MAX-PLANCK-GESELLSCHAFT

Object Recognition in Man and Machine

MPI FOR BIOLOGICAL CYBERNETICS

Heinrich H. Bülthoff

Max-Planck-Institut
für biologische Kybernetik
Tübingen

www.tuebingen.mpg.de

MPI for Biological Cybernetics
Tübingen, Germany
www.kyb.tuebingen.mpg.de

- Mission
  Understanding of biological information processing
  How does the brain work?
- Approach
  Study cognition at three levels of understanding
  - Biological **Hardware** in the Physiology Department
    - system physiology in primates
    - multi-electrode recordings and brain imaging (fMRI)
  - Behavior and **Algorithms** in the Psychophysics Department
    - perception experiments (human psychophysics)
    - behavioral experiments in closed action-perception-loop (VR)
  - Computational **Theory** in the Empirical Inference Department
    - statistical learning theory
    - computer vision and robotics

Overview

- What can we learn from biology to build intelligent machines?
  - from flies to autonomous robots
- Object Recognition & Categorization
  - How do we interpret the world through our senses?
    - with our eyes
    - with our hands
- Machine Recognition

What can we learn from biology

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

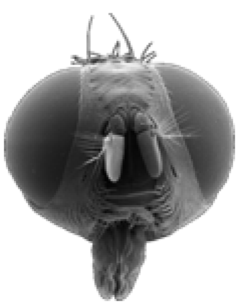
Top-Down Processing:

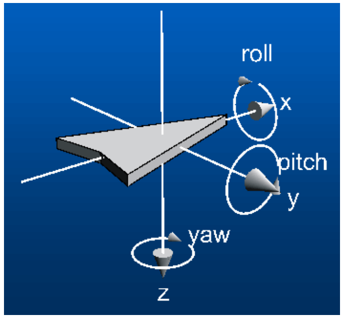
- cognitive, learned behavior
- memory-based computation
- learned behavior
- massive parallel processing: many feedback connections
- flexible, multi-purpose hardware
- adaptive sensor fusion
- attention
- awareness

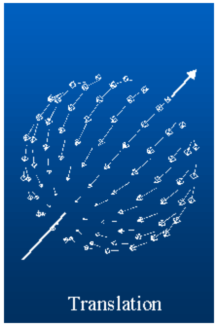
Insect Inspired Flight Control

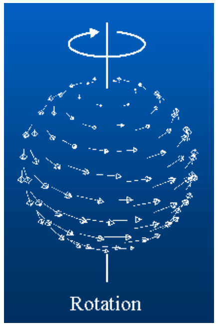
Titus R. Neumann

- Flies have minimal brain size and energy consumption
- Fast and robust behavior:
  - course stabilization
  - obstacle avoidance
- Minimalistic sensing and processing:
  - large field integrating neurons
  - for roll, pitch and yaw control
- Hard-wired statistical a priori knowledge about
  - environment
  - sensors, actuators and morphology of agent
- Applications:
  - robots for restricted space and energy consumption
  - micro-/nano-robotics
  - aerospace









Insect inspired flight control
obstacle avoidance & height control

Drosophila Vision
with 642 Photoreceptors

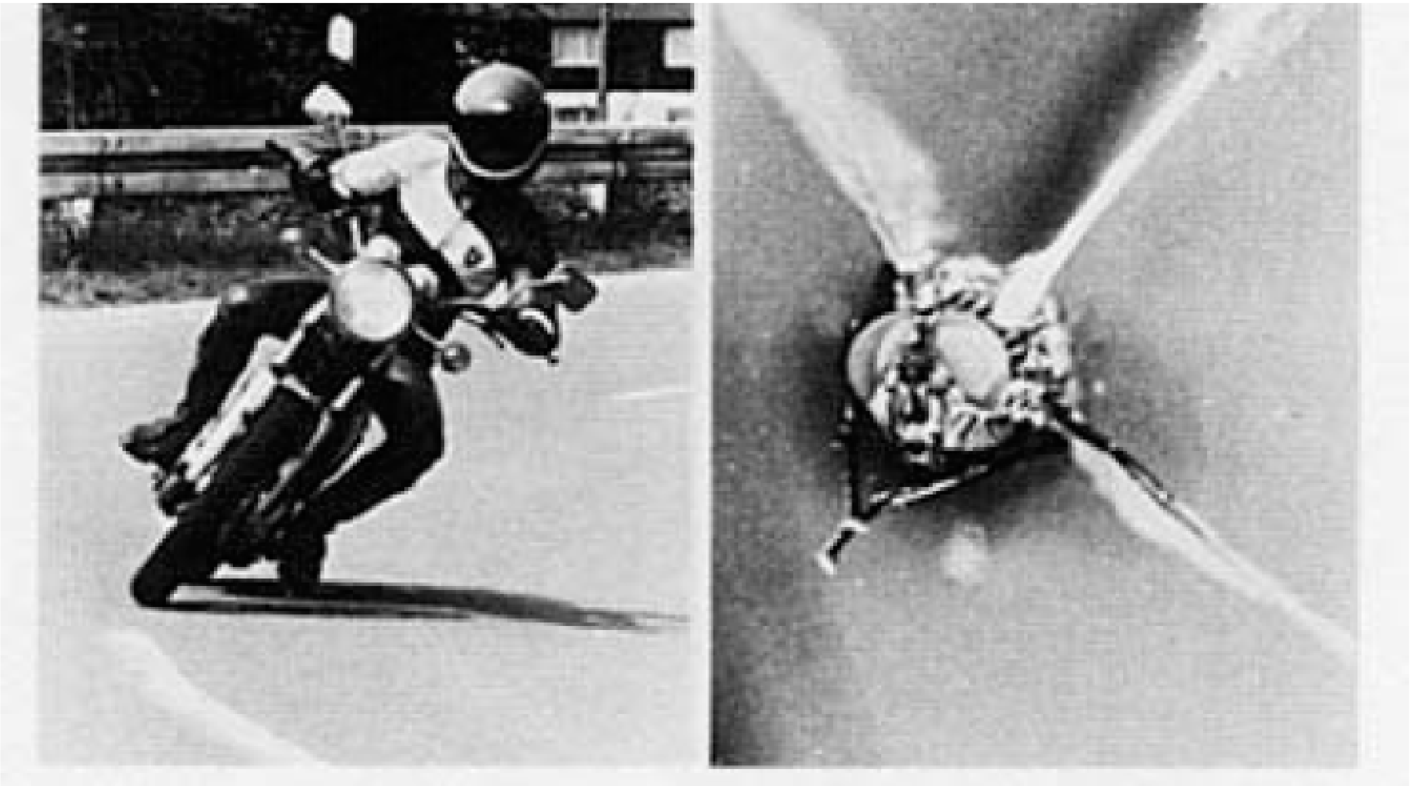
Autonomous Vehicles
Obstacle avoidance & height control

Head stabilization in man and fly

- Head stabilization with tiny gyrocopes

biker in Tübingen

fly in a fly flight simulator



R. Hengstenberg, Nature (1998)

Vision in Fly, Man and Machine

- Object Recognition
  - is poor in flies
  - is excellent in humans
  - is good and bad in machine vision
    - good for specific objects
      - machine inspection
    - bad for object categorization
      - classify everyday objects (chairs, tables, …)


Isabelle Bulthoff
MPI f. biol. Cybernetics
Tübingen, Germany

Shape Changes

By changing the shape of an object
we change the perceived category.

Viewpoint Changes

By changing the viewpoint we can also get a change in the perceived category.

Viewpoint Changes

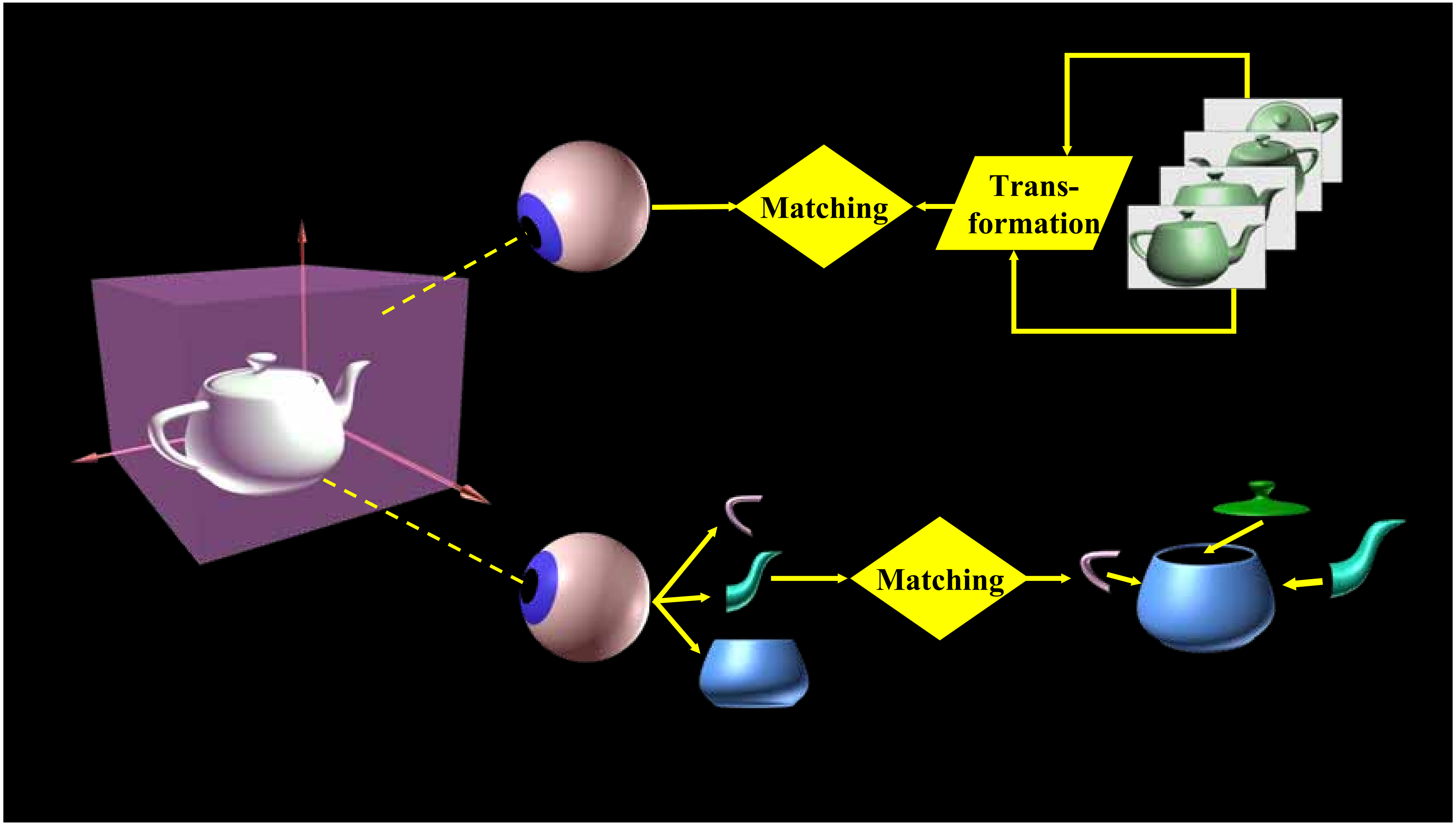
How does the brain disentangle viewpoint changes from shape changes?

Object Recognition

- In order to recognize objects we need an adequate representation of objects in the brain.
- Whether this representation is
  - image-based (appearance-based)
    - multiple views with low dimensional features
  - or based on structural descriptions (Geons)
    - relations between 3D parts

  is (sometimes passionately) debated.

How Does the Brain ReCOGnize Objects ?

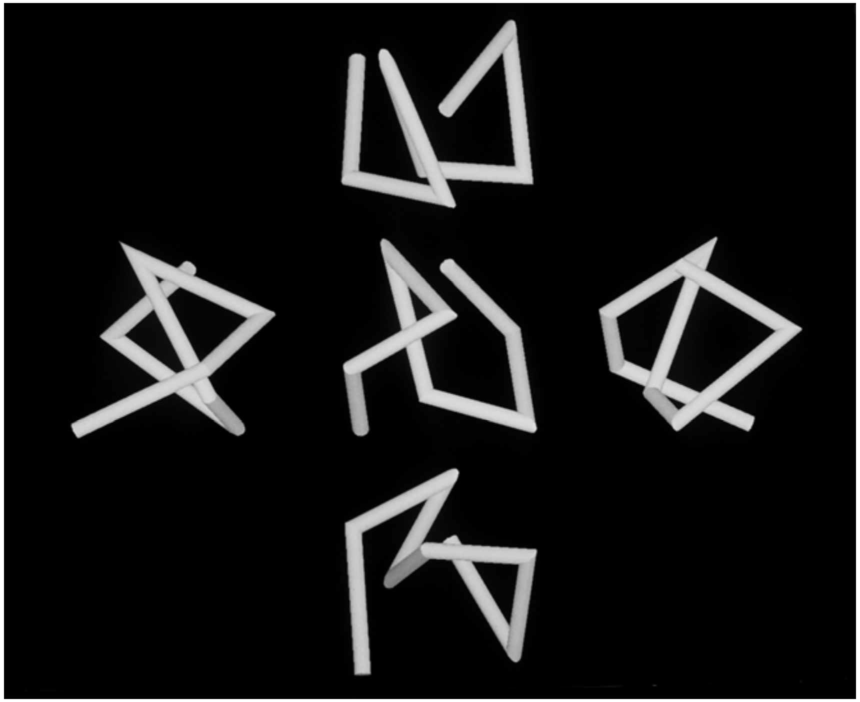


Our approach to test these models

- Test : viewpoint dependency of recognition
  - view-based models
    - depend on viewpoint
    - limited generalization to novel views
  - structural models
    - almost independent of viewpoint
    - good generalization to novel views
- Problem: for familiar objects all views are already known and therefore useless for view generalization experiments

