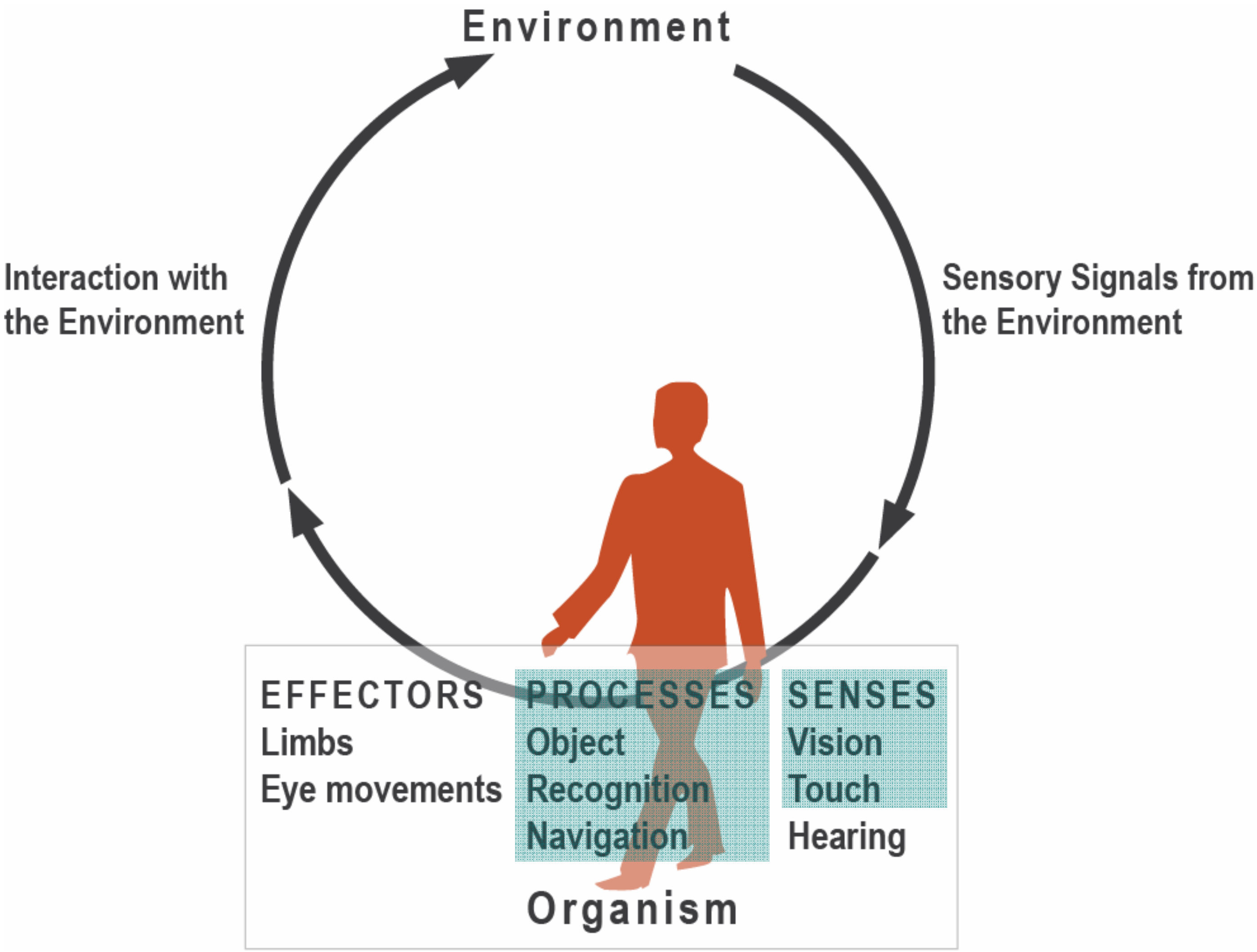
Perception and Action in Virtual environments – *an image-based approach*

Heinrich H. Bülthoff

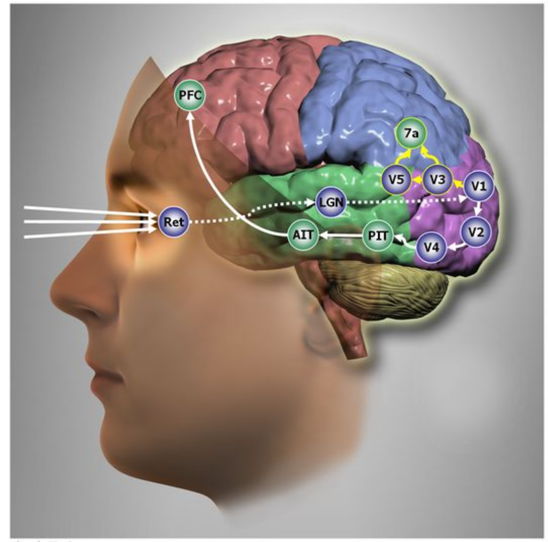
Cognitive & Computational Psychophysics
Max Planck Institute for Biological Cybernetics

The Human: a complex cybernetic system

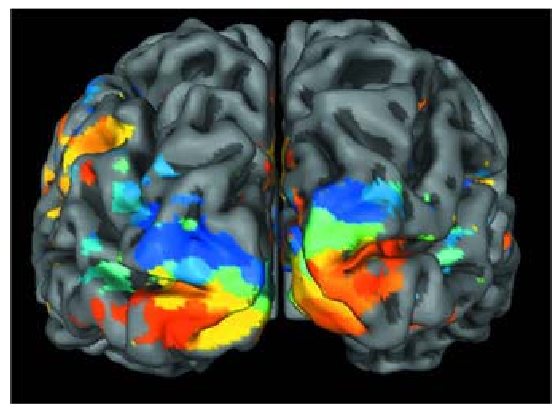


Vision: the view onto the world

- When lights hits the retina, vision starts – a process that takes up ~80% of the brain‘s resources
- Experimental and theoretical understanding of this perceptual process has come from many disciplines:
  - neurophysiology
  - psychophysics
  - cognitive psychology
  - computer vision
  - information theory
  - brain imaging

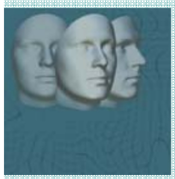


The visual object recognition pathway
(c) Christian Wallraven



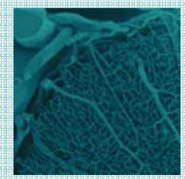
Visual areas in the brain (fMRI data)
from http://www.grp.hwz.uni-muenchen.de/

MPI for Biological Cybernetics



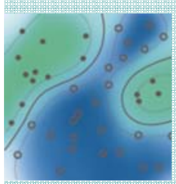
Psychophysics Dept.
Bülthoff, since 1993
algorithmic level

- human psychophysics and cognitive system technology
- perception-action experiments in virtual reality



Neurophysiology Dept.
Logothetis, since 1997
hardware level

- system physiology with primates
- multi-electrode recordings and functional imaging (fMRI)



Empirical Inference Dept.
Schölkopf, since 2001
theory level

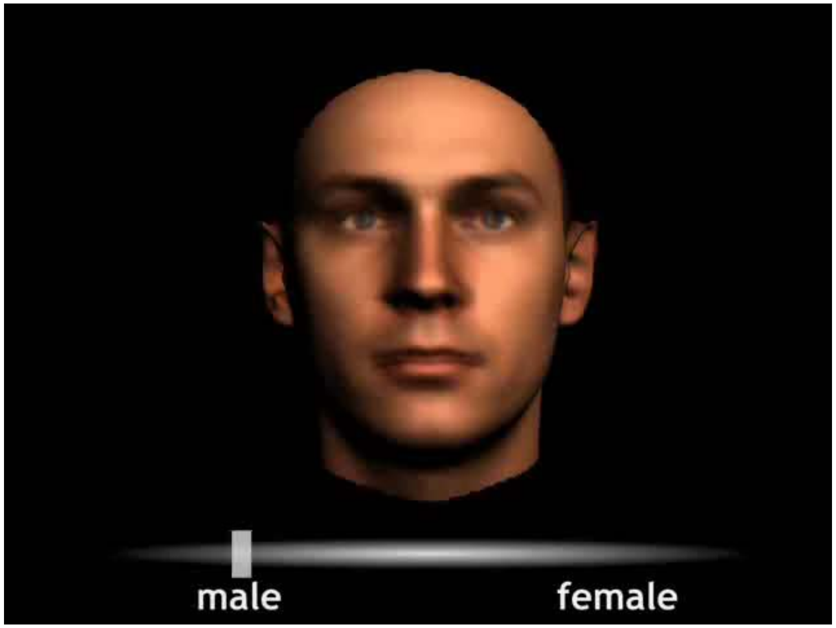
- statistical learning theory
- applications to data from vision, robotics and neurophysiology

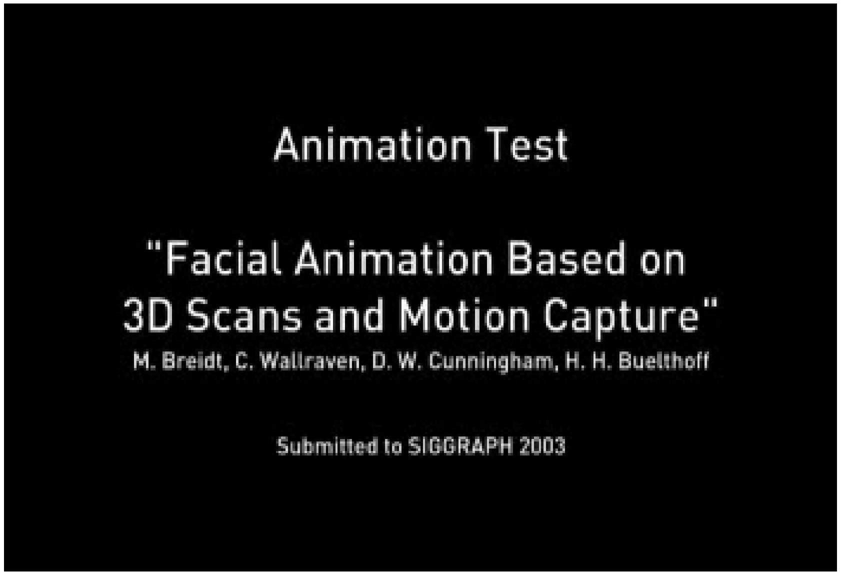
Magnetic Resonance Dept.
Ugurbil, since 2005
brain imaging

- new methods for high-field MR
- new contrast agents
- evaluation of biocompatibility, delivery, and localization

Research Paradigm

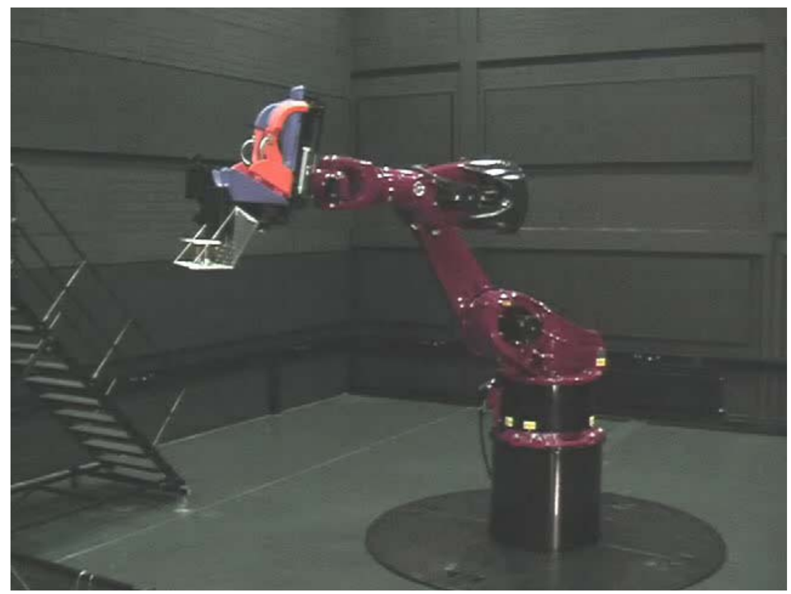
- Study perception and action with stimuli as close as possible to the real world, using
  - **Computer Graphics**
    to generate natural but well controlled stimuli of objects and scenes
  - MPI Face Database (open access)
    - faces.kyb.tuebingen.mpg.de
    - vdb.kyb.tuebingen.mpg.de
  - Database of High-Dynamic-Range Images (soon to come)
  - **Virtual Reality**
    to study perception and action in a closed loop





Research Paradigm

- Study perception and action with stimuli as close as possible to the real world, using
  - **Computer Graphics** to generate natural but well controlled stimuli of objects and scenes
  - **Virtual Reality**
    - www.cyberneum.de
    - motion simulators
    - haptic simulators
    - walking simulators
    - immersive environments
    - panoramic projections
    - EU-projects: JAST, BACS, CyberWalk, Immersense, Wayfinding



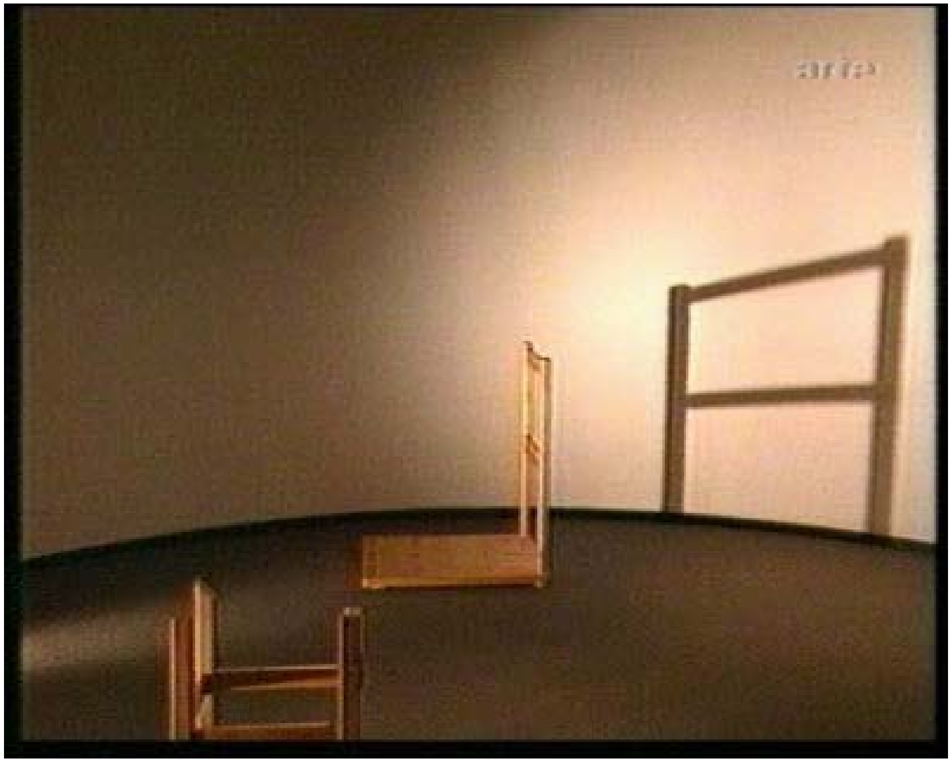
Our working hypotheses

- the brain does not need to build a full 3D representation of the world from the 2D images on our retina
- the brain adopts a more direct perception approach using an image-based strategy for perception & action (neo-Gibsonian)
- a simple demonstrations shows the importance of views

