding three resolution-levels of texture, respectively. Apparently, approximating the high spatial-frequency surface with an imperfect texture is better than no texture at all, for highly simplified geometries.

For the other geometries, which were much less simplified, adding a low-resolution texture (.0122 Mb) significantly decreased the perceived quality of the object. This may be because the simplified texture had a discriminably different (lower) spatial frequency than the perceived texture of the geometrically simplifed object. Adding additional texture resource increased the level of perceived quality.

4.3. Resource Trade-offs

Another question to ask of these data is what combination of texture and geometry gives the best perceptual effects? To answer this, we need to compare the perceptual ratings with the resources consumed in producing the stimuli. For example, in the above case, we see that for the crinkle spheres, the same perceived quality is obtained for low and high geometry, at .197 Mb texture resource. That is, a factor of 2 (7 Mb) is wasted by rendering the object with full geometry.

Figure 9 explores this issue more fully. Here the data for each of the four conditions are plotted in the same configuration as the plots in Fig. 8. Each quadrant plots the mean rating score across subjects as a function of the total resource, where total resource is the sum of the geometric and texture resource values, in megabytes. For the sphere front and sphere side stimuli, where there was a factor of 200 difference in resource, the values are plotted on a logarithmic scale. One approach to interpreting these data is to compare the perceived quality of the model when it is created with geometry alone (no texture), to the cases where texture has been added. The no texture case is shown as a full line. The horizontally shifted curves show the additional resource contributed when the three levels of texture are added. If the added texture enhances perceived quality, the geometry-plus-texture curve will lie above the dotted ”no-texture” curve. If adding the texture decreases perceived quality, then the geometry-plus-texture curve will lie below the no texture curve. The texture resources are indicated in the legend.

For the sphere-front stimulus (top left quadrant), the highest quality result is always obtained by geometry alone – for any resource level the dotted line is always at the top of the plot. That is, no expenditure of resource by adding texture improves the perceived quality of these stimuli. Furthermore, the poor quality texture (.0122 Mb) actually prevents any improved perceived quality when additional geometry is added, as indicated by the quality plot reaching a plateau. For the sphere-side stimulus (lower left quandrant) these general trends are repeated, with the gap between the quality of textured and untextured representations widening at each resource level.

For both the crinkle-front and crinkle-side stimuli (on the right) significant benefits are observed using the texture maps. Here the highest quality result is always obtained by the high-resolution texture mapped representation (.787 Mb). For the low-resolution geometry, applying any texture map results in an improved perceived quality, and the quality improves monotonically with the texture resolution. The resource required for the small geometry and full texture map is significantly less than the resource required for the medium geometry, without substantial loss in quality, particularly for the side lit case. The lines cross for the medium geometry case, reflecting the fact that only the high resolution texture is an improvement in this case. As in the sphere case, the flat line for the low resolution texture indicates that this level of texture mapping again actually prevents improvement by adding geometry.

5. DISCUSSION

In these experiments, we varied the texture and geometry of two models, a sphere with a low spatial frequency surface and a sphere with a crinkly, high spatial frequency surface. Each representation was compared with an original, "perfect" representation, providing a measure of the quality of the graphical representation.

4.1 Perceptual Discussion

In this section, we interpret the results in perceptual terms, exploring how our manipulations of geometry, texture, and lighting affected perceived fidelity for these two models. We consider two perceptual dimensions of these objects, their boundary contours and their texture. In this discussion, we discuss how our experimental manipulations affected these dimensions in a qualitative manner, but plan to make explicit measurements in the future.

Boundary Contour. Detecting small changes in an object's silhouette can be a very precise task for human observers, who can reliably discern variations on the order of seconds of arc, for example, seeing a broadcast spire against a sunset horizon. There has been a long tradition in the psychological literature regarding the role of boundary contours in the determination of object shape for objects with a countable number of contours. For complex silhouettes, the fractal dimension of the boundary contour has been shown to be important for object recognition (Rogowitz and Voss [15]). For simple shapes, Cortese and Dyre[16] have shown that shape discrimination depends on the frequency, amplitude and phase of the Fourier boundary contour. Reducing the geometry of an object makes the boundary contour more jagged, reflecting the fact that the object has been created with fewer polygons. In Fourier contour terms, the smooth sphere with full geometry has spatial frequency contours with zero amplitude. As the geometry is reduced, higher spatial-frequency contours are introduced, with increasing amplitude. The crinkly sphere has a high spatial-frequency contour. As the geometry is reduced, these high-spatial frequency components are replaced with successively lower spatial-frequency contours, with increasing amplitude.

The spatial resolution of the texture mapped onto the geometry can also affect an object's boundary contour. For the high-geometry crinkly sphere, the subsampled texture reduced the spatial-frequency of the boundary contour. For the low-geometry crinkly sphere, high-resolution texture increased the spatial frequency of the boundary contour.

Texture Discrimination. Another dimension along which these stimuli can be compared is the texture on the body of the object. Texture discrimination depends both on the frequency composition of the texture and the amplitude modulation of its components. For the smooth sphere, geometric simplification produces low spatial-frequency facets and contours. For the crinkly sphere, with a high spatial-frequency surface, geometric simplification reduces the spatial frequency of the object's surface, and can smooth out the crinkly surface altogether. Oblique illumination increases the contrast modulation of these crinkles and facets, which are made especially distinct as the object moves from full illumination to shadow, increasing texture discrimination.

In these experiments, several levels of texture were added to the geometric objects. In the case of the smooth sphere, this was a very low spatial-frequency texture, emulating the low-spatial frequency effect of light on a smooth sphere. Decreasing the resolution of the low spatial frequency texture produced some banding , contouring and worming, visible mostly in the oblique lighting condition. In the case of the crinkly sphere, this was a very high spatial-frequency texture, representing the effect of light on a crinkly sphere. At the lowest level of simplification, it produced a distinct, regular, pixellated pattern. With each

increase in texture resolution, the more the texture map emulated the spatial frequency of the crinkly surface. At the highest level of texture resource, the majority of the high spatial-frequency detail was represented.

Smooth Sphere. For the sphere-front stimulus, ratings of perceived quality seem to be based only on boundary contour. A reduction in quality was only observed when the geometric resource was reduced from .093 Mb to .047, reflecting a perceptible change in the boundary contour. At all levels of geometric simplification, quality judgements were independent of texture resource, possibly because the texture was not visible under direct lighting. For the sphere-side stimulus, simplifying the geometry increased the amplitude and the spatial frequency of the boundary contour; increasing the texture slightly decreased the amplitude of the contour modulation. If the perceptual judgments were based solely on characteristics of the boundary contour, we would expect perceived quality to decrease for greater degrees of geometric simplification, which it does, and to increase slightly with added texture, which it does not. Adding a very simplified texture dramatically decreased perceived quality, presumably because the texture was discriminably different from the original. Adding additional texture resource improved the perceived quality, presumably reflecting a decrease in sampling artifacts.

Crinkly Sphere. If boundary contour were the major factor in determining the perceived quality of the crinkly sphere, perceived quality would be highest for the highest-geometry shape and decrease monotonically with reductions in geometry, correlated with the introduction of high amplitude, low spatial freqeuency jagged contours. This is certainly the effect obtained for geometry alone; under both illumination conditions, perceived quality increases with each increase in geometric resource. When texture is introduced, however, the medium-geometry object appeared to have the same quality as the high-geometry object. This suggests that the texture may mask the imperfections in the medium geometry's boundary contour.

Texture, however, can also decrease perceived quality. The most simplified texture, for example, brings perceived quality to its lowest levels, independent of the underlying geometry. This may be because this pixelated texture has a much lower spatial-frequency than the original, and is easily discriminated. Also, for all geometries, and under both illuminations, perceived quality increased monotonically with increased texture resolution, suggesting that successive approximations to the high spatial-frequency of the "original" stimulus may be driving these results.

Illumination. For the smooth sphere, the object was judged to be of systematically higher quality when viewed under direct illumination than when viewed under oblique illumination. For the crinkly sphere, representations based on the most simplified geometry and the most simplified texture appeared to have lower quality under oblique illumination, where the oblique lighting emphasized the jagged underlying geometry or the pixelated texture.

In interpreting these data, however, it is important to keep in mind that the goal of the experiment was not to study images of objects, but to understand the perception of objects. Thus, the front-view and the oblique-view are both descriptions of the same object, just viewed under two illuminations. In future experiments, we plan to ask observers to make judgments while dynamically varying the illumination. We suspect that when the observer is forced to view these stimuli as one object, that the judgments will be dominated by the lower perceived quality of the oblique view.

4.2 Graphics Discussion

When does texture successfully substitute for geometry? For the smooth sphere, texture was not able to compensate for the decrease in perceived quality produce by reducing geometry. For the crinkly sphere, however, the opportunity to use low-resource texture to substitute for geometry was clear. For highly simplified high spatial-frequency objects, viewed under either illumination, adding even small amounts of texture increase perceived quality substantially, and perceived quality increases monotonically with texture resource. Under both illumination conditions, adding less than 1 Mb of texture to the 6.39 Mb object improved the mean rating score 25% giving it the same fidelity as the full-geometry object with 8 Mb more geometry. Clearly, texture can successfully trade for geometry when the geometry is complex and has been simplified by a factor of two or more. Since this is the arena of interest, these results are quite promising.

6. CONCLUSIONS

The framework and results presented here provide many useful insights into geometric simplification with texture substitution. The results of our experiment indicate that scaling experiments can produce consistent data regarding the perceived quality of object representations. The different results for front and side lighting for the smooth sphere indicate that lighting effects need to be accounted for in comparing objects, making this a fundamentally different problem from comparing images. The differing results for the smooth and crinkled sphere demonstrate that the benefit of expending resources substituting geometry with texture is object dependent. The results for varying texture resolution show that textures may be counterproductive if they are not of sufficiently high resolution. Practically speaking, if a system were forced to reduce texture map resolution because of a resource bottleneck, it may be preferable to not use texture at all.

These insights provide guidance for the development and testing of new simplification algorithms. Rather than using the same strategy for all objects, or all parts of a single object, it may be useful to analyze the nature of the object's geometry to test if it can be replaced by a texture map. We believe that our framework for measuring the perceptual consequences of different geometries, textures, and lighting conditions can be used to evaluate the success of such tests.

REFERENCES

1. E. Horvitz and J. Lengyel, "Perception, attention, and resources: A decision theoretic approach to graphics rendering," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997.
2. B. Watson, A. Friedman, and A. McGaffey, "Using naming time to evaluate quality predictors for model simplification," in *Proceedings of SIG CHI 2000*, p. to appear.
3. P. Cignoni, C. Montani, and R. Scopigno, "A comparison of mesh simplification algorithms," *Computers and Graphics* , 2000.
4. W. J. Schroeder, J. A. Zarge, and W. E. Lorensen, "Decimation of triangle meshes," in *Proceedings of SIGGRAPH 92*, E. Catmull, ed., pp. 65–70, ACM, 1992.
5. A. Gueziec, "Locally toleranced surface simplification," *IEEE Transactions on Visualization and Computer Graphics* **5**, pp. 168–169, 1999.
6. J. Rossignac and P. Borrel, "Multi-resolution 3D approximations for rendering complex scenes," in *Geometric Modeling in Computer Graphics*, B. Falcidieno and T. Kunii, eds., pp. 455–465, Springer-Verlag, 1993.
7. R. Klein and A. Schilling, "Efficient rendering of multiresolution meshes with guaranteed image quality," *The Visual Computer* **15**, pp. 443–451, 1999.
8. P. Heckbert, "Survey of texture mapping," *IEEE Computer Graphics and Applications* **6**(11), pp. 56–67, 1986.
9. F. Weinhaus and V. Devarajan, "Texture mapping 3D models of real-world scenes," *ACM Computing Surveys* **29**(4), pp. 325–365, 1997.
10. J. Shade, D. Lischinski, T. DeRose, J. Snyder, and D. Salesin, "Hierarchical image caching for accelerated walkthroughs of complex environments," in *Proceedings of SIGGRAPH 96*, H. Rushmeier, ed., pp. 75–82, ACM, 1996.
11. P. Maciel and P. Shirley, "Visual navigation of large environments using textured clusters," in *SIGGRAPH Symposium on Interactive Graphics*, pp. 95–102, 1995.
12. J. Torborg and J. Kajiya, "Talisman: Commodity realtime 3D graphics for the PC," in *Proceedings of SIGGRAPH 96*, H. Rushmeier, ed., pp. 353–364, 1996.
13. M. Soucy, G. Goudin, and M. Rioux, "A texture-mapping approach for the compression of colored 3D triangulations," *The Visual Computer* **12**, pp. 503–514, 1996.
14. J. Cohen, M. Olano, and D. Manocha, "Appearance-preserving simplification," in *Proceedings of SIGGRAPH 98*, M. Cohen, ed., pp. 155–122, ACM, 1998.
15. B. Rogowitz and R. Voss, "Shape perception and low-dimension fractal boundary contours," in *Proc. SPIE Human Vision and Electronic Imaging: Models, Methods, and Applications*, vol. 1249, pp. 387–394.
16. J. Cortese and B. Dyre, "Perceptual similarity of shapes generated from Fourier descriptors," *Journal of Experimental Psychology: Human Perception and Performance* **22**, pp. 133–143.

An Experimental Evaluation of Computer Graphics Imagery

GARY W. MEYER, HOLLY E. RUSHMEIER, MICHAEL F. COHEN, DONALD P. GREENBERG, and KENNETH E. TORRANCE
Cornell University

Accurate simulation of light propagation within an environment and perceptually based imaging techniques are necessary for the creation of realistic images. A physical experiment that verifies the simulation of reflected light intensities for diffuse environments was conducted. Measurements of radiant energy flux densities are compared with predictions using the radiosity method for those physical environments. By using color science procedures the results of the light model simulation are then transformed to produce a color television image. The final image compares favorably with the original physical model. The experiment indicates that, when the physical model and the simulation were viewed through a view camera, subjects could not distinguish between them. The results and comparison of both test procedures are presented within this paper.

Categories and Subject Descriptors: I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*intensity, color, photometry, and thresholding*; I.3.3 [**Computer Graphics**]: Picture/Image Generation—*display algorithms; viewing algorithms*; I.3.6 [**Computer Graphics**]: Methodology and Techniques—*ergonomics*; I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*color, shading, shadowing, and texture*; I.4.8 [**Image Processing**]: Scene Analysis—*photometry*

General Terms: Experimentation, Human Factors, Measurement, Verification

Additional Key Words and Phrases: Color science, image science, light reflection models, radiosity

1. INTRODUCTION

The creation of realistic images requires an accurate simulation of light propagation within an environment, as well as a perceptually accurate method for displaying the results of the simulation. The need for physically based illumination models and perceptually based imaging techniques means that the lighting calculations and the production of the final simulation are separate tasks, each having different objectives to be met. If a scientific basis for the generation of images is to be established, it is necessary to conduct experimental verification on both the component steps and the final simulation.

This research was funded in part by National Science Foundation grant DCR 8203979, "Interactive Computer Graphics Input and Display Techniques."
Authors' present addresses: G. W. Meyer, Department of Computer and Information Science, University of Oregon, Eugene, OR 97403; H. E. Rushmeier, M. F. Cohen, D. P. Greenberg, and K. E. Torrance, Program of Computer Graphics, 120 Rand Hall, Cornell University, Ithaca, NY 14853-5501.

Early realistic image synthesis techniques and the lighting models that they employed were severely limited by processing and storage constraints, as well as by the display hardware characteristics. The need for computational simplicity substantially influenced the illumination algorithms that were originally developed. The results were light models that did not make direct use of established physical behavior; reflection models arbitrarily assigned ambient, diffuse, and specular portions to the reflected light. The perceptual significance of the monitor's primaries was not recognized as colors were directly computed in terms of the RGB (red, green, blue) primaries. Given specific viewing parameters, the intensity of each picture element was determined only on the basis of the single surface "seen" through that pixel and its direct relationship to light sources. These approaches, which do not simulate the global illumination effects and the interreflections among surfaces in an environment, result in pictures that are obviously computer generated.

Recently, ray-tracing techniques, which attempt to model the global illumination effects of specular surfaces, have been introduced. Ray tracing is still a view- and resolution-dependent approach, but employs a more comprehensive lighting model. Each picture element can receive light directly from the surface immediately behind it and indirectly by ray reflection (and/or refraction) from other objects. However, each participating surface still receives its illumination only in a direct path from light sources or from an arbitrary constant ambient term. In most cases the light model is expressed in terms of the RGB primaries and is not based on sound physical principles. Although the technique is quite expensive computationally, the pictures produced can be impressive and are a substantial improvement over those generated by previous techniques.

The introduction of the radiosity method has led to a complete decoupling of the light reflection simulation from the final imaging technique. An illumination model based on energy conservation principles is used to account for all interreflection of light in an environment. The illumination calculations are independent of viewing parameters and can be performed on a wavelength basis rather than the particular red, green, and blue channels provided by the phosphors of a specific raster display device. The results of the global illumination calculations are used in conjunction with the principles of color science to convert the resulting spectral energy distributions to the RGB primaries of the display device.

What has emerged from this sequence of events is the need for a clear distinction between the physical and perceptual portions of the image synthesis process and the need for experimental verification of each of these steps. The first step in the image synthesis process should be to model correctly the transport of light in the environment. This is inherently a physically based step where the flow of energy is modeled as accurately as possible. To verify the light model, physical measurements should be made on a real scene and should be compared with the simulated values. The second step of the image synthesis process should be to use the results from the physical modeling of the propagation of light to produce the final simulation to be observed. This is inherently a perceptually based step, where the objective is to satisfy the final observer. To verify the final simulation and thereby the overall objective of realistic image synthesis, the simulation should be visually compared with the real scene.

In this paper a simple environment is used to demonstrate an approach to image synthesis that has distinct physical and perceptual portions and that employs experiments to verify both parts of the process. In Section 2, the radiosity method is used to do the light modeling, and the results are compared against physical measurements made on an actual model. In Section 3, the principles of color science are used to produce an image of the same model on a color television monitor, and this picture is visually compared with the real scene by a group of experimental subjects. The observations and conclusions of the paper, summarized in Section 4, indicate that by using a rigorous scientific methodology a good match can be obtained for both the physical and perceptual comparisons.

2. RADIOMETRIC COMPARISON

In this section an example is presented of the use of a physical experiment to verify the first part of the image synthesis process—the simulation of reflected light intensities. The distribution of radiation in simple scenes is considered, and the particular theoretical procedure for calculating the radiant transfer to be verified is outlined. This technique, known as the radiosity method, is used to generate all of the synthetic computer images in this paper. An experimental apparatus is also described. This apparatus allows simple, real-world scenes to be tested and is used for all of the scenes presented in this study. Measurements of radiant energy flux densities on a wall of the physical model are compared with the predictions of the radiosity method; a method for measuring the radiant flux densities, which are directly related to the light intensities, is detailed, and measurements on three environments of varying complexity are presented.

2.1 Overview of Experimental Design

In an ideal experiment for verifying the accuracy of light intensity calculations on an image plane, an instrument would be used that could be positioned at the "eye" position with respect to the real environment. This instrument would have an angular resolution that would allow it to measure the light energy reaching the "eye" through the solid angle subtended by each pixel in the image plane. This instrument would also have the ability to measure each wavelength band of light reaching the eye.

The instrument defined above would need precise angular and spectral resolution. The associated measurements would be geometrically difficult and time consuming, and would require high photon sensitivity under very carefully controlled lighting conditions. Since the present study is an initial effort to compare a real environment with a synthetic computer image, such a refined experimental study was not carried out. Indeed, a relatively inexpensive and simple radiation measuring instrument (a radiometer) was employed. The instrument gave a single reading corresponding to the hemispherically incident radiant flux over the range of visible wavelengths.

Measuring an entire environment or scene from the "eye" position with this instrument would yield a single reading on a meter. Since this single reading represents a spatial and spectral average of the flux incident on the radiometer, it would not be sufficiently discriminating to allow an evaluation of simulation methods. It represents a point measurement, which is indicative only of the

magnitude of the radiant field. A more discriminating approach requires measurements at several locations to assess the spatial distribution of light energy. Thus measurements would be needed for many different viewing locations.

Furthermore, the light received by a radiometer varies continuously with position and depends on the geometric and optical properties of the entire radiant environment. Although such measurements do not allow direct verification of the detailed predictions of a lighting model, they do allow verification of the integral predictions (i.e., integrated over wavelength and the incident hemisphere) of a lighting model. Such integral measurements at several locations are employed in this study to assess a particular lighting model. In general, a lighting model must be capable of predicting the relative values of these integral quantities if it is to be relied upon to simulate accurately the more detailed light intensities required for image synthesis.

2.2 Radiosity and Irradiation

The above radiometric method is used in the present article to evaluate the standard radiosity method and one variation of the radiosity method. The radiosity method is a theoretical procedure for predicting light intensities in a totally diffuse environment. The method was developed in the field of heat transfer to calculate the heat exchange by means of electromagnetic radiation in enclosures. It can also be applied to visible light. The method was first applied to synthetic image generation by Goral et al. [5], and extended by Cohen and Greenberg [2]. In this paper, the radiosity method is used to predict the light energy impinging on, and measured by, the radiometer. A brief summary of the radiosity method is included as background material for the experiments.