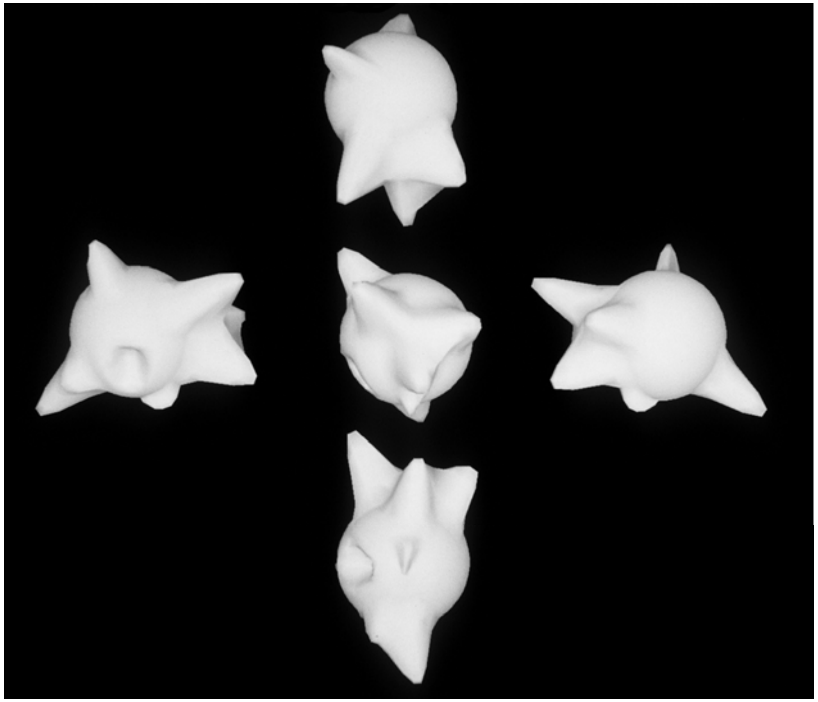
Recognition of unfamiliar objects

Paperclips



Bülthoff & Edelman, PNAS 1992

Amoebae



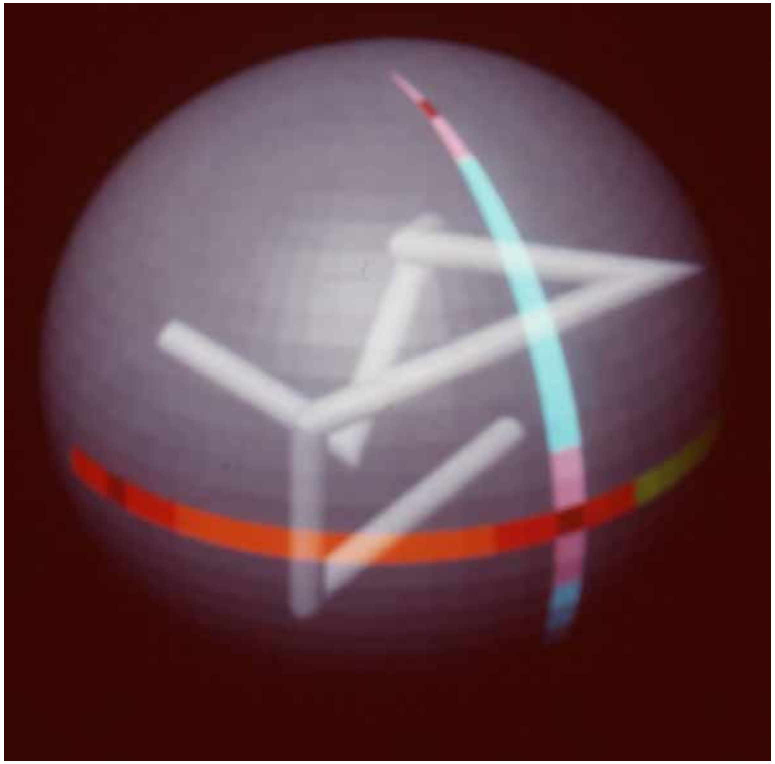
Edelman & Bülthoff , Vis. Res. 1992

Generalization Experiments

Bülthoff & Edelman

PNAS, 89, 60-64, 1992

- **Generalization:** better for views spanned by the training views than for orthogonal axis.
- **Conclusion:** "This is difficult to reconcile with any theory except the image combination approach." S. Ullman 1996.



Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Generalization Fields of Paperclips

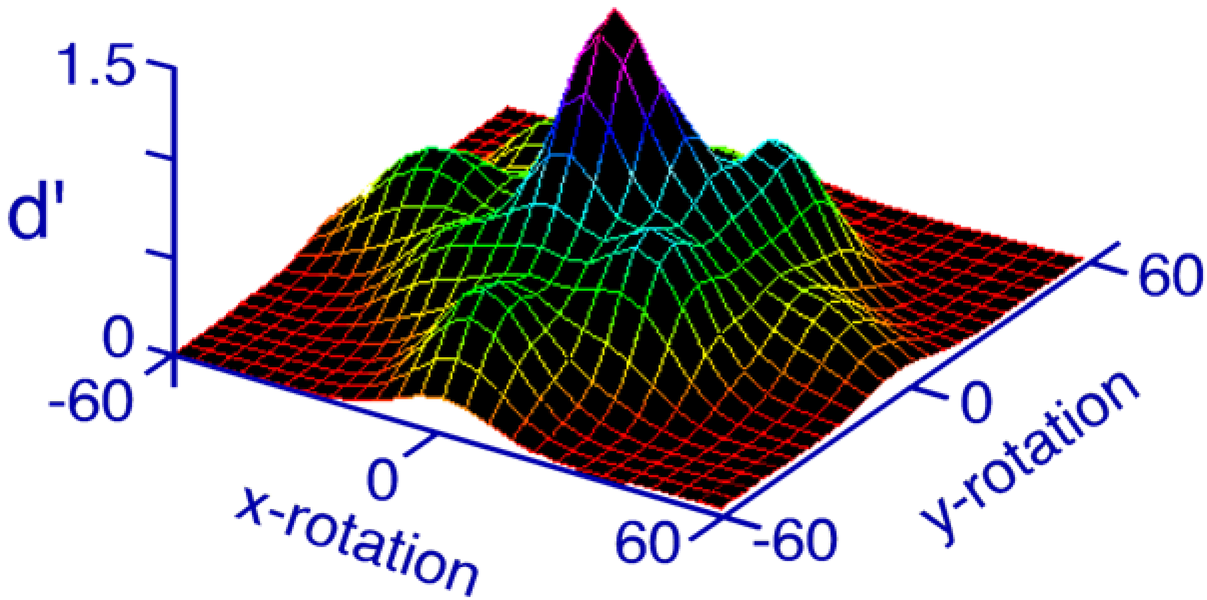
E. Bricolo

MIT PhD Thesis

stereo

only one test per target

distractor= target + noise



10 subjects

150 target objects

25 viewpoints

23% distractor noise

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

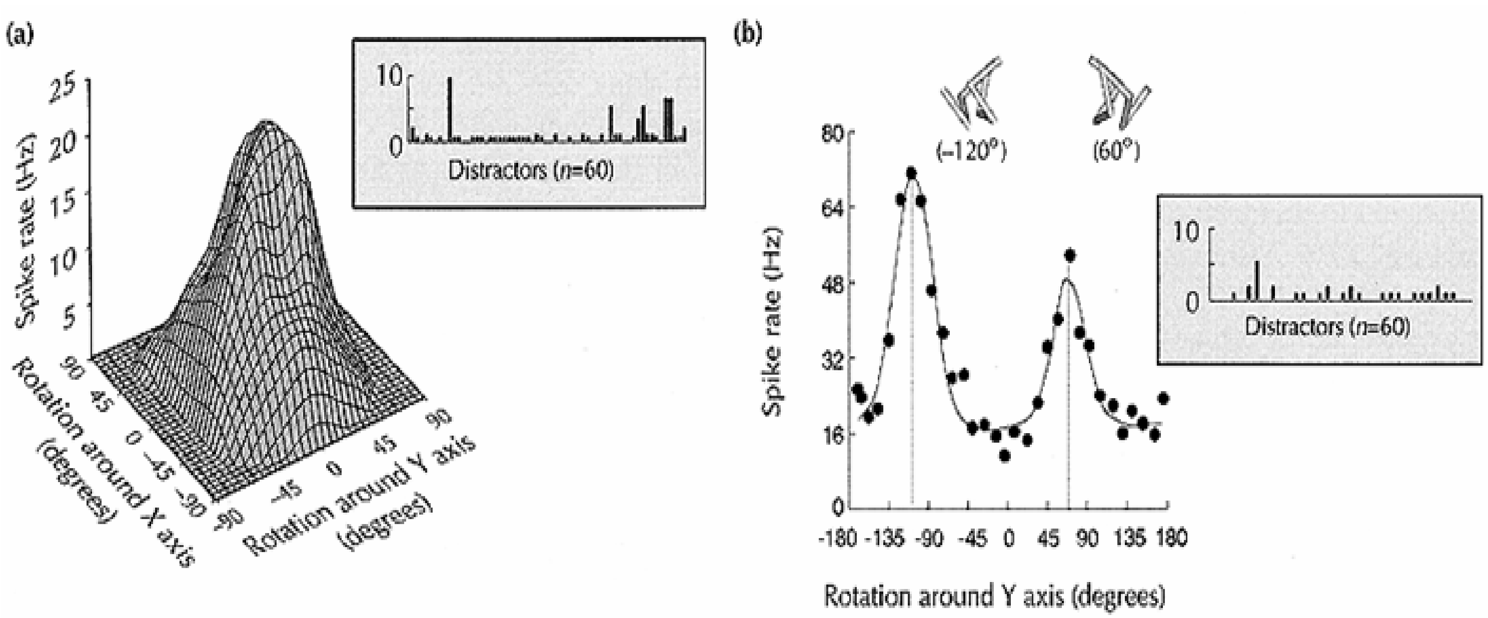
View-specific Paperclip Neurons

Logothetis, Pauls, Bülthoff, Poggio

Current Biology 4, 401-414 (1994); 5, 552-563 (1995)

Limited Generalization

Symmetric Views



Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

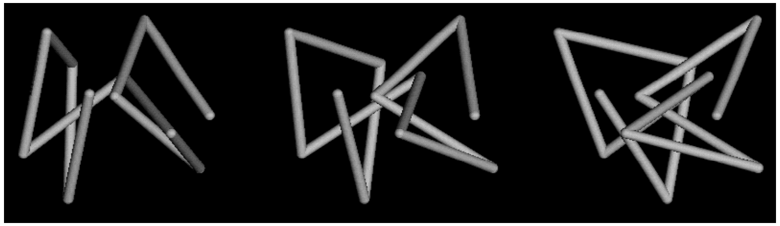
Temporal Association

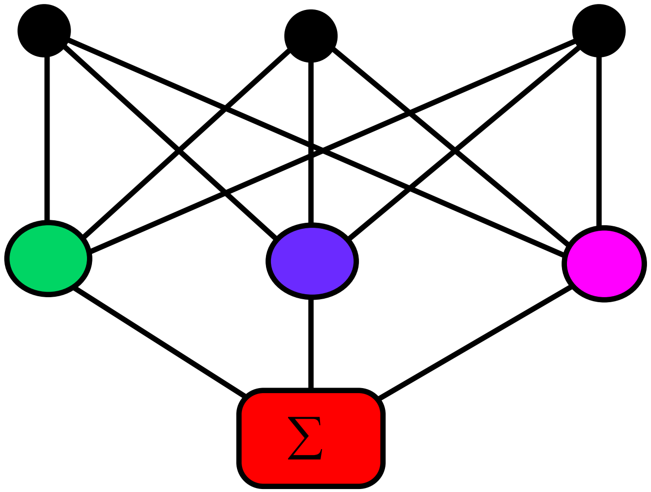
Computer Vision

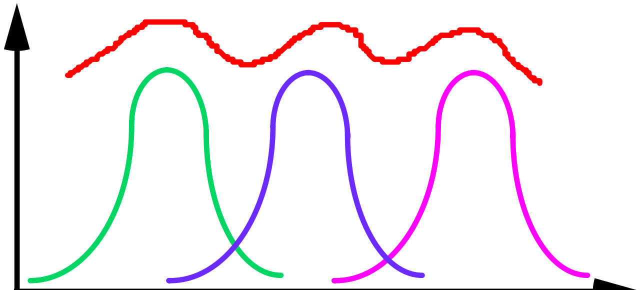
A View Interpolation Network

Poggio & Edelman

Nature, 343, 263-266, 1990







Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Demonstration for Image-based Recognition

What's in this picture?

A bit more information

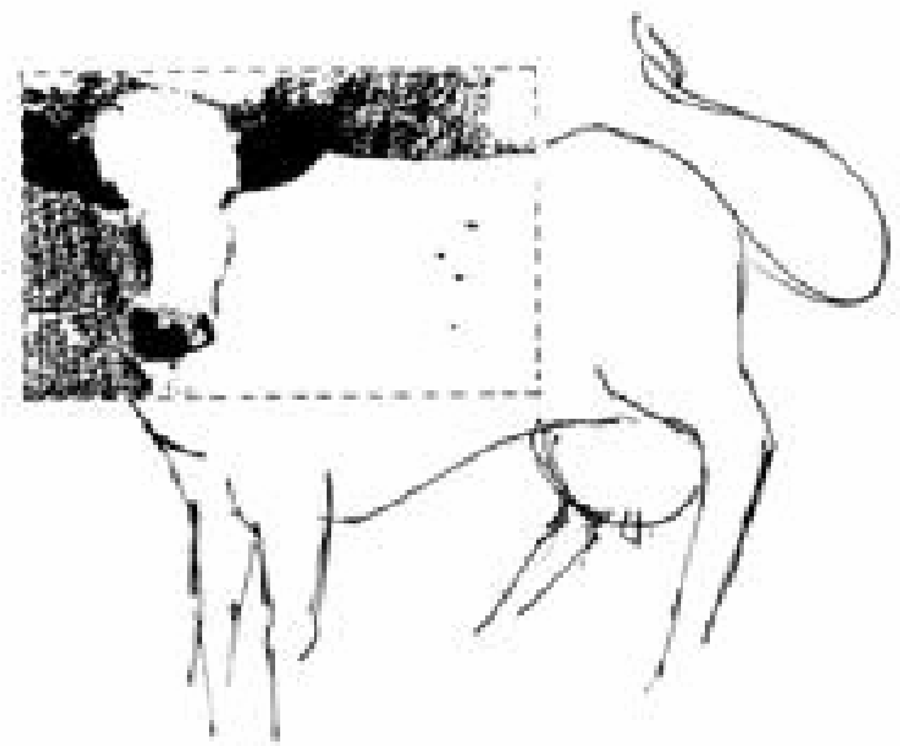


Image-based Recognition

What do you see now?