A simple demonstration (Beuchet Chair)

- the brain assumes that parts in close proximity belong together
- this assumption can be wrong in rare cases
- an “accidental view” leads to the wrong 3D interpretation

Beuchet Chair



- 2D images usually are sufficient for the correct interpretation of a scene
- only from a single viewpoint (accidental view) the *proximity assumption* leads to the wrong conclusion

Dwarfs and Giants

- high level interpretation (size, shadows) is ignored
- occlusions can solve the perceptual puzzle

Outline of the talk

- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based temporal information for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

Object Recognition Models

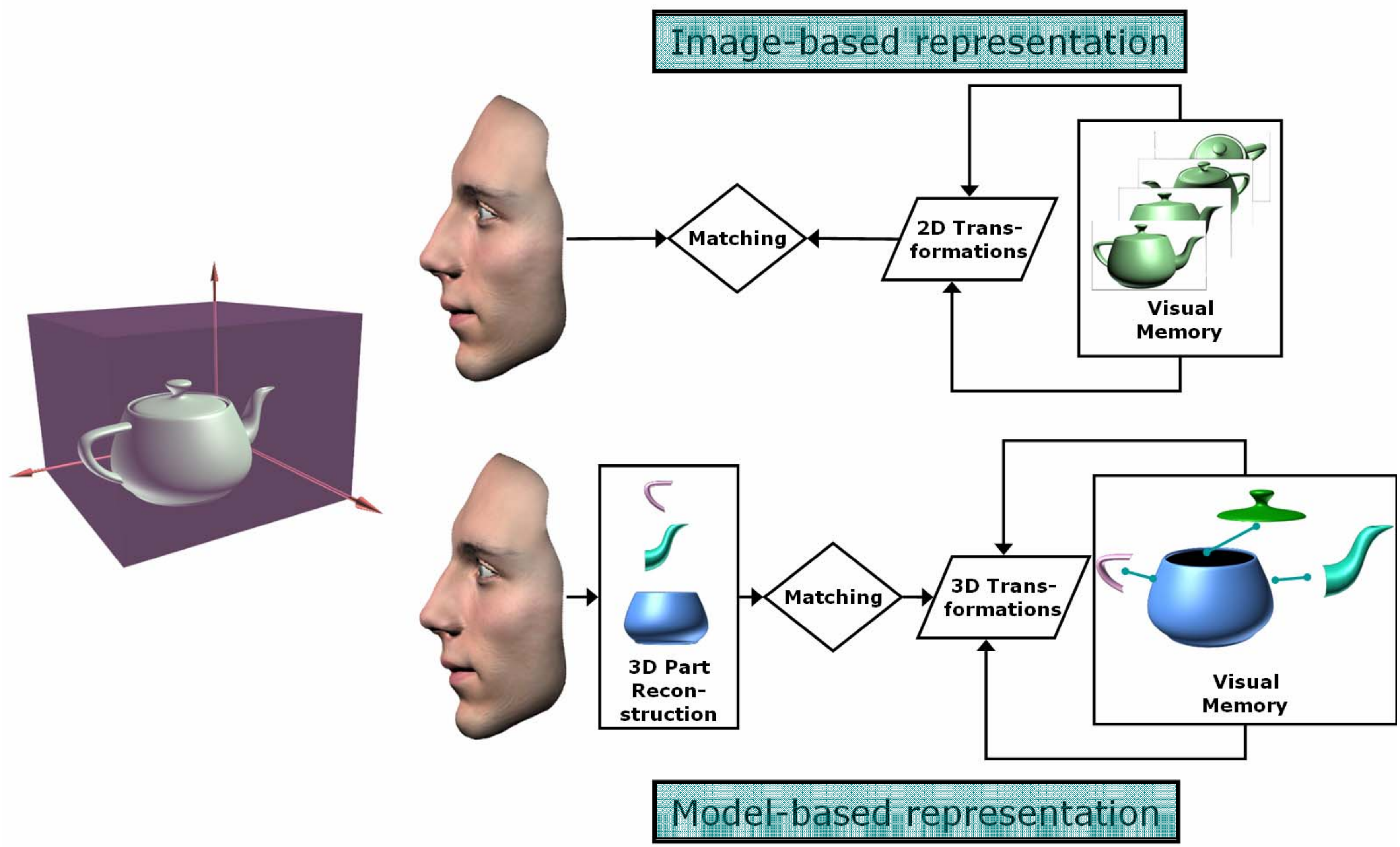
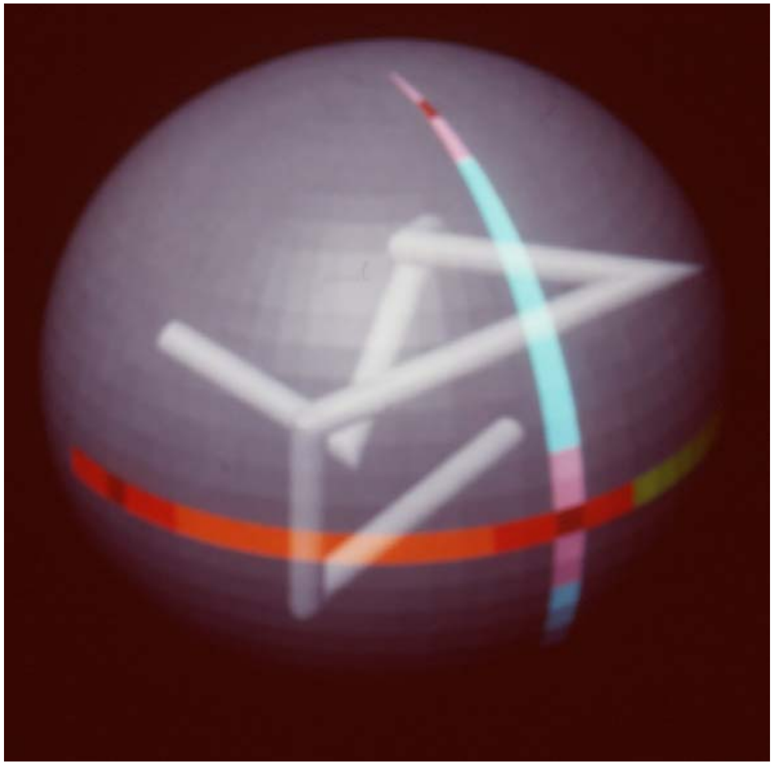


Image-based Object Recognition

Bülthoff & Edelman, *PNAS* (1992)

- Recognition better for views spanned by the training views than for orthogonal axis.
- "This is difficult to reconcile with any theory except the image combination approach."
  S. Ullman in *High-Level Vision (1996)*



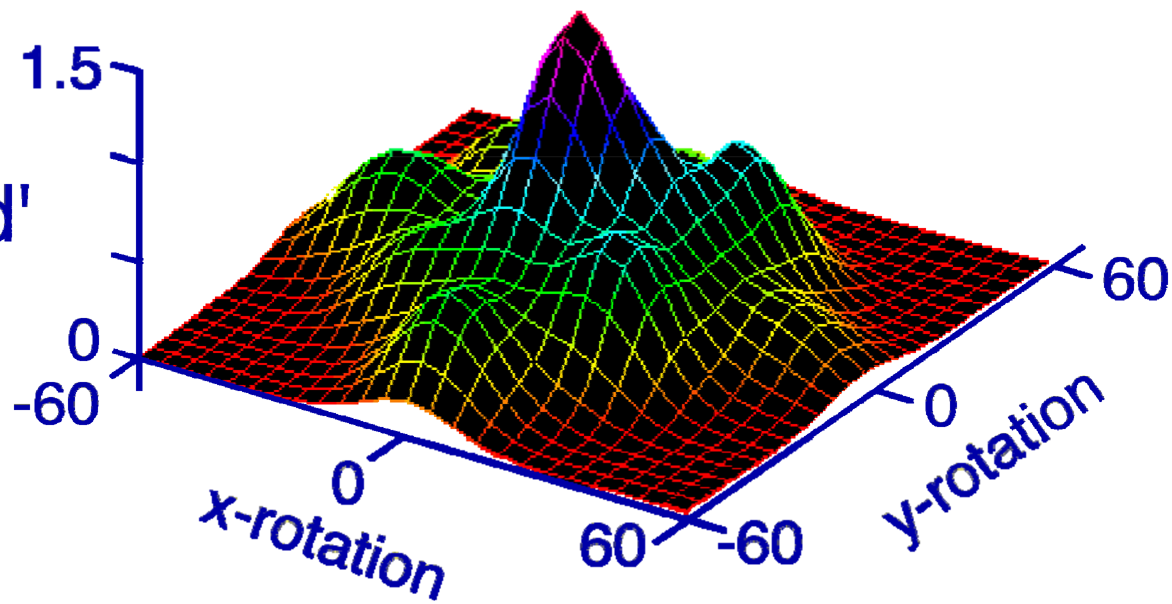
Generalization Fields

Bricolo, MIT PhD Thesis (1996)

stereo

only one test per target

distractor= target + noise



10 subjects

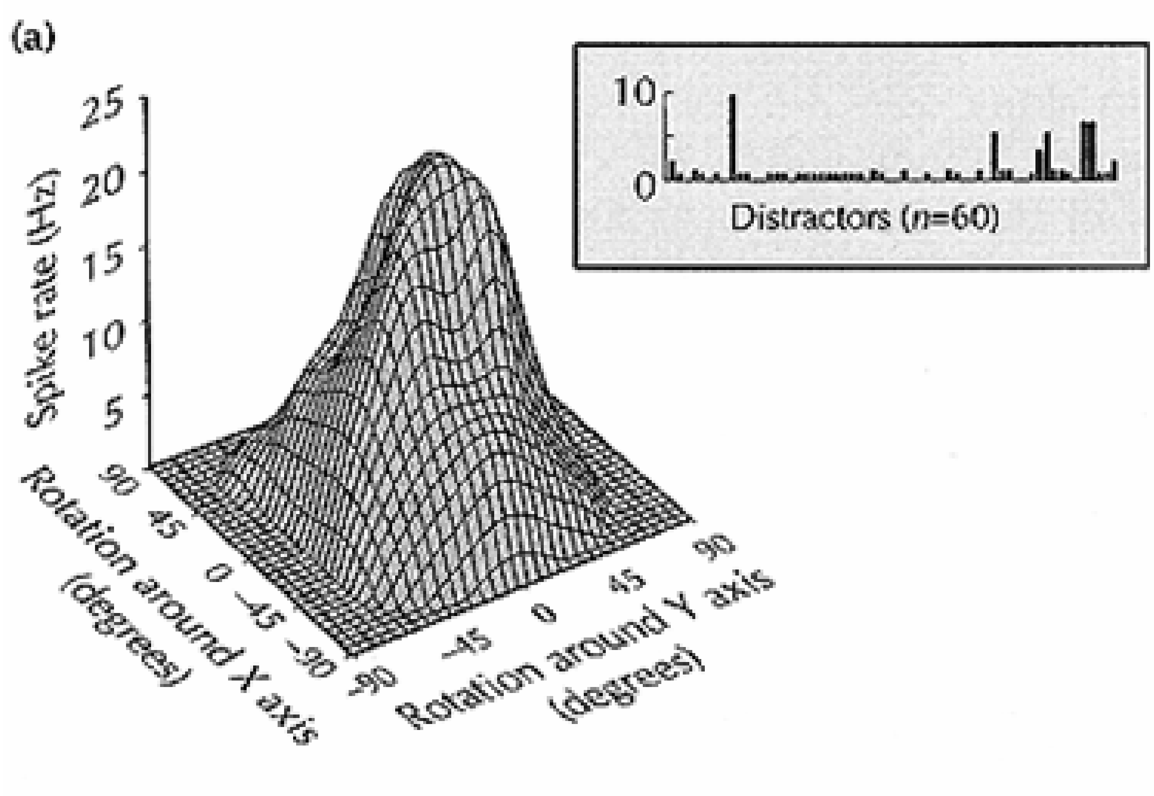
25 viewpoints

150 target objects

23% distractor noise

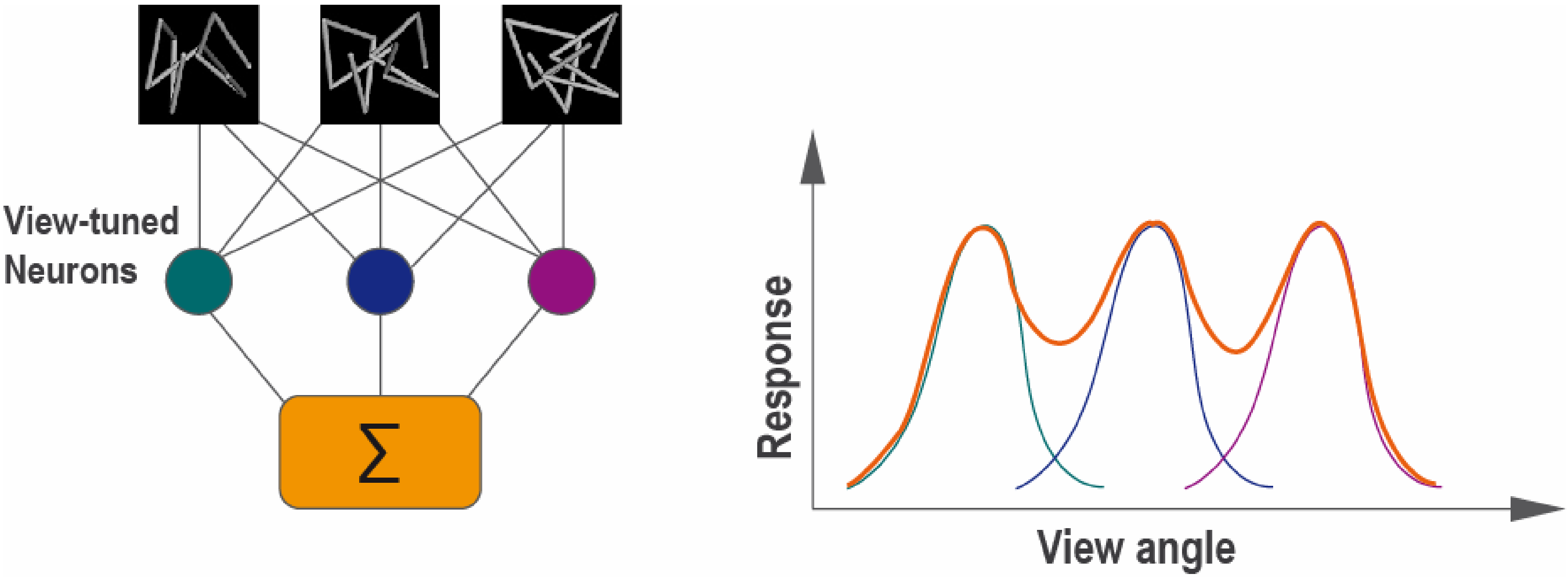
View-specific Recognition Neurons

Logothetis, Pauls, Bülthoff, Poggio, *Current Biology* (1994), (1995)



A View-interpolation Network

Poggio, Edelman, *Nature* (1990)



Demonstration for Image-based Recognition

More evidence in:
Object Recognition and Man, Monkey and Machine
Tarr & Bülthoff (eds.) MIT Press (1999)

What‘s in this picture?