In the radiosity method, all emission and reflection processes are assumed to be perfectly diffuse (Lambertian). The scene or enclosure is divided up into discrete surfaces, each of which is assumed to be of uniform radiant intensity. With these assumptions, the intensity of radiation leaving a particular surface is directly proportional to the radiant flux density (energy per unit area per unit time) or radiosity B leaving the surface. The radiosity of a surface i in an enclosure is related to the radiosities of all the surfaces in an enclosure by

$$B_{i\lambda} = E_{i\lambda} + \rho_{i\lambda} \sum_j F_{ij} B_{j\lambda}, \tag{1}$$

where λ denotes wavelength, $E_{i\lambda}$ denotes the energy emitted from the surface per unit time and area, $\rho_{i\lambda}$ is the diffuse reflectance of the surface, and the summation j is over all the surfaces in the enclosure. There is one such equation for each surface. The form factor F_{ij} depends only on geometry and represents the fraction of energy leaving surface i that arrives at surface j. The energy source term $E_{i\lambda}$ is zero for surfaces that are not light sources. Equation (1) holds for a particular wavelength. However, it also applies for discrete wavelength bands in which $B_{i\lambda}$, $E_{i\lambda}$, and $\rho_{i\lambda}$ are constant, as long as energy is not exchanged between the bands.

The basic radiosity method can be extended to account for directional variations in the light source. The extension is achieved by computing the amount of emitted light directly reaching the surfaces illuminated by the light source. If reflections off the light source are neglected, the equation corresponding to the

light source can be set aside. The radiosity equation for the other surfaces becomes

$$B_{i\lambda} = \rho_{i\lambda}\{\mathrm{df}\ F_{i,\mathrm{light}}\ \max[E_{\mathrm{light},\lambda}] + \sum_j F_{ij}B_{j\lambda}\}, \tag{2}$$

where the index i and the summation j do not include the light source, df is a light source directional factor, and the maximum directional radiosity of the light source is denoted by $\max[E_{\mathrm{light},\lambda}]$ and is found from measurements. The directional factor (df) is zero for surfaces not directly illuminated by the light source.

As discussed above, it is difficult to make spatially and spectrally detailed measurements of the radiant energy in an environment. In the heat transfer literature, direct measurements of radiosity (the energy leaving a surface) are rarely found [10]. The approach of measuring irradiation is more commonly used [9]. The irradiation $H_{i\lambda}$ incident on a surface i is given in terms of the radiosities of all the other surfaces by

$$H_{i\lambda} = \sum_j F_{ij}B_{j\lambda}. \tag{3}$$

For surfaces that are not light sources, comparison with eq. (1) shows that the radiosity of surface i is directly proportional to the irradiation onto the surface. The constant of proportionality is the reflectivity $\rho_{i\lambda}$ of the surface. Since intensity is proportional to radiosity, the intensity of the surface is directly proportional to the irradiation. Thus the relative spatial distributions of the incident irradiation and reflected intensity are the same.

Irradiation can be measured by a radiometric probe. If the sensitivity of the probe varies with direction, however, the probe response cannot be compared directly with eq. (3). Instead, the radiosity B at a particular angle of incidence must be multiplied by an angle-dependent correction factor cf. The predicted response of the probe is then given by

$$H_{i\lambda} = \sum_j \mathrm{cf}(\theta)F_{ij}B_{j\lambda}, \tag{4}$$

where θ denotes the angle of incidence on surface i of irradiation coming from surface j. The predictions of this equation are compared later with radiometric measurements.

2.3 Experimental Apparatus

The test environment was the five-sided cube shown in Figure 1. The dimensions of the cube are shown in Figure 2, as are the dimensions of two small boxes that were placed within the cube for some observations. All five sides of the cube could be removed independently so that the color of each side could be changed. All of the surfaces of the cube were painted with flat latex house paints, which are close to being ideal diffuse reflectors. The spectral reflectances of the paints were measured using a Varian Cary 219 spectrophotometer with an in-cell space diffuse reflectance accessory. These reflectances are shown in Figure 3a.

The light source consisted of a 150-watt incandescent flood light mounted at the top of a 15-inch-high metal cone. The interior of the cone was covered with a flat white paint. The light shone through a piece of 4.5 by 3.5-inch flashed opal

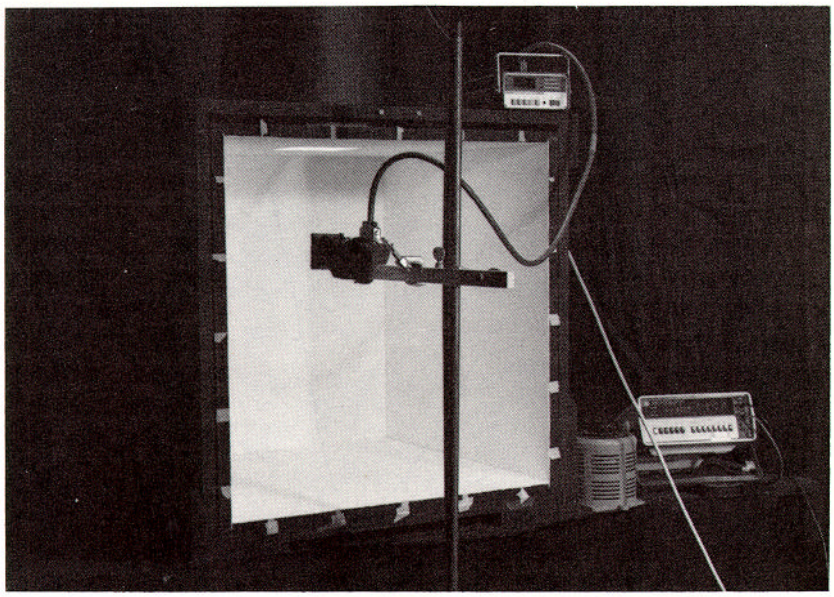


Fig. 1. Experimental setup used to make radiometric measurements.

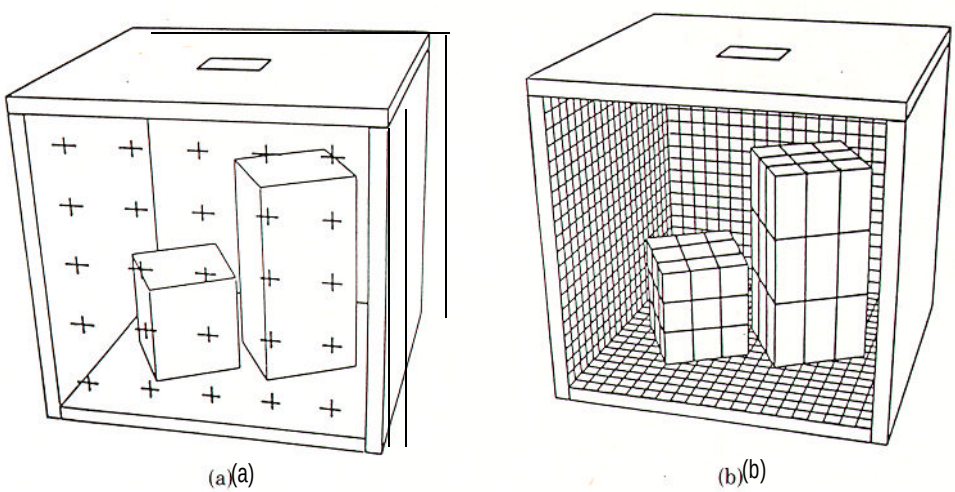
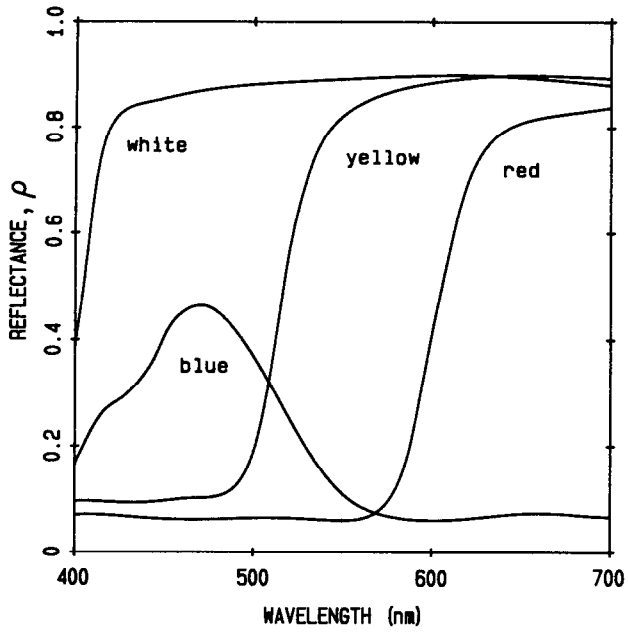
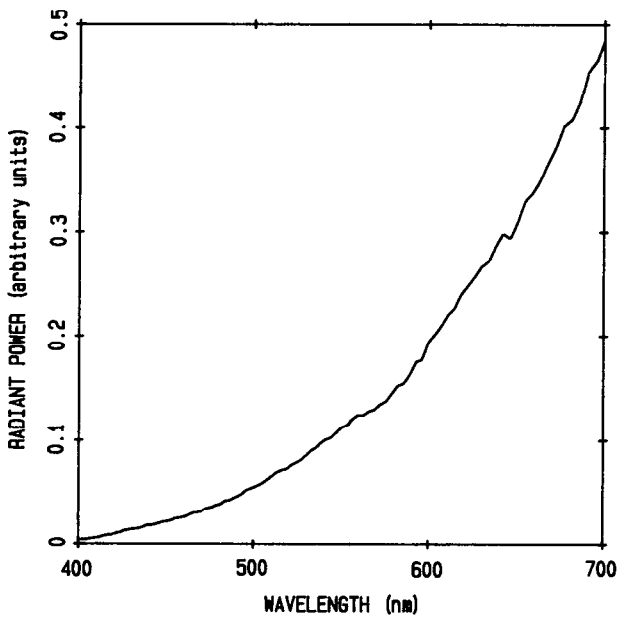


Fig. 2. Schematic of test environment. (a) The crosses indicate positions for radiometric measurements. (b) The surfaces are discretized for radiosity calculations.

Dimensions:	Width (inches)	Height (inches)	Depth (inches)	Color
Enclosure	21.6	21.5	22.1	White; extra red/blue walls
Large box	6.5	13.0	6.5	White
Small box	6.5	6.5	6.5	Yellow
Light	4.5	—	3.5	See Figure 3b.



(a)



(b)

Fig. 3. (a) Reflectances of paints used to paint cube and small boxes. (b) Spectral energy distribution of light after passing through opal glass.

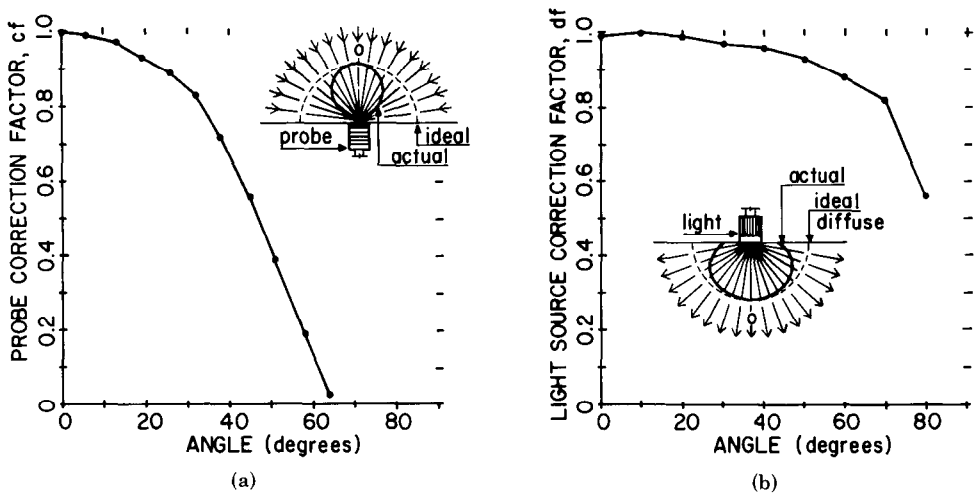


Fig. 4. (a) Probe directional sensitivity. (b) Normalized light source intensity versus angle.

glass, which was mounted in the ceiling of the cube. An autotransformer and digital voltmeter were used with the light source to maintain a constant 115 volts. The spectral energy distribution of the entire light source assembly was measured using equipment described by Imhoff [6] and is shown in Figure 3b.

The enclosure was placed on a flat black table in a small room. The walls of the room were covered with black fabric so that essentially no visible radiation entered the cube through the open side. From the inside of the enclosure, the open side appeared as a black wall.

Irradiation was measured using a Tektronix J16 photometer with a J6502 irradiance probe. This probe has a flat spectral sensitivity in the visible and near infrared ranges. A Corning Glass 1-56 filter was placed on the front of the probe to filter out the infrared energy emitted by the light source. The directional sensitivity of the probe (and filter) was determined by rotating the probe while illuminating it with collimated light. The correction factor cf as a function of incident angle is shown in Figure 4a.

To complete the apparatus specification, an additional measurement of the light source strength is required. This was measured for visible light by holding the J6502 probe flush against the opal glass. For a perfectly diffuse light source, the measured irradiation H can be related to the total light source emission E_{light} by

$$E_{\text{light}} = \frac{H}{F_{\text{sensor,light}}}. \tag{5}$$

The form factor between the probe sensor and the light, $F_{\text{sensor,light}}$, was estimated to be 0.52 [11, p. 826].

To use eq. (2), the directionality of the light source df is also needed. The light intensity at various angles from the normal was measured by using the photometer with the irradiance probe. The probe was fitted with a long tube to restrict

the acceptance angle of the probe. The results of these measurements are shown in Figure 4b and indicate that the light source is not perfectly diffuse. A perfectly diffuse light source would have a df of unity.

2.4 Procedure

Measurements of irradiation were made at 25 locations in the plane of the open face of the cube (shown in Figure 2a) and compared with the simulations. Measurements were made for three scenes: the empty white cube, the empty white cube with the left panel replaced by a blue panel, and the all-white cube with the large white box inside it.

The measurement locations were chosen for two reasons: (1) to maximize the light energy incident at any point, and thus to minimize the uncertainty in each reading, and (2) to minimize the effect of the probe on the environment. The probe should cast no shadows and should reflect little light back into the environment. The foregoing criteria are satisfied by placing the probe at the open side of the cube.

Tests were made to examine the potential sources of error in the measurements. The light source voltage could be controlled so that variations in light source emission changed by less than 1 percent. Movement of objects within the room surrounding the cube and changing the position of the cube within the room had no measurable effect on the irradiation at the open face of the cube. Doubling or tripling the size of cracks between the panels had no measurable effect.

Another potential source of error in the measurement was the position of the probe. The three-dimensional position and angular orientation of the probe were carefully controlled. Very small variations in these parameters could result in large variations in the measured irradiation. This error was estimated by repositioning the probe at each measurement location several times and recording the measured irradiation. The maximum deviation at each point was approximately ±5 percent of the mean value of the readings.

The error in the photometer itself is given by the manufacturer as less than ±5 percent. Combining the estimated error due to all of the foregoing factors leads to a total root mean square estimated error of ±7 percent.

2.5 Results

For each of the three scenes, measurements were made at the 25 test locations and compared with theoretical predictions based on the radiosity method.

For the radiosity calculations, the form factors were determined as described by Cohen and Greenberg [2]. In every case, each of the five walls of the cube was divided into 225 elements of equal area (see Figure 2b), the light source was divided into nine elements, and the sides of the large rectangular box in the third scene were each divided into nine elements. The reflectances used for the walls were averages of the measured spectral reflectance curves in Figure 3a. The open side of the cube was modeled as a black wall with zero reflectance. This side was divided into 25 surfaces, the center of each surface corresponding to a measurement position. The average measured irradiation when the probe was held flush against the light source was 240 microwatts per square centimeter. Using eq. (5), the total emission of the light source was estimated to be 460 microwatts/per square centimeter. This value, $\max[E_{\text{light}}]$, was used for all calculations. The

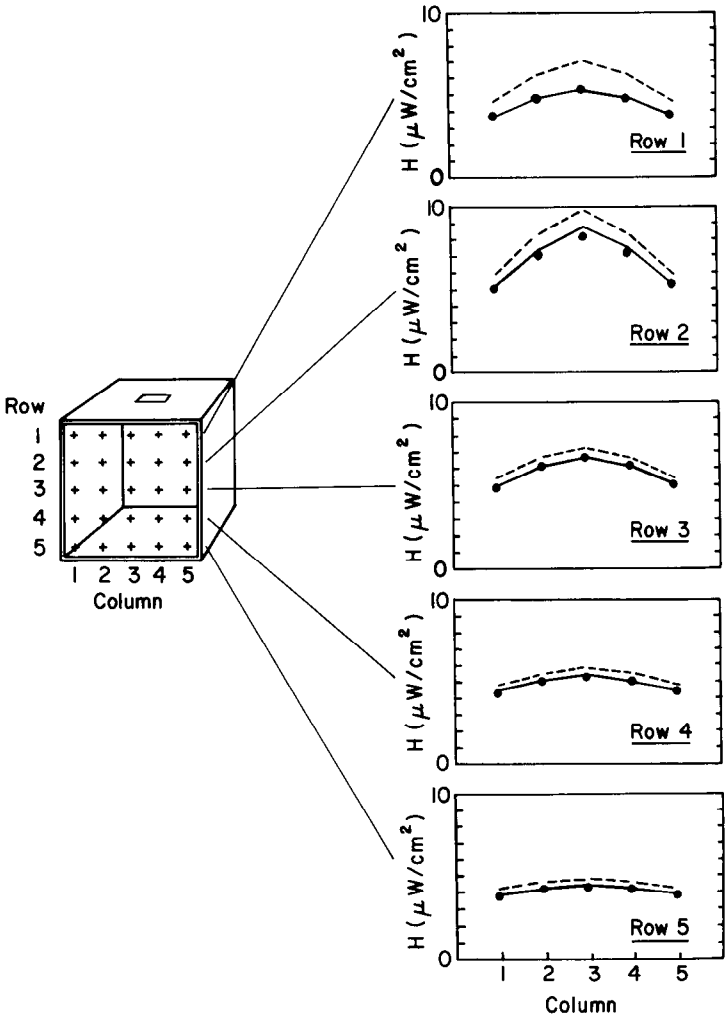


Fig. 5. Comparison of the measured and calculated irradiation at the open side of the empty white cube: ----, radiosity calculation with diffuse light source; ———, radiosity calculation with directional light source; •, radiometer measurement.

spectral distribution, $\max[E_{\text{light},\lambda}]$, was obtained from Figure 3b. The results predicted by the radiosity calculations were converted to an equivalent probe response by using eq. (4).

The measurements made with the empty all-white cube are shown in Figure 5 (filled circles). Each of the plots is for one horizontal row of locations. In general, the results show that the irradiation is highest near the center of the open side of the cube. This area has the best view of the light source and the other walls.

Figure 5 also shows the results of calculations using a completely diffuse light source (dashed lines). These calculations were made using eqs. (1) and (4). The calculated values are much higher than the measurements.

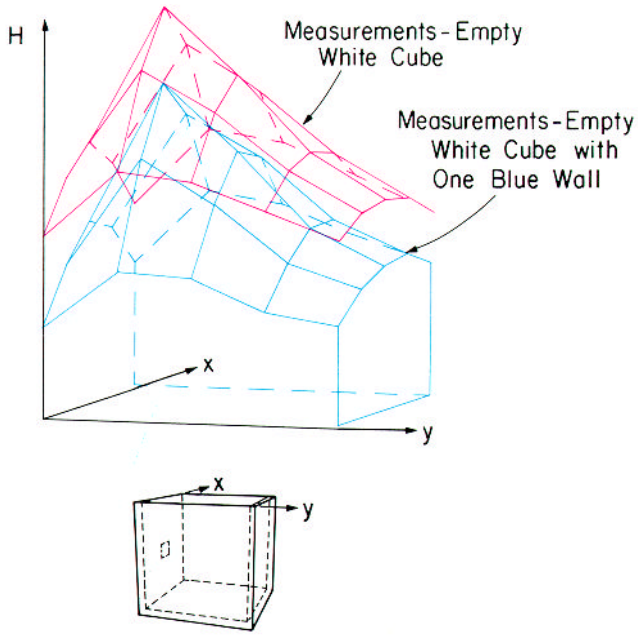


Fig. 6. Irradiation H emerging from the open side of the cube. The sketch shows the cube with the open side facing upward and the top side to the left; the blue wall is on the side closest to the reader.

The major source of discrepancy between the measured results and those calculated for the purely diffuse environment is the directionality of the light source. This causes a difference in both curve shape and overall illumination level. Calculations using the diffuse radiosity method extended to include the directionality of the light source, as described by eq. (2), are also included in Figure 5 (solid lines). The calculated values are obviously lower than those for a purely diffuse light source, since less energy reaches each surface directly, and less energy is interreflected within the cube. The calculated results for the upper row show the largest changes since this row has the largest angle of incidence, and thus the greatest deviation, with respect to the light source emission. When light source directionality is accounted for, the root mean square difference between the calculated and measured results at the open face of the cube is less than 4 percent and the root mean square difference between normalized results is less than 3 percent. These values compare with 18 and 7 percent, respectively when the light source directionality is not accounted for, and are both less than the estimated measurement error of 7 percent. Thus the calculations made by assuming a directional light source are significantly more accurate than those made by assuming a perfectly diffuse light source. The rest of the calculated results presented in this section assumed a directional light source.