- Recognition is not bottom-up
- 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

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Dalmatian Dog

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

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

Provocative question

- All the information for recognition is in the image, so why do we need 3D structural descriptions for recognition?
- The answer is:
- We don't, if we use image-like representations and trade memory for computation.

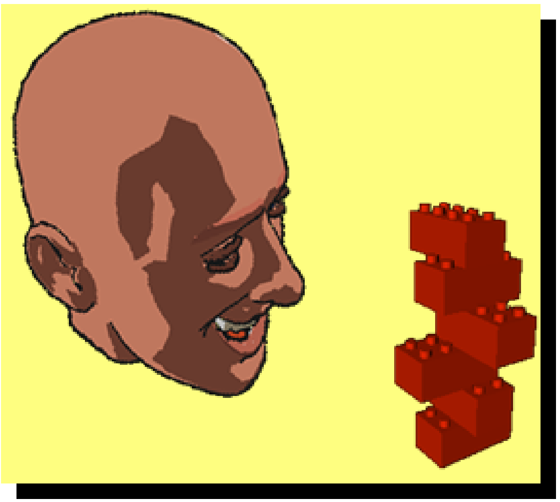
Experimental Evidence for Image-based Object Recognition

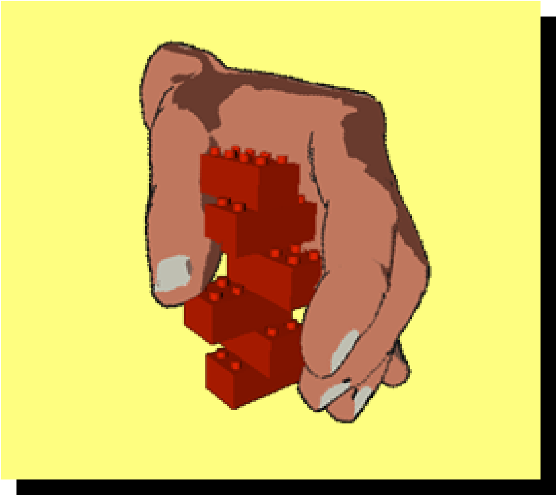
- Psychophysics
  - limited generalization to novel views in humans despite full 3D information
- Physiology
  - view-specific neurons in monkeys
- Theory
  - learning from examples
  - appearance-based computer vision
- More evidence for image-based recognition
  - *Object Recognition in Man, Monkey & Machine Tarr & Bülthoff, MIT Press*

Visual and Haptic Recognition

Newell, Ernst, Tjan & Bülthoff

Psychological Science, 12, 37-42, 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
- Common representation?
  - Cross-modal transfer ?

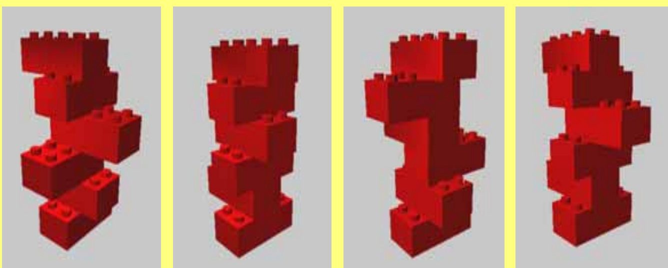
Haptic Recognition of 3D Lego™ Objects

- Objects made from identical Lego™ parts
  - 6 red bricks (8-dot)
  - 32 different objects
- Discriminable only by the configuration of the bricks
  - No advantage of color, texture or weight for vision or haptic
- Test in different orientations

Old/new Recognition Paradigm

Learning Session

Target Objects

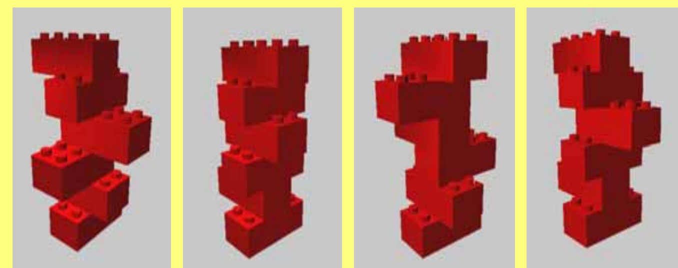


>> 4 Target Objects presented sequentially

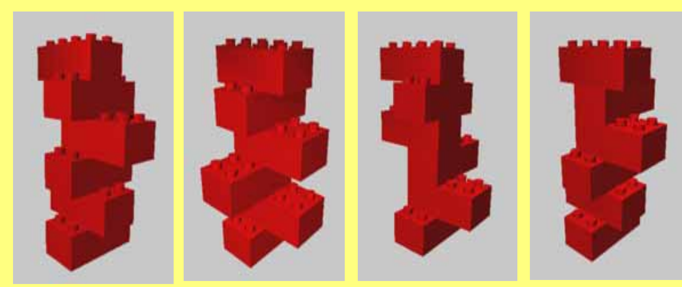
>> Familiarization time
30 s for visual learning
60 s for haptic learning

Testing Session

Target Objects



+ Distracter Objects



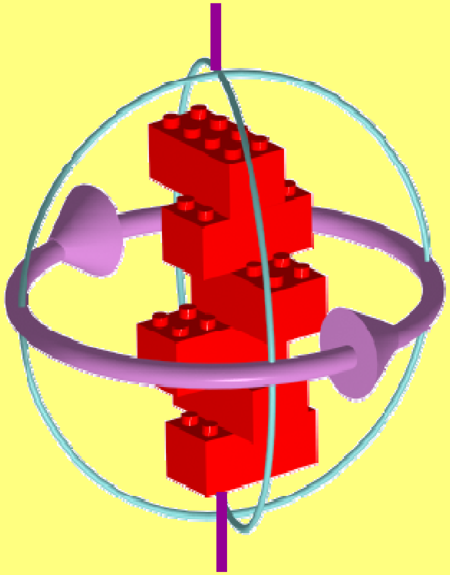
>> presented in random order
>> unlimited presentation time
>> 12 trials per block

Experimental Design

Testing Conditions

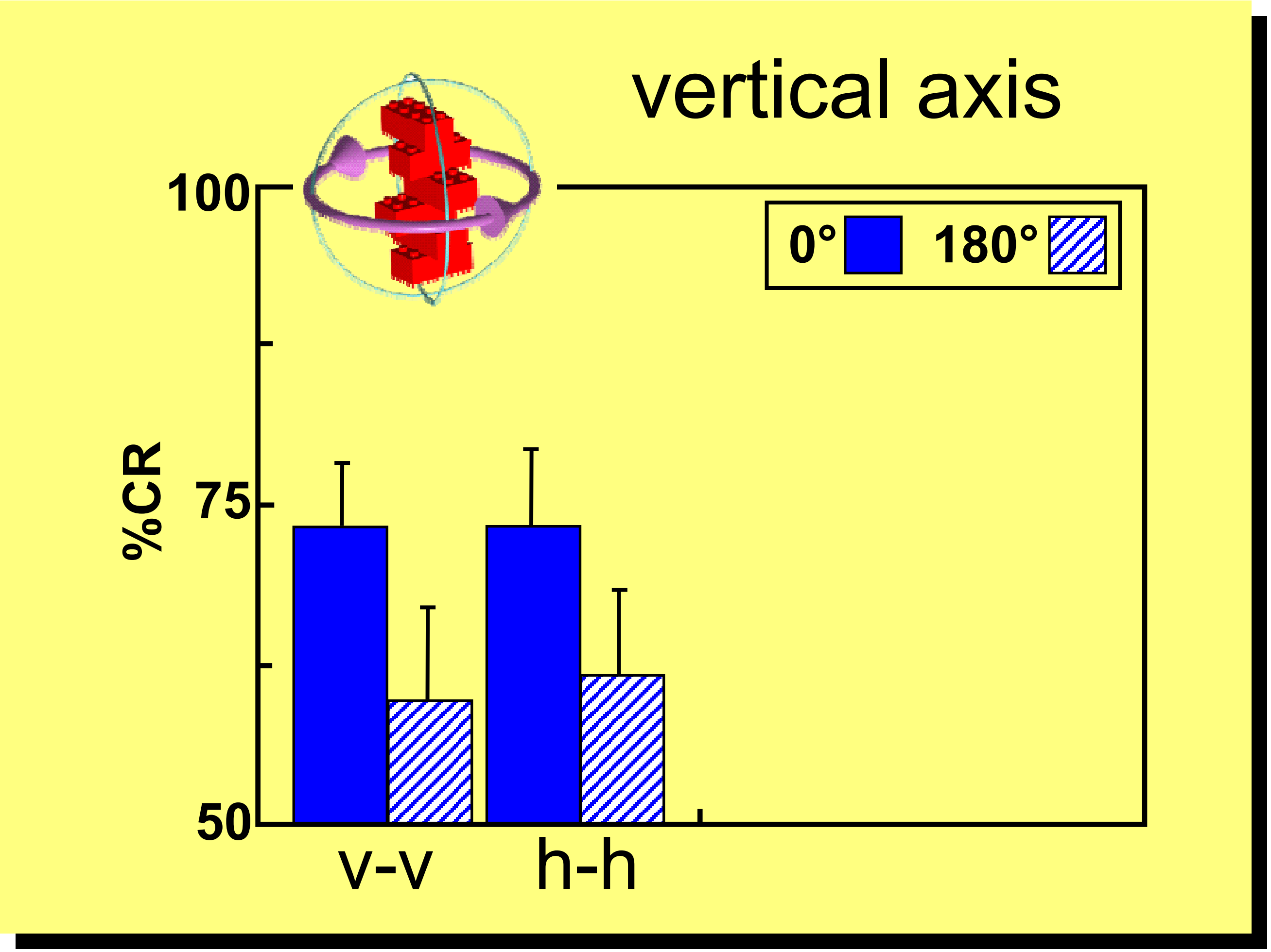
Learning	-	Test		
1. visual	-	visual	(v-v)	Within Modal
2. haptic	-	haptic	(h-h)	
3. visual	-	haptic	(v-h)	Cross Modal
4. haptic	-	visual	(h-v)	