A bit more information

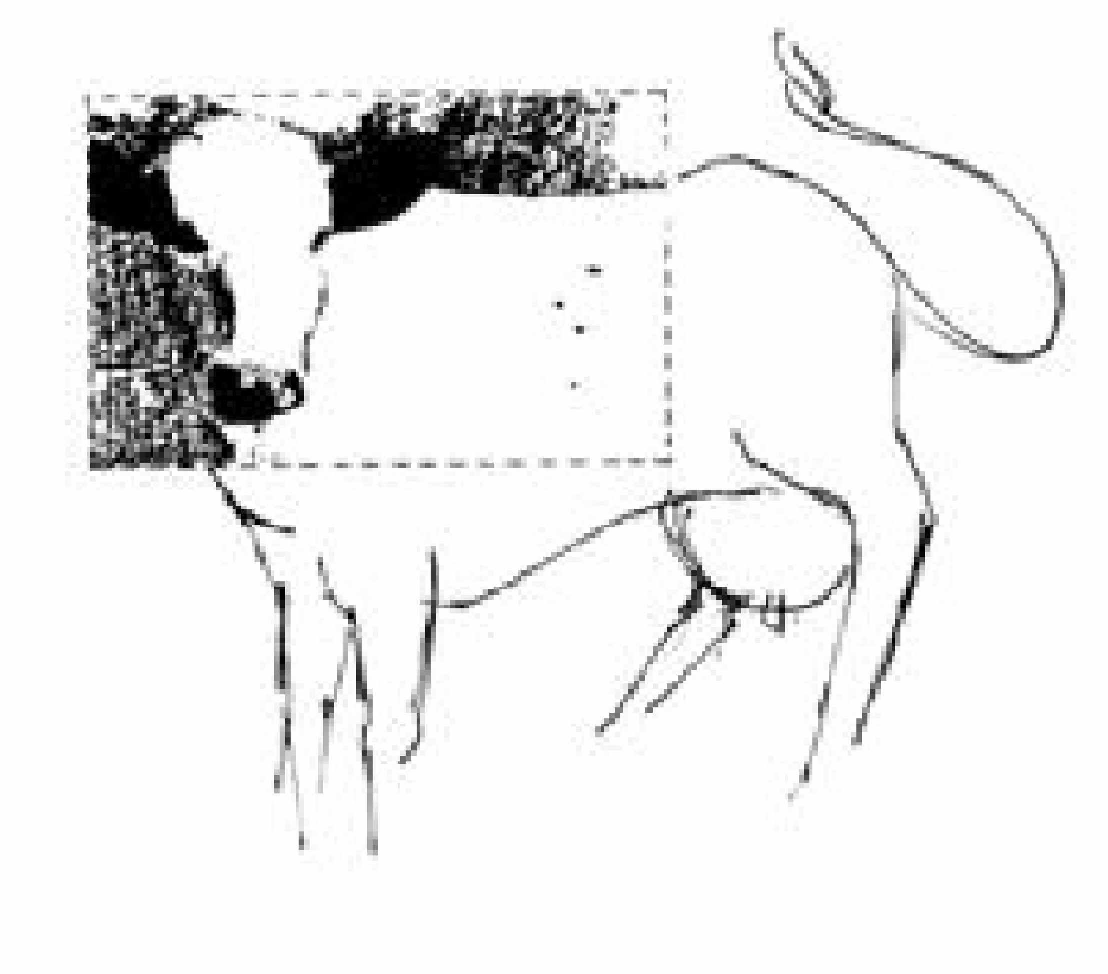


Image-based Recognition

What do you see now?

- Recognition is not bottom-up
- You need to have seen it before
- Recognition is matching to image-like representations
- Recognition memory for pictures
  - Roger Shepard (1967): 700 pictures
    even after a week still over 90% correct recognition
  - Standing, Conezio and Haber (1970) 2500 pictures
  - Standing (1973) 10 000 pictures

Dalmatian Dog

Dalmatian Dog

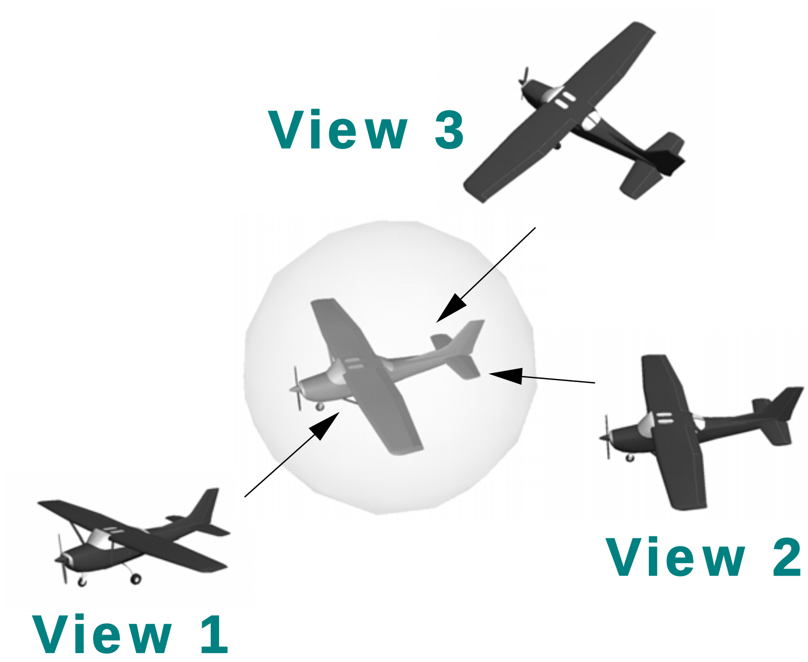
Dalmatian Dog

Where is the Dog?

- P. Sinha & T. Adelson Perception 26, 667, 1997
- Some people have too much top-down processing…
  they hallucinate the dog

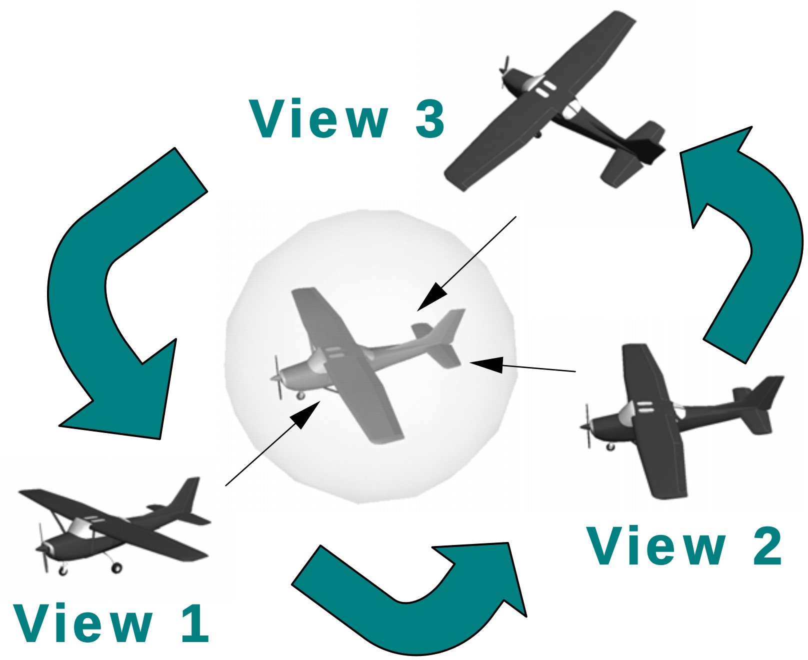
The binding problem

- Physical similarity can account for recognition with small viewpoint changes (image-based recognition)
- How does the brain know that different views of an object belong to the same object?



The binding problem

- Physical similarity can account for recognition with small viewpoint changes (image-based recognition)
- How does the brain know that different views of an object belong together?
- Solution: **temporal association** of contiguous views



Outline of the talk

- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based **temporal information** for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

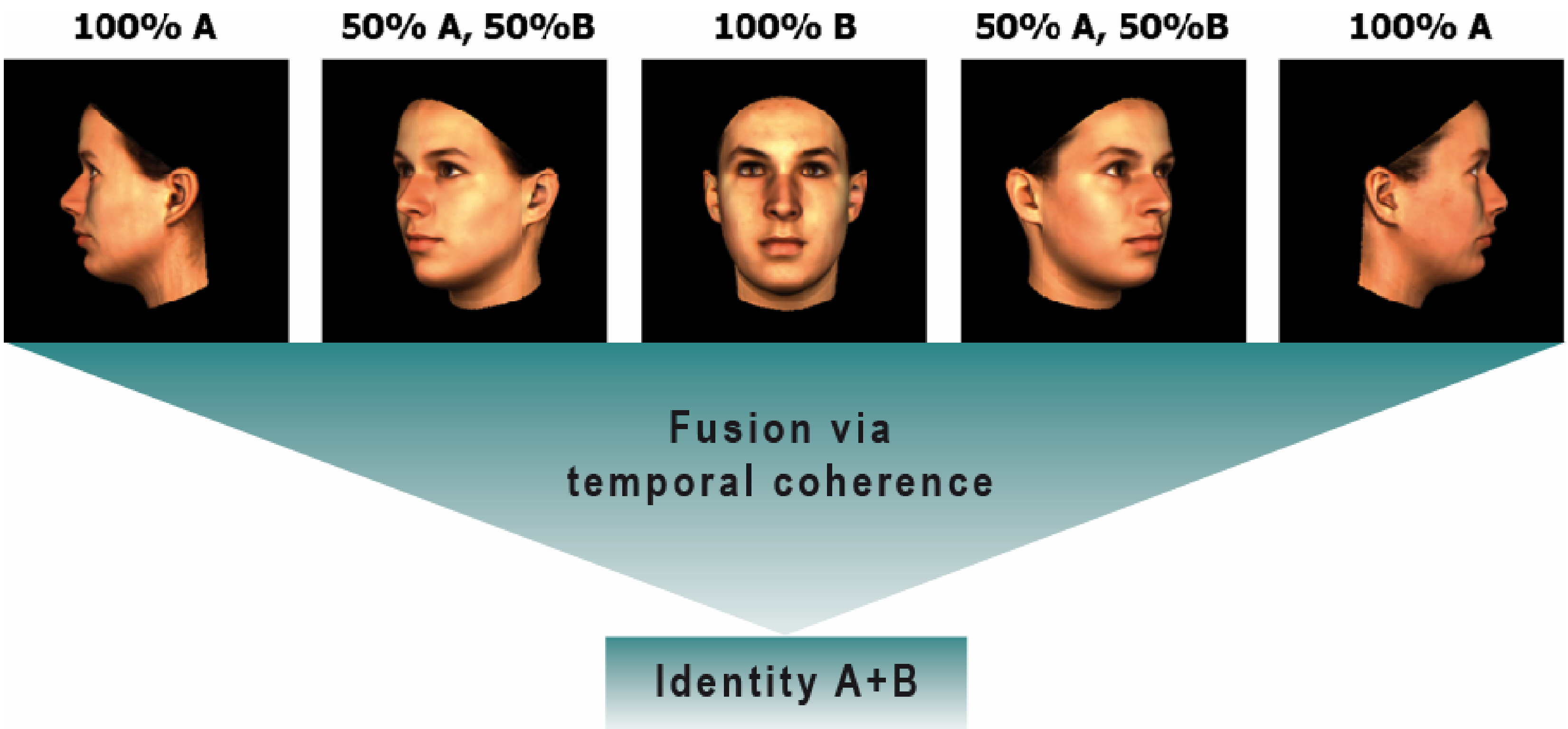
Object recognition: The role of time

Wallis, Bülthoff, *PNAS* (2001)

- Humans make active use of the temporal dimension for **learning** and recognition of objects

Object recognition: The role of time

Wallis, Bülthoff, *PNAS* (2001)

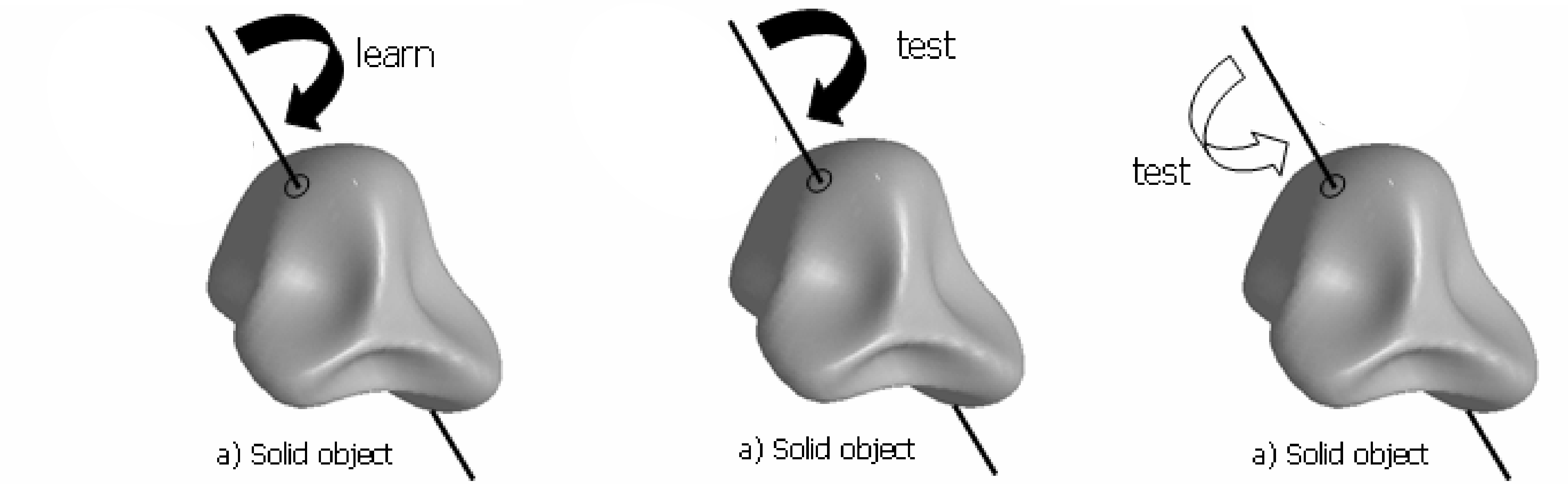


- By seeing views from two different people in a contiguous temporal sequence we bind all these views into the presentation of one person
- Humans use the temporal dimension for solving the binding problem

Object recognition: The role of time

Stone, *Vision Research* (1998)

- The same temporal sequence direction between learning and testing is important for recognition performance.



