Evaluating Human Motion

Jessica Hodgins
Robotics Institute and Computer Science
jkh@cs.cmu.edu
www.cs.cmu.edu/~jkh

How to evaluate?

- Looks good...
- Side-by-side comparisons (~1995)
- Perceptual Studies (~1998)
  - Hodgins, O'Brien and Tumblin
  - Harrison, Rensink, van de Panne
- fMRI (recently)
- Behavioral studies
  - Immersion (used successfully in VR)
  - Enactment
  - Interference

Looks Good?

Ron Fedkiw, Robert Bridson, and John Anderson

Looks Good?

Ron Fedkiw, Eran Guendelman, Andrew Selle and Frank Losasso

Looks Good?

Wes Fesler Kicking a Football, 1934
Dr. Harold Edgerton

Looks Good?

Movie
Side-by-side Comparison

Side-by-side Comparison

Side-by-side Comparison

Side-by-side Comparison

Or force plate data?

Perceptual Studies
Which motion is more natural?
Which motion is wrong?
Can you detect a change?

Reitsma and Pollard, SIGGRAPH 2003

Harrison, Rensink, van de Pannen SIGGRAPH 2004

Hypotheses:
Simple representations $$\rightarrow$$ fine distinctions
Complex, “accurate” representations $$\rightarrow$$ fine distinctions
Equally fine distinctions independent of model

The sensitivity measure, $d'$, is defined as:
$$d' = \frac{\log H/(1 - H) - \log F/(1 - F)}{\frac{1}{2}}$$
where $H$ is the fraction of pairs in a set that were different and which the subject labeled correctly, and $F$ is the fraction of pairs in a section that were the same and which the subject labeled incorrectly.

Fig. 8. Sensitivity scores by experiment averaged over subject groups. Skilled subjects are those who achieved a sensitivity score of $d' \geq 1.0$ on either the polygonal or the stick figure portion of the test. Note that sensitivity scores are consistently higher with the polygonal model.
Conclusions

Full model allowed finer distinctions for all three of our tests.

Different models allow different distinctions to be made. The graphics community should have standards for results to be compared.

Strengths? Weakness?

First study that looked at this question. Confirmed several times since in similar but different experiments.

None of the running motions looked natural? Did we span the space of variations? Only tested two models (both fairly crude). Subdivision into skilled and not skilled subjects (post-hoc)

Follow-on Studies?

Camera motion? Clothing, Hair motion? Breathing, facial expressions?

Perceptual Studies

Conclusions

Numbers showing change in limb length that should not be perceivable:
- 3% with full attention
- 20% when not the focus of attention
Sensitivity to growing higher than to shrinking (why?)
Slower changes are less noticeable
Changes are less noticeable during fast motions

Strengths? Weakness?

Distractor task is a good experimental design. Explored space where one or both segments changed, fast/slow velocities, duration of change.

Study somewhat distant from real question—if you don’t see it on the line drawing does that really mean that you won’t see it on the cute little kid?

Is perceivable or not the right question? With the little kid, the question we really care about is whether it looks natural or not?

Follow-on Studies?

Is change in limb length of benefit even if it is noticeable? Makes the kid look like he is trying harder?

Sub-threshold effects? Higher LOD in soccer players increases rating of skill.

Response to Model

Experimental paradigm

Collaborators: Thierry Chaminade, Mitsuo Kawato, ATR

Response to Model

Motion capture

Keyframed motion
Behavioral Experiments

Analysis

Do models influence perceived naturalness of motion?
→ test effect of model on proportion of biological responses.

Results: biological responses

Dot models cause increased biological response rates when compared to others (all pairwise comparisons \( p < 0.05 \)). Ellipses and Robots different from Alien, Clown and Humans (all pairwise comparisons \( p < 0.05 \) except Human vs Robot, \( p = 0.11 \)); Within groups comparisons are not significant.

Experiments

Relationship between the model (rendering style) and the perception of motion. More complex/anthropomorphic models are less likely to be perceived as being biological motion

Reinforces common wisdom in animation community – motion must be fully rendered to be assessed
Now repeating experiments with fMRI. Preliminary results indicate that model has an effect on STS activity

fMRI

• What is measured?
  – Blood flow to areas of the brain
  – About 2 seconds after event
  – Scan completed every ~2 seconds

• Data processing
  – Align brain scan with “typical” brain
  – Look for differences in activation between regions for various stimuli
  – Running the machine costs $600/hour
    • few subjects

• Conclusions
  – X area lights up when we show human motion but not when we show similar frequency non-human motion
  – X area has known to be associated with y so it’s interesting that it also turns up in our study of z

Perceptual experiments tell us what we can perceive—but not necessarily what makes a compelling character.

We really want to know how the audience will respond to a character—maybe behavioral metrics get closer to that?
  enactment
  interference
  imitation

Behavioral Studies
Behavioral Studies—Immersion in VR

Use heart rate, galvanic skin response to measure immersion. Test frame rate, lag, walking vs. flying, and other factors.
http://www.cs.unc.edu/~eve/walk_exp/

Method: Use enactment as a metric
Extensively studied behavior
Classic experiment:
- hear, see, or perform ~50 phrases like “carry the suitcase”
- delay or distracter task
- tested with recall or recognition
- measure percentage correct and reaction time

<table>
<thead>
<tr>
<th>13%</th>
<th>27%</th>
<th>46%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal</td>
<td>Experimenter performed</td>
<td>Subject performed</td>
</tr>
</tbody>
</table>

Data from an experiment in the literature

Method: one verb, multiple objects
30 video + audio phrases
30 audio-only phrases
Pull the handle
Pull the cord
Pull the rope
15 minute delay
Free verbal recall test
Complete recall
Partial recall

Enactment (video) improves complete recall by 94%

Enactment (video) improves partial recall by 133%

What to test?
Animated character pantomime

What to test?
Animated character with object

What to test?
Object
What to test?

- Animated character pantomime
- Rendering styles
- Degraded motion
- Different characters

How might we fail?

- Hard to create good animations of these phrases
- Might not be a fine enough discriminator
- Only have n% to work with
  - More phrases
  - Recall rather than old/new
  - Longer delay

What else might work?

- Imitation behaviors
  - Yawning
  - Walking in step
- Interference behaviors
  - Performing one arm motion while watching another
  - Harder for human arm motion
  - Not for automation robot?

Are any of these really measuring what we care about in animated characters?