Rotations about vertical-axis

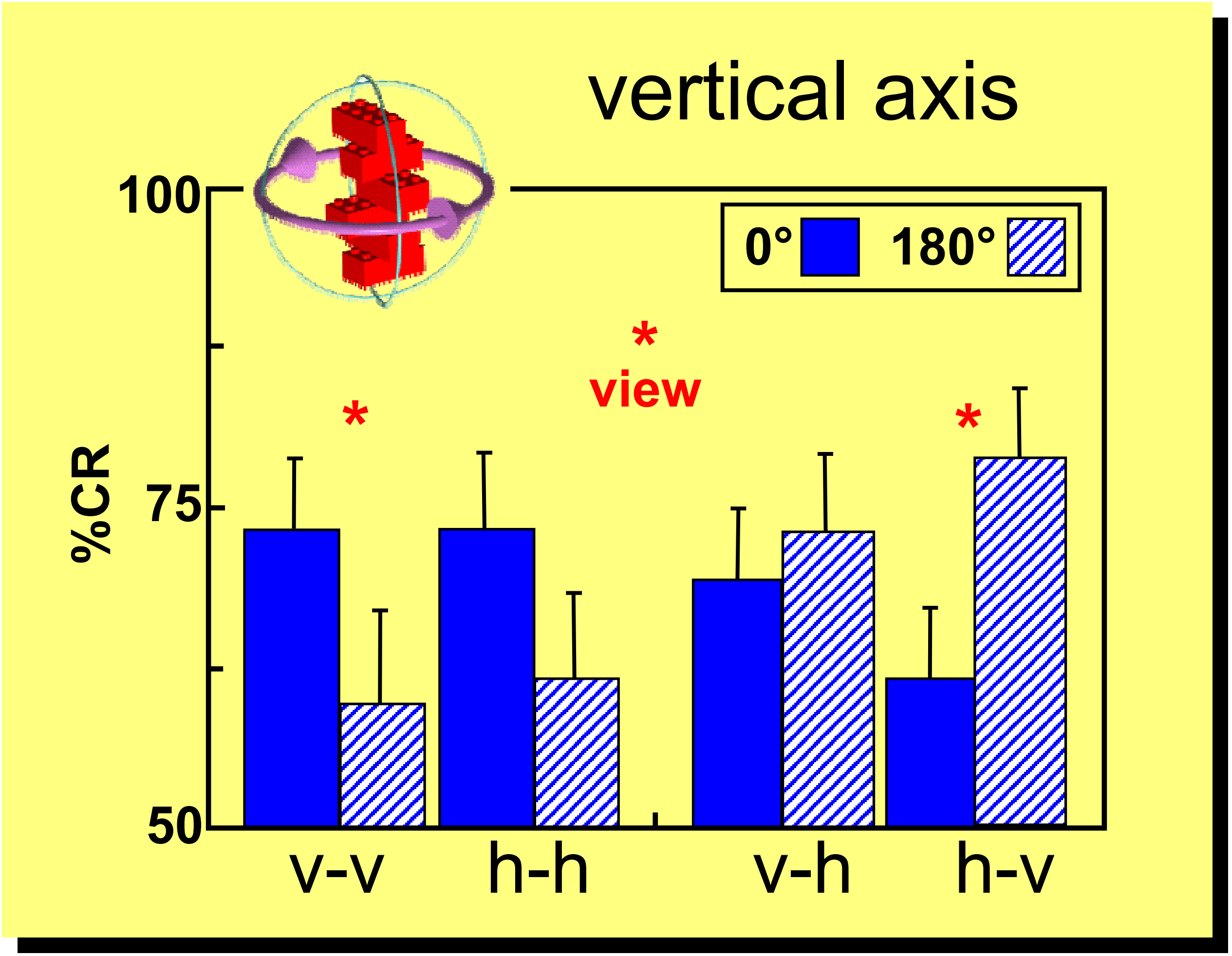


Test angles:
0° and 180°

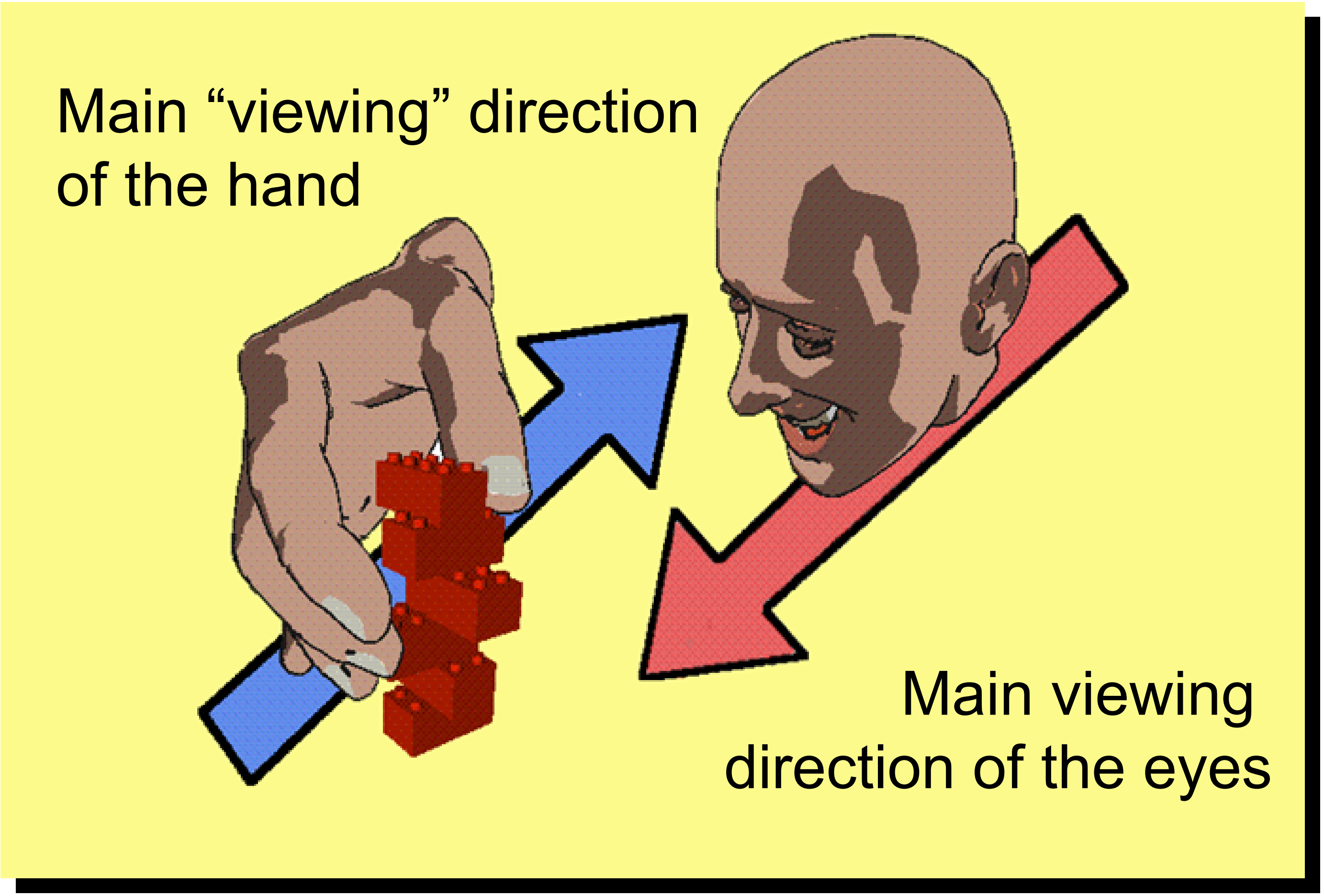
Rotation Around Vertical Axis
visual-visual & haptic-haptic



Cross-modal Transfer



The Visual and Haptic "View"

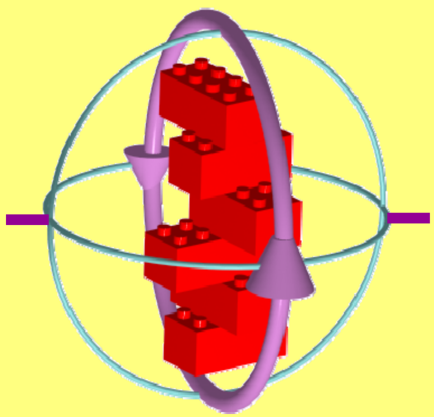


Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Testing the "Haptic View"

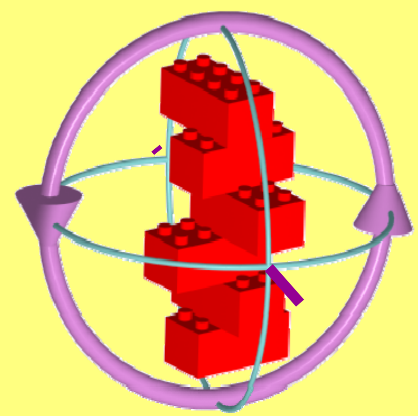
- Prediction: Whenever rotations involve a front/back change, the cross modal performance is better.

Horizontal axis



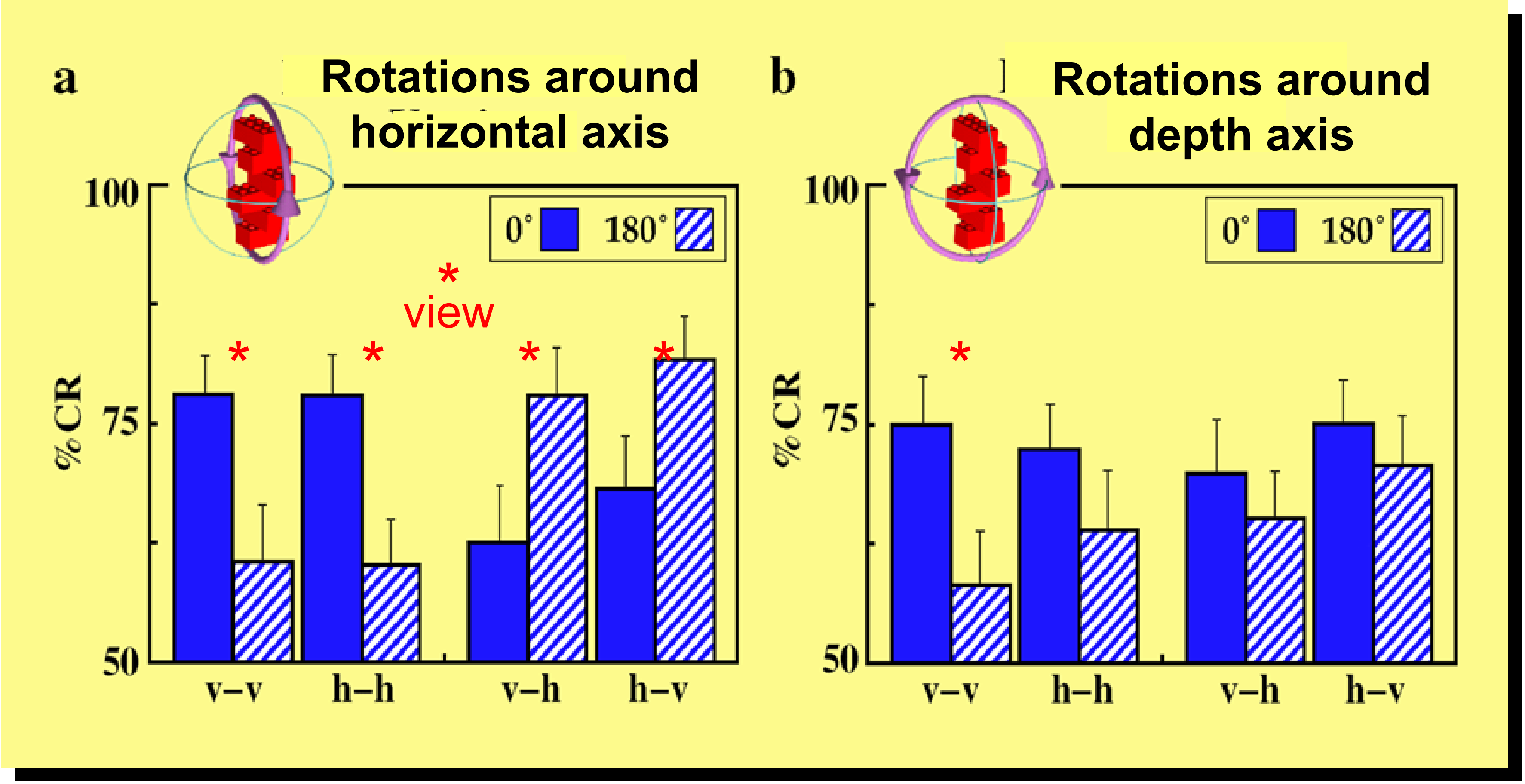
180° Rotations → **Front/ Back Change** Up/Down Change

Depth axis



180° Rotations → Left/Right Change Up/Down Change

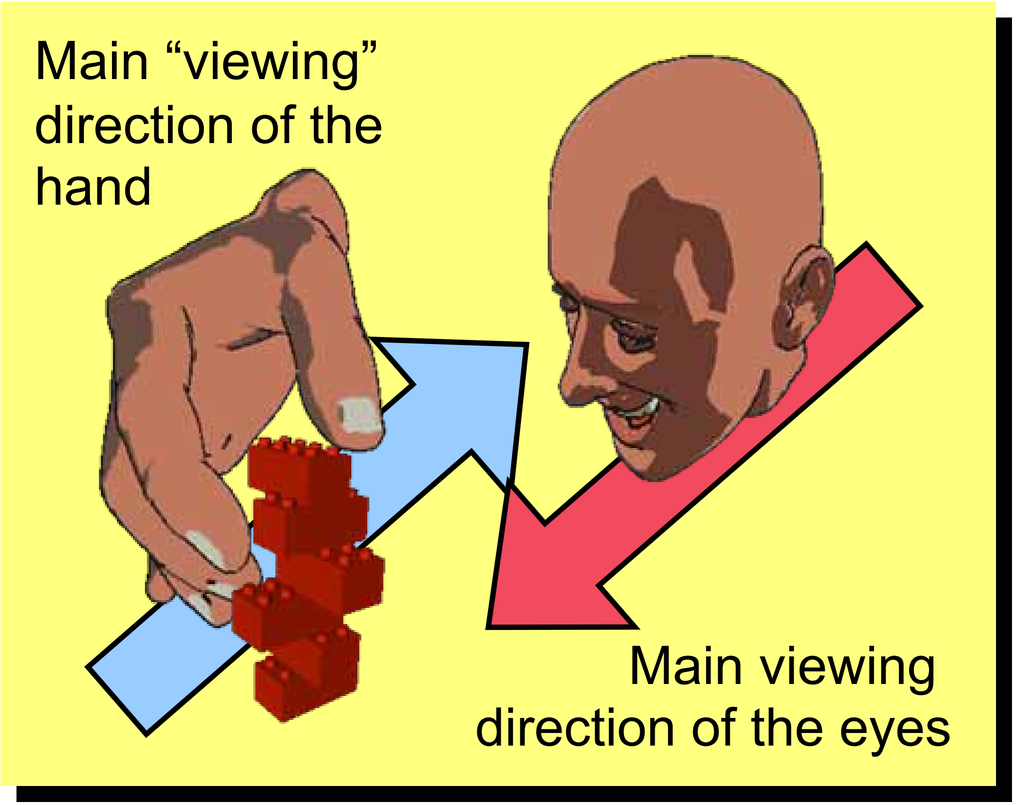
Recognition Performance:
Other Axes



Human Body Anatomy

Why are our hands linked to the arms
this way and not this way?

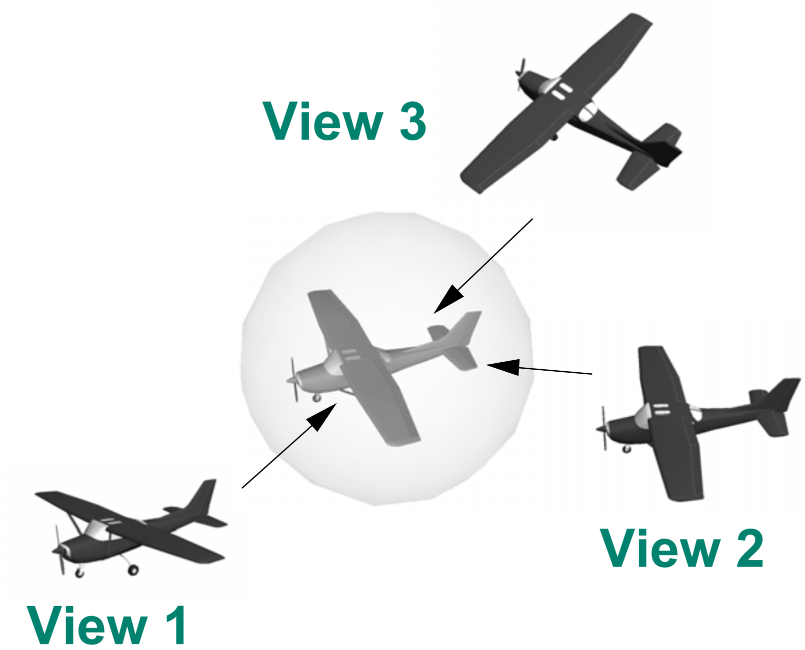
Integration of Information



Recognition becomes less viewpoint dependent, if the visual and haptic information is integrated into a common representation with cross-modal access.