Learning

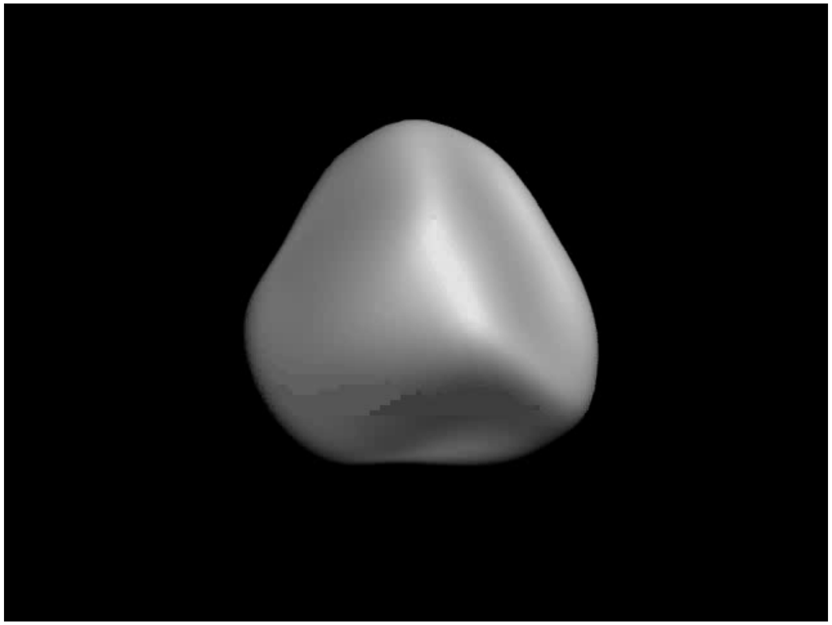
Test phase I
Recognition good

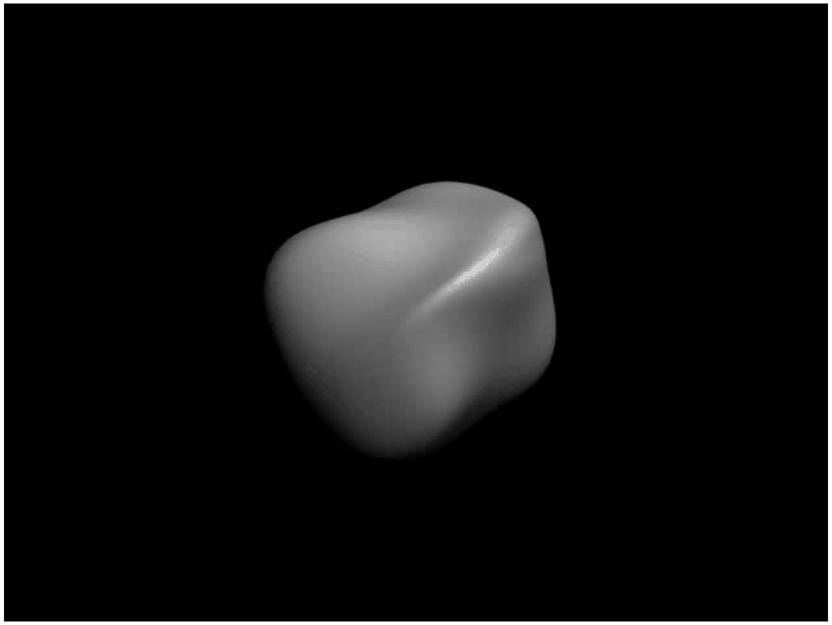
Test phase II
Recognition worse

Object recognition: The role of time

Chuang, Vuong, Thornton and Bülthoff, *Visual Cognition* (2006)

- The temporal dimension for learning and **recognition** of objects is important even for non-rigidly deforming objects.





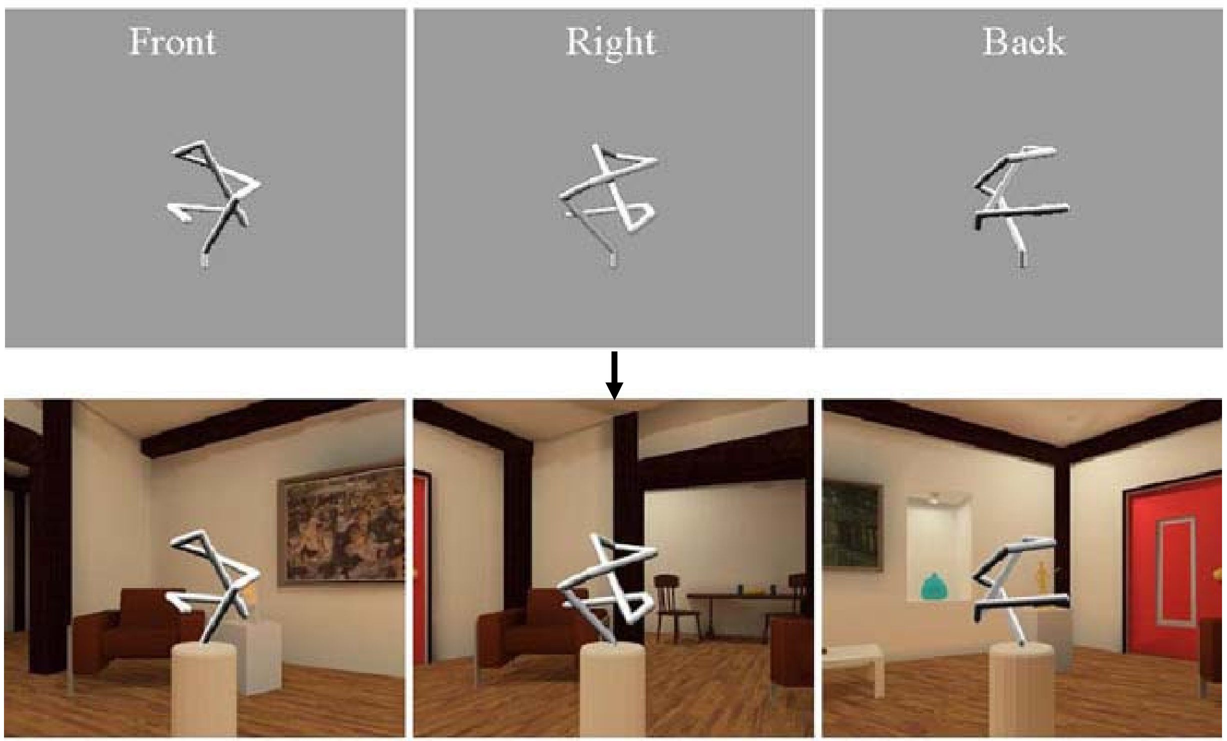
Outline of the talk

- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based temporal information for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

Scene information for object recognition

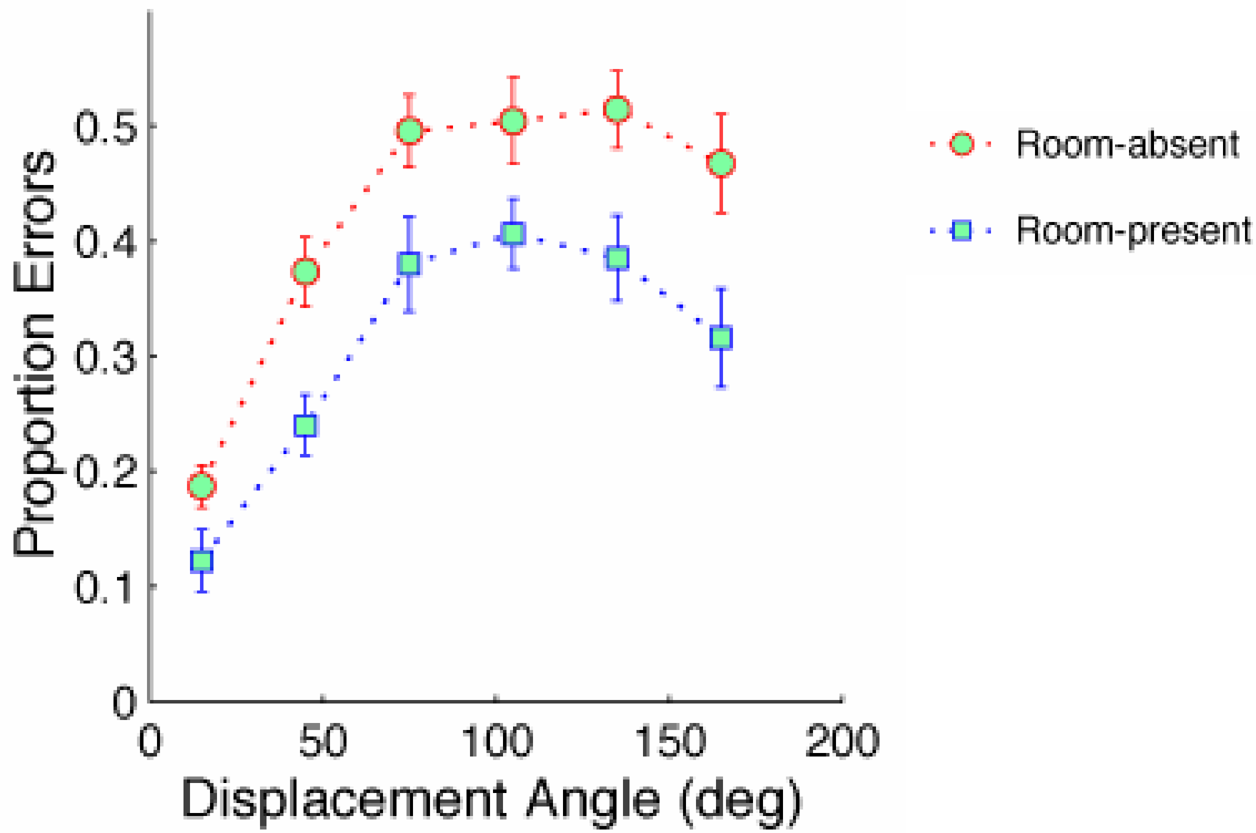
Christou, Bülthoff, *Journal of Vision* (2003)

- Image-based recognition has been investigated with isolated objects
- In real-life, however, we rarely see objects in isolation
- Putting an object into a scene can provide external cues to its orientation
- Does the human visual system exploit this information?



Scene information for object recognition

Christou, Bülthoff, *Journal of Vision* (2003)



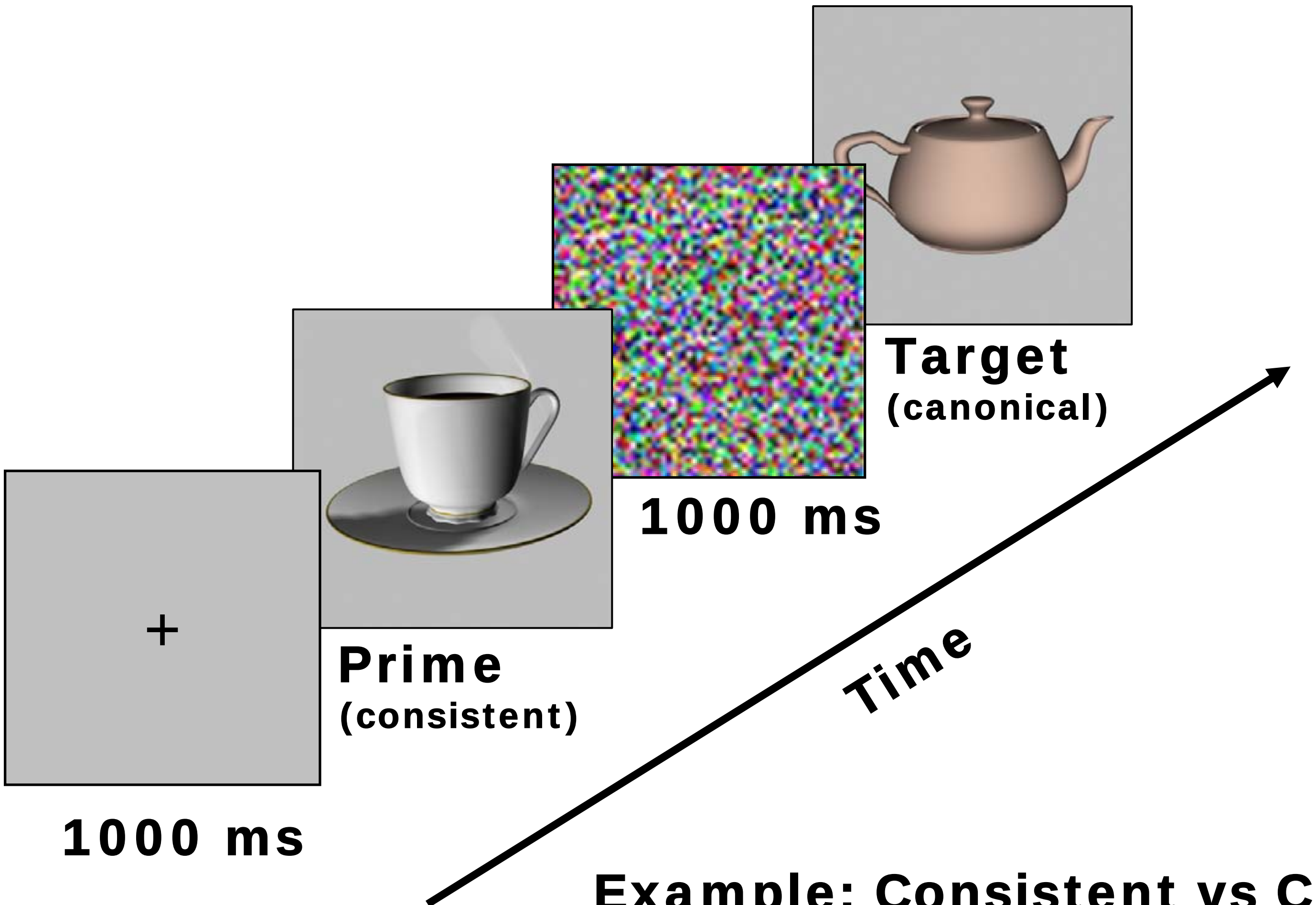
- Knowledge of viewpoint via room context improved recognition
- This suggests an ego-centric, image-based encoding of objects

Contextual information for object categorization

- Most current object recognition models are **bottom-up**
- **But**: Objects tend to co-occur in certain object contexts
- Coffee cups tend to be near coffee pots, wine glasses near wine bottles etc.
- Does the human visual system **use** such contextual information also for categorization?