Figures 6-9 provide a comparison of the three scenes that were considered. In each figure the irradiation H is shown as a function of measurement position. Figure 6 provides a comparison of the measurements on the empty white cube with measurements on an empty cube with one blue wall, the other walls being

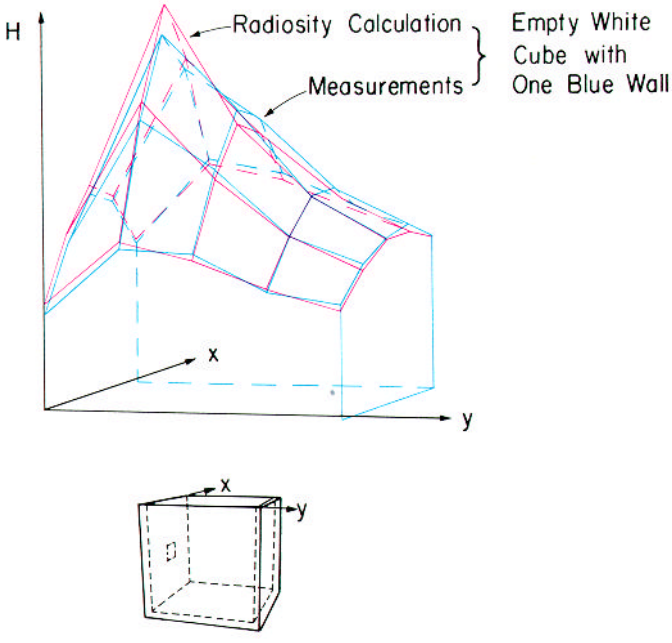


Fig. 7. Irradiation H emerging from the open side of the cube. The sketch shows the cube with the open side facing upward and the top side to the left; the blue wall is on the side closest to the reader.

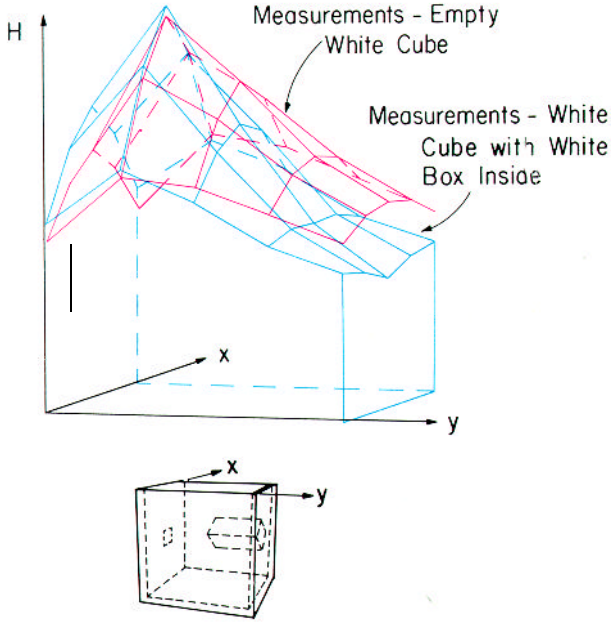


Fig. 8. Irradiation H emerging from the open side of the cube. The sketch shows the cube with the open side facing upward and the top side to the left; the large white box is on the floor opposite the light source.

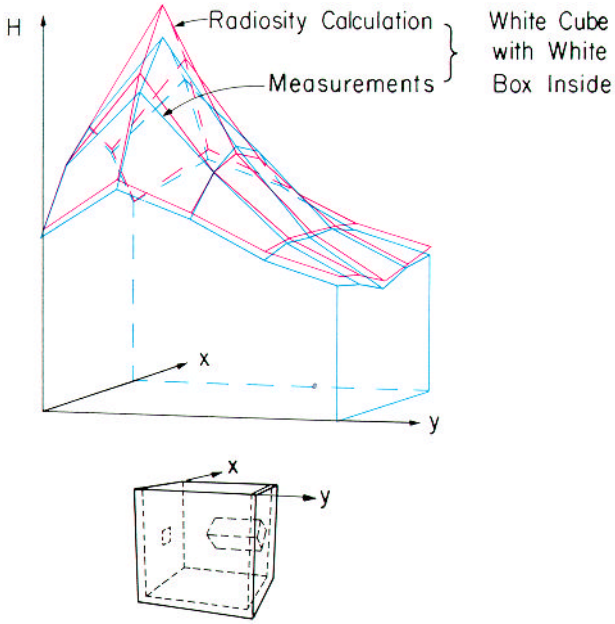


Fig. 9. Irradiation H emerging from the open side of the cube. The sketch shows the cube with the open side facing upward and the top side to the left; the large white box is on the floor opposite the light source.

white. The low reflectance of the blue wall (Figure 3a) reduces the overall illumination in the cube. Clearly, this reveals that a large proportion of the light incident on any surface in the cube is due to reflection from other surfaces rather than direct illumination from the light source. Furthermore, measurements along the leftmost column of measuring positions (see Figure 2a) are only about 75 percent of the corresponding measurements on the right side of the cube. Thus small surfaces located on the left side of the open side of the cube would appear darker than surfaces located on the right side. This influence of the neighboring surfaces on the intensity of a point is related to the color bleeding effect discussed in earlier work [2, 5].

Figure 7 shows a comparison of measured and calculated results for an empty cube with one blue wall and four white walls. The calculated results differ from the measured results by a root mean square difference of less than about 7 percent. On comparing relative values, the results differ by 4 percent. The calculated and measured results are lower for the left side than for the right side of the cube. A method for calculating light intensities that only takes into account the location of the light source, and not surface interreflections, would have given equal reflected intensities on the left and right sides.

A comparison of measurements in an all-white cube with and without an internal all-white box is shown in Figure 8. The large white box was placed in the center of the floor below the light source and was turned at a 45° angle to the walls of the cube. The intensity of the top two horizontal rows is higher for

the case with the internal box because light is reflected from the top of the white box to the upper edges of the cube. The bottom two rows of the open side are less intense than those for an empty cube. These rows see the dark sides of the internal box, which receive relatively little energy since they, in turn, face a black wall and have no direct view of the light source.

Figure 9 shows a comparison of measured and calculated results for an all-white cube with an internal white box. The calculated and measured results have a root mean square difference of about 4 percent and a root mean square difference in relative values of 3 percent. The presence of hidden surfaces did not appear to reduce the accuracy of the calculations. The calculated results follow the same trend as the measured results in having values for the top row that are higher than those for the open white cube, and values for the bottom two rows that are lower. Calculation methods that do not account for diffuse interreflections would not predict the increase in intensity near the top caused by reflection off the box.

In summary, there is good agreement between the radiometric measurements and the predictions of the lighting model. A full summary of results is deferred until Section 4. The perceptual experiments are described in the next section.

3. PERCEPTUAL COMPARISONS

Given the experimentally verified output of a light model, the next step in the image synthesis process is to use this information to produce the final simulation. In this section color science methods are used to create a color television image of the simple cubical environment from the output of the radiosity method. This picture is then compared by a group of experimental subjects against a real model as seen through the back of a view camera. This step is taken to evaluate the simulation and thereby to determine whether the overall objective of realistic image synthesis has been achieved.

There is some precedent for performing comparisons between pictures and reality. O. W. Smith constructed an experiment to study depth perception in which a subject viewed a picture and a real scene through a peephole [12]. In computer graphics, comparisons have been made between photographs of reality and photographs of computer-generated images [5, 8], the value of synthetic images for interior illumination design has been studied by an indirect comparison against a real scene [4], and two computer-generated pictures have been compared in order to determine how many polygons are necessary to represent a surface [1].

This section begins with a discussion of the rationale for viewing the model through a view camera while making the comparisons. Next, the experimental apparatus is described, and the procedures that were used to compute the color television picture and compare it against a view of the real model are discussed. Finally, the results of having a group of human observers make the comparison are presented.

3.1 Selecting the View of the Real Model

Although the field of view is restricted and some perceptual cues are eliminated, the view camera has been selected for several reasons: (1) it allows simultaneous side-by-side comparisons to be made without introducing the effect of the

observer's memory (a factor that is unavoidable in an alternative viewing scheme such as a pinhole), (2) it corresponds closely to the "synthetic camera" approach employed in computer graphics, (3) it is an experimental setup that can easily be controlled, (4) it is a starting point that must be mastered before other standards can be evaluated, and (5) the degree of restriction is relatively unimportant once the unavoidable step of limiting the view has been taken. In order to present the real and synthetic scences to the observer in the same way, the color television picture was also observed through a view camera.

3.2 Apparatus

The simple cubical enclosure that was described in the previous section was used for the perceptual experiments (see Figure 2). In this test case, the model was set up to have a blue wall on the right and a red wall on the left, with the rest of the walls white. The small yellow block was placed on the left and the large white block on the right. The blocks were turned at a slight angle with respect to one another.

The imaging media consisted of a frame buffer and a color television monitor. The frame buffer (Grinnel Systems GMR-27) had a resolution of 480 vertical by 512 horizontal pixels, had eight bits of intensity information in each of its three channels, and produced an interlaced video signal with a frame rate of 30 hertz and a field rate of 60 hertz. The monitor (Barco CTVM 3/51) had a 20-inch display tube with phosphor chromaticity coordinates:

$$x_R = 0.64, \quad x_G = 0.29, \quad x_B = 0.15,$$
$$y_R = 0.33, \quad y_G = 0.60, \quad y_B = 0.06.$$

The individual brightness and contrast controls for each of the monitor guns were adjusted to yield a D6500 white point, and the individual gamma correction functions were measured for each of the guns. The luminance ratios necessary to set the white point were found to be

$$Y_R : Y_G : Y_B = 0.3142 : 1.0 : 0.1009.$$

By determining the proportional relationship between luminance and radiance for each of the guns, these luminance ratios were converted to radiance ratios and were used to balance the guns over their entire dynamic range. The luminance of the white point was set to 24 foot lamberts.

Two Calumet 4 × 5 view cameras were used to view the model and the monitor. The two lenses used were a Schneider–Kreuznach Symmar f1 : 5.6/150mm and a Schneider–Kreuznach Symmar-S f5.6/150mm. Fresnel lenses ruled with 110 lines to the inch and with 10-inch focal length were placed in front of the ground glass of each camera to act as image intensifiers. The combination of the Fresnel lenses and the ground glass introduced some image degradation that made construction and imaging artifacts in both images less obvious.

The positions of the view cameras, the model, and the monitor are shown in Figure 10. The cameras were positioned so that the images were identical in size ($3\frac{5}{8}$ inches by $3\frac{5}{8}$ inches), and the f-stop settings of each camera were adjusted so that the intensities were the same. The combination of f-stop setting and camera-to-model distance were such that the entire depth of the model was in focus,

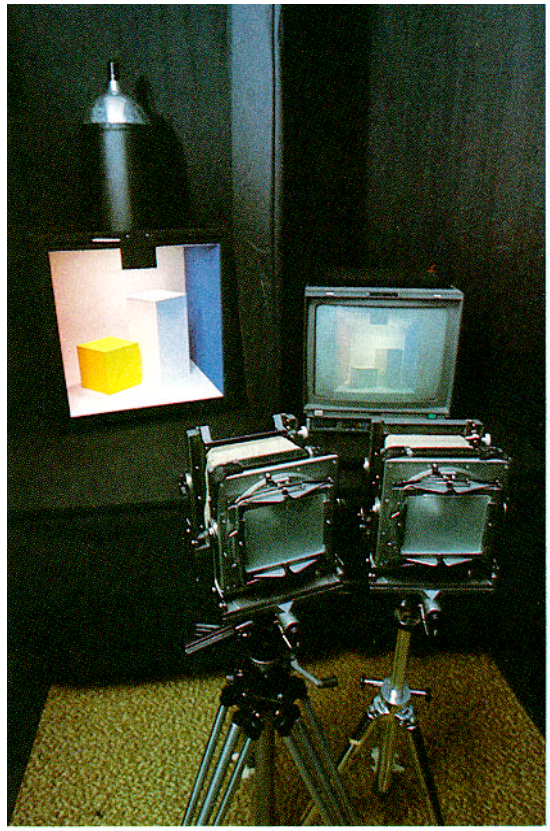


Fig. 10. Experimental setup with the partitioning curtains removed.

thereby minimizing depth of field problems. To minimize reflections, all of the walls were draped in black, and a black curtain (only part of which can be seen in Figure 10) split the room lenghtwise and separated the two view cameras. Another black curtain was hung across the width of the room to separate the subject from the model and the monitor, and the view cameras protruded through holes cut in this curtain.

Figure 11 shows the experimental setup with the widthwise curtain in place and an experimental subject evaluating the view-camera images. The centers of the view-camera backs were $8\frac{1}{2}$ inches apart and were 44 inches off the ground. The subjects were positioned so that their eyes were 25 inches from the view cameras and 48 inches off the ground. The scene, as viewed by the subjects, was inverted, and observations were made under dark ambient conditions.

3.3 Procedure

The image was computed using the radiosity software described above [2]. The frustum angle and eye-point position were selected to properly simulate the 150-millimeter lens on the 4 X 5 view camera. Radiosity computations were performed in 15 evenly spaced wavelength bands between 400 and 700 nanometers, and the resulting spectral energy distributions were converted to CIE *XYZ*

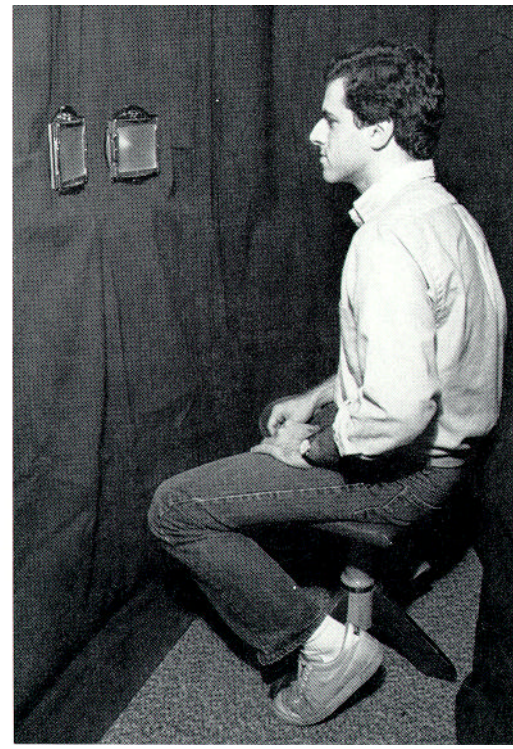


Fig. 11. A subject comparing the real and the simulated images.

tristimulus values. The RGB triplets were found by applying a matrix based on the chromaticity coordinates of the monitor phosphors and the monitor white point [3, 7]. These RGB triplets were subsequently gamma corrected and loaded into the frame store.

Preliminary observations indicated that the limited dynamic range of the monitor would not allow the light source in the ceiling to be rendered convincingly. To avoid this problem, it was decided to alter the experimental design by adding a 4.5 x 2.75-inch opaque flap at the top of the open side in order to obscure the light source when viewing the cube. This minimized the range of light intensities in the scene.

Figure 12 is a black and white picture taken from the position of the observer. It gives an approximate idea of what was seen. Further documentation was obtained by exposing color negative film in the view cameras and producing the color prints shown in Figure 13. No attempt was made to compensate for distortions caused by the photographic process or for the fact that reflection prints seen under bright ambient conditions present an entirely different mode of viewing than self-luminous images seen in a dark ambient environment. Thus these photographs should not be used to evaluate the responses given by the subjects during the comparison experiment.

The subjects for this test consisted of 10 members of the Cornell University Program of Computer Graphics Laboratory who had extensive experience evaluating computer graphics images, and 10 people with little or no experience with

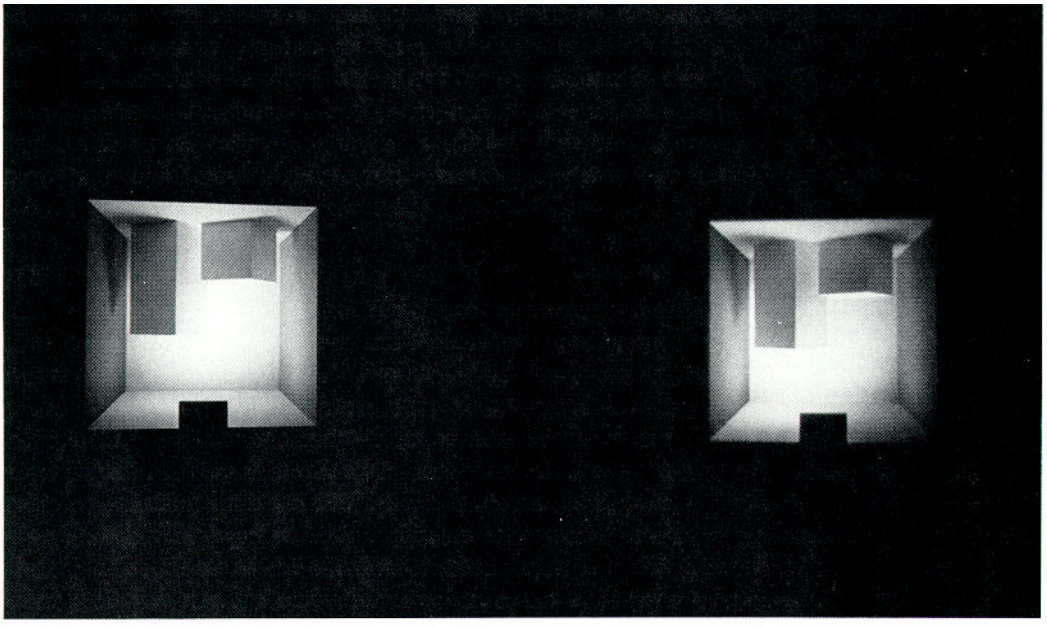


Fig. 12. Photograph taken from the position of the observer that gives an approximate idea of what was seen. The real scene is on the left and the simulation is on the right.

computer-generated pictures. In order to factor out the possible effect of differences between the two lenses, five of each group did the experiment with the lenses on particular sides and five of each group did the experiment with the lenses switched. Because no color vision test was available, the subjects were all taken at their word regarding the normalcy of their color vision.

3.4 Results

In trying to decide which was the picture of the model and which was the computer-generated picture, 9 out of 20 people, or 45 percent, selected the wrong answer. The subjects did no better than they would have by guessing.

In all cases, the subjects considered the match between the model and the simulation to be quite good. Specifically, the overall match was rated as being between good and excellent, the color match was rated as being slightly better than good, and the shadow correspondence was rated as being slightly less than good.

Two differences between the pictures were pointed out quite frequently in the written comments. The shadows were described as being fuzzy in the computer-generated image but distinct in the image of the model. This may be due to not discretizing finely enough the surfaces in the environment. It was also noted that the ceiling corners were brighter in the computer-generated image than in the image of the model. Given the results of the radiometric study, where it was discovered that the actual light emits more radiation downward than it does to the sides, this comment is not surprising, since the image was computed with the assumption that the light source emits evenly in all directions.

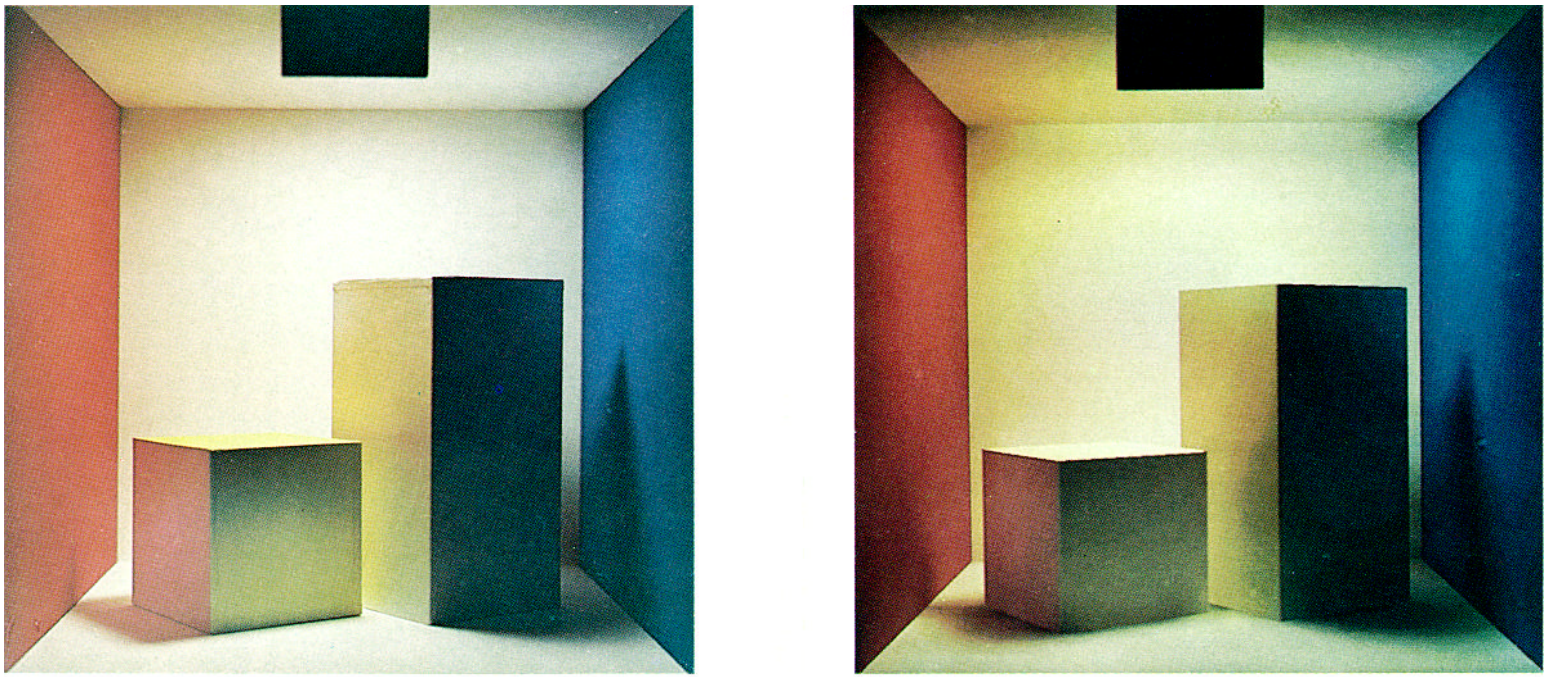


Fig. 13. Photographs taken by exposing color negative film in the view cameras. The real scene is on the left and the simulation is on the right.