The binding problem

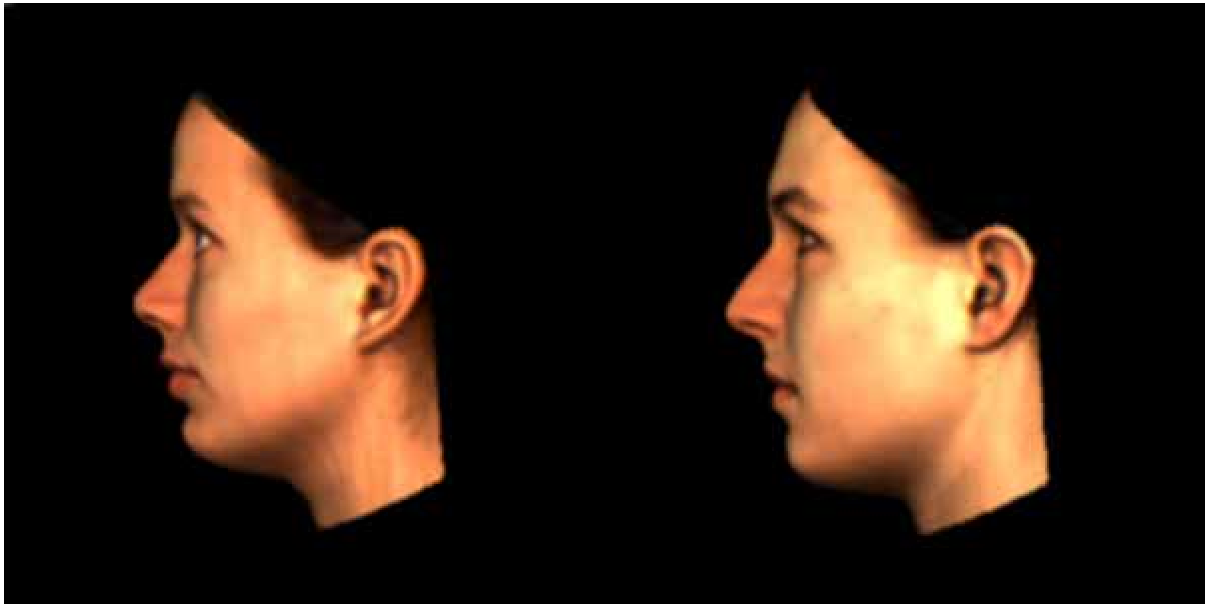
- Physical similarity can account for recognition with small viewpoint changes (view-based recognition)
- How does the brain know that different views of an object belong together?



Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Temporal Association Hypothesis

Temporal similarity can link many views to one object identity, because different views of objects are usually seen in close succession.



Same or different?

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

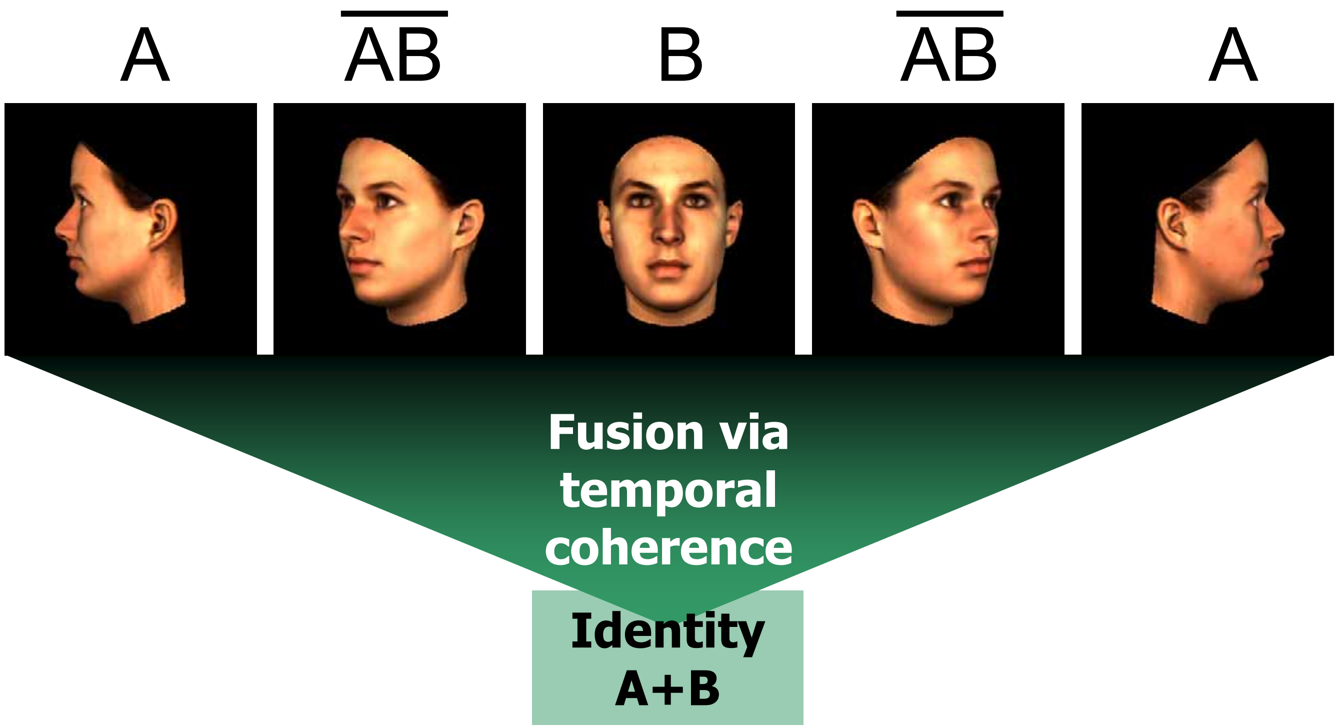
Computer Vision

Rotation with Identity Change

Wallis & Bülthoff

PNAS, 98(8), 4800-4804, 2001

The temporal association hypothesis predicts that morph sequences of a rotating head which changes identity from A to B should bind all images to one single person.



Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Tübingen 3D Face Database

Troje, Vetter, Blanz, I. Bülthoff, Knappmeyer, Kleiner

http://faces.kyb.tuebingen.mpg.de

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

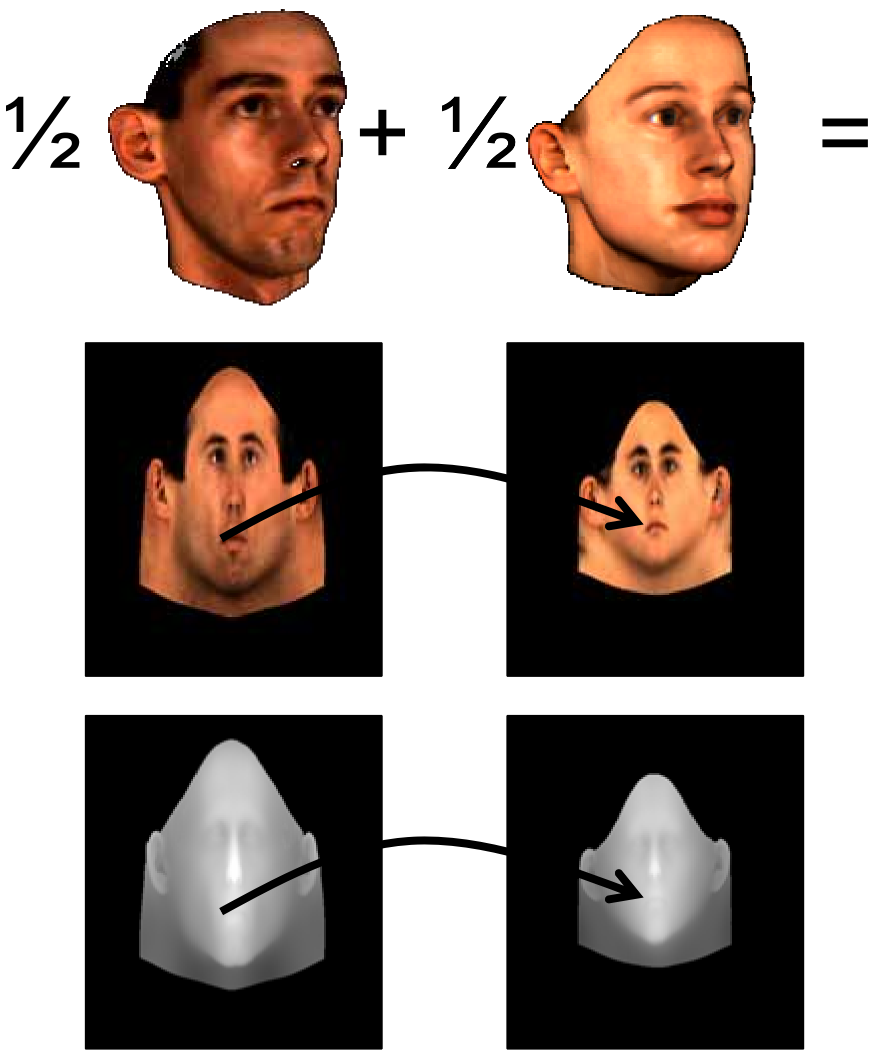
Experiments

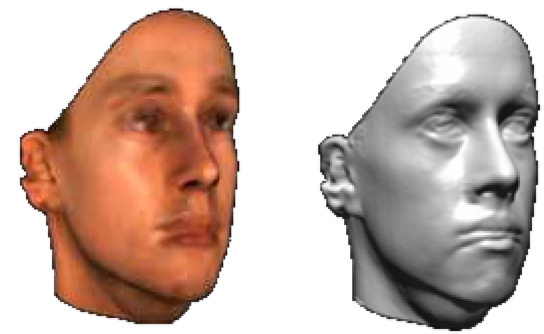
Binding Problem

Temporal Association

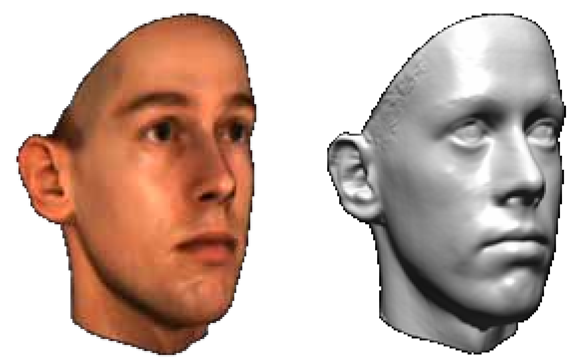
Computer Vision

Morphing of 3D Faces
Correspondence Problem





without correspondence

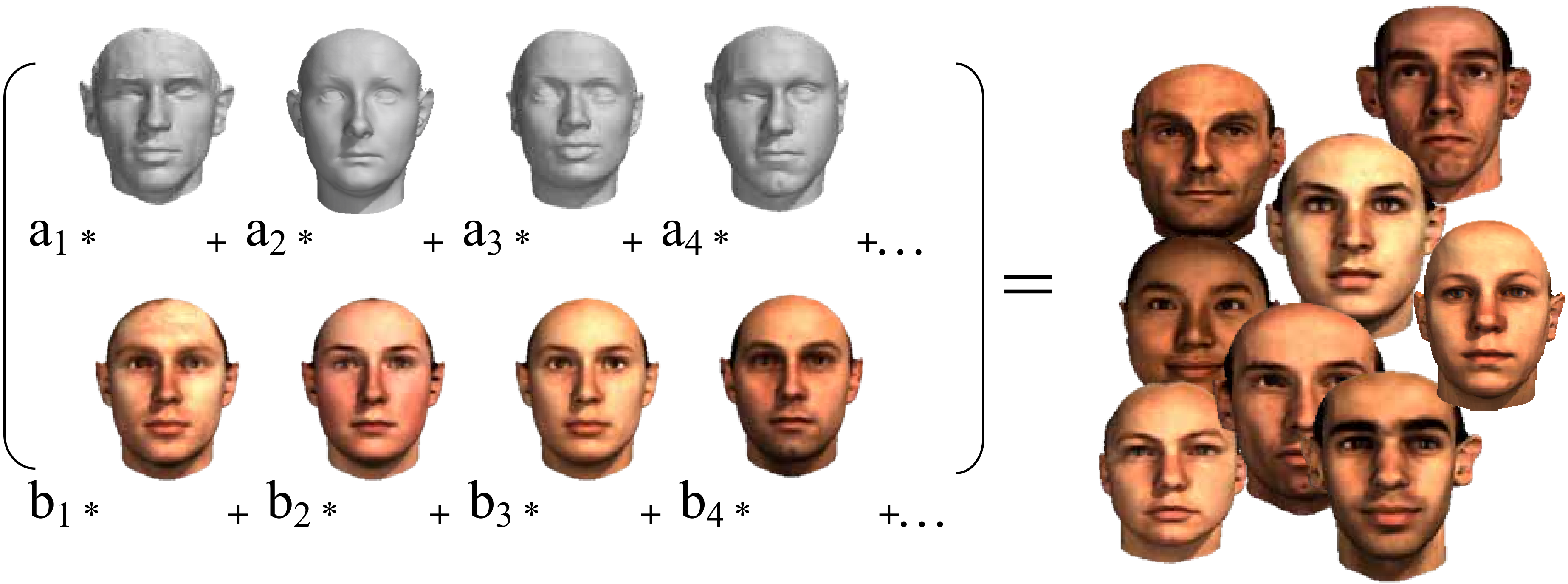


with 3D correspondence

Vector space of 3D faces

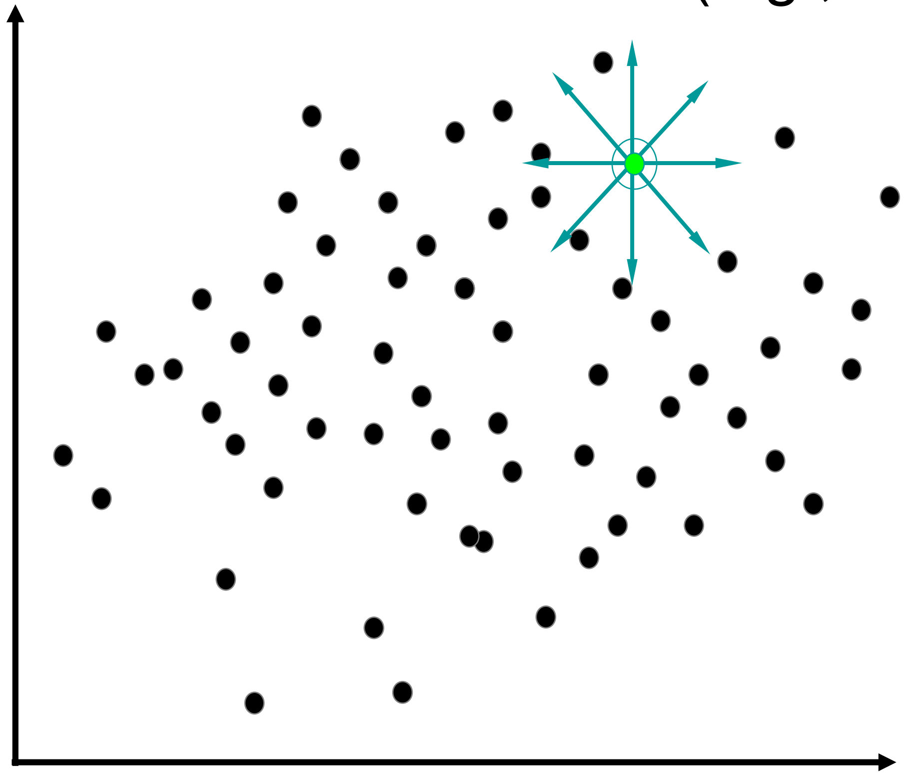
Thomas Vetter & Volker Blanz

- A Morphable Model can generate new faces and facial expressions.