Contextual information for object categorization

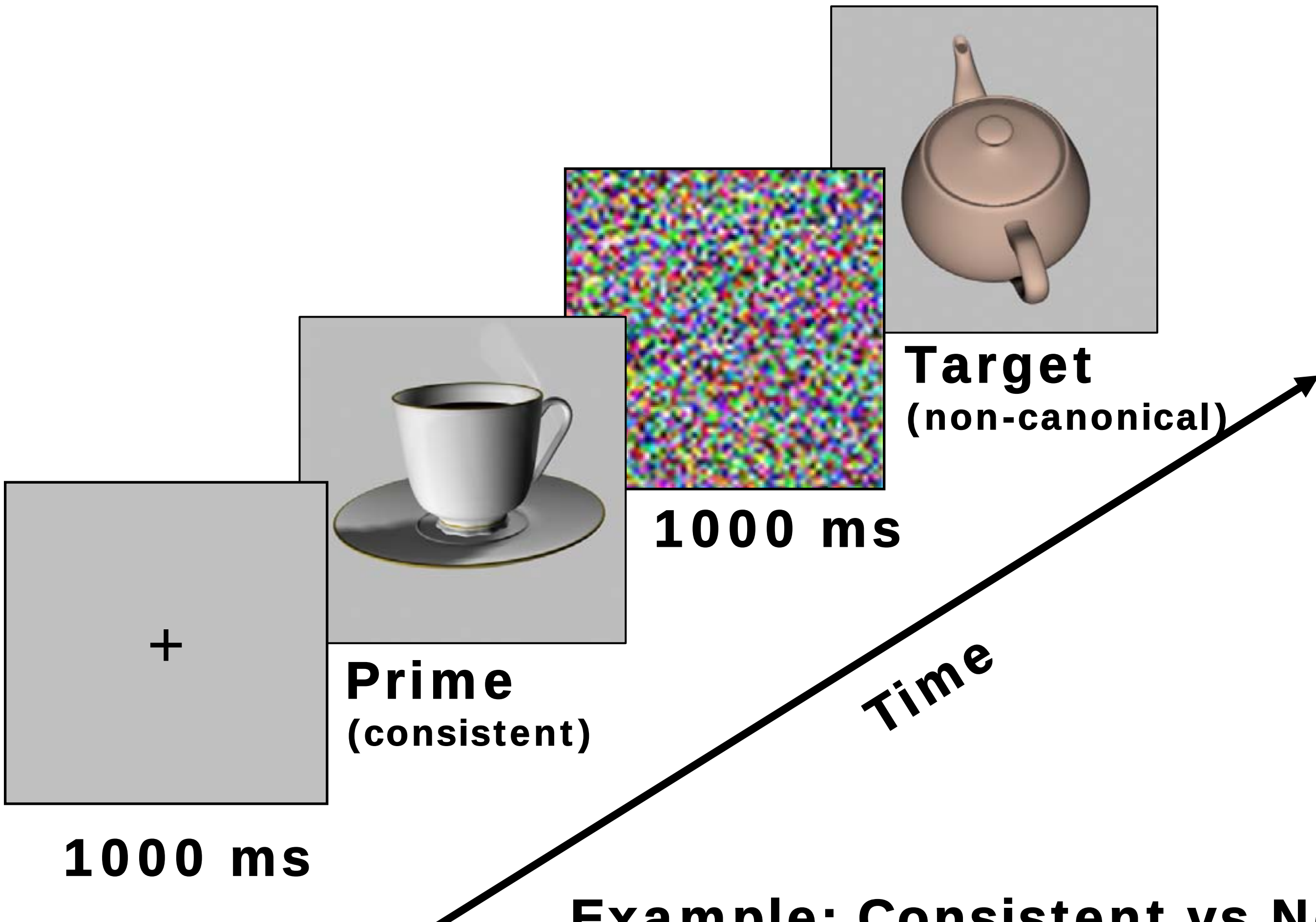
Schwaninger et al., (in prep.)



Example: Consistent vs Canonical

Contextual information for object categorization

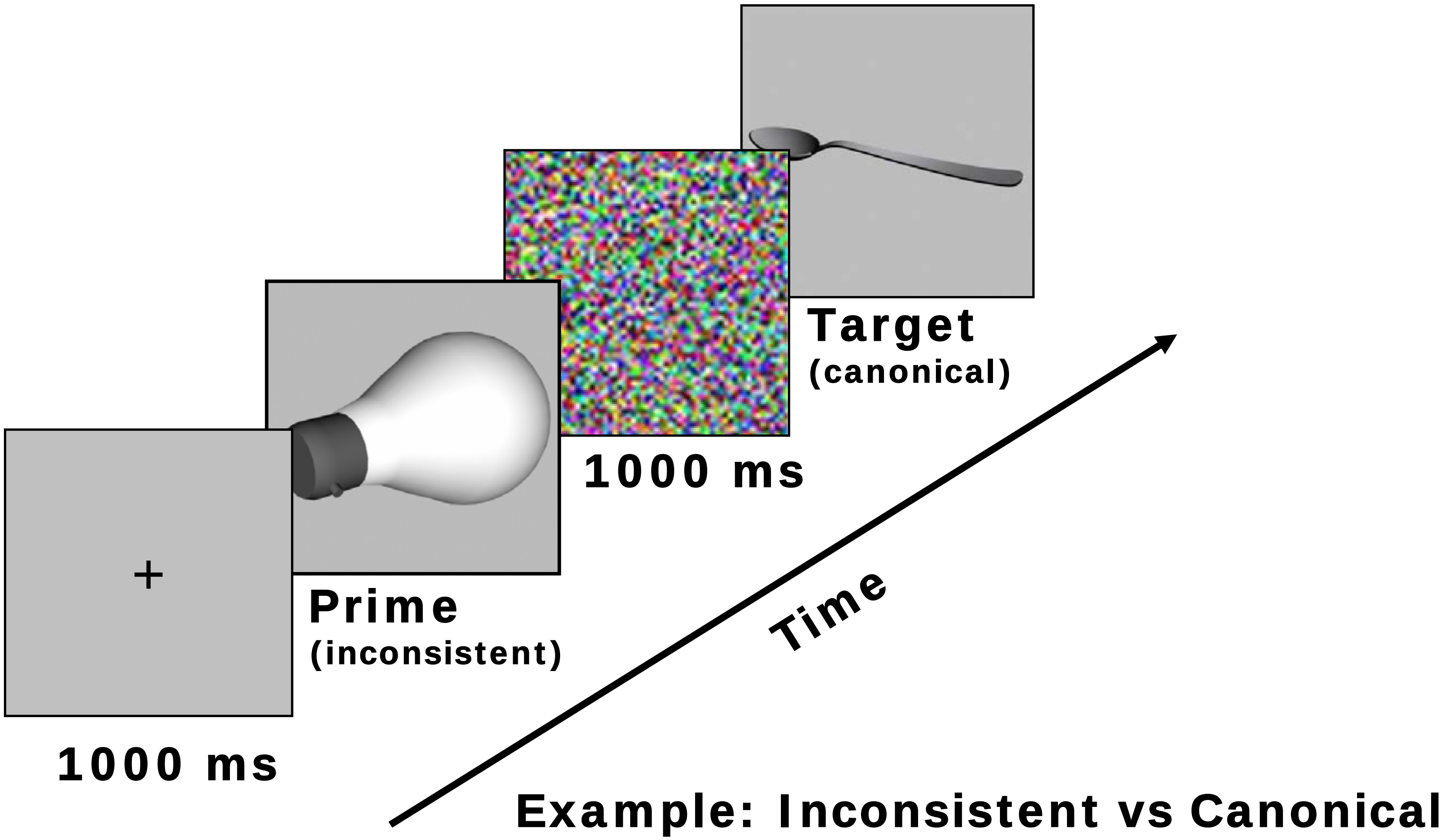
Schwaninger et al., (in prep.)



Example: Consistent vs Non-canonical

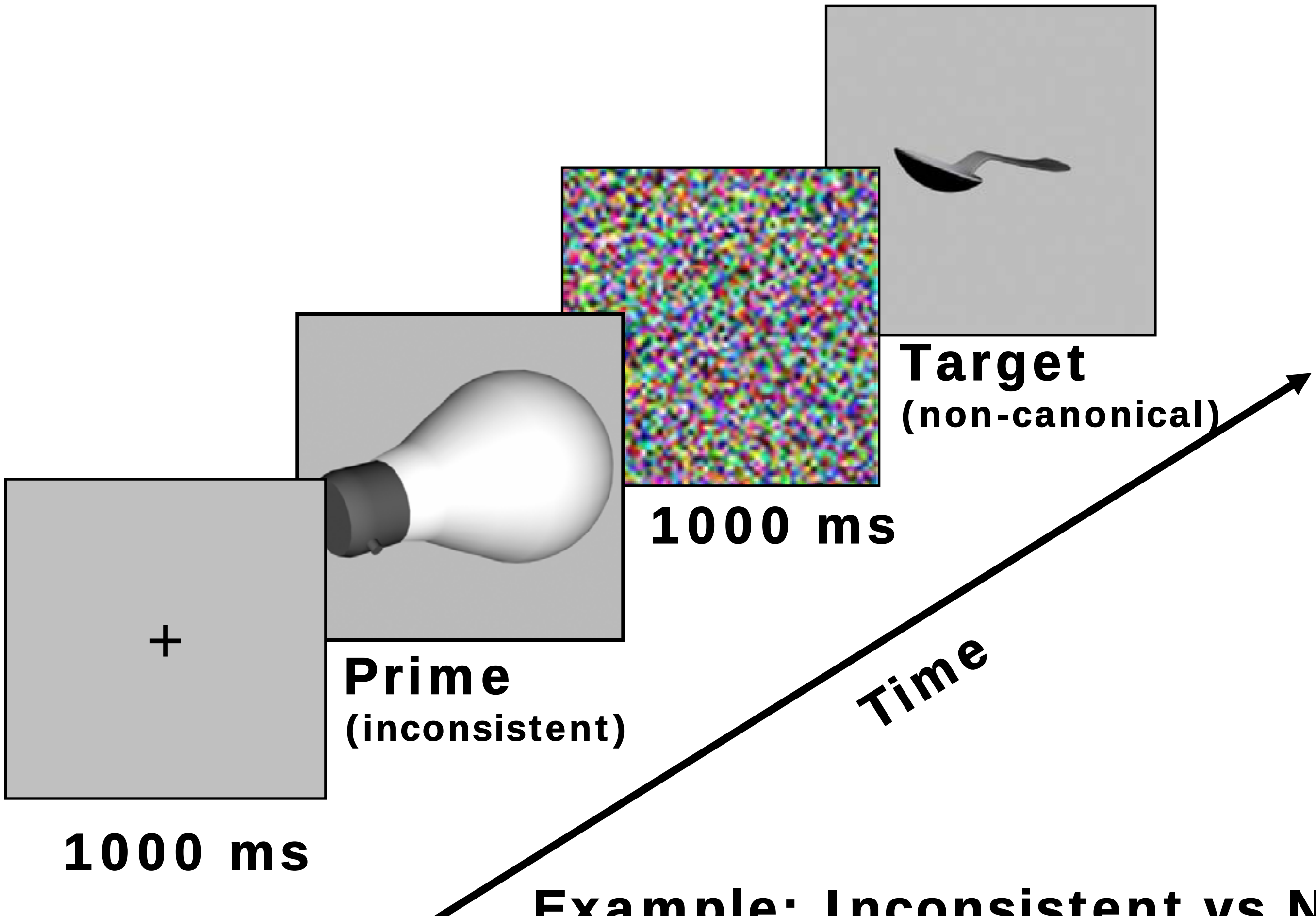
Contextual information for object categorization

Schwaninger et al., (in prep.)



Contextual information for object categorization

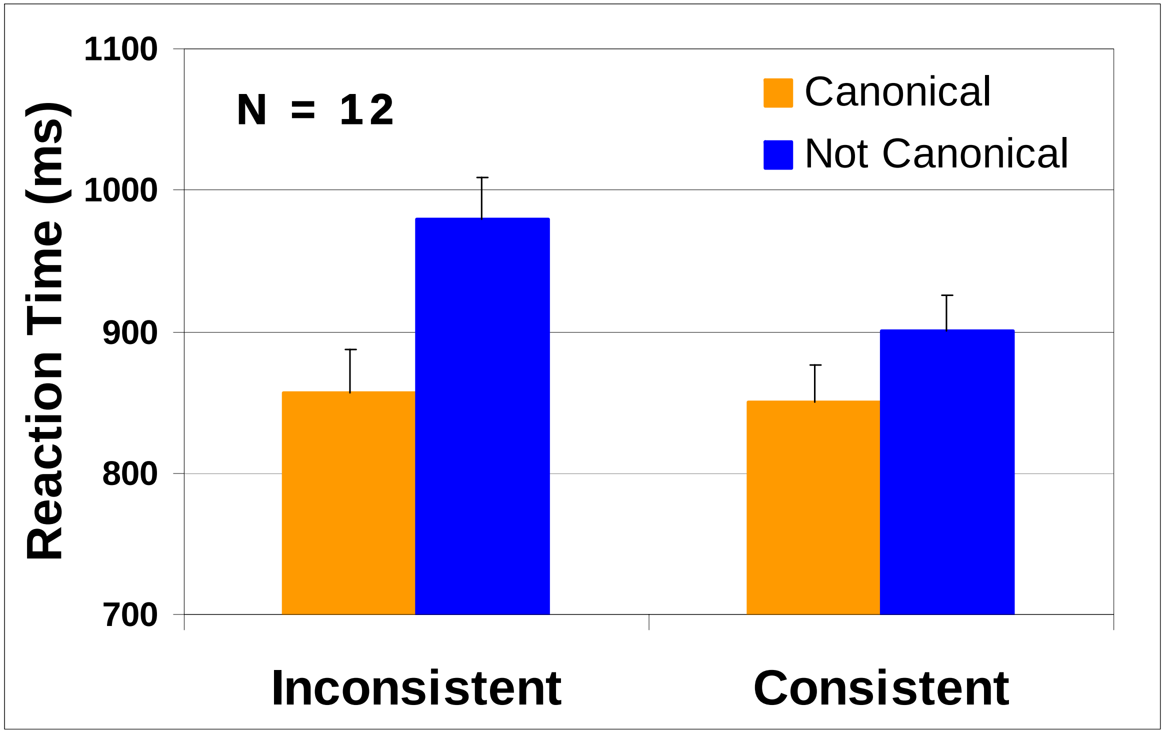
Schwaninger et al., (in prep.)



Example: Inconsistent vs Non-canonical

Contextual information for object categorization

Schwaninger et al., (in prep.)