4. SUMMARY AND CONCLUSIONS

Two experimental studies were carried out to assess the physical and perceptual aspects of the image synthesis process. A physical model was built using diffusely reflecting materials. The physical model was compared in two different ways to the predictions of a diffuse lighting model (the radiosity method). The first study allowed a test of the physical aspects of the lighting model (energy transfer rates were compared). The second study allowed a perceptual test of a rendered image against the physical model.

In the first study, radiometric measurements were made on the physical model and were compared with the predictions of the radiosity method. A simple, hemispherically and spectrally integrating radiometer was used. Three different environments were considered. Some general guidelines emerged for creating a lighting model that accurately describes light (and energy) transport processes in the physical scene. First, the spectral reflectance of materials in the scene must be measured and used as input to the lighting model. Similarly, the spectral and directional characteristics of the light source must be measured and used as input. It is especially important that any directionality in the light source be accounted for. If the light source is not ideal diffuse, an extension to the radiosity method as described in the paper can be used. In the experimental measurements it is necessary to account for the spectral and directional characteristics of the radiometer. With the foregoing factors accounted for, and with care in conducting the experiments, the radiometric measurements and lighting model predictions were found to be in good agreement (see Figures 5, 7, and 9). This agreement lends strong support for the radiosity method as an accurate simulation of the light transfer processes that occur in diffuse environments.

In the second study, the physical model was compared with an image on a color television monitor. The image was synthesized by applying the radiosity method on a spectral basis, selecting the viewing direction, and then converting the predicted spectral energy distributions to XYZ tristimulus values and rendering the image. A single scene was considered (see Figure 10). A perceptual experiment was carried out by asking a group of experimental subjects to compare the simulated image against the physical model. A restricted mode of viewing was employed by asking the subjects to observe the scenes through two view cameras (see Figures 11–13). In comparing the physical scene against the monitor, the subjects did no better than they would have by simple guessing. Although they considered the overall match and the color match to be good, some weaknesses were cited in the sharpness of the shadows (a consequence of the discretization in the simulation) and in the brightness of the ceiling panel (a consequence of the directional characteristics of the light source). The overall agreement lends strong support to the perceptual validity of the simulation and display process.

The present experiments provide a first step in assessing the physical and perceptual aspects of the image synthesis process. Future work should be directed to refining these comparisons. Possible steps include radiometric measurements with high directional and spectral resolution, perceptual tests with alternative modes of viewing, and extensions to more complex environments and to other lighting models.

ACKNOWLEDGMENTS

We express our appreciation to the people from both inside and outside the Computer Graphics Laboratory who donated their time to be experimental subjects. We also acknowledge the assistance of several departments and individuals at Cornell: Physics (David Weidman), Design and Environmental Analysis (Yarrow Namaste), and Mechanical and Aerospace Engineering. Photographic assistance was provided by Emil Ghinger and Rebecca Slivka. The anonymous referees contributed many constructive and clarifying suggestions for which we are grateful. All calculations were performed on VAX computers provided by a generous grant from the Digital Equipment Corporation.

REFERENCES

1. ATHERTON, P., AND CAPOREAL, L. A subjective judgement study of polygon based curved surface imagery. *CHI'85 Conference on Human Factors in Computing Systems* (San Francisco, Calif., Apr. 14–18). ACM/SIGCHI, New York, 1985.
2. COHEN, M., AND GREENBERG, D. P. The hemi-cube: A radiosity solution for complex environments. *Comput. Graph. 19*, 3 (July 1985), 31–40.
3. COWAN, W. B. An inexpensive scheme for calibration of a colour monitor in terms of CIE Standard coordinates. *Comput. Graph. 17*, 3 (July 1983), 315–321.
4. DAVIS, R. G., AND BERNECKER, C. A. An evaluation of computer graphic images of the lighted environment. *J. Illum. Eng. Soc. 14*, 1 (Oct. 1984), 493–514.
5. GORAL, C., TORRANCE, K. E., GREENBERG, D. P., AND BATTAILE, B. Modeling the interaction of light between diffuse surfaces. *Comput. Graph. 18*, 3 (July 1984), 213–222.
6. IMHOFF, E. A. Raman scattering and luminescence in polyacetylene during the *cis–trans* isomerization. Ph.D. dissertation, Physics Dept., Cornell Univ., Ithaca, N.Y., May 1983, pp. 92–95 and Appendix A.
7. MEYER, G. W. Colorimetry and computer graphics. Rep. 83-1, Program of Computer Graphics, Cornell Univ., Ithaca, N.Y., Apr. 1983.
8. MILLER, N. J., NGAI, P. Y., AND MILLER, D. D. The application of computer graphics in lighting design. *J. Illum. Eng. Soc. 14*, 1 (Oct. 1984), 6–26.
9. SCHORNHORST, J. R., AND VISKANTA, R. An experimental examination of the validity of the commonly used methods of radiant heat transfer analysis. *J. Heat Transfer 90* (Nov. 1968), 429–436.
10. SHIH, S. H., LOVE, T. J., AND FRANCIS, J. E. Direct measurement of the radiosity of a nonisothermal hemispherical cavity. In *AIAA Progress in Astronautics and Aeronautics: Heat Transfer, Thermal Control, and Heat Pipes*, vol. 70, W. B. Olstad, Ed. American Institute of Aeronautics and Astronautics, New York, 1980.
11. SIEGEL, R., AND HOWELL, J. R. *Thermal Radiation Heat Transfer*. Hemisphere Publishing, Washington D.C., 1981.
12. SMITH, O. W., AND GRUBER, H. Perception of depth in photographs. *Percept. Motor Skills 8*, (1958), 307–313.

Received May 1985; accepted April 1986

Toward a Psychophysically-Based Light Reflection Model for Image Synthesis

Fabio Pellacini* James A. Ferwerda* Donald P. Greenberg*
Program of Computer Graphics
Cornell University

ABSTRACT

In this paper we introduce a new light reflection model for image synthesis based on experimental studies of surface gloss perception. To develop the model, we've conducted two experiments that explore the relationships between the physical parameters used to describe the reflectance properties of glossy surfaces and the perceptual dimensions of glossy appearance. In the first experiment we use multidimensional scaling techniques to reveal the dimensionality of gloss perception for simulated painted surfaces. In the second experiment we use magnitude estimation methods to place metrics on these dimensions that relate changes in apparent gloss to variations in surface reflectance properties. We use the results of these experiments to rewrite the parameters of a *physically-based* light reflection model in *perceptual* terms. The result is a new *psychophysically-based light reflection model* where the dimensions of the model are perceptually meaningful, and variations along the dimensions are perceptually uniform. We demonstrate that the model can facilitate describing surface gloss in graphics rendering applications. This work represents a new methodology for developing light reflection models for image synthesis.

Keywords

I.3.7 Three-Dimensional Graphics and Realism, Human Factors, Experimentation, Light Reflection Models, Gloss, Visual Perception.

1. INTRODUCTION

Color and *gloss* are two fundamental visual attributes used to describe the appearances of objects in synthetic images. In a typical graphics rendering application a user specifies an object's color as an RGB triple and describes its gloss in terms of the parameters of a light reflection model such as Phong [Phon75].

In addition to RGB, many rendering applications allow users to describe color in more perceptually meaningful color spaces such as HSV, Munsell, or CIELAB, that have grown out of the science of colorimetry [Wysz82]. Working in these spaces makes it easier to specify color, because the dimensions of the spaces are representative of our visual experience of color, and the scaling of the dimensions is perceptually uniform.

Unfortunately similar perceptually-based spaces for specifying surface gloss do not yet exist. At the present time the parameters used to describe gloss are either based on ad-hoc lighting models such as Phong, or are motivated by research into the physical aspects of light reflection [Blin77, Cook81, He91, Ward92, Schl93, LaFo97, Stam99]. In either case, the visual effects of the parameters are relatively unintuitive and interactions among different parameters make it difficult to specify and modify surface gloss properties. A light reflection model grounded in the visual psychophysics of gloss perception would greatly facilitate the process of describing surface gloss properties in computer graphics renderings, and could lead to more efficient and effective rendering methods.

Figure 1: Coffee mugs with different gloss attributes.

In this paper we introduce a new light reflection model for image synthesis based on experimental studies of surface gloss perception. To develop the model, we have conducted two psychophysical studies to explore the relationships between the physical parameters used to describe the reflectance properties of glossy surfaces and the perceptual dimensions of glossy appearance. We use the results of these experiments to rewrite the parameters of a *physically-based* light reflection model in *perceptual* terms. The result is a new *psychophysically-based light reflection model* where the dimensions of the model are perceptually meaningful, and variations along the dimensions are perceptually uniform. We demonstrate that the model is useful for describing and modifying surface gloss properties in graphics rendering applications. However, the long-term impact of this work may be even more important because we present a new methodology for developing psychophysical models of the goniometric aspects of surface appearance to complement widely used colorimetric models.

2. BACKGROUND

To develop a psychophysically-based light reflection model for image synthesis we first need to understand the nature of gloss perception.

In his classic text, Hunter [Hunt87] observed that there are at least six different visual phenomena related to apparent gloss. He identified these as:

* 580 Rhodes Hall, Ithaca NY, 14853
http://www.graphics.cornell.edu/-{fabio,jaf,dpg}@graphics.cornell.edu

specular gloss – perceived brightness associated with the specular reflection from a surface

contrast gloss – perceived relative brightness of specularly and diffusely reflecting areas

distinctness-of-image (DOI) gloss – perceived sharpness of images reflected in a surface

haze – perceived cloudiness in reflections near the specular direction

sheen – perceived shininess at grazing angles in otherwise matte surfaces

absence-of-texture gloss – perceived surface smoothness and uniformity

Judd [Judd37] operationalized Hunter's definitions by writing expressions that related them to the physical features of surface reflectance distribution functions (BRDFs). Hunter and Judd's work is important, because it is the first to recognize the multidimensional nature of gloss perception.

In 1987 Billmeyer and O'Donnell [Bill87] published an important paper that tried to address the issue of gloss perception from first principles. Working with a set of black, gray, and white paints with varying gloss levels, O'Donnell collected ratings of the apparent difference in gloss between pairs of samples and then used multidimensional scaling techniques to discover the dimensionality of perceived gloss. He concluded that for his sample set and viewing conditions (flat samples, structured/direct illumination, black surround) the appearance of high gloss surfaces is best characterized by a measure similar to distinctness-of-image gloss, while the appearance of low gloss surfaces is better described by something like contrast gloss.

In the vision literature, studies of gloss have focused primarily on its effects on the perception of shape from shading. Todd and Mingolla [Todd83, Ming86] found that gloss generally enhances the perception of surface curvature. Blake [Blak90] found categorical changes in surface appearance and shape depending on the 3d location of the specular highlight. Braje [Braj94] found interactions between apparent shape and apparent gloss, showing that a directional reflectance pattern was perceived as more or less glossy depending on the shape of its bounding contour. More recently Nishida [Nisi98] also studied interactions between shape and gloss, and found that subjects are poor at matching the Phong parameters of bumpy surfaces with different frequency and amplitude components.

Finally, in computer graphics, while there has been extensive work on developing physically-based light reflection models, there has been relatively little effort to develop models whose dimensions are perceptually meaningful. One exception is Strauss's model [Stra90], a hybrid of Phong and Cook-Torrance, that describes surface properties with five parameters: color, smoothness, metalness, transparency, and refractive index. He reports that users find it much easier to specify surface gloss with this model than with others.

There is still much work to be done in this area. First, with the exception of Billmeyer and O'Donnell's work there has been little investigation of the multidimensional nature of glossy appearance from first principles. Hunter's observations about visual gloss phenomena are insightful but we need studies that quantify these different appearance dimensions and relate them to the physical properties of materials. Second, all previous gloss studies have looked exclusively at locally illuminated surfaces in uniform surrounds. This practice is understandable given the difficulty of controlling complex environments, but it's strange considering that one of the most salient things about glossy surfaces is their ability to reflect their surroundings. To really understand how we perceive surface gloss, we need to study three-dimensional objects in realistically rendered environments. Fortunately, image synthesis gives us a powerful tool to study the perception of surface gloss. Physically-based image synthesis methods let us make realistic images of three-dimensional objects in complex, globally-illuminated scenes, and gives us precise control over object properties. By using image synthesis techniques to conduct psychophysical experiments on gloss perception we should be able to make significant progress toward our goal of developing a psychophysically-based light reflection model that can describe the appearance of glossy materials.

3. EXPERIMENTS

3.1 Motivation

In many ways the experiments that follow are analogous to early research done to establish the science of colorimetry. In that work, researchers wanted to understand the relationships between the physical properties of light energy, and our perception of color. Many of the earliest experiments focused on determining the *dimensionality* of color perception, culminating with Young's trichromatic theory [Helm24]. Following this, further experiments were done to find *perceptually meaningful axes* in this three-dimensional color space. Hering's work [Heri64] on opponent color descriptions, falls into this category. Finally, many experiments have been done to scale these axes and create *perceptually uniform* color spaces. Munsell, Judd, and MacAdam's efforts to develop uniform color scales are good examples (see [Wysz82] for a review).

Although we recognize the great effort involved in the development of color science, our overall goals with respect to understanding gloss are similar: we are conducting experiments to understand gloss perception with the goal of building a psychophysical model of gloss that relates the visual appearance of glossy surfaces to the underlying physical properties of the surfaces.

- In Experiment 1 we will use multidimensional scaling techniques to reveal both the *dimensionality* of gloss perception, and to suggest *perceptually meaningful axes* in visual "gloss space"
- In Experiment 2 we will use magnitude estimation techniques to place quantitative metrics on these axes and create a *perceptually uniform* gloss space.
- Finally we will use these results to develop a psychophysically-based light reflection model for image synthesis.

Gloss is a visual attribute of a wide variety of materials including plastics, ceramics, metals, and other man-made and organic substances. Eventually we would like to develop a model that can explain the appearances of all these kinds of materials, but initially we need to restrict our studies to a manageable subclass. To start, we've chosen to study a set of achromatic glossy paints. We chose paints because they exhibit a wide variety of gloss levels from flat to high gloss; their reflectance properties have been measured extensively so there are good models to describe their physical characteristics, and they are widely used in art and industry, so hopefully our findings will be immediately useful.

3.2 Experiment 1: Finding the perceptual dimensions of gloss space

3.2.1 Purpose

The purpose of Experiment 1 is to determine the dimensionality

of gloss perception for painted surfaces in synthetic images and to find perceptually meaningful axes in this visual gloss space. To do this we've designed an experiment based on multidimensional scaling techniques.

3.2.2 Methodology: Multidimensional scaling

Multidimensional scaling (MDS) is statistical method for finding the latent dimensions in a dataset [Borg97]. Multidimensional scaling takes a set of measures of the distances between pairs of objects in a dataset and reconstructs a space that explains the dataset's overall structure. This concept is best illustrated by example.

Table 1 shows a matrix of the distances between a number of U.S. cities. This matrix indicates how far one city is from another but gives no sense of their spatial relations. If this *proximity matrix* is used as input to the PROXSCAL MDS algorithm [Busi97], it attempts to reconstruct the spatial positions of the cities to best explain the proximity measures.

The two-dimensional MDS solution produced by the algorithm is shown in Figure 2, where you can see that MDS has recovered the true spatial layout of the cities (the outline of the U.S. map is overlaid for reference). Since distances in a space are unaffected by rotations or inversions, MDS solutions are only specific up to these transformations, and it is the experimenter's job to find meaningful axes in the solution.

Although a two-dimensional MDS solution is shown in Figure 2, MDS can produce solutions in any number of dimensions to try to achieve the best fit to the data. The goodness of the fit is known as the *stress* of the solution. The stress formula used in the example is:

$$stress = \sum_{i,j} \left[\delta_{i,j} - d(x_i, x_j)\right]^2 \qquad (1)$$

where $\delta_{i,j}$ are the input proximities, x_i and x_j are the recovered locations in the n^{th} dimensional solution, and d is a measure of the distance between them. The MDS algorithm attempts to minimize the stress for each of the solutions.

Figure 3 plots the stress values for solutions running from 1 to 5 dimensions. The stress curve will drop sharply as dimensions are added that explain more of the data and will decline more slowly as further superfluous dimensions are added. Standard practice is to choose the dimensionality indicated by this inflection point in the stress curve. The stress curve in Figure 3 indicates that a two-dimensional solution provides the best fit to the data, but this is to be expected since the dataset is inherently two dimensional, and error in the proximity measures is negligible, providing a perfect two-dimensional fit. In typical experimental datasets, noise in the data results in a stress curve that drops then asymptotes as greater-than-necessary dimensions are added.

MDS algorithms come in a variety of flavors that depend on the form of the stress function the algorithm uses. In our work we use a variant called *weighted Euclidean non-metric MDS* [Borg97] that allows us to combine data from multiple subjects, compensate for individual differences, and analyze datasets where the proximities may only reflect ordinal rather than interval relations in the data. We also use a second variant called *confirmatory* MDS [Borg97] which let us test hypotheses about the functional forms of the dimensions and their orthogonality.

	Atl	Chi	Den	Hou	LA	Mia	NYC	SF	Sea	DC
Atlanta	0									
Chicago	587	0								
Denver	1212	920	0							
Houston	701	940	879	0						
LA	1936	1745	831	1374	0					
Miami	604	1188	1726	968	2339	0				
NYC	748	713	1631	1420	2451	1092	0			
SF	2139	1858	949	1645	347	2594	2571	0		
Seattle	2182	1737	1021	1891	959	2734	2406	678	0	
DC	543	597	1494	1220	2300	923	205	2442	2329	0

Table 1: Proximity matrix of distances between U.S. cities.

Figure 2: MDS reconstruction of the U.S. map.

3.2.3 Experimental Procedure

3.2.3.1 Stimuli

To apply MDS to the problem of finding the dimensionality of gloss perception, we first need to construct a stimulus set with objects that vary in gloss, and then collect measures of the apparent differences in gloss between pairs of objects in the set. These apparent gloss differences then serve as the proximities that the MDS algorithm uses to construct a representation of visual "gloss space".

A composite image of the stimulus set used in Experiment 1 is shown in Figure 4. The environment consisted of a sphere enclosed in a checkerboard box illuminated by an overhead area light source. Images were generated using a physically-based Monte Carlo path-tracer that used an isotropic version of Ward's [Ward92] light reflection model:

$$\rho(\theta_i, \phi_i, \theta_o, \phi_o) = \frac{\rho_d}{\pi} + \rho_s \cdot \frac{\exp[-\tan^2 \delta / \alpha^2]}{4\pi\alpha^2 \sqrt{\cos\theta_i \cos\theta_o}} \qquad (2)$$

where $\rho(\theta_i,\phi_i,\theta_o,\phi_o)$ is the surface BRDF, θ_i,ϕ_i, and θ_o,ϕ_o are spherical coordinates for the incoming and outgoing directions, and δ is the half-angle between them. Ward's model uses three parameters to describe the BRDF: ρ_d – the object's diffuse reflectance; ρ_s – the energy of its specular component, and α – the spread of the specular lobe. Our reason for choosing Ward's model is that we wanted the objects in the stimulus set to be representative of the gloss properties of real materials, and Ward gives parameters that represent measured properties of a range of glossy paints. The parameters used in our stimulus set span this range. Each parameter was set to three levels. ρ_s values were (0.033, 0.066, 0.099), α values were (0.04, 0.07, 0.10), and ρ_d was set to (0.03, 0.193, 0.767) which are the diffuse reflectance factors corresponding to Munsell values (N2, N5, and N9). The

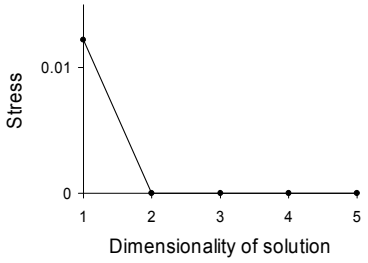


Figure 3: Stress vs. dimensionality graph for MDS solution.