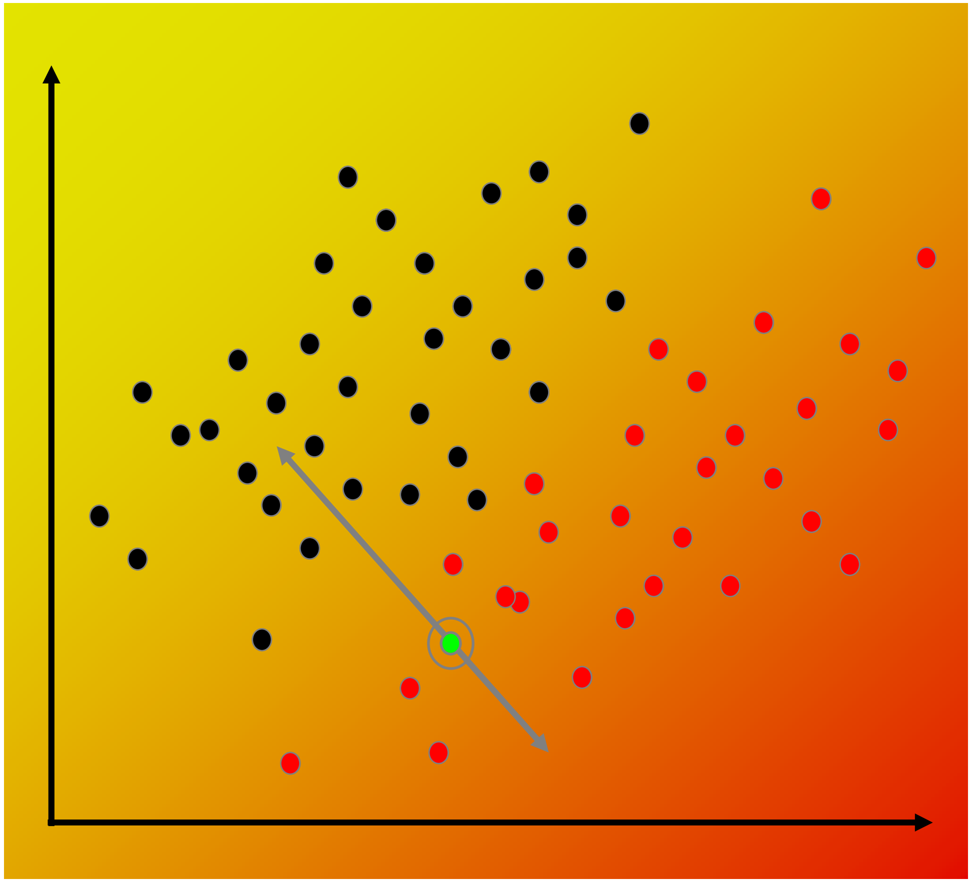
Modeling the Appearance of Faces

- Face is represented as a point in Face Space
- Direction codes for Face Attributes (e.g., Gender)



Learning from Labeled Faces

Fitting a regression function

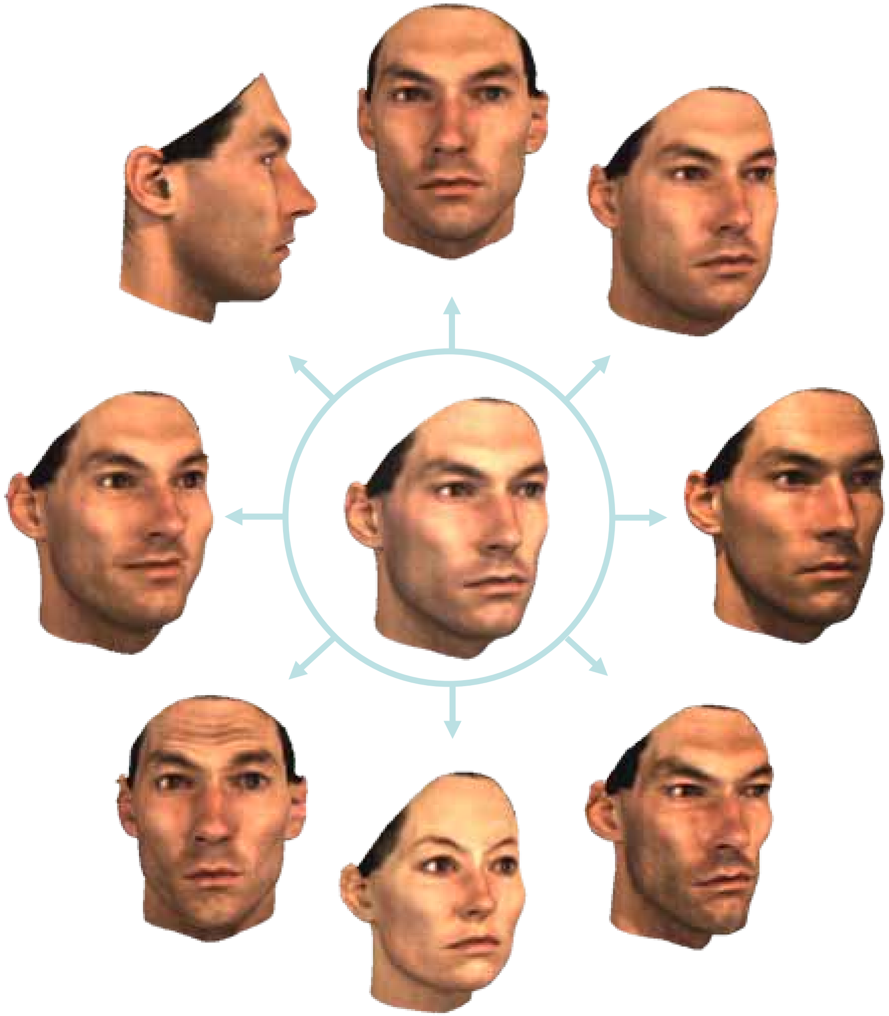


Background: Morphable 3D Faces

Thomas Vetter & Volker Blanz

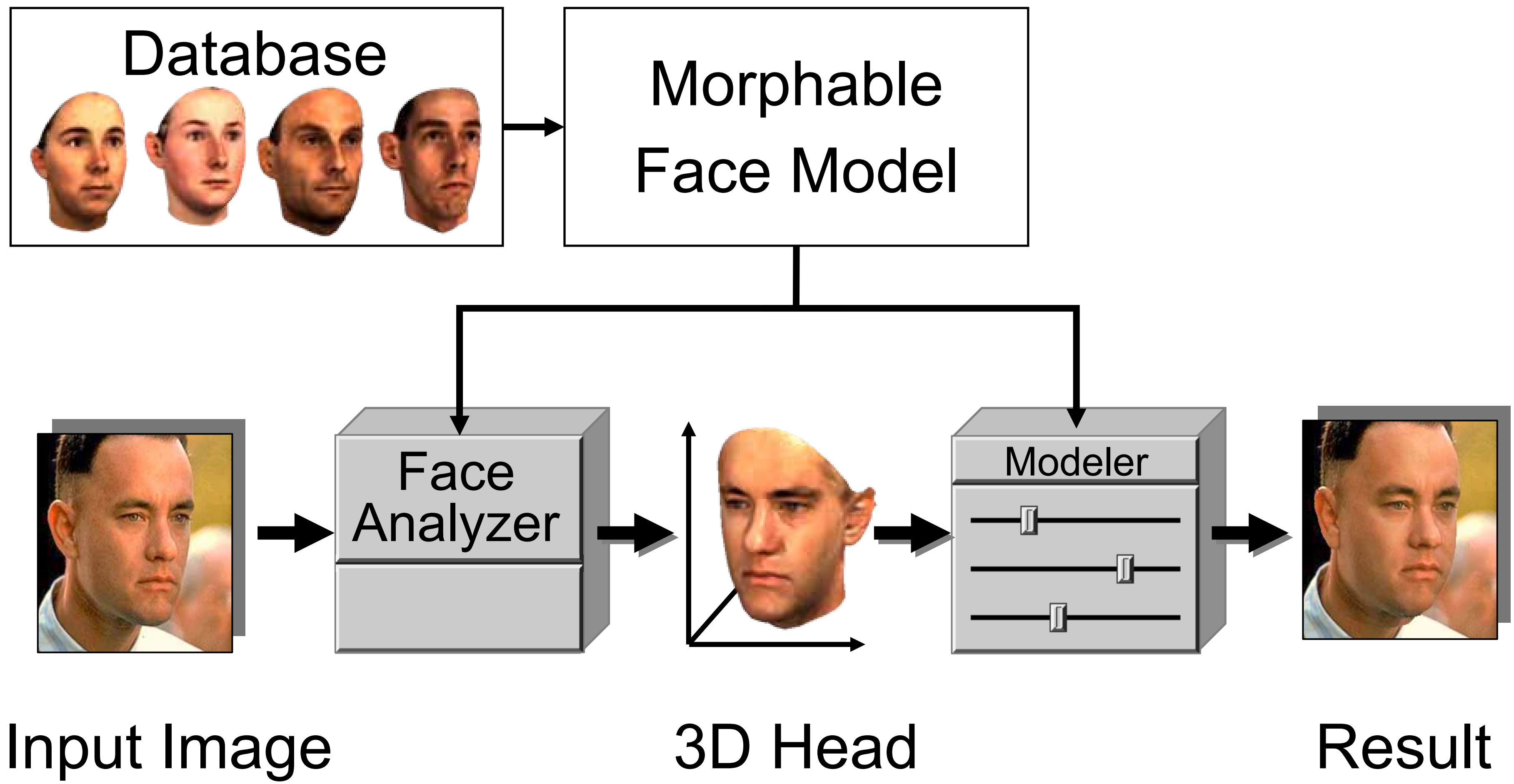
➢ From a single image

- Novel views
- Novel expressions
- Synthesis of siblings
- Change of illumination
- Variations of body weight