- Consistent is faster than inconsistent
- Canonical is faster than non-canonical
- These top-down effects are especially important when **non-canonical views** have to be processed
  - this advantage thus holds both for recognition and categorization of objects

Outline of the talk

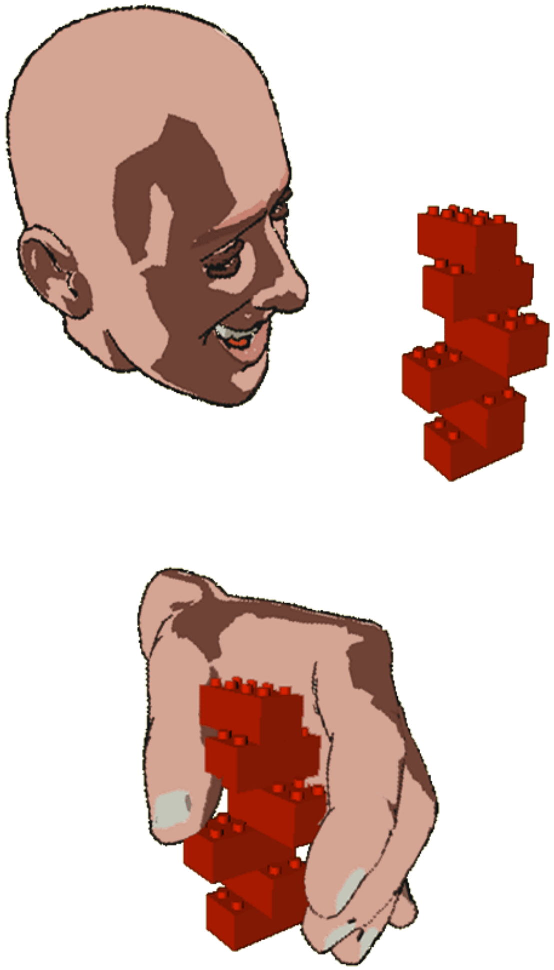
- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based temporal information for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

Multisensory object representations

- The world is coming into our head not only via our eyes
- Particularly important for object processing is the combination of visual and haptic information
  - provides information about material
  - solves visual size ambiguity by implicitly providing a scale
- Some important questions for visuo-haptic processing:
  - How is the information integrated?
  - Are there haptic views?
  - Are there common object representations?

Visual and Haptic Recognition

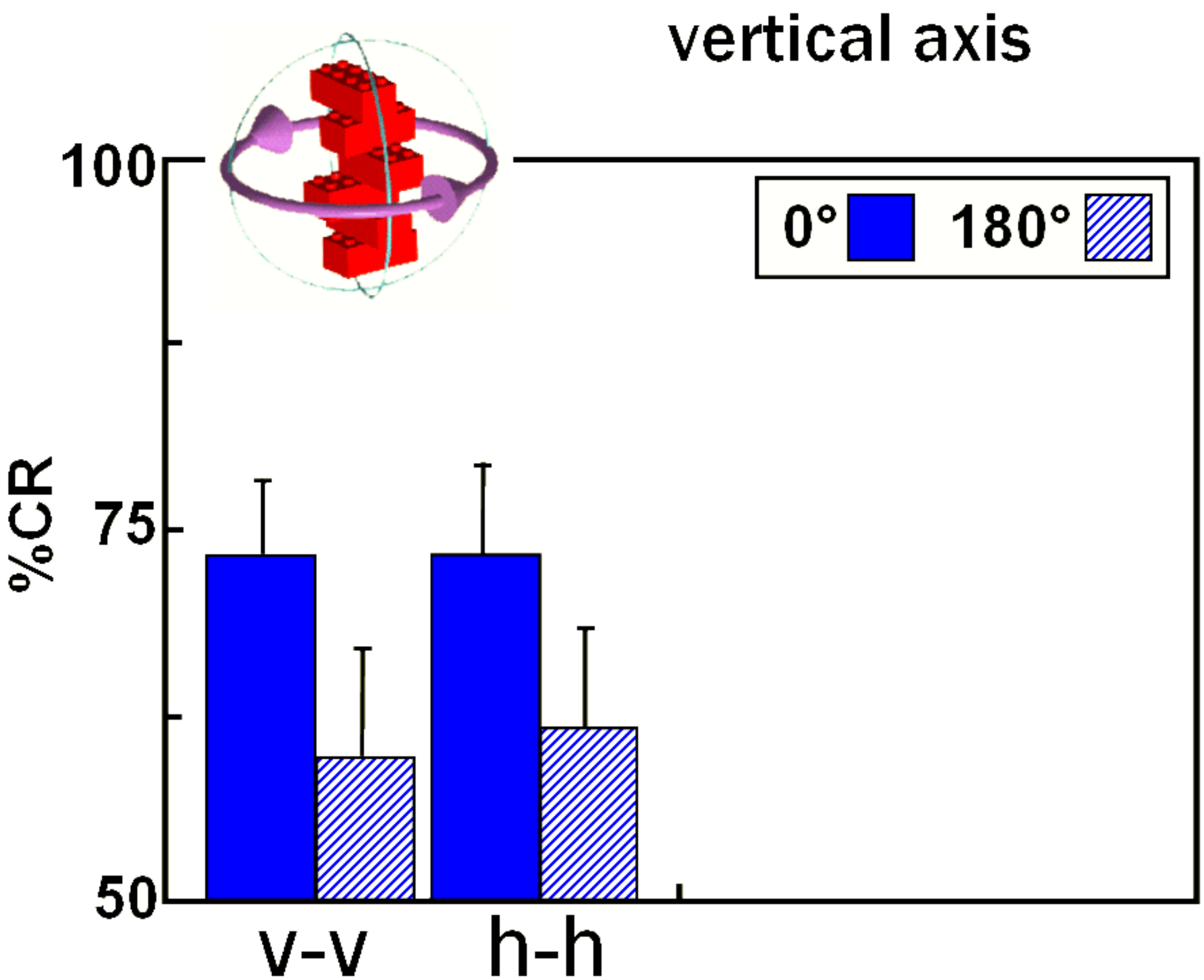
Newell, Ernst, Tjan & Bülthoff, *Psychological Science*, 2001



- Visual object recognition
  - 2D input
  - image-based recognition
  - egocentric encoding
- Haptic object recognition
  - 3D input
  - 2D or 3D representation?
  - only few reports
    - Lederman & Klatzky, 1987
    - Easton, Srinivas & Green, 1997
- Open questions?
  - How is the information integrated?
  - Are there haptic views?
  - Are there common object representations?

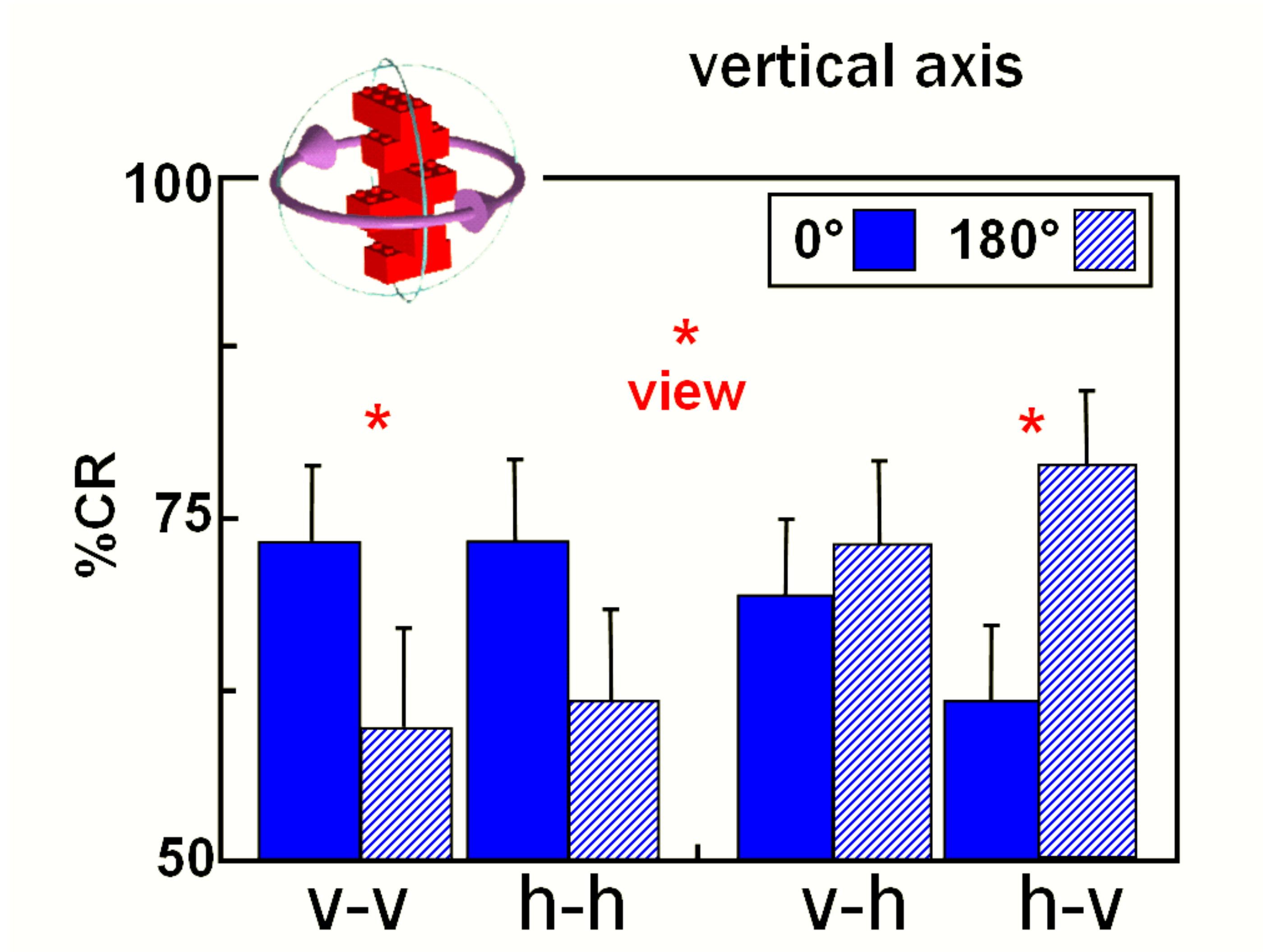
Rotation Around Vertical Axis

Newell, Ernst, Tjan & Bülthoff, *Psychological Science*, 2001



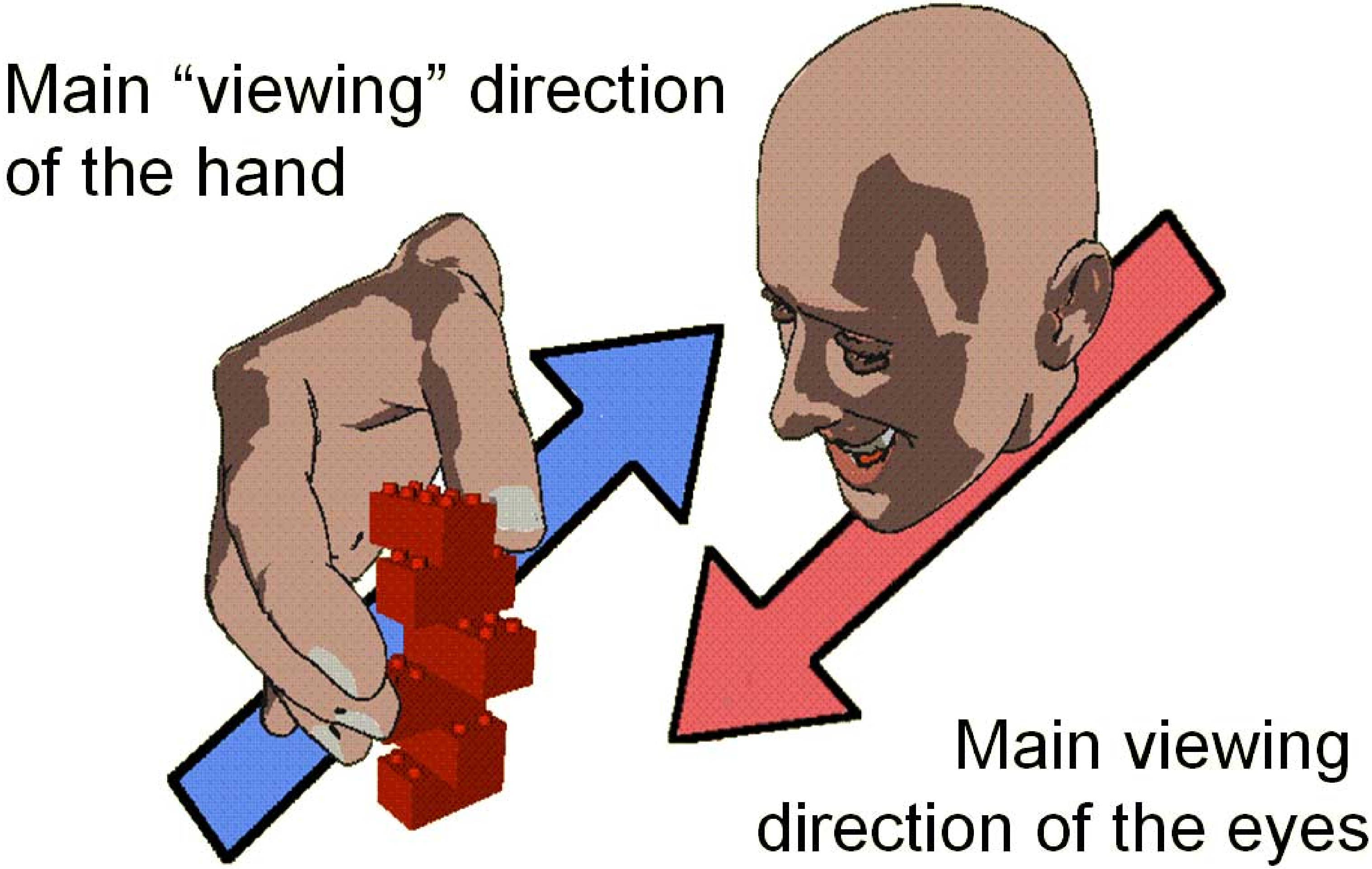
Cross-modal Transfer

Newell, Ernst, Tjan & Bülthoff, *Psychological Science*, 2001



The Visual and Haptic "View"

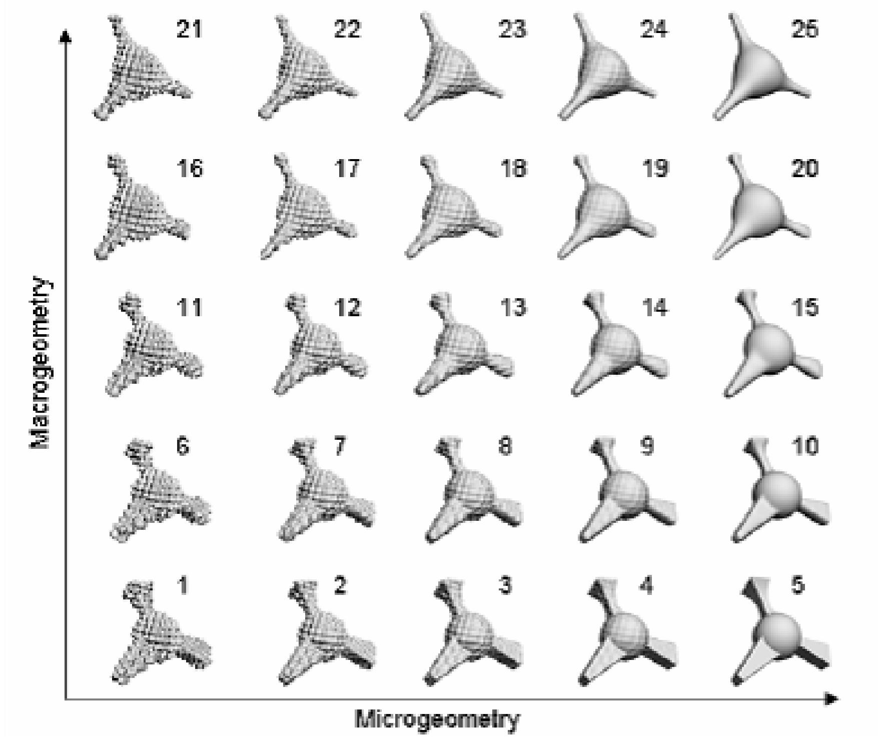
Newell, Ernst, Tjan & Bülthoff, *Psychological Science*, 2001