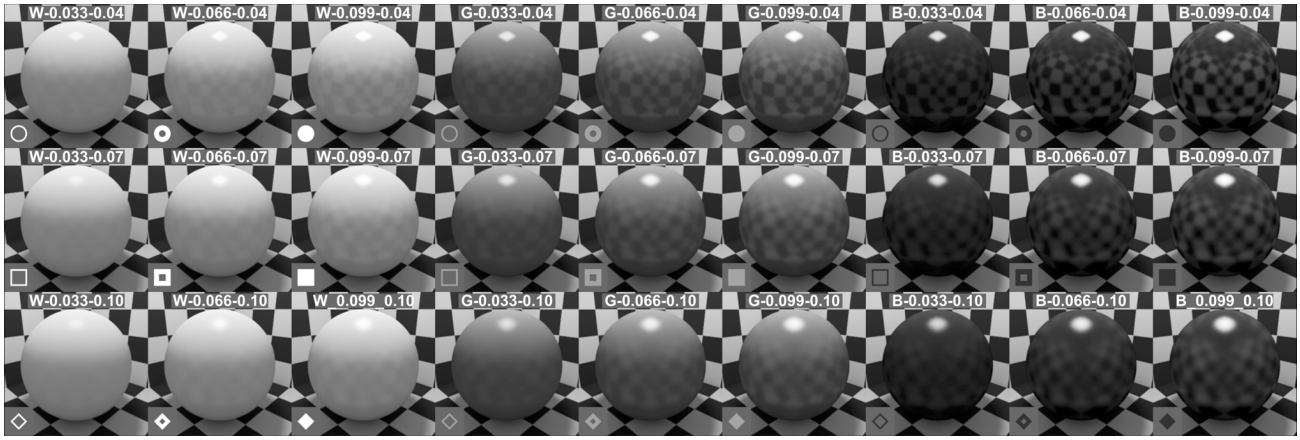


Figure 4[†]: Composite image of the stimulus set used in Experiment 1. Labels indicate the diffuse color (white, gray, black), and ρ_s and α values. Symbols are included as an aid for interpreting subsequent figures.

black and white checks in the checkerboard surround were completely diffuse and had ρ_d's of 0.03 and 0.767 respectively. By using all combinations of the ρ_d, ρ_s, and α parameters for the sphere objects, we produced the 27 images shown in Figure 4.

Choosing a tone reproduction operator to map from calculated image radiances to display values presented a challenge because the images had high dynamic ranges caused by the visible reflection of the light source. We experimented with a number of tone reproduction operators including simple clipping and gamma compression as well as Pattanaik [Patt98] and Ward-Larson's [Ward97] high dynamic range operators but we abandoned these methods because they produced objectionable artifacts such as halos and banding. We settled on Tumblin's [Tumb99] Rational Sigmoid function which compresses the light source highlight without abrupt clipping and allows all other scene values to be directly mapped to the display.

One of the consequences of the limited dynamic range of display devices is that any gloss attribute related to the absolute intensity of a highlight is not likely to play much of a role in how glossy surfaces appear in images. Given the amount of effort that has gone into developing physically accurate light reflection models for realistic image synthesis, addressing the particular dynamic range problems caused by trying to display images of glossy surfaces is certainly a subject that merits future work.

3.2.3.2 Procedure

Nine subjects participated in Experiment 1. The subjects were the first two authors and seven graduate and undergraduate Computer Science students. All had normal or corrected to normal vision. With the exception of the authors, all were naïve to the purpose and methods of the experiment.

In the experimental session, the subjects viewed pairs of images displayed on a calibrated SXGA monitor. Minimum and maximum monitor luminances were 0.7 and 108 cd/m^2 and the system gamma was 2.35. The images were presented on a black background in a darkened room. The monitor was viewed from a distance of 60 inches to ensure that the display raster was invisible. At this viewing distance each image subtended 3.2 degrees of visual angle.

Subjects were asked to judge the apparent difference in gloss between the pair of objects shown in the images. They entered their responses using a mouse to vary the position of a slider that was displayed below the images. The ends of the slider scale were labeled "0, small difference" and "100, large difference". A readout below the slider indicated the numeric position along the scale.

Subjects judged the apparent gloss differences of all 378 object pairs in the stimulus set. The pairs were presented in random order. For each subject, the apparent gloss differences measured in the experiment were used to fill out a 27 x 27 proximity matrix. All nine proximity matrices were used as input to the PROXSCAL MDS algorithm using the weighted Euclidean non-metric stress formulation.

3.2.4 Analysis/Discussion

Recall that our goal in this experiment is to discover the dimensionality of gloss perception for the painted surfaces and to find perceptually meaningful axes in this gloss space. To do this we observed how the stress varied with the dimensionality of the MDS solution. Figure 5 plots stress values for solutions running from 1 to 5 dimensions. The stress value drops significantly with the change from a 1-dimensional to a 2-dimensional solution, but declines more slowly with the addition of higher dimensions which are probably only accommodating noise in the dataset. *From this pattern of results we infer that under these conditions apparent gloss has two dimensions.*

The two-dimensional gloss space recovered by MDS is shown in Figure 6. In the Figure, MDS has placed the objects at locations that best reflect the differences in apparent gloss reported by the subjects.

As stated earlier, since distances in this space are invariant under rotation, inversion or scaling, it is our job to look for perceptually meaningful axes in the space. The cross in the lower right corner of the diagram indicates two important trends in the data that are related to properties of the reflected images formed

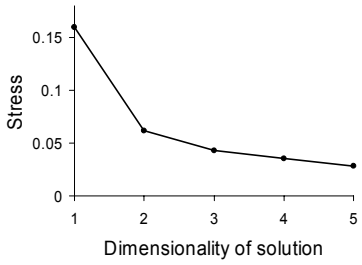


Figure 5: Dimensionality vs. stress graph for Experiment 1.

[†]Gloss appearance parameters are specified for the display conditions described in the experiments. Appearance in the printed images is subject to the limitations of the printing process.

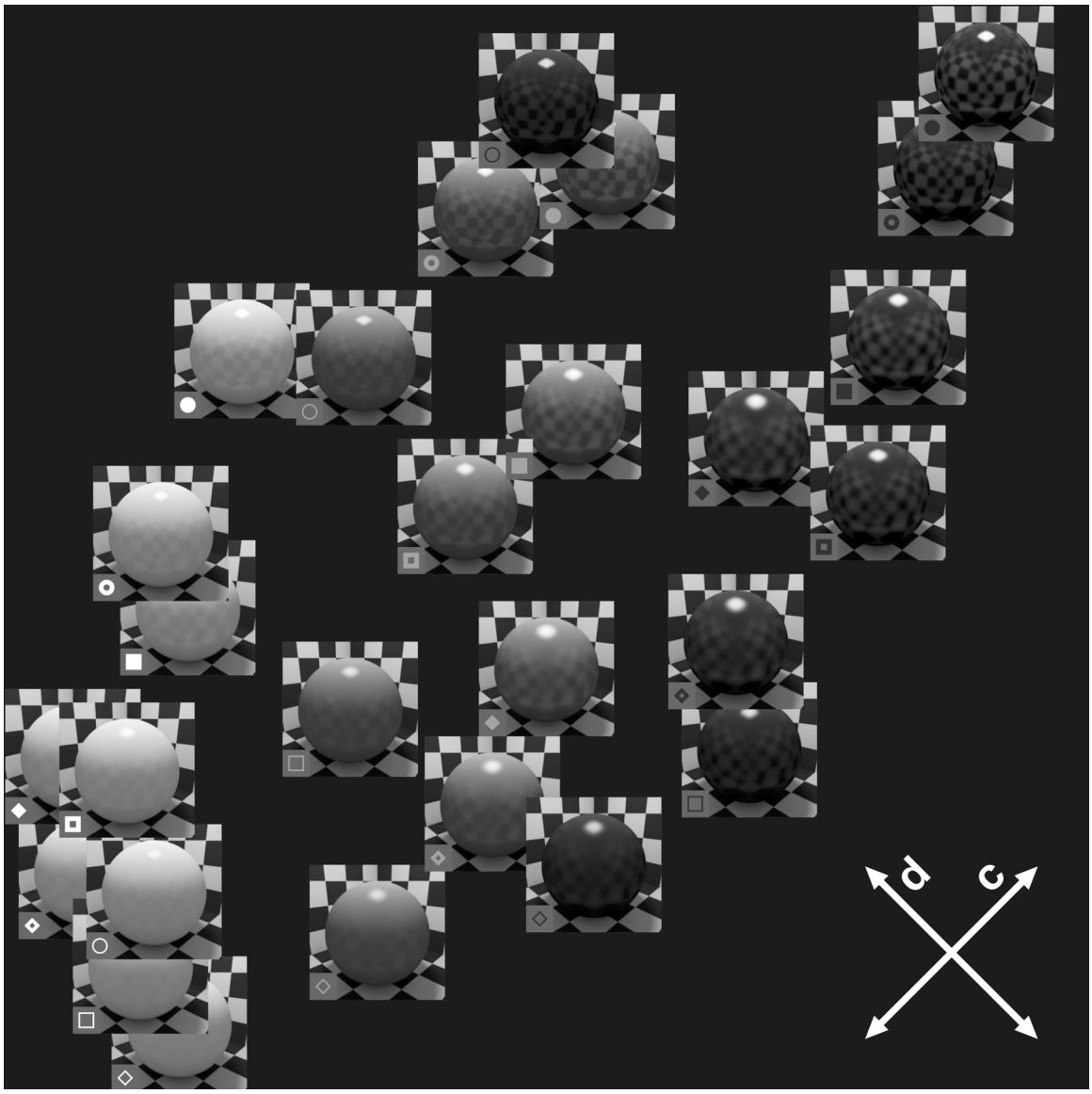


Figure 6[†]: Two-dimensional MDS solution for Experiment 1.

by the surfaces. First, the *apparent contrast of the reflected image* increases from the lower left to the upper right of the diagram. Second, the apparent sharpness or *distinctness of the reflected image* increases from lower right to upper left. We believe these dimensions are qualitatively similar to the *contrast gloss* and *distinctness-of-image (DOI) gloss* attributes Hunter observed and so we will name our dimensions *c* for contrast gloss and *d* for DOI gloss. However, to foreshadow the results of the next experiment, we will differ significantly from Hunter (and Judd) in the quantitative formulation of relationship between these perceptual dimensions and the physical dimensions used to describe surface BRDFs.

3.3 Experiment 2: Creating a perceptually uniform gloss space

3.3.1 Purpose

In Experiment 1 we discovered the dimensionality of gloss perception and identified perceptually meaningful axes in visual gloss space for painted surfaces in synthetic images. The purpose of Experiment 2 is to place psychophysical metrics on these axes and rescale them to create a perceptually uniform gloss space. To do this we've designed an experiment based on magnitude estimation techniques.

3.3.2 Methodology: Magnitude estimation

Magnitude estimation is one of a family of psychophysical *scaling* techniques designed to reveal functional relationships between the physical properties of a stimulus and its perceptual attributes [Torg60]. In the basic magnitude estimation procedure, subjects are presented with a random sequence of stimuli that vary along some physical dimension, and they are asked to assign a number to each stimulus that indicates the apparent magnitude of the corresponding perceptual attribute. Magnitude estimates are then used to derive a psychophysical scale.

3.3.3 Experimental Procedure

3.3.3.1 Stimuli

Two magnitude estimation studies were performed in Experiment 2 to scale the perceptual gloss dimensions found in Experiment 1. In both cases the stimuli used were subsets of the stimuli used in Experiment 1, supplemented by new stimuli with

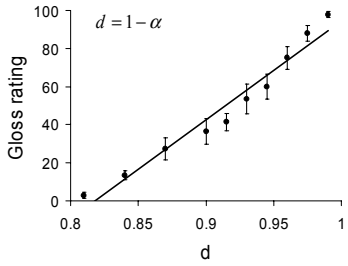


Figure 7: Magnitude estimates and fit for DOI gloss *d*.

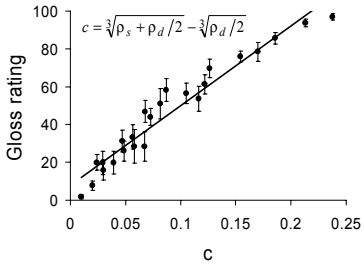


Figure 8: Magnitude estimates and fit for contrast gloss *c*.

parameters intermediate to those in the original set. In the *contrast gloss* scaling study 24 images were used, showing objects with combinations of ρ_d levels of (0.03, 0.087, 0.193, 0.420, 0.767) (black, dark/medium/light gray, white) and ρ_s levels of (0.017 0.033, 0.050, 0.066, 0.083 0.099) (low to high specular energy), the α parameter was fixed at 0.04 (small spread) to make variations along the contrast gloss dimension as salient as possible. In the *DOI gloss* scaling study, α was varied in 11 levels from 0.01 to 0.19 (small to large spread), and the ρ_d and ρ_s parameters were fixed at 0.03 (black) and 0.099 (high specular energy) to make variations along the DOI gloss dimension as salient as possible.

3.3.3.2 Procedure

The subjects in Experiment 2 were the same as those in Experiment 1, and the same display techniques, viewing conditions, and data gathering methods were used.

In each magnitude estimation study, subjects viewed single images from the new stimulus sets. Images were presented in a random sequence and each sequence was repeated three times. On each trial subjects were asked to judge the apparent glossiness of the object in the image on a scale from 0 to 100 by adjusting the on-screen slider.

3.3.4 Analysis/Discussion

Our goal in these experiments is to derive psychophysical scaling functions that relate changes in apparent gloss along the perceptual dimensions we discovered in Experiment 1 to variations in the parameters of the physical light reflection model. To achieve this goal we tested various hypotheses about functional relationships between the physical and perceptual dimensions, first with least squares fitting techniques on the magnitude estimation data and then with confirmatory MDS on the full dataset from Experiment 1. This approach allowed us to verify that the scaling functions are task independent and to determine whether the perceptual dimensions are orthogonal.

First we examined the *d* (DOI gloss) dimension. Our hypothesis was that *d* is inversely related to the α parameter. In Figure 7 subjects' gloss ratings are plotted versus the function $d = 1 - \alpha$. The line was obtained through linear regression and the r^2 value of the fit was 0.96. Polynomial fits only increased r^2 by less than 0.01 so we concluded that the relationship is linear.

Interpreting the *c* (contrast gloss) dimension was less straightforward. In the MDS solution from Experiment 1 (Figure 6) it is clear that *c* varies with diffuse reflectance, since the white, gray, and black objects form distinct clusters that occupy different ranges along the *c* dimension. Our first hypothesis was that *c* is a simple function of the physical contrast (luminance ratio) of the black and white patches in the reflected image but this provided a very poor fit to the data ($r^2 = 0.76$). Our second hypothesis was that "contrast" in this situation is a function of the *difference in apparent lightness* of the two patches, where lightness is defined as in CIELAB [Fair98]. This second formulation provided a much better fit to the magnitude estimation data ($r^2 = 0.87$). However when we tested this second hypothesis on the full dataset from Experiment 1 using confirmatory MDS, we found that the fit was poor for surfaces with large α values where the physical contrast in the image plane drops as the reflected image gets blurrier. We then tested a third hypothesis that subjects' lightness judgments are based on inferred object-space reflectance values rather than image-space intensity values (i.e. subjects show lightness constancy [Fair98], compensating for blur-related image contrast losses). This hypothesis is formalized in Equation 4 which we derived using standard integration techniques under the assumption of small α values and high environmental contrast.

Figure 8 plots the data from the contrast gloss scaling study, which shows how subjects' gloss ratings relate to this final formulation for the *c* dimension. The line was obtained through linear regression and is a good fit to the data with an r^2 value of 0.94. This result shows that subjects appear to be compensating for the decrease in physical image contrast caused by blurring in making their judgments of the lightnesses of the reflected patches. Using this formulation also decreased the stress value in a subsequent confirmatory MDS test on the full dataset, which indicates that the *c* and *d* axes are independent, and therefore orthogonal in gloss space.

Equations 3 and 4 show the final formulas for the *c* and *d* axes. These formulas define psychophysical metrics that relate changes in apparent gloss along these two axes to variations in the physical parameters of the light reflection model.

$$d = 1 - \alpha \tag{3}$$

$$c = \sqrt[3]{\rho_s + \rho_d/2} - \sqrt[3]{\rho_d/2} \tag{4}$$

These axes are perceptually linear, but to make the space perceptually uniform, we need to find weighting factors for the axes so that distances in the space can be measured. These weights are given as a byproduct of the confirmatory MDS tests we ran which lets us write the distance as:

$$D_{ij} \propto \sqrt{[c_i - c_j]^2 + [1.78 \cdot (d_i - d_j)]^2} \tag{5}$$

Figure 9 shows a visualization of the perceptually uniform gloss space with the stimuli from Experiment 1 placed at their predicted locations. The Figure shows the contrast gloss (*c*) and DOI gloss (*d*) dimensions form a two-dimensional space, (which is also shown in the inset), and surface lightness (*L*) (which we will incorporate in the following section) is an orthogonal third dimension.

Like perceptually uniform color spaces, this perceptually uniform gloss space has a number of important properties. For example, it allows us to:

- predict the visual appearance of a glossy paint from its physical reflectance parameters

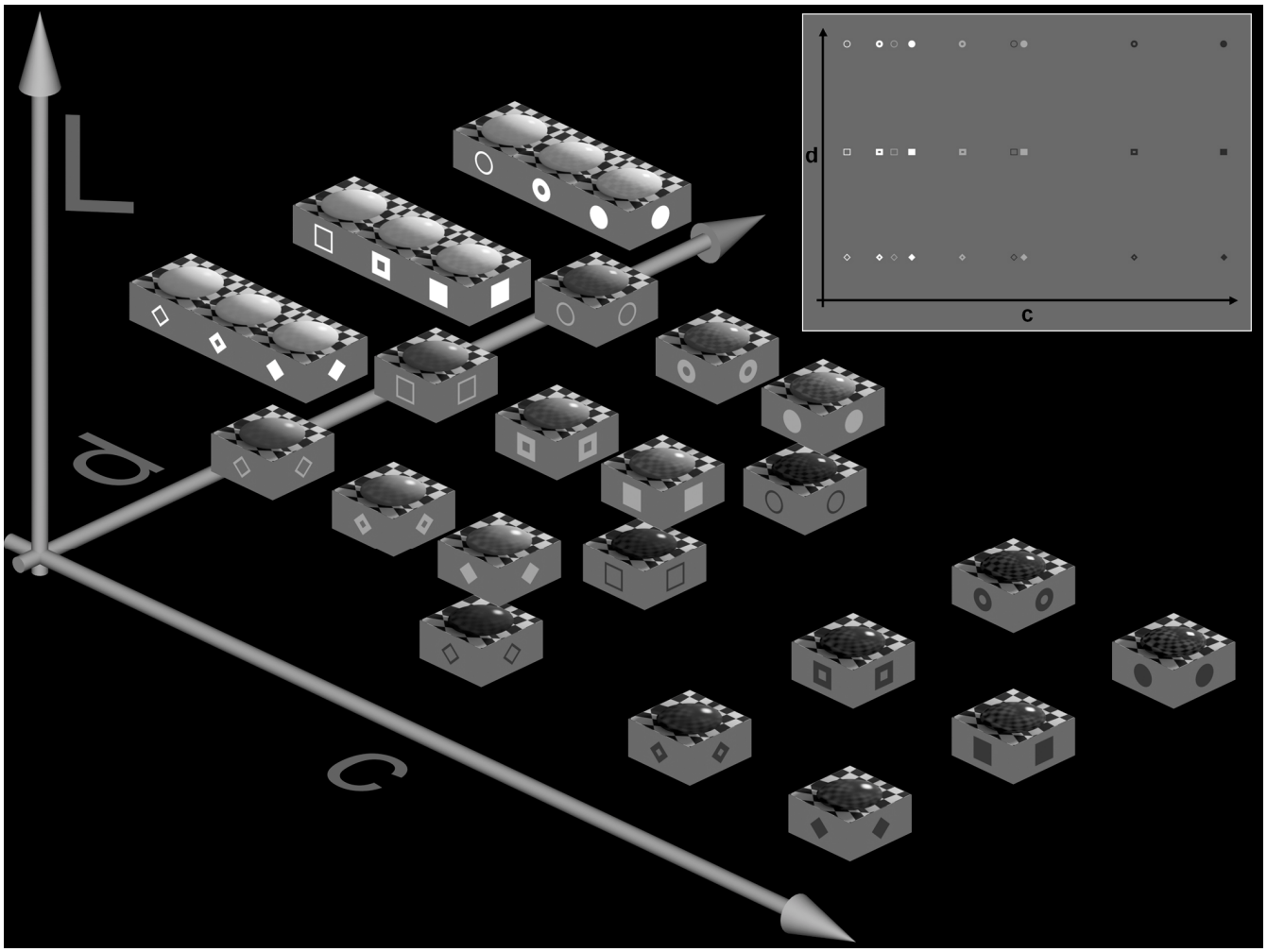


Figure 9[†]: The perceptually uniform gloss space derived from Experiment 2.

- compare two paints with respect to the two visual gloss dimensions
- produce paints with different physical reflectance values that match in terms of apparent gloss
- calculate isogloss contours that describe paints that differ equally in apparent gloss from a standard.