Synthesis of Faces

Volker Blanz & Thomas Vetter



Synthetic Actors

Tom Hanks

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

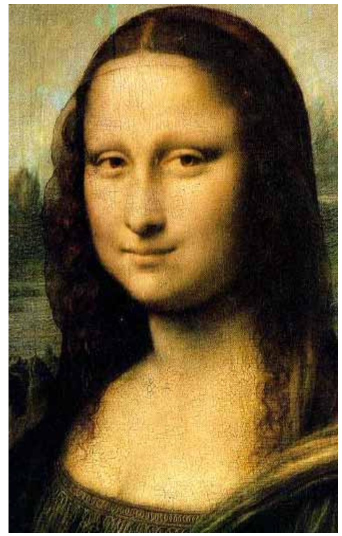
Binding Problem

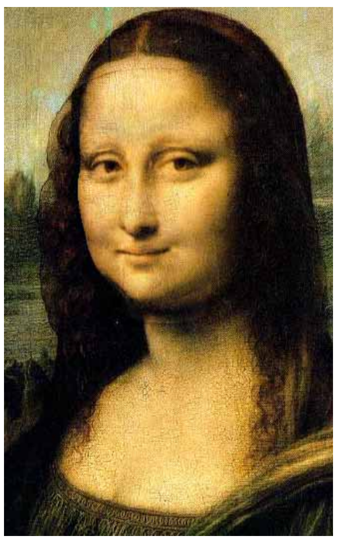
Temporal Association

Computer Vision

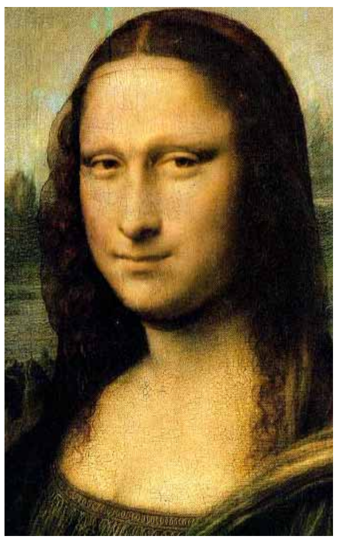
Mona Lisa Variations

Volker Blanz

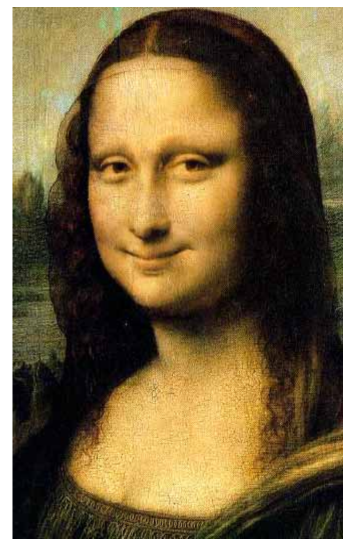




more female



more male



more friendly

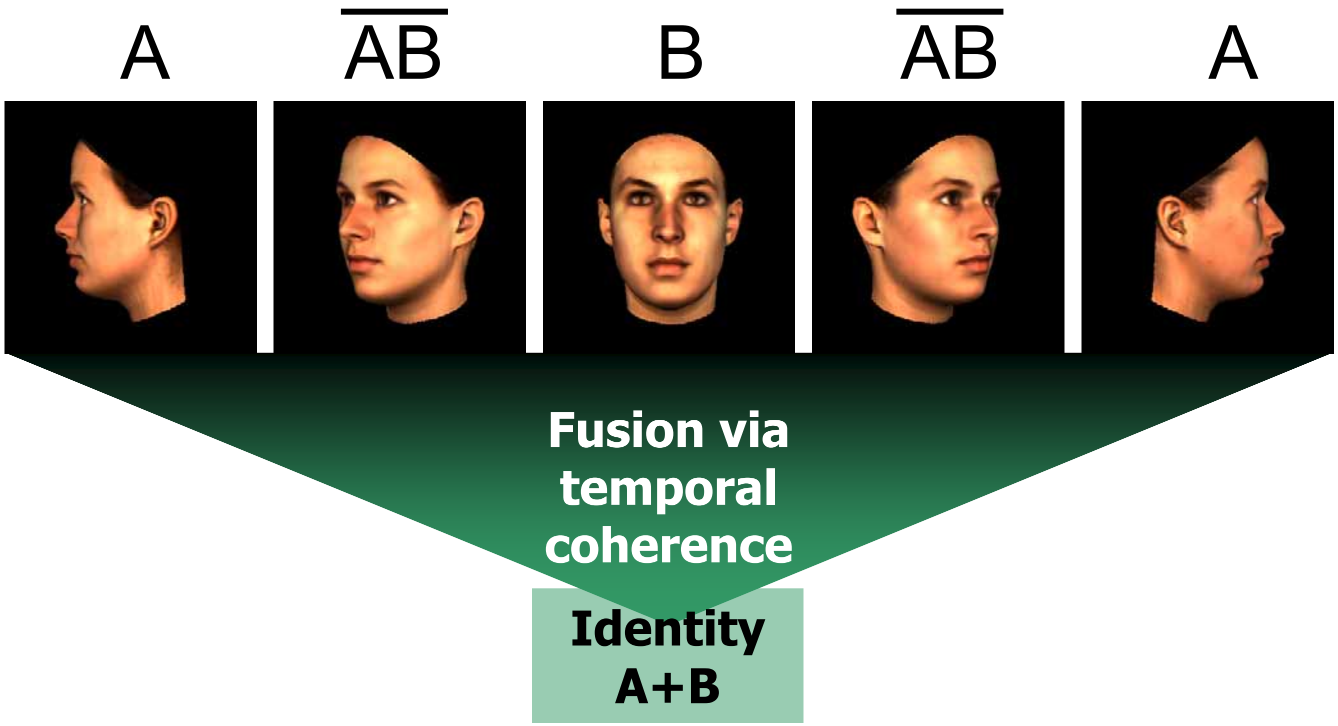
more attractive

Rotation with Identity Change

Wallis & Bülthoff

PNAS, 98(8), 4800-4804, 2001

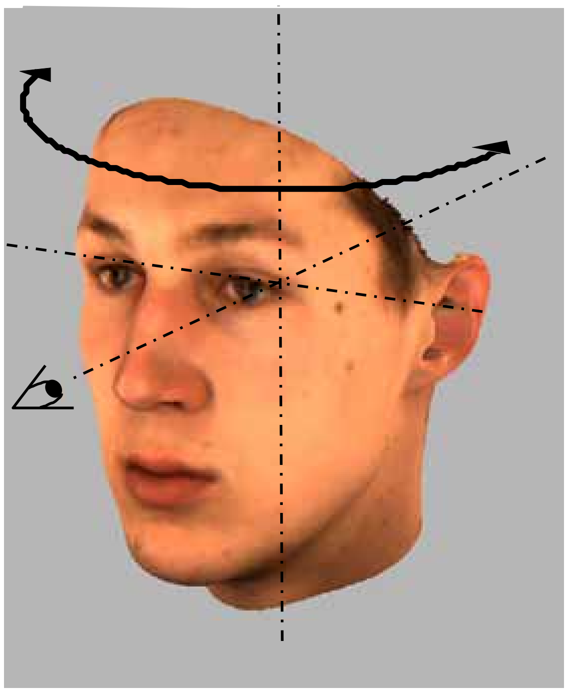
The temporal association hypothesis predicts that morph sequences of a rotating head which changes identity from A to B should bind all images to one single person.



Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Training: with Morph Sequences

TRAINING



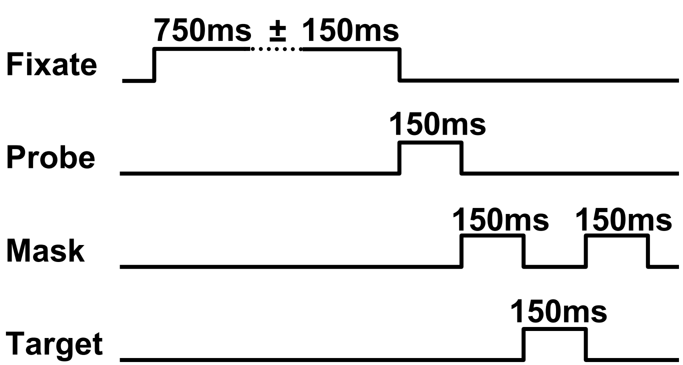
- 36 morph sequences
- each image shown for 300ms and
- immediately replaced by next image
- rotation from left to right profile and back
- two back and forth rotations per sequence

Testing: Same/Different Task

TESTING

Probe

Target



Match-to-sample testing with two groups

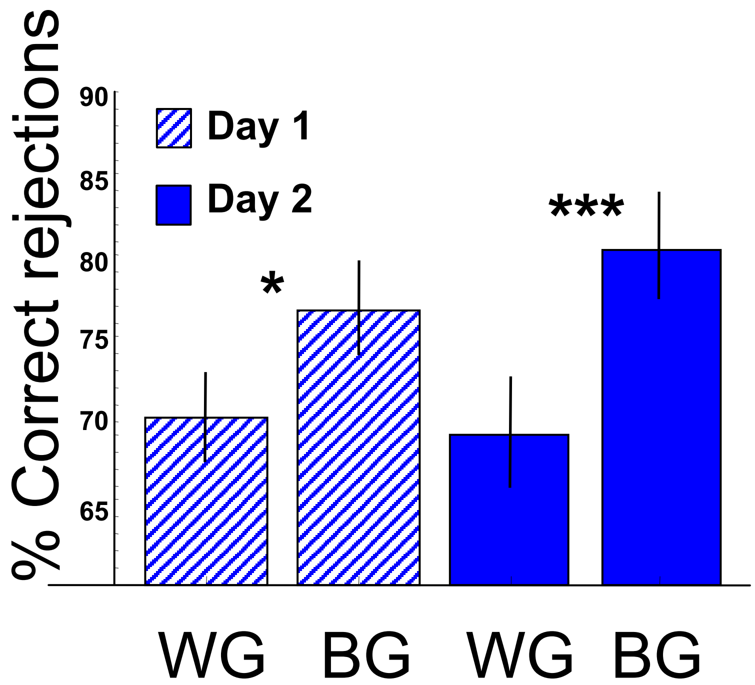
Within Group (WG) faces have been seen within a morphing sequence

Between Group (BG) faces have not been seen within a morphing sequence

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Wrong associations

- WG faces which have been seen together but belong to different persons are classified as different (correct rejections) less often.



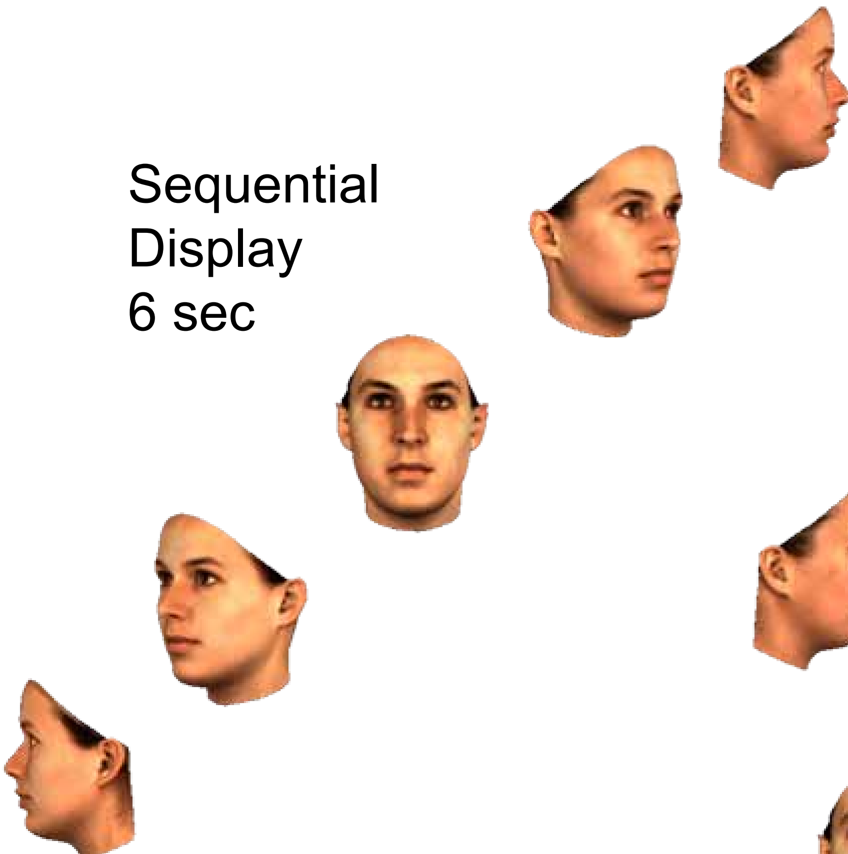
Temporal association ?

- Maybe seeing the intermediate morphed faces confused already the identity of WG faces.
- A further test of the temporal association hypothesis compared static with dynamic displays

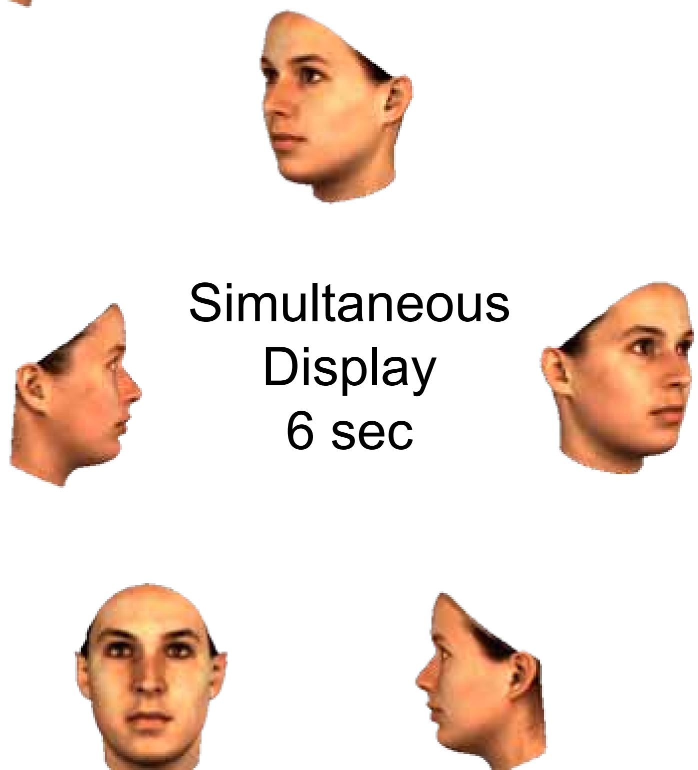
Sequential vs. simultaneous

Experiment 2

Sequential
Display
6 sec



Simultaneous
Display
6 sec



No confusion for static displays
Experiment 2

- There was no significant effect of group (WG or BG) for static displays
  $F(1,9)=0.133$, $P=0.724$.
- The simultaneous appearance and scanning of 5 faces was not sufficient to associate the WG faces

Randomized temporal order
Experiment 3

- There was also no significant effect of group (WG or BG) for randomized display order during training F (1,9)=0.044, P=0.839.
- The continuous but spatiotemporal disrupted presentation rendered training ineffectual (compared to the ordered spatiotemporal presentation).

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Facial Distortion
Experiment 4

- If one sees the frontal view of face **A** turn to the profile view of face **B**, there will be an associated subjective impression of a change in identity, and of facial distortion during rotation.
- Conversely, if no such change is detected then presumably the frontal and profile views must appear to belong to the same face.

Facial Distortion

Experiment 4

- Training: with 5 morph sequences.
- Testing: with these 5 sequences and 5 new sequences with true appearance of 5 of the 10 faces seen during training.
- Task: Report if heads changed form during rotation.
- Prediction: Morph sequences perceived as single face should appear rigid.

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Facial Distortion
Experiment 4

TRAINED

STIMULUS \ RESPONSE	DEFORMING	NON DEFORMING
DEFORMING	0.29	**0.71**
NON DEFORMING	**0.85**	0.15

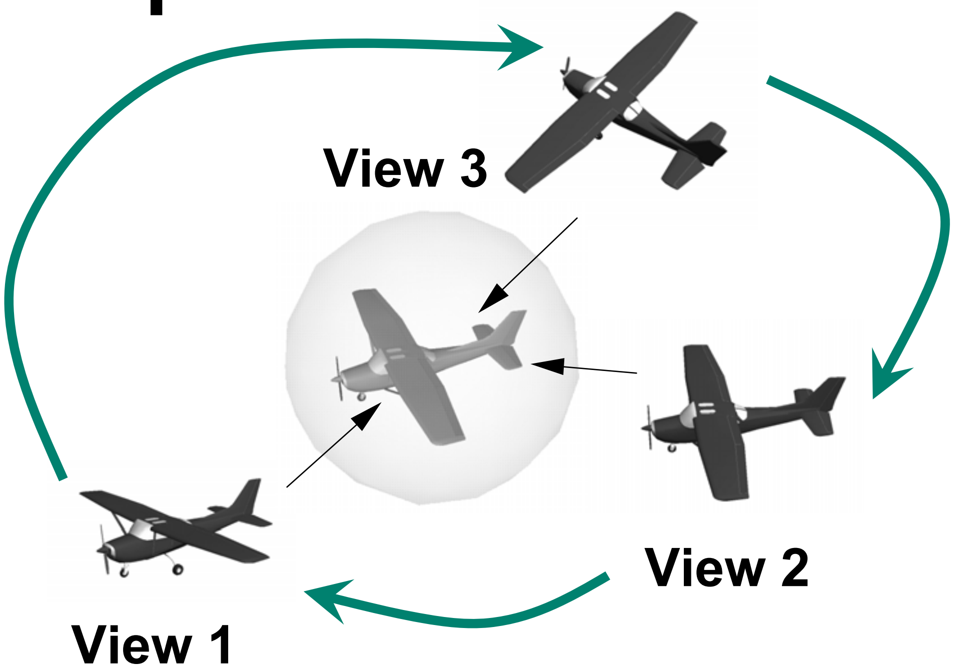
UNTRAINED

STIMULUS \ RESPONSE	DEFORMING	NON DEFORMING
DEFORMING	**0.63**	0.37
NON DEFORMING	0.34	**0.66**

- Original faces are deforming while faces from a morphing sequence appear rigid.
- The opposite is true for untrained faces.