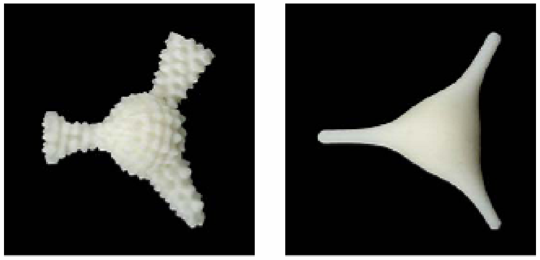
Multi-modal similarity and categorization of novel, 3D objects

Cooke, Jäkel, Wallraven, Bülthoff, *Neuropsychologia* (2007)

- Develop framework for understanding multi-sensory (visuo-haptic) object perception
- Controlled space of visuo-haptic stimuli printed in 3D
- Multi-Dimensional-Scaling for finding perceptual space for haptic, visual and bimodal exploration



Parametric stimulus space varying in shape and texture

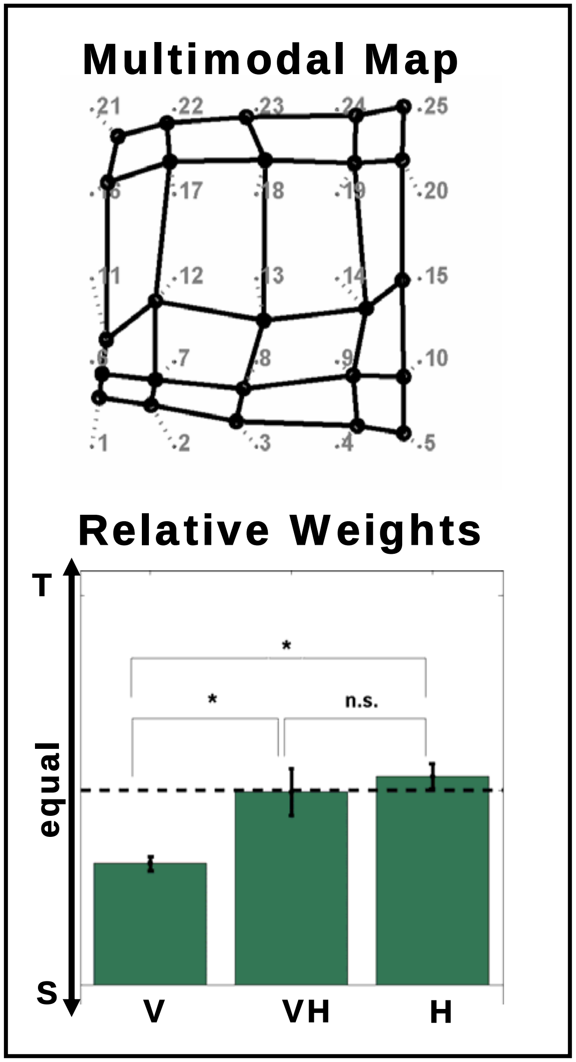


Photographs of printed 3D objects

Multi-modal similarity and categorization of novel, 3D objects

Cooke, Jäkel, Wallraven, Bülthoff, *Neuropsychologia* (2007)

- Main results from similarity ratings:
  - The perceptual map along with the two dimensions of shape and texture is recovered remarkably well
  - Humans weight object properties differently when they explore objects using different modalities
  - **This is a good indication that, indeed, object representations might be shared across modalities**



Outline of the talk

- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based temporal information for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

Image-based navigation – from flies to humans

Insects

Bottom-Up Processing:

- very fast, reactive behavior
- (almost) no memory
- hard-wired reflexes
- massive parallel processing: feed forward processing
- task-specific hardware, adapted to environment
- simple sensor fusion

Humans

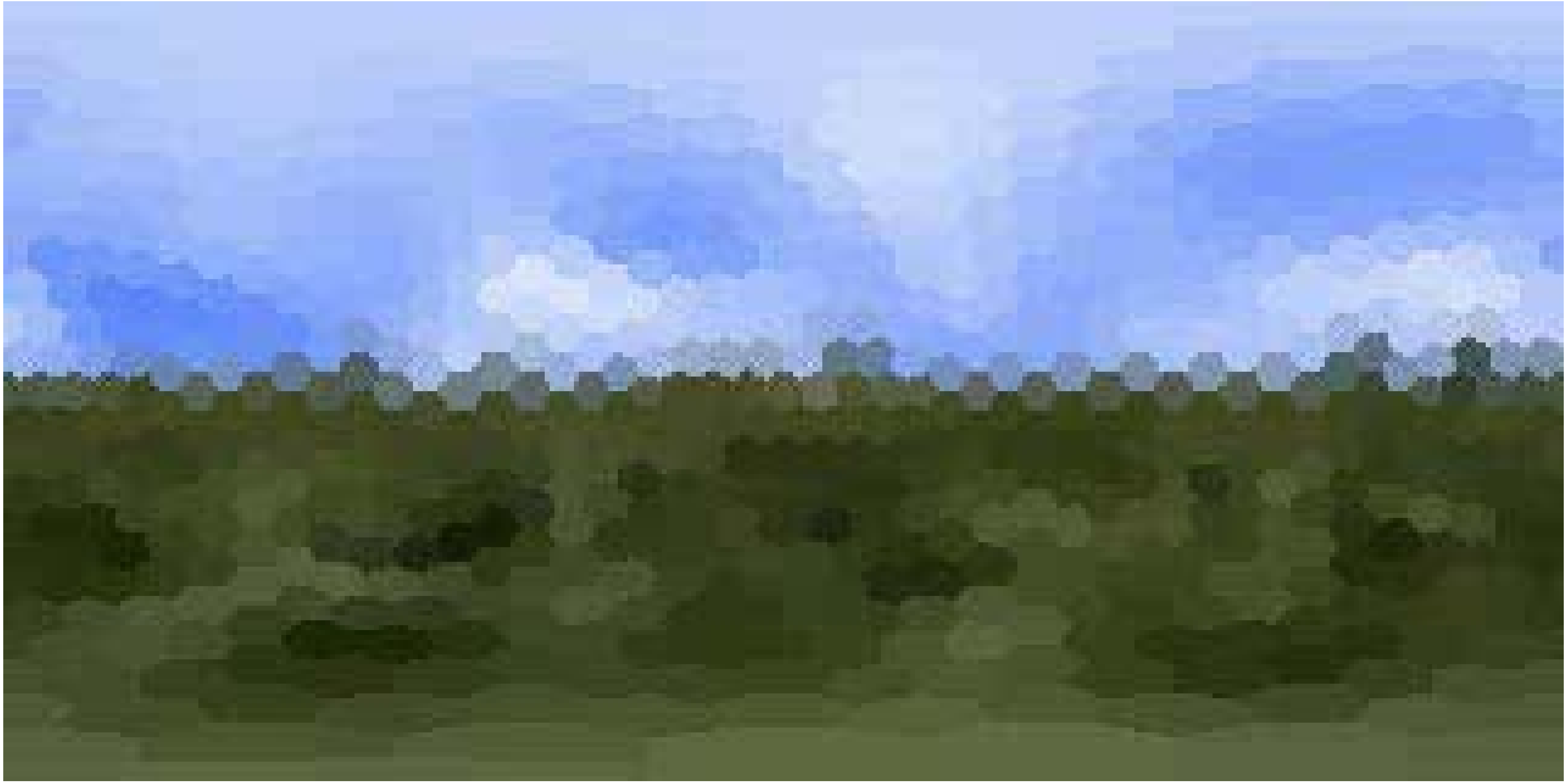
Top-Down Processing:

- cognitive, learned behavior
- memory-based computation
- learned behavior
- massive parallel processing: many feedback connections
- flexible, multi-purpose hardware
- adaptive sensor fusion
- attention
- awareness

Image-based flight control

Titus Neumann, Dissertation (2003)

Drosophila Vision
only 642 Photoreceptors

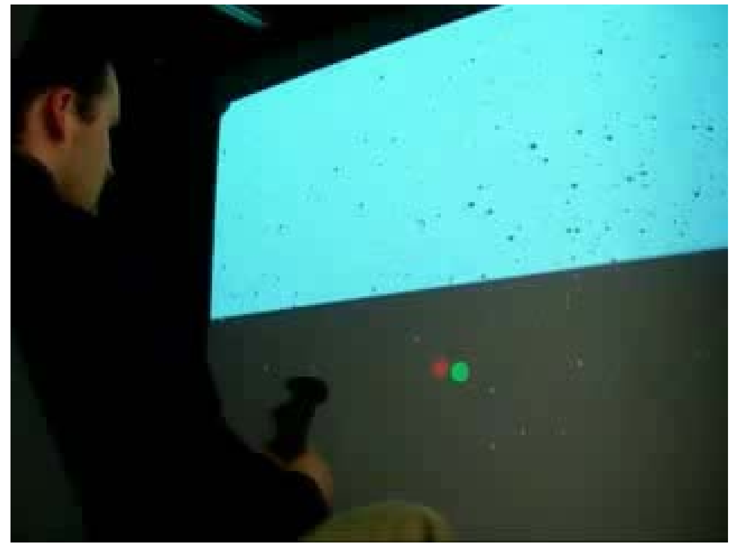


Insect inspired autonomous vehicles
The view from of the cockpit of the fly

Beyond image-based flight control

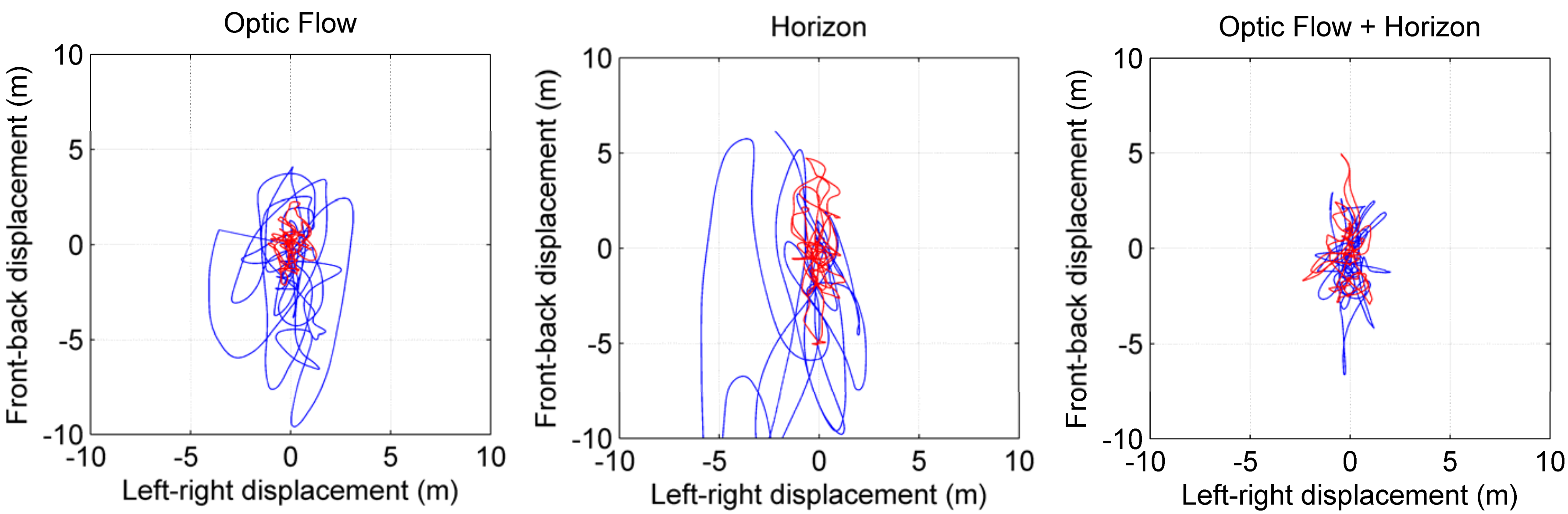
Berger, Terzibas, Bülthoff, *Proc. TWK*, 2006

- Current simulators do not seem to support a realistic feel for pilots
  - hover performance is weak
  - experienced pilots suffer from simulator sickness
- Which visual cues and how much motion is necessary to built a realistic helicopter simulator?
  - only position markers
  - optical flow
  - horizon position
  - horizon orientation
  - platform motion



Beyond image-based flight control

Berger, Terzibas, Bülthoff, *Proc. TWK*, 2006



- Platform motion is essential for good hover performance
- Horizon only is almost impossible
- Integration of all cues gives best performance

Outline of the talk

- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based temporal information for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

Application: Image-based material editing

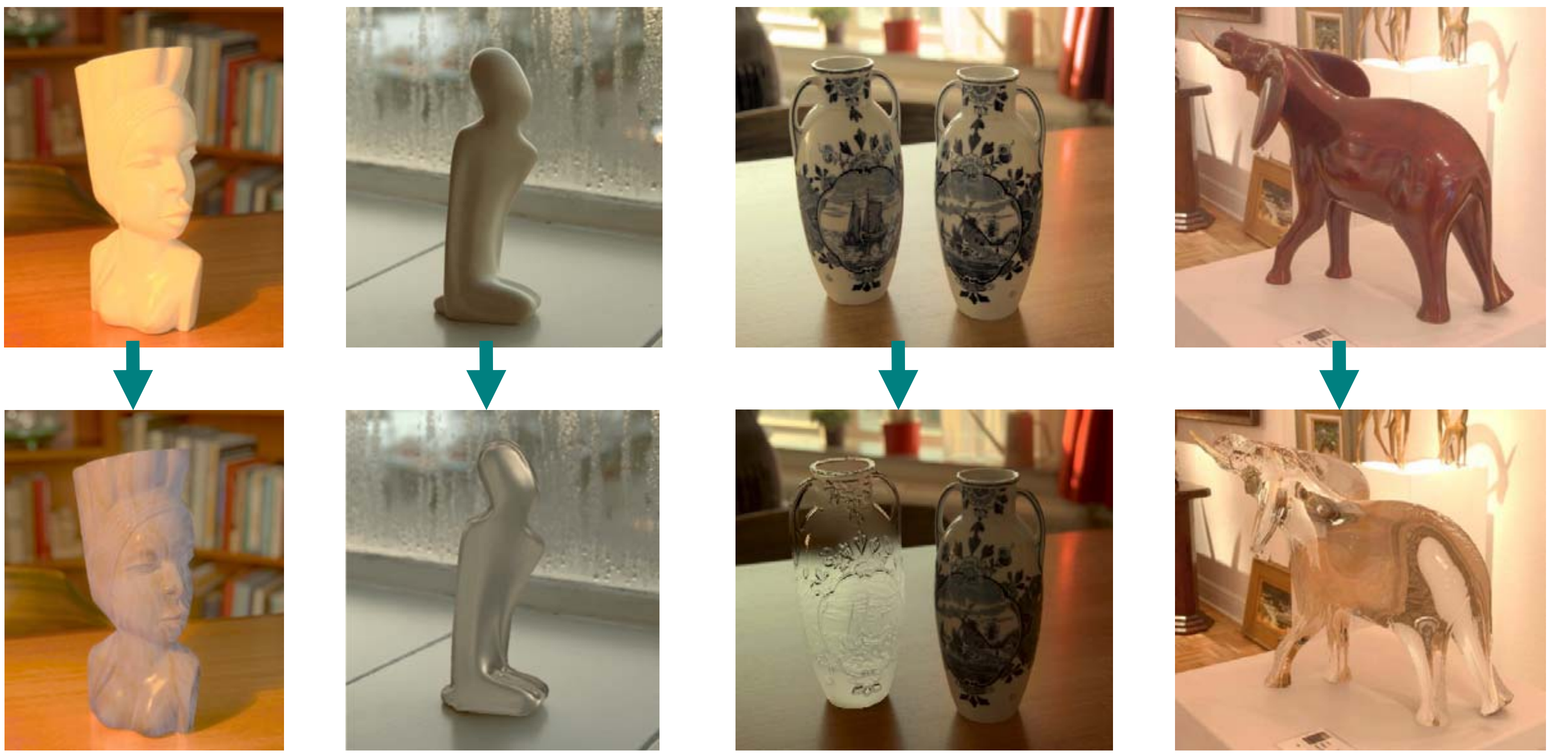
Kahn, Reinhard, Fleming, Bülthoff, *ACM Transactions on Graphics* (2006)