4. A PSYCHOPHYSICALLY-BASED LIGHT REFLECTION MODEL

To take full advantage of this new space, we are going to rewrite the parameters of the physically-based light reflection model (Equations 6,7,8) in perceptual terms to create a psychophysically-based light reflection model that can be used to describe both the physical and visual characteristics of the paints we studied. To do this, we need to introduce a perceptually linear parameter related to diffuse reflectance. For compatibility with perceptually uniform color spaces we chose CIELAB lightness (L). This final addition allows us to express the physical parameters in terms of the perceptual ones through the following equations:

$$\rho_d = f^{-1}(L) \tag{6}$$

$$\rho_s = \left(c + \sqrt[3]{f^{-1}(L)/2}\right)^3 - f^{-1}(L)/2 \tag{7}$$

$$\alpha = 1 - d \tag{8}$$

where f is the CIELAB lightness function normalized in [0,1].

Figure 10 illustrates the influence of the lightness of the diffuse component on perceived gloss. Here the solid curve plots the maximum contrast gloss c achievable for different lightness values (derived by enforcing energy conservation of the BRDF). This defines the envelope of gloss space with respect to lightness. We also plotted how contrast gloss varies with lightness for a fixed energy of the specular lobe. This curve shows that for the same specular energy, contrast gloss is smaller for lighter objects. That is to say, if two surfaces are painted with black and white paints having the same physical formulations, the black surface will appear glossier than the white one.

Strictly speaking, the model we've developed is only predictive

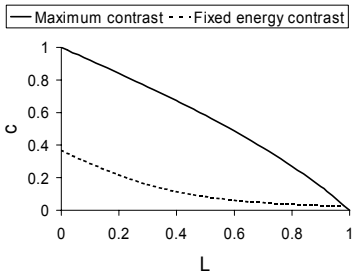


Figure 10: Effect of surface lightness on apparent gloss.

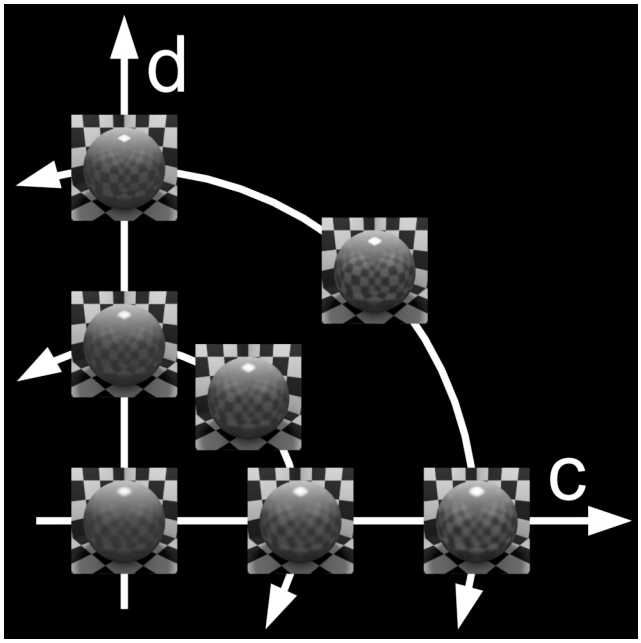


Figure 11†: Isogloss difference contours.

within the range of our stimuli, which covers a substantial range of measured glossy paints. However we feel confident that the model can be applied outside this range to cover the space of physically plausible BRDFs expressible using the Ward model, but we believe that the physical parameters should be maintained in the range of the ones measured for real materials. In particular, the α value should not be much larger than 0.2 since the specular lobe of the BRDF is not normalized for larger values [Ward92].

5. APPLYING THE MODEL

In the previous section we used the results of our gloss perception studies to develop a psychophysically-based light reflection model for image synthesis where the dimensions of the model are perceptually meaningful and variations along these dimensions are perceptually uniform. In this section we demonstrate the power of the model by showing how it can be used to facilitate the process of describing surface appearance in graphics rendering applications.

5.1 Describing differences in apparent gloss

One of the benefits of working in a perceptually uniform description space is that steps along the dimensions produce equal changes in appearance. This is true of uniform color spaces such as CIELAB where equal numerical steps in lightness (L) or chroma (a,b) produce perceptually equal changes in color appearance.

The perceptually uniform gloss space our light reflection model is based on has similar properties. Figure 11 shows *isogloss difference contours* with respect to the object in the lower left corner of the diagram (c = 0.087, d = 0.93). According to the model, the objects falling on the circular contours are equally different in apparent gloss from the reference object. The concentric circles show two degrees of isogloss difference (Δc = 0.04, Δd = 0.22 = 0.04/1.78).

It's important to observe that because the gloss space is two-dimensional (c,d), objects equidistant from a reference object may have different reflectance properties even though they will be judged to be equally different in gloss from the reference. For example, the two objects at 12 and 3 o'clock in Figure 11 have very different reflectance properties: the one at 12 o'clock produces a sharp but low contrast reflection, while the one at 3 o'clock makes a blurry but high contrast reflection, still the model predicts that they will be judged to be equally different in gloss from the reference object. This prediction was supported by an informal ranking study we ran using the stimulus set from Experiment 1. Objects whose parameters fell along isogloss contours with respect to a low gloss reference object received similar rank values implying that they appeared equally "glossy" but in different ways.

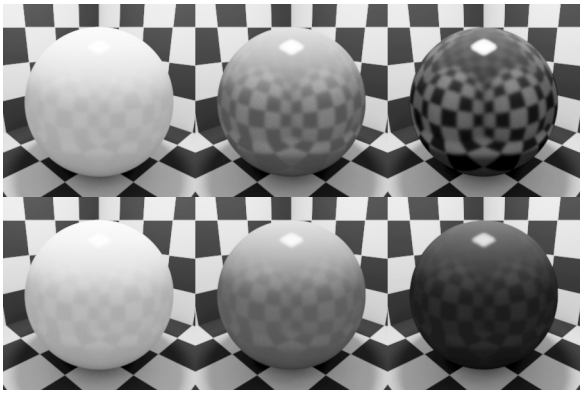


Figure 12†: Matching apparent gloss: white, gray, and black objects having the same physical gloss parameters (top row) and perceptual gloss parameters (bottom row).

This demonstration shows that our model provides the ability to specify differences in apparent gloss. This should make it much easier to modify object gloss properties in controlled ways in graphics rendering applications.

5.2 Matching apparent gloss

Many studies of gloss perception [Hunt87, Bill87] have noted that apparent gloss is affected by the diffuse reflectance of a surface, with light colored surfaces appearing less glossy than dark ones having the same finish. This effect is illustrated in the top row of Figure 12 where the white, gray and black objects have the same physical gloss parameters (ρ_s = 0.099, α = 0.04) but differ in apparent gloss with the white sphere appearing least glossy and the black sphere appearing most glossy. This phenomenon makes it difficult to create objects with different lightnesses that match in apparent gloss. The bottom row of Figure 12 shows the results produced with our psychophysically-based gloss model. When the objects are assigned the same perceptual gloss values (c = 0.057, d = 0.96) they appear to have similar gloss despite differences in their lightnesses. This property of the model should make it much easier to create objects that have the same apparent gloss, since the parameters that describe object lightness (L) and gloss (c,d) have been decoupled.

5.3 A new tool for modeling surface appearance in computer graphics

In the previous subsections we have demonstrated that our new model has two important features: it allows us to describe differences in apparent gloss, and it lets us make objects match in apparent gloss. These features should make it much easier to specify surface appearance in graphics rendering applications. To demonstrate how the model might be used, Figure 13 shows a prototype of a perceptually-based color/gloss picker for painted surfaces that could be incorporated into an application. We add color to the model by assuming (as suggested in [Astm89] and [Aida97]), that surface chromaticity and apparent gloss are

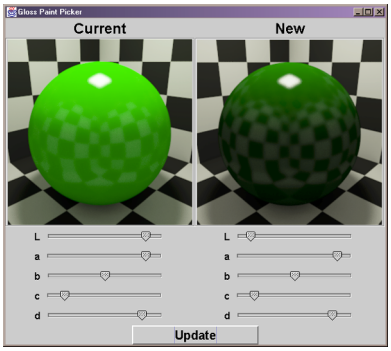


Figure 13†: Prototype of a perceptually-based color/gloss picker for painted surfaces. Surface appearance is specified by three color parameters CIELAB lightness (*L*) and chroma (*a*,*b*) and two gloss parameters (*c*) contrast gloss and (*d*) distinctness-of-image gloss.

relatively independent. For consistency with the lightness parameter *(L)* we use CIELAB chroma (*a*,*b*) to specify color. In the interface, surface appearance is specified by these three color parameters and by the two gloss parameters (c,*d*).

Figure 14 shows an image where this five parameter color/gloss description has been used to match the apparent gloss of the dark red and light blue mugs. Notice that the glossy appearance of the mugs is similar even though they differ significantly in lightness and color. This image suggests that psychophysically-based light reflection model we have developed through our experiments may be usefully applied under more general conditions, however further testing and validation are clearly necessary.

6. CONCLUSIONS/FUTURE WORK

In this paper we've introduced a new light reflection model for image synthesis based on experimental studies of surface gloss perception. To develop the model we conducted two experiments that explored the relationships between the physical parameters used to describe the reflectance properties of glossy surfaces and the perceptual dimensions of glossy appearance in synthetic images. We used the results of these experiments to develop a psychophysically-based light reflection model where the dimensions of the model are perceptually-meaningful and variations along the dimensions are perceptually uniform. We've demonstrated that the model can facilitate the process of describing surface appearance in graphics rendering applications. Although we feel that these results are promising, there is clearly much more work to be done.

First, we want to make clear that strictly speaking, the model we've developed only accurately predicts appearance within the range of glossy paints we studied, under the viewing conditions we used. Although we believe our results will generalize well, if the goal is to develop a comprehensive psychophysically-based light reflection model for image synthesis, many more studies need to be done: 1) to investigate different classes of materials like plastics, metals, and papers (possibly requiring different BRDF models); and 2) to determine how object properties like shape, pattern, texture, and color, and scene properties like illumination quality, spatial proximity, and environmental contrast and texture affect apparent gloss. Additionally, even though in our experiments we found that apparent gloss has two dimensions, we fully expect that for other materials and under other conditions different gloss attributes such as sheen and haze may play a greater role. Finally, we feel that a very important topic for future work is to develop better tone reproduction methods for

Figure 14†: Demonstration that the model can be used effectively in a typical rendering application (3D Studio MAX™). The model was used to make the dark red and light blue mugs match in apparent gloss.

accurately reproducing the appearance of high dynamic range glossy surfaces within the limited ranges of existing display devices.

By using physically-based image synthesis techniques to conduct psychophysical studies of surface appearance, we should be able to make significant progress in these areas. This will allow us to develop models of the goniometric aspects of surface appearance to complement widely used colorimetric models.

7. ACKNOWLEDGMENTS

Thanks to Steve Westin and our anonymous reviewers for their helpful comments on the preparation of this paper. Thanks to Will Alonso, Steve Berman, Reynald Dumont, Bill Feth, Suanne Fu, Clint Kelly, Rich Levy, and Corey Toler for serving as subjects in the experiments. Special thanks to James Cutting for his useful comments throughout this research project and for his help with the experimental design and data analysis.

8. REFERENCES

[Aida97] Aida, T. (1997) Glossiness of colored papers and its application to specular glossiness measuring instruments. Systems and Computers in Japan, 28(1), 1106-1118.

[Astm89] American Society for Testing and Materials. (1989) Standard practice for establishing color and gloss tolerances (Designation: D3134-89). Annual Book of ASTM Standards, 324-329.

[Bill87] Billmeyer, F.W. and O'Donnell, F.X.D. (1987) Visual gloss scaling and multidimensional scaling analysis of painted specimens. Color Res. App. 12(6), 315-326.

[Blak90] Blake, A. and Bulthoff, H. (1990) Does the brain know the physics of specular reflection? Nature, 343, 165-168.

[Blin77] Blinn, J.F. (1977) Models of light reflection for computer synthesized pictures. Computer Graphics (SIGGRAPH 77 Conference Proceedings), 11(4), 192-198.

[Borg97] Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling: Theory and Applications*. Springer: New York.

[Braj94] Braje, W. L. and Knill, D. C. (1994) Apparent surface

shape affects perceived specular reflectance of curved surfaces. Invest. Ophth. Vis. Sci. Suppl. 35(4), 1628.

[Busi97] Busing, F., Commandeur, J., and Heiser, W. (1997) PROXSCAL: a multidimensional scaling program for individual differences scaling with constraints. In W. Bandilla and Faulbaum (Eds.), *Advances in Statistical Software, 6*, Lucius & Lucius: Stuttgart, 67-73.

[Cook81] Cook, R.L. and Torrance, K.E.. (1981) A reflectance model for computer graphics. Computer Graphics (SIGGRAPH 81 Conference Proceedings), 15(4), 187–196.

[Fair98] Fairchild, M.D. (1998) *Color Appearance Models*. Addison-Wesley, Reading, MA.

[He91] He, X.D., Torrance, K.E., Sillion, F.X., and Greenberg, D.P. (1991) A comprehensive physical model for light reflection. Computer Graphics (SIGGRAPH 91 Conference Proceedings), 25(4), 175–186.

[Helm24] Helmholtz, H. von (1924) *Treatise on Physiological Optics* (vol. II), (Trans. by J.P. Southhall). Optical Society of America.

[Heri64] Hering, E. (1964) *Outlines of a Theory of the Light Sense*, (Trans. by L. Hurvich and D. Jameson). Harvard University Press: Cambridge, MA.

[Hunt87] Hunter, R.S. and Harold R.W. (1987) *The Measurement of Appearance* (2nd edition). Wiley, New York.

[Judd37] Judd, D.B. (1937) Gloss and glossiness. Am. Dyest. Rep. 26, 234-235.

[Lafo97] Lafortune, E.P., Foo, S.C., Torrance, K.E., and Greenberg, D.P. (1997) Non-linear approximation of reflectance functions. SIGGRAPH 97 Conference Proceedings, 117-126.

[Ming86] Mingolla, E. and Todd, J.T. (1986) Perception of solid shape from shading. Bio. Cyber. 53(3), 137-151.

[Nish98] Nishida, S. and Shinya, M. (1998) Use of image-based information in judgements of surface reflectance properties. J. Opt. Soc. Am., 15(12), 2951-2965.

[Patt98] Pattanaik. S. Ferwerda, J.A., Fairchild, M.D. and Greenberg, D.P. (1998) A multiscale model of adaptation and spatial vision for realistic image display. SIGGRAPH 98 Conference Proceedings, 287-298.

[Phon75] Phong B.T. (1975) Illumination for computer generated pictures. Comm. ACM 18(6), 311–317.

[Schl93] Schlick, C. (1993) A customizable reflectance model for everyday rendering. Proc. 4th Eurographics Workshop on Rendering, 73–83.

[Stam99] Stamm, J. (1999) Diffraction shaders. SIGGRAPH 99 Conference Proceedings, 101-110.

[Stra90] Strauss, P. S. (1990) A realistic lighting model for computer animators. IEEE Comp. Graph. & Appl. 10(6), 56-64.

[Todd83] Todd, J.T. and Mingolla, E. (1983) Perception of surface curvature and direction of illumination from patterns of shading. J. Exp. Psych.: Hum. Percept. and Perf. 9(4), 583-595.

[Torg60] Torgerson, W.S. (1960) *Theory and Methods of Scaling*. Wiley: New York.

[Tumb99] Tumblin, J., Hodgins J.K., and Guenter, B.K. (1999) Two methods for display of high contrast images. ACM Trans. on Graph., 18(1), 56-94

[Ward92] Ward, G.J. (1992) Measuring and modeling anisotropic reflection. Computer Graphics (SIGGRAPH 92 Conference Proceedings), 26(2), 265–272.

[Ward97] Ward-Larson, G., Rushmeier H., and Piatko, C. (1997) A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Trans. on Vis. and Comp. Graph., 3(4):291-306.

[Wysz82] Wyszecki, G. and Stiles, W.S. (1982) *Color Science: Concepts and Methods, Quantitative Data and Formulae* (2nd ed.), Wiley: New York.

Measuring and Predicting Visual Fidelity

Benjamin Watson
watson@northwestern.edu
Dept. Computer Science
Northwestern University
1890 Maple Ave
Evanston, IL 60657 USA
+1 847 491 3710

Alinda Friedman
alinda@ualberta.ca
Dept. Psychology
University of Alberta
Edmonton, Alberta
Canada T6G 2E9
+1 780 492 2909

Aaron McGaffey
mcgaffey@gpu.srv.ualberta.ca
Dept. Psychology
University of Alberta
Edmonton, Alberta
Canada T6G 2E9
+1 780 492 2909

ABSTRACT

This paper is a study of techniques for measuring and predicting visual fidelity. As visual stimuli we use polygonal models, and vary their fidelity with two different model simplification algorithms. We also group the stimuli into two object types: animals and man made artifacts. We examine three different experimental techniques for measuring these fidelity changes: naming times, ratings, and preferences. All the measures were sensitive to the type of simplification and level of simplification. However, the measures differed from one another in their response to object type. We also examine several automatic techniques for predicting these experimental measures, including techniques based on images and on the models themselves. Automatic measures of fidelity were successful at predicting experimental ratings, less successful at predicting preferences, and largely failures at predicting naming times. We conclude with suggestions for use and improvement of the experimental and automatic measures of visual fidelity.

CR Categories: I.3.7 Three-Dimensional Graphics and Realism, I.3.5 Computational Geometry and Object Modeling

Keywords: visual fidelity, model simplification, image quality, naming time, human vision, perception

1 INTRODUCTION

Polygonal models, images and the techniques for rendering them are growing steadily in complexity, and with this growth comes a need for visual quality control. For interactive computer graphics applications, fidelity of displayed scenes must be adjusted in real time [Lueb97, Lind96, Redd98]. In many other less interactive applications, models must be simplified to contain fewer polygons, while preserving visual appearance [Garl97, Garl99, Hink93, Ross93, Turk92]. Image generators must determine where and if to add additional image detail [Boli98, Rama99]. Finally, image compression algorithms must preserve appearance while reducing image size [Cosm93, Gers92].

How can visual quality and fidelity be measured? This paper focuses on this question. Ultimately, visual quality can only be assessed by human observers. We compare and contrast three different experimental measures of visual quality: naming times [Wats00], ratings [Cosm93, Mart93] and forced choice preferences. However, the interactive demands of many applications requiring control of visual fidelity do not allow experimentation, which has led many researchers to develop automatic measures of visual fidelity [Boli98, Cign98, Daly93, Lubi93, Rama99]. These measures have then been incorporated into image generation and simplification algorithms [Lind00, Vole00]. We evaluate some of these automatic measures by comparing their results to those of the experimental measures studied herein.

In the following sections, we review the rating, preference, and naming time experimental fidelity measures; present a brief survey of existing automatic fidelity measures; and discuss the small body of computer graphics literature that uses experimental fidelity measures or evaluates automatic fidelity measures. We then present our comparisons and evaluations of several experimental and automatic fidelity measures in the context of model simplification.

2 EXPERIMENTAL FIDELITY MEASURES

Ratings and preferences have been widely used in the experimental sciences to obtain relative judgments from human participants. With ratings, observers assign to a stimulus a number with a range and meaning determined by the experimenter. With preferences, observers simply choose the stimulus with more of the experimenter identified quality. Both represent conscious decisions, and so both have proven useful in a wide array of settings, including discomfort ratings in psychiatry, political and popular polling, and the social sciences. With regard to visual fidelity, the experimentally defined meaning or quality of the underlying scales used usually references "quality" or "similarity".

Naming time, the time from the appearance of an object until an observer names it, has a long history of use in cognitive psychology. Existing research has already shown that naming time indexes a number of factors that affect object identification, including the frequency of an object's name in print, the proportion of people who call the object by a particular name and the number of different names in use for it [Vitk95]. Factors of interest to computer graphics researchers include viewpoint [Joli85, Palm81], familiarity [Joli89] and structural similarity [Bart76, Hump95]. In work of particular interest for this study, researchers have shown repeatedly that natural objects take longer to name than manmade artifacts [Hump88]. They hypothesize that natural objects are structurally more similar to one another, requiring more disambiguation than artifacts.

3 AUTOMATIC FIDELITY MEASURES

Although these experimental measures of visual fidelity can be quite effective, time or resources often do not allow their use. In such cases researchers and application builders often turn to automatic measures of visual fidelity.

For level of detail control, researchers estimate error by tracking the deviation of geometry in the image plane [Lueb97, Lind96], and possibly modulating the importance of this error with

ACM SIGGRAPH 2001, 12-17 August 2001, Los Angeles, CA, USA

knowledge of human perception [Redd98]. On the other hand, model simplification researchers have long used three dimensional (3D) measures of distance [Ross93], curvature [Hink93, Turk92], or volume [Lind99] since one typically does not know what part of the model an user may be observing, and these measures are view independent. Lindstrom [Lind00] has measured fidelity for simplification by taking virtual snapshots of the model being simplified from several view points, and then measuring the difference between the snapshots taken before and after the simplification with mean squared error (MSE) (see below). Although it was not used in actual simplification, Cignoni, Rocchini, and Scopigno [Cign98] have offered the Metro tool, which allows users to evaluate the quality of already simplified models with 3D measures of distance and volume.