- Changing the material appearance of an object
- given a **single photograph**
- without 3D reconstruction
- using “*perceptual tricks*“

Application: Image-based material editing

Kahn, Reinhard, Fleming, Bülthoff, *ACM Transactions on Graphics* (2006)

- Method for changing the material appearance of an object given a **single photograph** as input and no 3D reconstruction
- Uses "*cheap tricks*" that exploit assumptions of the human visual system to achieve illusion of material transformation



re-textured new BRDF transparency

Application: Image-based material editing

Kahn, Reinhard, Fleming, Bülthoff, *ACM Transactions on Graphics* (2006)

Outline of the talk

- **Main working hypothesis:**
  - the brain adopts an image-based strategy for perception & action
- **Support for this hypothesis comes from:**
  - Image-based object recognition
  - Image-based temporal information for object learning and recognition
  - Scene and Contextual information for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- **Application examples:**
  - Image-based heuristics for material perception
  - An image-based, multisensory robot

Application: An image-based, multisensory robot

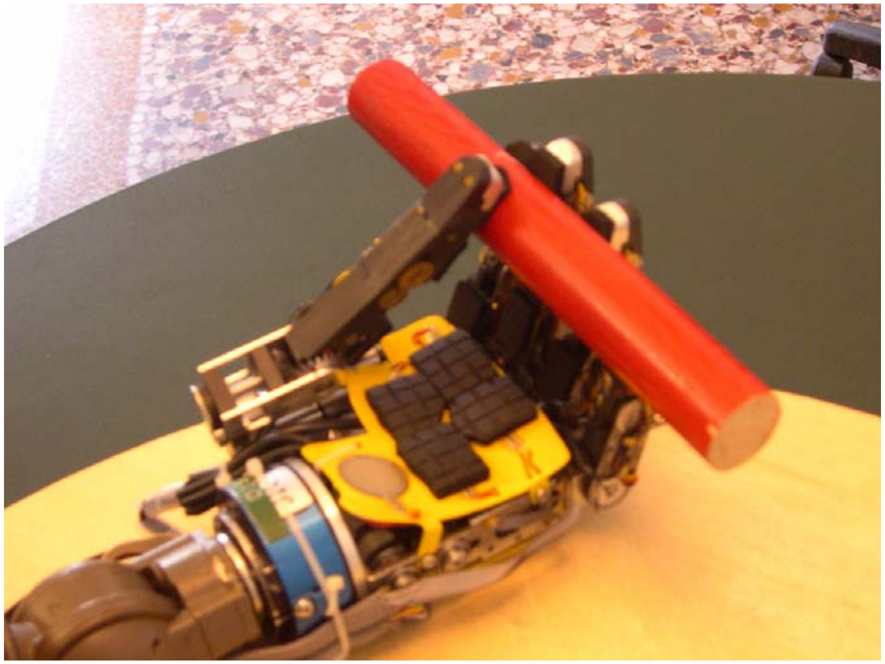
Wallraven, Bülthoff, *Object Recognition, Attention and Action*, (in press)

- Framework for integration of proprioceptive and visual information:
  - image-based
  - integrates temporal information
  - common representation for proprioceptive and visual information

Application: An image-based, multisensory robot

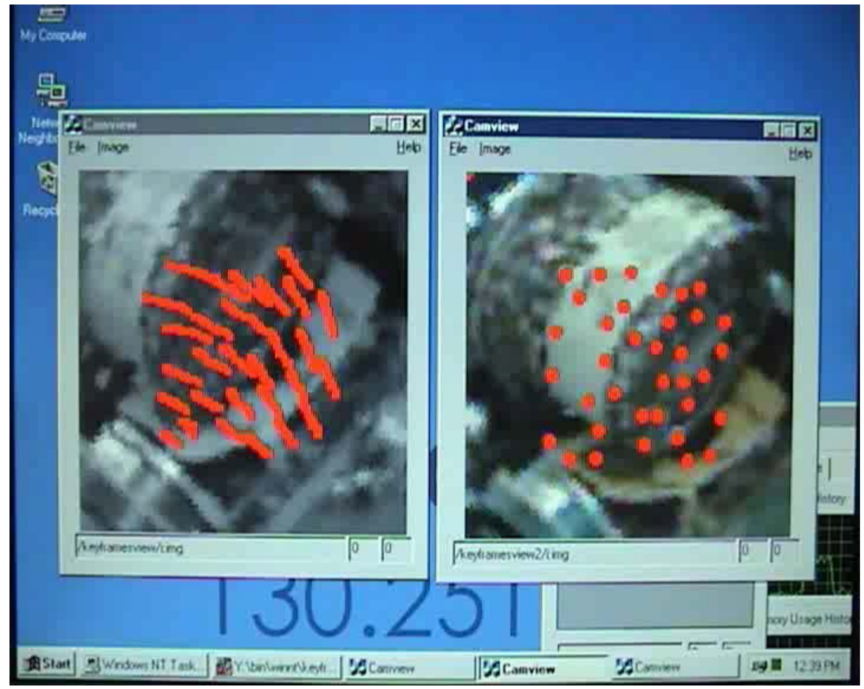
Wallraven, Bülthoff, *Object Recognition, Attention and Action*, (in press)

- The robot learns an object by manipulating it according to a pre-programmed motor program
- The visual input is used to extract **keyframes**
- Every time a keyframe is found, the proprioceptive information of the robot hand is saved alongside
- This information is used to create a **multisensory** object representation



Application: An image-based, multisensory robot

Wallraven, Bülthoff, *Object Recognition, Attention and Action*, (in press)

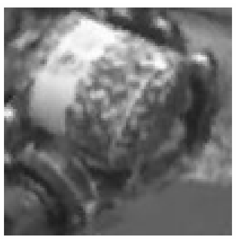


External View Keyframes Tracking

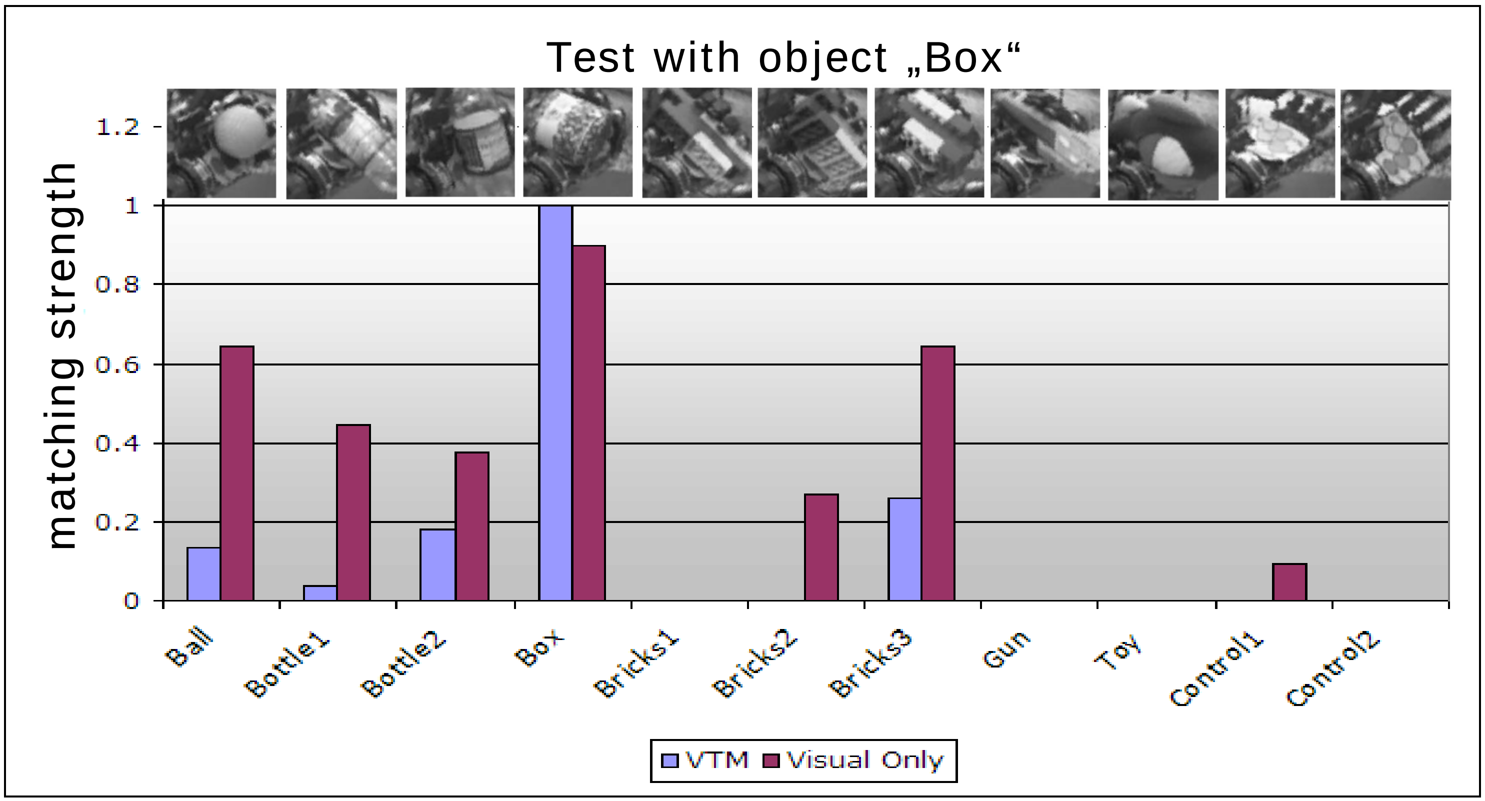
Application: An image-based, multisensory robot

Wallraven, Bülthoff, *Object Recognition, Attention and Action*, (in press)

- Recognition was shown to be much more **discriminant** using multisensory information (cf. **red bars (visual)** vs. **blue bars (multisensory)**)



Test-object



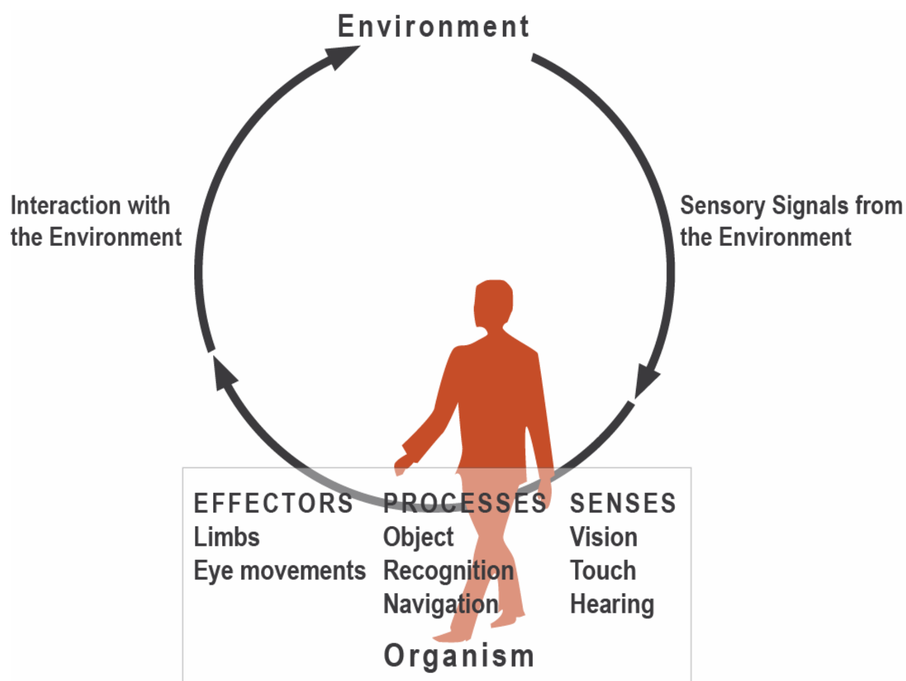
Application: An image-based, multisensory robot

Wallraven, Bülthoff, *Object Recognition, Attention and Action*, (in press)

- From action to views
  - Learn and recognize object representations by interaction
  - Execute movements that take you to informative views
- From views to action
  - Given a view, select an appropriate action
  - Important for manipulation, e.g., inserting an object into a hole
- Extensions
  - Generalizes also to other sensory channels

Summary

- We have discussed the following problems:
  - Image-based temporal information for object learning and recognition
  - Image-based heuristics for material perception
  - Contextual information from a scene for object recognition
  - Multi-sensory object processing
  - Image-based flight control
- Two applications from computer graphics and robotics have demonstrated the usefulness of the image-based approach



These examples support our philosophy that the brain adopts an image-based strategy for perception & action

The 2D image not the 3D structure is the key to recognition

One Object –Two Views
Man or Hare ?

Markus Raetz

Raetz explained
by Isabelle Bülthoff

Credits

Daniel Berger
Visuo-vestibular interaction

Lewis Chuang
Recognizing deformable objects

Theresa Cooke
Visuo-haptic integration

Chris Cristou
Scene recognition

Roland Fleming
Material perception

Titus Neumann
Visual flight control

Fiona Newell
Visuo-haptic recognition

Adrian Schwaninger
The role of context

Guy Wallis
Temporal assocation

Christian Wallraven
Perceptual computer vision

Open Questions

- next 10 years:
  - face recognition in airport terminals
- next 10-20 years:
  - Categorization in real world situations Turing Test for Recognition (*Chair Award*)
- next 20-30 years:
  - child-like one-shot learning of categories


Isabelle Bülthoff
MPI f. biol. Cybernetics
Tübingen, Germany