In the fields of image generation and compression, researchers have focused on view dependent automatic fidelity measures that compare the quality of images. The MSE measure simply finds the mean of the squared pixel by pixel differences between the original and approximate images (often the differences are normalized by the squared value of the pixel in the original image). However, recently several shortcomings of MSE were noted [Giro93] and more complex measures were built based on numerical models of the early stages of the human visual system [Boli98, Daly93, Lubi93]. These were then used to evaluate image compression algorithms and incorporated into image generation algorithms [Boli98, Rama99, Vole00].

4 PREVIOUS FIDELITY EXPERIMENTS

The study of visual fidelity measures and their is are just beginning to make their way into computer graphics research. Rushmeier, Rogowitz and Piatko [Rush00] used a fine grained, one dimensional experimental rating (or scaling) system to evaluate the effects on fidelity of approximations in geometry and texture. They found indications that the ability of texture to hide approximations in geometry depends on the coarseness of the original geometry. Pellachini, Ferwerda and Greenberg [Pell00] used similarity ratings combined with multidimensional scaling and magnitude estimation to derive a perceptually equidistant gloss space.

An initial perceptual evaluation of the automatic fidelity measure designed by Daly was performed by Martens and Myszkowski [Mart93]. They found a high correlation between the Daly measures and observer ratings of texture masked objects. Previously we [Wats00] used naming times as an experimental fidelity measure to examine the effects of model simplification. After duplicating the natural/manmade effect discussed above and confirming that naming times were sensitive to simplification, we turned to an evaluation of several automatic fidelity measures. We found that at severe simplifications, the automatic measure designed by Bolin (BM) [Boli98] was the most reliable, with MSE and maximum 3D distance also fairly reliable. However, at more moderate simplifications none of the automatic measures reliably modeled naming time.

5 EXPERIMENTAL MEASURE STUDY

Our evaluation of the naming time, rating, and preference fidelity measures took the form of an experiment using these measures as the dependent outcomes. This experiment had two goals: to learn about the relative strengths and weaknesses of these measures in their responses to model and image fidelity, and to provide an experimental test bed for our evaluation of automatic fidelity measures in the following section.

5.1 Methods

Here we outline experimental methodology and detail our experimental stimuli. For full detail, please see the appendix.

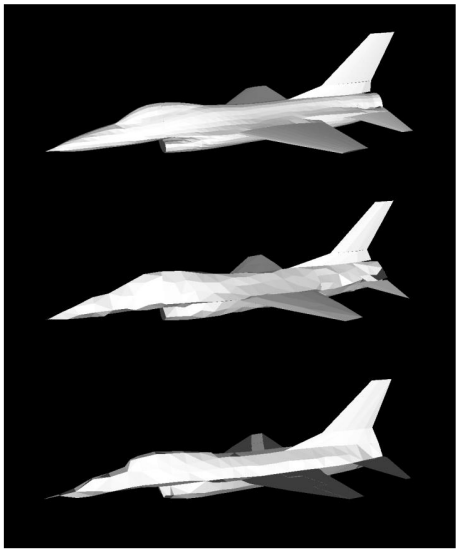


Figure 1: One stimuli from the experimental set. At the top is the original, the middle Qslim 80%, at the bottom Vclust 80%.

Stimuli were created from 36 3D polygonal models (31 in the public domain; 5 from a commercial source). None contained color, texture, material, or vertex normal information. Half the models represented manmade artifacts and the other half were representations of animals. Each of these models was simplified using two simplification algorithms (Vclust [Ross93] and Qslim [Garl97]) resulting in two levels of simplification each. We chose these algorithms because they are widely used and according to prevailing opinion, produce models differing widely in visual fidelity. Thus this experiment had three independent variables: *simplification type* (Vclust vs. Qslim), *simplification level* (three levels including unsimplified), and *object type* (animals vs. artifacts). These were varied within participants.

Models were simplified in two stages. First, Qslim was used to simplify all models to the number of polygons contained in the smallest model in the set (3700 ±50). We refer to these as the "standards" (0% simplification), and label a member of this set *s*. Second, the standards were simplified using Qslim and Vclust by removing 50% and 80% of the original 3700 polygons. We refer to members of the resulting four model sets as *q5*, *q8*, *v5* and *v8*. There were thus five examples of each of the 36 objects, for a total of 180 stimuli.

Each stimulus image was uniformly scaled to 591 pixels in width and displayed in the center of the screen. The rating and preference task stimuli each consisted of two exemplars of a single object model that were scaled to 400 pixels in width and displayed side-by-side, each centered within a 512(w) x 768(h) pixel space. Figure 1 shows a stimulus simplified at 80% by Qslim and Vclust.

Naming task. Participants were asked to name each object as quickly and accurately as they could. They were told that some pictures would be simplified representations and were shown printed examples.

Rating task. All four simplified exemplars of an object were rated against the standard ((*s*, *q5*), (*s*, *q8*), (*s*, *v5*) and (*s*, *v8*)). Each participant rated all 36 objects once at each simplification type and level. Stimuli were presented in a random order.

Participants were told that on each trial their task was to rate the likeness of the picture on the right against the standard picture on the left, using a 7-point scale. They had four practice trials.

Preference task. Exemplars of both simplification types were compared at the same simplification level (e.g. (*q5*, *v5*) or (*q8*,

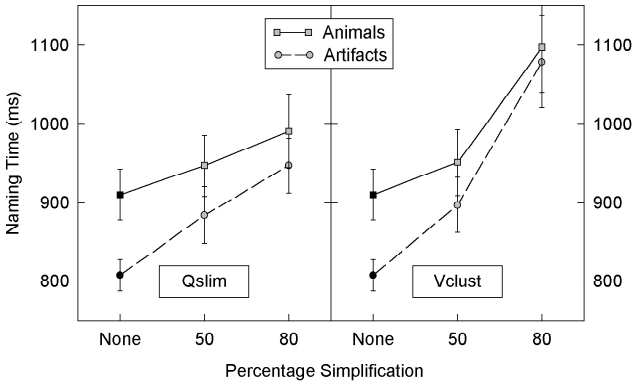


Figure 2: Naming times as a function of simplification type, simplification level, and object type.

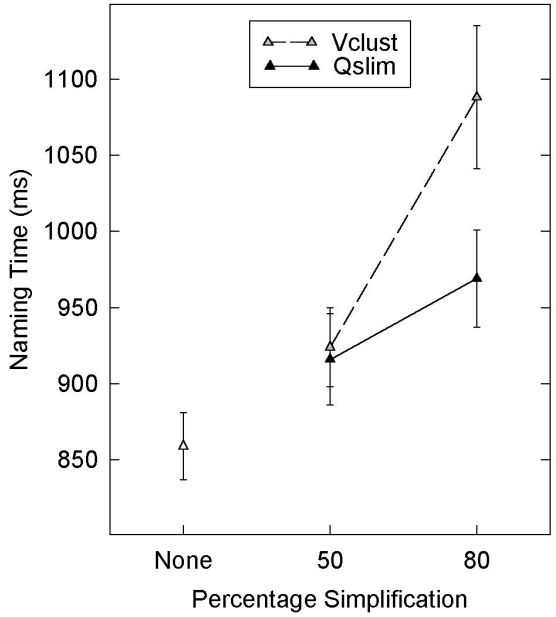


Figure 3: Naming times averaged across object type as a function of simplification type and level.

Variable	Avg By	ANOVA
object type	participants	F(1,35) = 10.24
simp level	participants	F(1,35) = 13.59
simp level	objects	F(1,33) = 13.80

Table 1: 2 way analysis on naming times averaged over simp type. All effects p<.05.

Variable	Avg By	ANOVA
simp type	participants	F(1,35) = 5.29
simp level	participants	F(1,35) = 13.59
simp level	objects	F(1,33) = 13.80
stype x slevel	participants	F(1,35) = 4.70

Table 2: 3 way analysis on naming times without standard models. All effects p<.05.

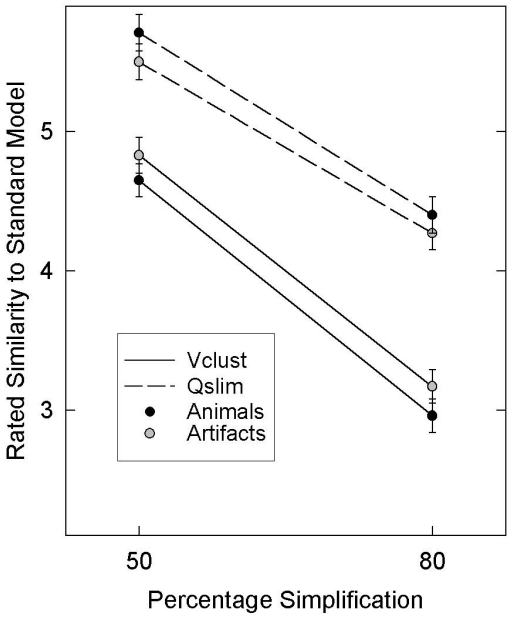


Figure 4: Ratings by simplification type, simplification level, and object type.

v8)). There were 36 objects, each with two simplification types and two simplification levels, for a total of 72 comparisons. The left-right position of the Qslim (Vclust) example was distributed evenly throughout the trials. Participants had four practice trials. Participants were asked to choose which picture in the set was a better example of each object.

5.2 Results

5.2.1 Naming Times

Figure 2 shows the mean naming times as a function of object type, simplification algorithm, and simplification level. It can be seen that all three factors affected performance: animals were named more slowly than artifacts, naming times were longer with increasing simplification, and naming times were longer with Vclust (see Table 1). There were no interactions between object type and simp level. Reassuringly, this replicates the main trends of our earlier study [Wats00]. In the only interaction, the effect of simp type varied with simp level (see Table 2). Figure 3 shows the data averaged over type of model. Clearly the Vclust algorithm was much more devastating to naming times than Qslim at the higher levels of simplification.

We corroborated these observations with analyses of variance (ANOVAs) on the naming time means averaged two ways. For details on these analyses please see the appendix.

5.2.2 Ratings and Preferences

Rating results are shown in Table 3, averaged two ways. Figure 4 shows the average similarity ratings as a function of object type, simplification type, and simplification level. Participants were sensitive to simplification level and rated the 50% simplified objects closer to the "ideal" than the 80% simplified objects (5.2 versus 3.7). Second, they also clearly thought that the Qslim-simplified objects were closer to the ideal than were the Vclust-simplified objects (5.0 versus 3.9). Third, simplification type interacted with simplification level; similar to the naming time data, there was less of a difference between the algorithms when objects had been simplified to 50% (5.6 versus 4.7 for Qslim and Vclust, respectively) than when they had been simplified to 80% (4.3 versus 3.1).

Variable	Avg By	ANOVA
simp type	participants	F(1,35) = 243.56
simp type	objects	F(1,33) = 100.97
simp level	participants	F(1,35) = 264.29
simp level	objects	F(1,33) = 388.86
stype x slevel	participants	F(1,35) = 32.23
stype x slevel	objects	F(1,33) = 11.75
stype x otype	participants	F(1,35) = 29.51

Table 3: 3 way statistical analysis on ratings. All effects $p<.05$.

Variable	Avg By	ANOVA
object type	participants	F(1,35) = 79.68
object type	objects	F(1,33) = 5.25
simp level	objects	F(1,35) = 18.20

Table 4: 2 way analysis on preferences. All effects $p<.05$.

In all of these respects, ratings results were similar to naming time results. However, ratings and naming times differed in their response to object type. Ratings did not respond simply to object type, and in fact there was an interaction between object type and simp type: the animal models were rated closer to the standard when they had been simplified using the Qslim algorithm (5.1 versus 4.9 for animals versus artifacts, respectively), but the artifacts were rated as being closer to the standard when they had been simplified using Vclust (3.8 versus 4.0 for animals versus objects).

In the preference results, there were main effects for both object type and simplification level (see Table 4 and Figure 5). Essentially, the preference for Qslim-simplified stimuli was greater for the animal models than for the artifact models (90.1% versus 77.0%), and it was greater for 80% objects than for the 50% objects (86.5% versus 80.6%).

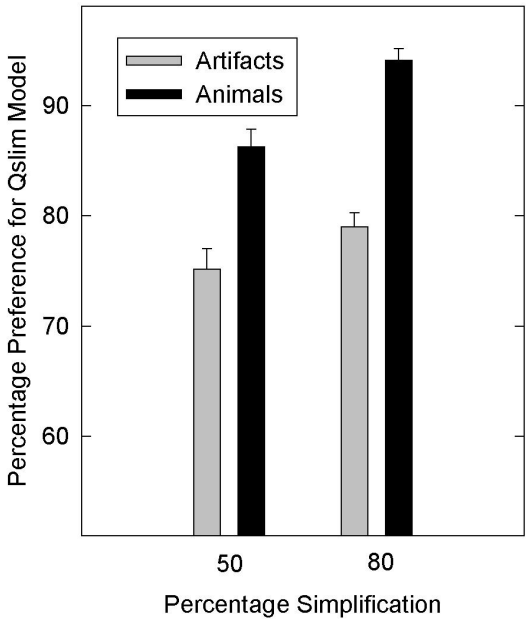


Figure 5: Preferences for Qslim by simp level and object type.

6 AUTOMATIC MEASURE STUDY

We now turn our attention to automatic measures of visual fidelity, and their ability to predict experimental measures of fidelity provided by human observers. Such automatic measures, if effective, could be quite useful in evaluating the effectiveness of various algorithms -- and if efficient enough, might even be incorporated into the algorithms themselves. We examine three tools for measuring fidelity: an implementation of the image comparison algorithm described by Bolin and Meyers [Boli98] (BM), mean squared image error (MSE), and the Metro tool from Cignoni, Rocchini and Scopigno [Cign98].

6.1 Methods

Both BM and MSE accept as input an ideal image and an approximate image, and return summary measures of the difference between these images. MSE returns a single number as its estimate. BM returns a difference image, with the value at each image location estimating the ability of viewers to perceive the local difference between the images. Since we require a single value summarizing image fidelity, we use the average of all the local values contained in the difference image. For both MSE and BM, the images used were the same images used in experimentation.

Metro accepts as input two similar 3D polygonal models, and as a result is not sensitive to viewpoint. It returns rough estimates of the difference in volume between the two models. It also samples the two surfaces at multiple points, and measures the distance from each point on the first (pivot) model to the surface of the other model. It returns three summaries of these distance measures. The first is the mean of these distances, obtained by normalizing the sum of the distances with the surface area of the pivot model. The second simply squares each of the summed distances before normalization. The third is the maximum of the measured distances. Metro returns its summaries in model coordinates, as well as in coordinates normalized by the diagonal of the pivot model bounding box and the diameter of the smallest sphere that encloses the pivot model.

Our evaluation of Metro's fidelity measures includes the volume difference (MetroVol), as well as each of the mean (MetroMn), mean squared (MetroMSE) and maximum (MetroMax) summaries. All three distance summaries were normalized by the diagonal of the pivot model bounding box. For the maximum summary, we used a Metro option that returned the Hausdorf distance, that is, that found the maximum of two-sided distance measurements both from the first model to the second, and the second to the first.

We found four sets of automatic fidelity measures for each of the 36 models in the experimental set. If, for a given model, *s* is the standard, *q5* and *q8* are versions of *s* simplified by Qslim 50% and 80% respectively, and *v5* and *v8* are versions of *s* simplified by vertex clustering 50% and 80% respectively, then we found sets of fidelity measures for each of the following four model pairs: (*s*,*q5*), (*s*,*q8*), (*s*,*v5*) and (*s*,*v8*). For each of these pairs, the set of fidelity measures included BM, MSE, MetroMn, MetroMSE, MetroMax and MetroVol. For Metro, we always used *s* as the pivot.

To each fidelity measure in each model pair set we compared experimental measures. For naming times, we used the time for the non-standard model in the pair (e.g. for (*s*,*q5*), we used the time it took to name *q5*, or *name*(*q5*)). For ratings, we used the rating of the non-standard model in comparison to the standard (e.g. *rate*(*s*,*q5*)).

Automatic measures compared to experimental preference measures took a special form. Typically persons will compare

Automatic Measure	Naming Times						Ratings					
	All Models		Animals		Artifacts		All Models		Animals		Artifacts	
	Qslim	Vclust	Qslim	Vclust	Qslim	Vclust	Qslim	Vclust	Qslim	Vclust	Qslim	Vclust
BM	-0.07	***0.30***	-0.07	0.21	-0.03	***0.41***	***-0.62***	***-0.60***	***-0.43***	***-0.54***	***-0.72***	***-0.67***
MSE	0.07	***0.31***	0.02	0.14	0.18	***0.48***	***-0.67***	***-0.71***	***-0.68***	***-0.71***	***-0.74***	***-0.77***
MetroMn	0.03	***0.31***	0.00	0.24	0.10	***0.38***	***-0.65***	***-0.77***	***-0.77***	***-0.78***	***-0.66***	***-0.77***
MetroMSE	-0.04	***0.25***	-0.20	0.27	0.06	0.22	***-0.46***	***-0.55***	-0.21	***-0.53***	***-0.56***	***-0.60***
MetroMax	-0.05	***0.27***	-0.16	0.26	0.04	0.28	***-0.60***	***-0.73***	***-0.52***	***-0.75***	***-0.66***	***-0.72***
MetroVol	0.19	0.14	-0.07	0.08	***0.41***	0.19	-0.21	-0.13	***-0.58***	***-0.34***	0.00	-0.04

Table 5: Correlations of naming times and ratings to automatic fidelity measures.

Fidelity Measures	Simp Type	Simp Level	SType x SLevel	SType x OType	Three Way
Naming	*5.29*	13.80	*4.70*		
Rating	100.97	388.86	11.75	*29.51*	
BM	11.73	108.08	6.31		
MSE	78.31	100.12	37.55		
MetroMn	56.48	192.71	32.27	8.02	8.18
MetroMSE	23.58	135.08	14.72	8.67	7.03
MetroMax		32.86			
MetroVol		6.68	7.82		

Table 6: Significant ANOVAs for naming times, ratings and automatic fidelity measures. Italics represent participant analyses.

two stimuli for quality by judging which of the two is closer to a visually presented or completely cognitive ideal. Therefore the automatic measures we compared to experimental preferences were constructed from the previous measured pairings, and took the form $p5 = (meas(s,q5) - meas(s,v5))$ and $p8 = (meas(s,q8) - meas(s,v8))$, where *meas* is one of the six measures we evaluated. *p5* and *p8* predict preference among the 50% and 80% simplified models, respectively, with a positive result predicting a preference for Vclust, a negative result for Qslim. We also compared naming times and ratings to *p5* and *p8*. These comparisons used the differences in naming times and ratings across simplification type (e.g. $(name(q5) - name(v5))$ and $(rate(s,q5) - rate(s,v5))$).

6.2 Results

Table 5 shows automatic fidelity measure correlations to naming times and ratings used to judge quality with (at least implicit) reference to an ideal. Each correlation measure reflects comparisons for both simplification levels within a simplification type. Where correlations are presented in bold, the associated automatic measure accounts for a marginally significant ($p < 0.1$) proportion of the variation in the experimental measure. Where they are also italicized, the automatic measure accounts for a significant ($p < 0.5$) proportion of experimental variation.

All automatic measures with the exception of MetroVol were very successful predictors of quality as judged by ratings. Correlations were quite high, with ANOVAs indicating that a statistically significant portion of experimental variance was accounted for. Note that correlations are consistently negative, since low automatically measured error correlates consistently with high experimental ratings. Correlations are slightly worse for animals as opposed to artifacts, and for Qslim as opposed to Vclust.

The automatic measures were much less successful at predicting quality as judged by naming times. Correlations were in this case generally positive, since low automatically measured error correlates to short naming times. The most successful automatic fidelity measures were BM, MSE and MetroMn. The striking failures here are the consistently low correlations for Qslim, and to a lesser degree, the lower correlations for animals, echoing the same trends in the ratings correlations.

We performed in-depth analyses of the automatic measures by treating their results as dependent variables in ANOVAs much like those used for the experimental measures, with simplification type, simplification level, and object type as independent variables. We present these results in Table 6. Table values in italics represent F values from analyses averaged across objects for each participant, rather than averaged across participant for each object. We graph the means for two of the better measures, BM and MetroMn, by objects in Figures 6 and 7, and show for comparison naming time and ratings graphs averaged over participants for each object. All measures, whether automatic or experimental, were significantly affected by simplification level. Most measures were significantly affected by simplification type and the interaction of simplification type and level. The effect of object type, however, differed greatly across the measures, whether experimental or automatic.

Table 7 shows automatic fidelity measure correlations to preferences and naming time and rating differences used to judge which of two stimuli has superior quality. Each correlation measure again reflects comparisons for both simplification levels. Where correlations are presented in bold, the automatic measure accounts for a marginally significant ($p < 0.1$) proportion of the variation in the experimental measure, where they are also italicized, the automatic measure accounts for a significant ($p < 0.5$) proportion of experimental variation. The automatic measure differences are negative if Qslim has less error, while rating differences and preferences are positive if Qslim is rated more highly or preferred, giving negative correlations. Since naming time differences are negative if Qslim produces the more recognizable model, correlations to it are largely positive. In general, the automatic measures correlated quite well to experimental preferences, less well to differences in ratings, and quite poorly to differences in naming times. Again correlations were worse for animals than for artifacts.

7 DISCUSSION

In this section we review our experimental and automated findings, make some recommendations on the use of fidelity measures, and provide some suggestions as to how automatic fidelity measures and the applications that use them might be improved.

7.1 Limitations

However, before we do so, we should note the limitations of our studies. We begin with a consideration of our stimuli. First, we have limited ourselves to the study of one almost optimal view of each object. Second, this study has focused on approximations made in model geometry, rather than in the illumination model, model texture, or in attributes such as color or per-vertex normal vectors. In addition, we have focused on recognition of objects presented in isolation, rather than in more natural scenes

Table 7: Correlations of preferences, naming time differences, and rating differences to automatic fidelity measures.

Automatic Measure	Naming Diffs			Rating Diffs			Preferences		
	All	Anims	Artifs	All	Anims	Artifs	All	Anims	Artifs
BM	0.21	0.23	0.23	***-0.36***	-0.23	***-0.38***	***-0.37***	-0.27	***-0.35***
MSE	***0.26***	0.15	***0.37***	***-0.44***	-0.25	***-0.54***	***-0.33***	***-0.42***	-0.27
MetroMn	0.18	0.20	0.21	***-0.42***	-0.21	***-0.47***	***-0.42***	***-0.41***	-0.32
MetroMSE	0.04	0.17	-0.03	-0.21	-0.25	-0.15	***-0.27***	***-0.42***	-0.16
MetroMax	0.13	0.19	0.14	***-0.41***	-0.16	***-0.45***	***-0.43***	***-0.40***	***-0.34***
MetroVol	-0.06	0.17	-0.17	-0.05	0.19	-0.15	-0.04	0.16	-0.11

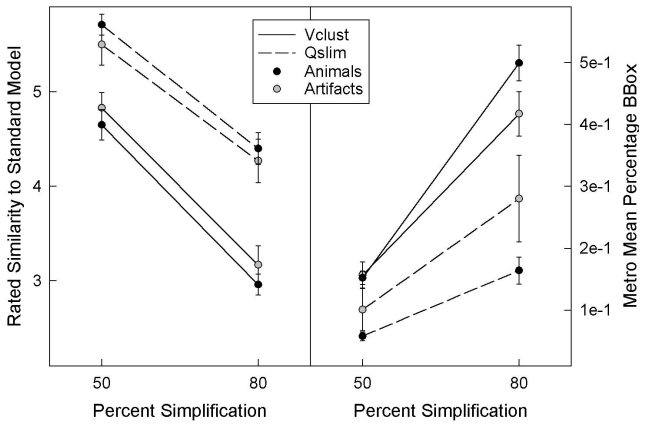


Figure 6: MetroMn response to simplification type, simplification level, and object type, vs. ratings by object.

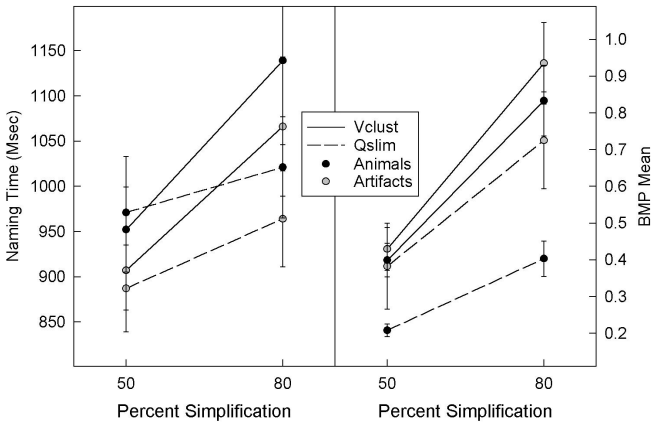


Figure7: BM response to simplification type, simplification level, and object type, vs. naming time by object.

containing several objects in their context. Finally, models are often used in interactive applications with viewpoint and model motion, while all of the stimuli presented and studied here were static. Removing these limitations in further studies would certainly increase the generality of our results. At the same time, such changes would increase variation in those results, and stiffen the challenge posed to the various automatic measures of visual fidelity, which would have to model a more complex experimental response. For example, the introduction of viewpoint and non-geometric experimental factors would certainly reduce the effectiveness of the Metro measures, at least in their current form.

In order to limit the scope of our experimentation, we also made choices in the use of our automatic measures. In particular, we chose BM as a quickly executing representative of those measures that model the early stages of human vision. But BM is a specialization of other slower measures ([Daly93, Lubi93]) that might be more effective (though BM was proved very effective in these results). BM and related measures were also developed for stimuli more complex (and more challenging) than those used here. Difference image summarizations other than the averaging used here might also increase measure effectiveness.

7.2 Confirmations

As we have noted above, our naming time results were largely in agreement with the results we obtained earlier in [Wats00]. We also found that simp level has the effect one would intuitively expect on the rating and preference measures. In agreement with prevailing opinion, Qslim was by all measures a more effective simplification tool than Vclust. Many have conjectured that simplification techniques show their mettle at low polygon counts. These results are in agreement with that hypothesis, with a simp level and simp type interaction showing that there is little difference between Qslim and Vclust at 50% simplification, a large difference at 80%.

7.3 Surprises

Ratings and preferences indicated that though Qslim is generally a better simplifier than Vclust, it simplifies animals most effectively. This may indicate that a specialization of Qslim for more regularly curved (or planar) surfaces is possible. On the other hand, Vclust is more effective when simplifying artifacts – a hint that Vclust's regular sampling approach is most effective when used with models typical of CAD/CAM applications, which contain many coplanar polygons and regularly curved surfaces.

Naming times did not respond to object type with the same complexity as ratings and preferences, instead, they were uniformly longer for animals. This is most likely a clue to the different natures of these experimental measures: while naming times probe subconscious perception from the low through the higher cognitive levels, ratings and preferences seem to sample very low level processes, avoiding the natural/manmade effect. (However, ratings and preferences are notoriously vulnerable to higher level, conscious cognitive qualities assigned to the axis of comparison).

In line with these differences in the experimental results, the automatic measures were poor predictors of naming times, but excellent predictors of experimental ratings, preferences, and to a lesser extent, differences in ratings. BM, MSE, and MetroMn were particular success stories in this respect. Obviously the differing experimental responses to object type played a role in these correlative trends.

However, correlations were low even within simplification and object type, where only simplification level was varying. We see two possible reasons for this. First and most simply, naming times are very variable, much more so than ratings and preferences. Obtaining good correlations to them within simplification and object type may require the use in experiments of larger model stimuli libraries and more participants. Second and more provocative, although it is certain that in general, increasing simplification increases naming times; we noticed that

for several models, increasing simplification *reduced* naming times. We will call this the *distillation effect*. There are precedents for this in the psychology literature [Ryan56, Edel99]. The basic notion is that by removing detail that allows fine grained identification of an object, the speed of coarse grained, categorizing identification is improved. The distillation effect seems to occur particularly often for animal models simplified with Qslim, and may explain some of the negative correlations (again, within simplification and object type) in Table 5. Automatic measures do not model this effect, reducing correlations.

7.4 Implications

For simplification. Our results indicate that simplification effectiveness varies by all experimental measures as a function of object type. This suggests the possibility of simplification algorithms that specialize in, or adapt to, different types of models. As simplification researchers continue their work, they should pay particular attention to the quality of their models at low polygon counts. Our results also suggest that mean distance is a more important heuristic for simplification than maximum distance. The rating and preference measures are well modeled by the automatic measures reviewed here, which should prove useful when comparing algorithms, or even during the process of simplification itself.

For use of experimental measures. All of the measures vary in the degree of explicit visual comparison they require of the viewer. With preferences, this comparison is very explicit. With ratings, the comparison is to some (at least cognitively imagined) visual ideal. Comparison may be involved in the process probed by naming times, but it is certainly not an explicit comparison between two visual images. It may be telling that this least comparative of experimental measures was also most poorly modeled by the automatic measures.

What is fidelity? Is it visual similarity to the original? Or is it successful communication of the original concept? The notion of fidelity most relevant in the current application should indicate the experimental measure that is most appropriate. During the processes of simplification, image compression and generation, the goal is typically one of appearance preservation in the face of each of a long series of minor alterations. Preferences, ratings, and their correlating automatic measures are probably the most appropriate indices for these applications. However, when making cross algorithm comparisons, the compared images or models are the result of very different, very long series of these alterations, and the appropriate notion of fidelity is less clear.

It is intriguing to note that in almost all computer graphics applications, users *never* make an explicit visual comparison. At most, they compare a currently displayed example with a previously displayed one. In highly interactive applications, this comparison, if it indeed occurs, is certainly cursory at best. In these sorts of settings, the naming time measure might be most appropriate, and the distillation effect, if it indeed exists, most effectively exploited. It is also intriguing to imagine a non-photorealistic pursuit of the distillation effect in its extreme.

For automatic measures. Many of these measures can be used for purely numerical ends, ensuring for example that a given approximation does not deviate from the original by more than some constant error. We do not consider such applications here.

Our results indicate that MetroVol is a poor predictor of visual fidelity and quality as indexed by any of our experimental measures, at least at the levels of geometric simplification (3700 polygons and below) examined here. BM, MSE, and MetroMn were excellent predictors of fidelity as measured by ratings and to a lesser extent preferences and rating differences. Unfortunately, we found no fully reliable predictors of the conceptual sort of fidelity measured by naming times. For now, the best automatic predictor of naming times and their differences is MSE, with BM and MetroMn coming very close behind. Given the poor correlations of all three of these measures with Qslim, these naming time predictors must be used with extreme skepticism, if at all.

For future work. These results raise many intriguing questions. First, do they generalize? We are currently investigating how well these results hold across differ viewpoints, and would like to examine the effects of both background and interactive motion. The element of comparison embodied both by these measures and typical graphics applications clearly needs further research, as does the hypothesized distillation effect. Obviously our automatic measures must improve their ability to model naming times. This will require understanding and modeling object type effects. In the long run, research into the object type and distillation effects may lead to new simplification algorithms.

8 CONCLUSION

This paper described our research into the experimental and automatic measurement of visual fidelity. Measuring visual fidelity is fast becoming crucial in the fields of model simplification, level of detail control, image generation, and image compression. In our study, we manipulated fidelity by applying two different model simplification algorithms to 36 polygonal models, divided into models of animals and manmade artifacts, producing approximating models at two different polygon counts. We examined the visual fidelity of these models with three different experimental measures: naming times, ratings, and preferences. All the measures were sensitive to the type of simplification algorithm used and the amount of simplification, however they responded differently to model type. We then analyzed model visual fidelity with several automatic measures of visual fidelity. These automatic measures proved to be good predictors of ratings and preferences, but only mediocre predictors of naming times.

9 ACKNOWLEDGEMENTS

Oscar Meruvia wrote indispensable 3D viewing software. Josh Anon helped with simplifications and quality predictors. We thank Greg Turk for his models and geometry filters. Peter Lindstrom participated with a thought provoking correspondence, and by assisting in finding relevant code. Oleg Veryuvka and Lisa Streit provided useful implementations of simple image quality metrics. Stanford University was the source of the bunny model. This research was supported by NSERC grants to the first two authors.

10 APPENDIX: EXPERIMENTAL DETAILS

Experimental methodology. Participants performed the naming, rating, and forced-choice preference tasks on the same set of items during a single session. All participants completed the naming task first because seeing a stimulus once reduces its subsequent naming time [Joli85]. Similarly, all participants performed the rating task prior to the preference task because it was possible that performing the preference task first could contaminate rating judgments by increasing the subjective distance between the less preferred object and the standard.

The virtual field of view used in forming stimuli was always 40 degrees and the virtual eye point was always at a distance that was twice the length of the bounding box. Views were generally directed towards the mean of a model's vertices, but 14 models required centering corrections because vertex distributions were not uniform. Each model was interactively rotated so that it was displayed in a canonical 3/4 view that revealed a reasonable level of detail across the models [Palm81]. Each model was illuminated with one white (RGB=[1,1,1]) light located at the eye point. All models were assigned the same white color and flat shaded, and displayed on a black background.

The images were displayed on a 17-inch Microscan CRT, with participants sitting approximately 0.7 m from the display. Participants performed the naming task by speaking into a hand-held microphone. Responses for the rating and preference tasks were entered on the

computer keyboard. Thirty-six undergraduate volunteers from the University of Alberta pool participated in the experiment.

For the naming task, stimuli were organized into six groups of six stimuli each. There were three groups for each simplification algorithm, and within algorithms, one group for each level of simplification (0%, 50%, and 80%). Each group contained three animal models and three artifact models. Stimuli were cycled through the groups such that across participants, each stimulus appeared once in each of the six experimental conditions (2 simp type x 3 simp level).

Each participant saw all 36 models only once; 12 were standards, 12 were simplified using Qslim, and 12 were simplified using Vclust; 6 of each of the simplified models were seen at 50% simplification and 6 were seen at 80% simplification. There were eight practice trials. On each trial, the experimenter pressed the space bar, a fixation cross appeared for 750 ms, the picture appeared on the screen, the participant named the picture, and the picture disappeared as soon as a name was said. Naming times were recorded from stimulus onset to the participant's response.

For the rating task, on each trial the participant pressed the space bar, and after a delay of 250 ms the standard and comparison pictures appeared on the screen and disappeared as soon as a rating was entered.

For the preference task, subjects pressed the "A" and "K" keys to choose the left and right stimuli, respectively. The participant pressed the space bar; after a delay of 250 ms the pictures appeared and then disappeared as soon as a preference was entered.

The models: ant, bear, bicycle, blender, bunny, camera, car, chair, cow, dinosaur, dog, dolphin, dump truck, elephant, fighter jet, fish, helicopter, horse, kangaroo, lion, microscope, motorcycle, piano, pig, plane, raven, rhino, sandal, shark, ship, skateboard, snail, sofa, spider, tank, tomgun.

Analysis of experimental measure results. Three kinds of trials were excluded from naming time analyses. First, we excluded naming times measured during spoiled trials (e.g., trials in which participants failed to trigger the microphone with their first vocalization – 4.6% of all trials). Second, we excluded naming times from trials in which a participant's response was an error (e.g., calling a picture of a sandal a "rocket" – 0.3% of all trials). Finally, we computed the overall mean of the remaining naming times and excluded trials that were more than 3 standard deviations longer than this average. These outliers comprised only 1.5% of the remaining trials.

For naming times and ratings, examining the relationship of object type to simp level required averaging over simp type for a two way analysis, because unsimplified objects were necessarily the same for both the Qslim and Vclust. Additional three way analyses were performed by excluding the unsimplified objects.

Most analyses used two ANOVAs, one averaged over objects (the participant analysis) and one averaged over participants (the object analysis). For the participant ANOVA on the preference data, we counted the frequency of times that each participant chose the Qslim-simplified model in each of the four object type and simplification level conditions, and converted the results to percentages. For the item analysis, we counted the frequency of participants who chose the Qslim model for each of the objects in each of the conditions.

11 REFERENCES

[Bart76] Bartram, D.J. (1976). Levels of coding in picture-picture comparison tasks. *Memory and Cognition, 4*, 593-602.

[Boli98] Bolin, M. & Meyer, G. (1998). A perceptually based adaptive sampling algorithm. Proc. of SIGGRAPH 98. In *Computer Graphics* Proceedings, Annual Conference Series, 1998, ACM SIGGRAPH, 299-309.

[Cign98] Cignoni, P., Rocchini, C. & Scopigno, R. (1998). Metro: measuring error on simplified surfaces. *Computer Graphics Forum, 17*, 2, 167-174. Available at: http://vcg.iei.pi.cnr.it/metro.html.

[Cosm93] Cosman, P., Gray, R. & Olshen, R. (1993). Evaluating quality of compressed medical images: SNR, subjective rating and diagnostic accuracy. *Proc. IEEE, 82*, 6, 919-932.

[Daly93] Daly, S. (1993). The visible differences predictor: an algorithm for the assessment of image fidelity. In Watson, A. B. (ed.). *Digital Images and Human Vision*. MIT Press, Cambridge, MA, 179-206.

[Edel99] Edelman, S. (1999). Representation and recognition in vision. Cambridge, MA: MIT Press.

[Garl97] Garland, M. & Heckbert, P. (1997). Surface simplification using quadric error metrics. Proc. SIGGRAPH 97. In *Computer Graphics* Proceedings, Annual Conference Series, ACM SIGGRAPH, 209-216. Available at: http://www.cs.cmu.edu/~garland/quadrics/qslim.html

[Garl99] Garland, M. (1999). Multiresolution modeling: survey and future opportunities. Eurographics '99, State of the Art Report (STAR).

[Gers92] Gersho, A. & Gray, R. (1992). Vector quantization and signal compression. Kluwer Academic Publishers, Norwell, MA.

[Giro93] Girod, B. (1993). What's wrong with mean-squared error? In Watson, A. B. (ed.). *Digital Images and Human Vision*. MIT Press, Cambridge, MA. 207-220.

[Hink93] Hinker, P. & Hansen, C. (1993). Geometric optimization. *Proc. IEEE Visualization '93*, 189-195.

[Hump88] Humphreys, G. W., Riddoch, M. J., & Quinlin, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology, 5*, 67-103.

[Hump95] Humphreys, G. W., Lamote, C., & Lloyd-Jones, T. J. (1995). An interactive activation approach to object processing: Effects of structural similarity, name frequency, and task in normality and pathology. *Memory, 3*, 535-586.

[Joli85] Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition, 13*, 289-303.

[Joli89] Jolicoeur, P., & Milliken, B. (1989). Identification of disoriented objects: Effects of context of prior presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 200-210.

[Lueb97] Luebke, D. & Erikson, C. (1997). View dependent simplification of arbitrary polygonal environments. Proc. of SIGGRAPH 97. In *Computer Graphics* Proceedings, Annual Conference Series, ACM SIGGRAPH, 199-208.

[Lind00] Lindstrom, P. & Turk, G. (2000). Image-driven simplification. *ACM Trans. Graphics, 19*, 3, 204-241.

[Lind96] Lindstrom, P., Koller, D., Ribarsky, W., Hodges, L., Faust, N., & Turner, G. (1996). Proc. of SIGGRAPH 96. In *Computer Graphics* Proceedings, Annual Conference Series, ACM SIGGRAPH, 109-118.

[Lind99] Lindstrom, P. & Turk, G. (1999). Evaluation of memoryless simplification. *IEEE Trans. Visualization and Computer Graphics*, 5, 2, 98-115.

[Lubi93] Lubin, J. (1993). A visual discrimination model for imaging system design and evaluation. In Peli, E. (ed.). *Vision Models for Target Detection and Recognition*, World Scientific, New Jersey, 245-283.

[Mart93] Martens, W. & Myszkowski, K. (1993). Psycho-physical validation of the visible differences predictor for global illumination applications. *IEEE Visualization '93*, Late Breaking Topics, 49-52. Also available at: http://wwwsv1.u-aizu.ac.jp/labs/csel/vdp/.

[Palm81] Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddelay (Eds.), *Attention & Performance IX*, Hillsdale, NJ : Erlbaum, 135-151.

[Pell00] Pellachini, F., Ferward, J. & Greenberg, D. (2000). Toward a psychophysically-based light reflection model for image synthesis. Proc. of SIGGRAPH 00. In *Computer Graphics* Proceedings, Annual Conference Series, ACM SIGGRAPH, 55-64.

[Rama99] Ramasubramanian, M., Pattanaik, S. & Greenberg, D. (1999). A perceptually based physical error metric for realistic image synthesis. Proc. of SIGGRAPH 99. In *Computer Graphics* Proceedings, Annual Conference Series, ACM SIGGRAPH, 73-82.

[Redd98] Reddy, M. (1998). Specification and evaluation of level of detail selection criteria. *Virtual Reality: Research, Development and Application, 3*, 2, 132-143.

[Ross93] Rossignac, J. & Borrel, P. (1993). Multi resolution 3D approximations for rendering complex scenes. In Falcidieno, B. & Kunii, T. (eds.), *Geometric Modeling in Computer Graphics*. Springer Verlag, 455-465.

[Rush00] Rushmeier, H., Rogowitz, B. & Piatko, C. (2000). Perceptual issues in substituting texture for geometry. *Human Vision and Electronic Imaging V*, Bernice E. Rogowitz, Thrasyvoulos, N. Pappas, Editors, Proc. of SPIE Vol. 3959, 372-383.

[Ryan56] Ryan, T., & Schwartz, C. (1956). Speed of perception as a function of mode of representation. *American J. Psychology, 69*, 60-69.

[Turk92] Turk, G. (1992). Re-tiling polygonal surfaces. Proc. of SIGGRAPH 92. In *Computer Graphics, 26*, 2 (July), ACM SIGGRAPH, 55-64.

[Vitk95] Vitkovitch, M., & Tyrell, L. (1995). Sources of name disagreement in object naming. *Quarterly Journal of Experimental Psychology, 48A*, 822-848.

[Vole00] Volevich, V., Myszkowski, K., Khodulev, A. & Kopylov, E. (2000). Using the visual differences predictor to improve performance of progressive global illumination computation. *ACM Trans. Graphics, 19*, 2, 122-161.

[Wats00] Watson, B., Friedman, A. & McGaffey, A. (2000). Using naming time to evaluate quality predictors for model simplification. *Proc. ACM Computer Human Interaction* (CHI), 113-120.