Human Recognition Summary

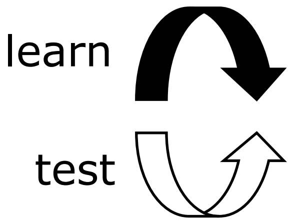
Our psychophysical experiments suggest that objects are represented as **collections of views** linked by **temporal association**



Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Additional Evidence for Temporal Associations

Stone 1998, 1999



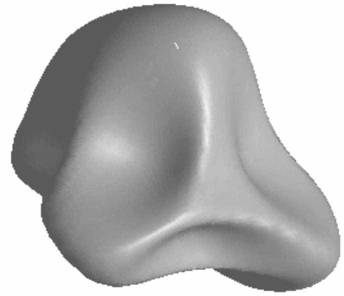


Figure 1: Example of a grey-level object.

Figure 2: Example of a textured object.

- Studies by Stone 1998,1999
  - Subjects learn objects rotating in one direction
  - During test, objects are either displayed as learned or in **reverse order**
  - Large performance loss for reverse condition
- Temporal characteristics form an integral part of learned **object representations**

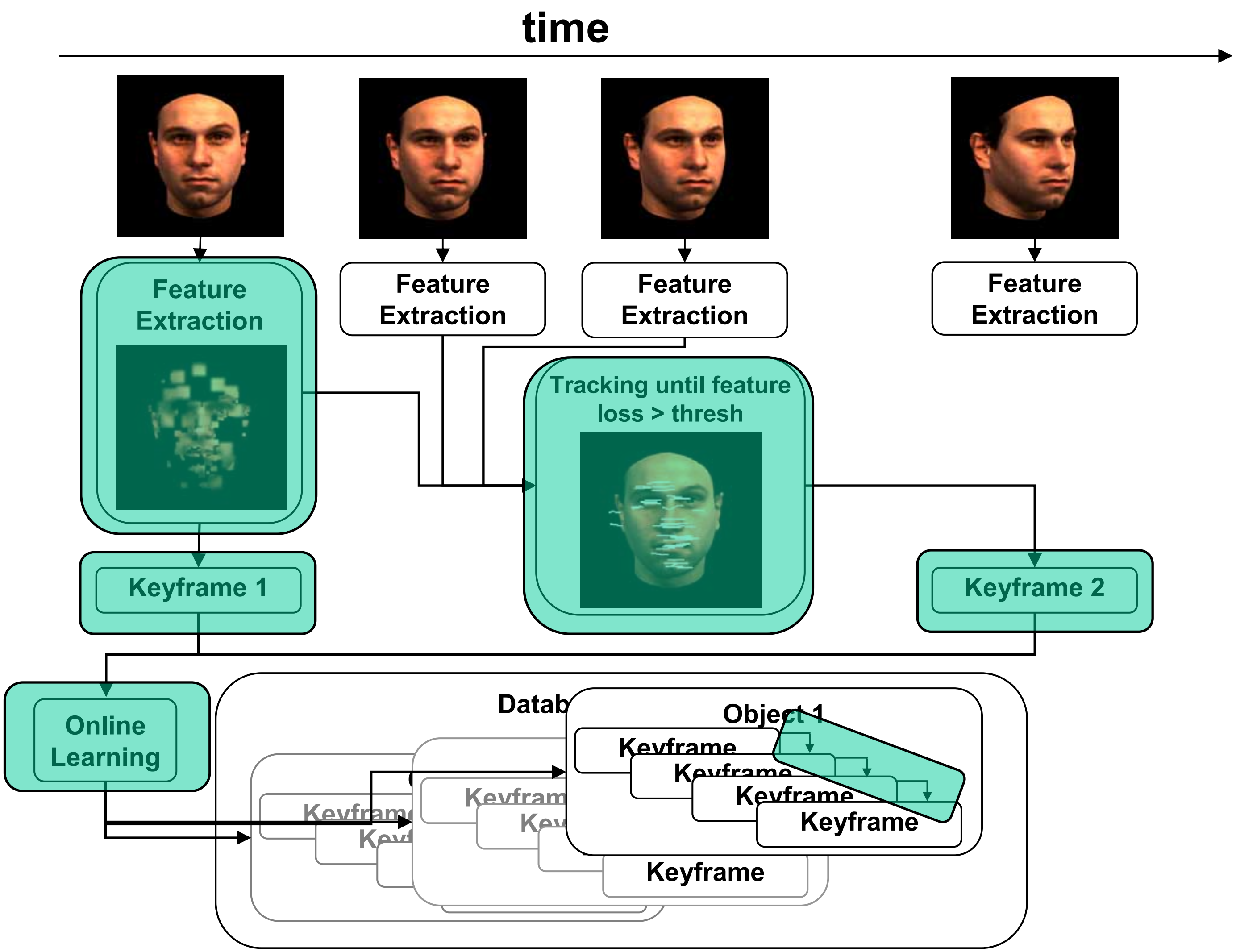
Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Further evidence for temporal association

- There is also good evidence for view-based recognition based on temporal association from single cell recordings
  - (e.g., Miyashita, 1998)
- automatic recognition system based on temporal association of views
  - CogVis project (FP5)
  - PhD project C. Wallraven

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Keyframe-based recognition system
Christian Wallraven



Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Feature matching
(motivated by Pilu 1997)

- Used both for tracking *and* recognition
- Constructs similarity matrix A with:

$$\mathbf{A}_{\mathbf{ij}} = \exp\left(-\frac{1}{\sigma_{dist}^{2}}\operatorname{dist}(i,j)\right)\cdot\exp\left(-\frac{1}{\sigma_{NCC}^{2}}\operatorname{NCC}(i,j)\right)$$

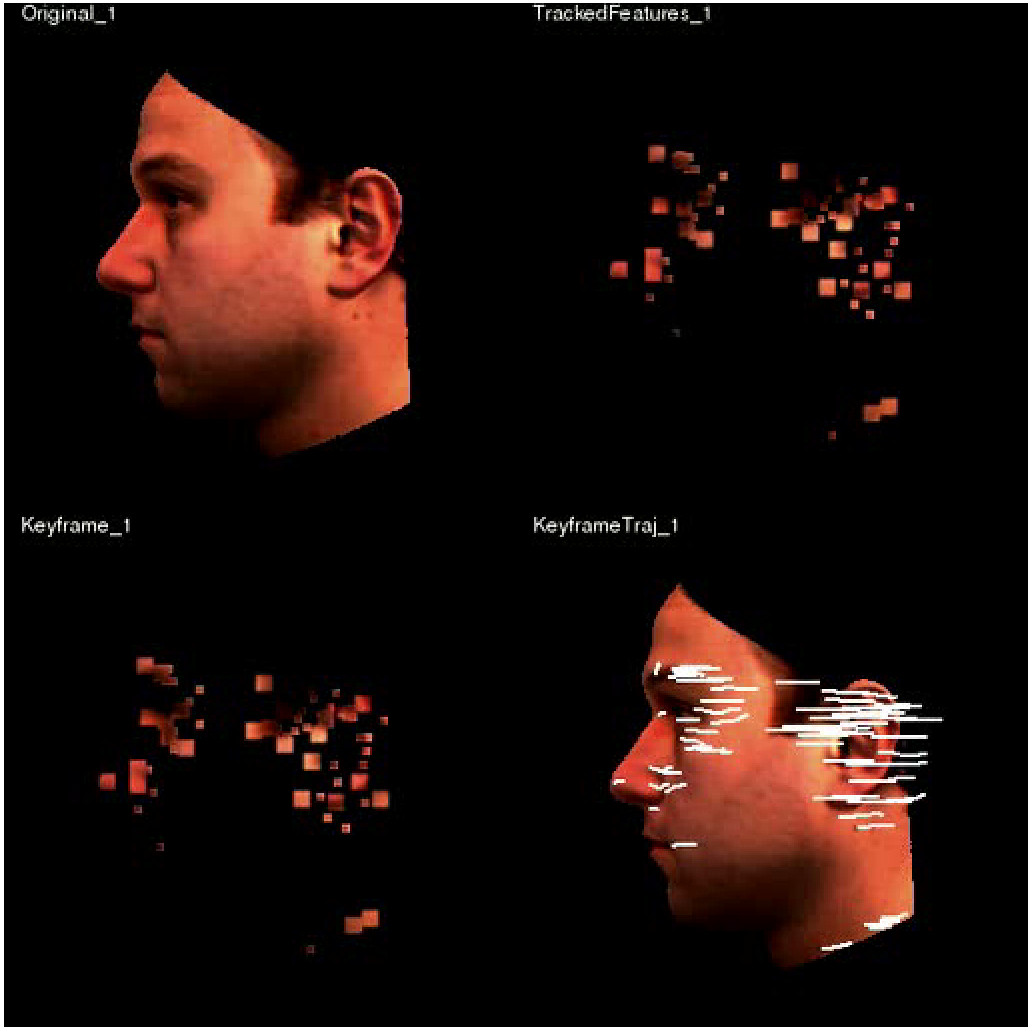
- Modified SVD of A provides a one-to-one mapping between features
- Tries to find matches maximizing *both*
  - configuration and image similarity
  - similarity to learned feature trajectories can be incorporated
- Both tracking and recognition is correspondence based

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Keyframes in Motion

Original images

Keyframes



Tracked features between keyframes

Feature trajectories between keyframes

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Keyframes as View-Graphs

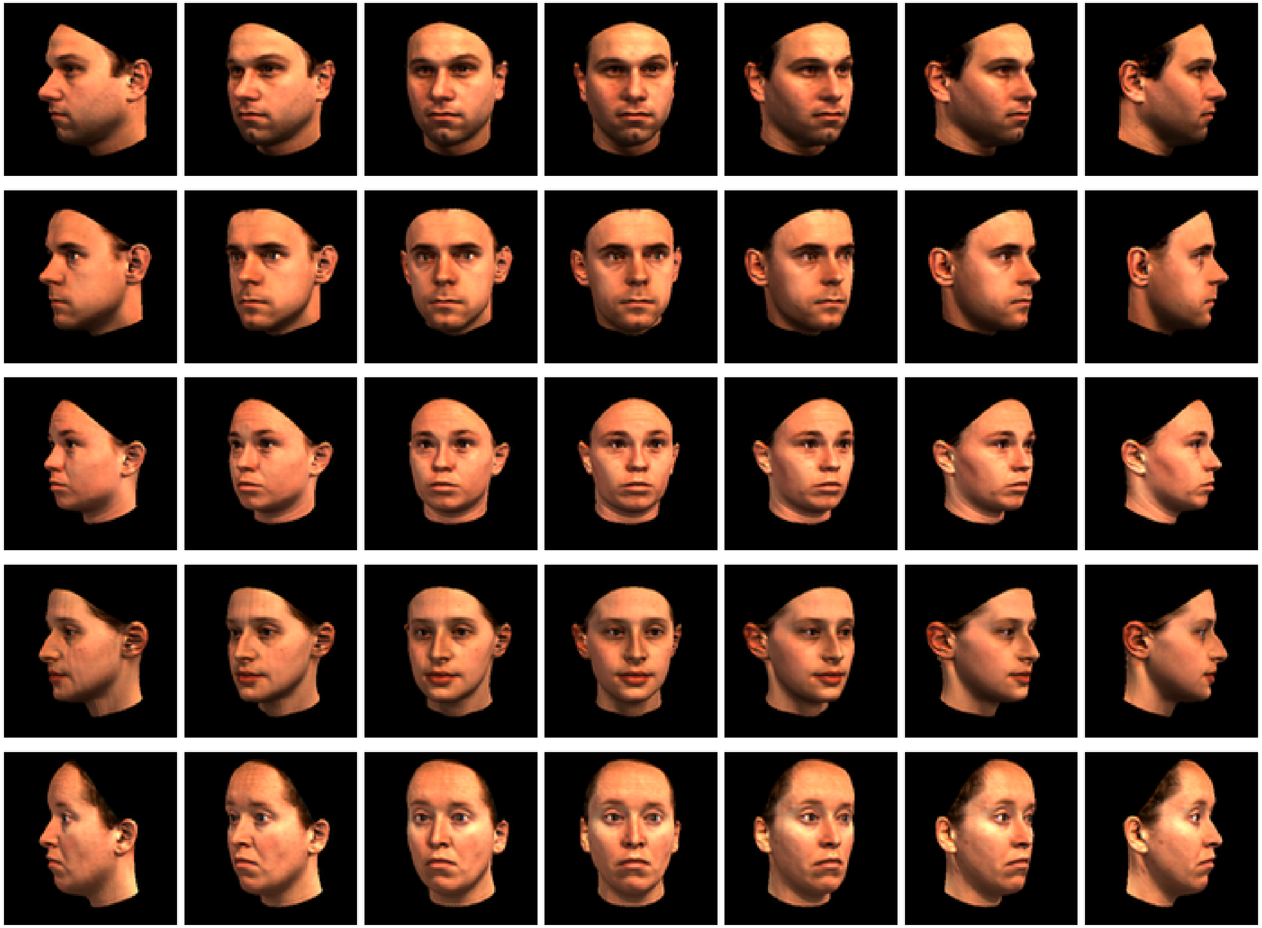
- Properties of keyframes
  - Selected in a bottom-up fashion (assumes no motion model)
  - Segments the sequence into temporally continuous chunks according to *visual events*
    - Capture motion complexity of the sequence
  - Forms a directed and connected view-graph
    - Extensible, view-based representation
    - Possible to model canonical views

Keyframes with Local Features

- Properties of tracked local features
  - In general: provide good compression
  - Access to feature *trajectories*
    - Allows analysis of image motion
    - Can be used as spatio-temporal matching priors
- Both keyframes and tracked features form a fingerprint of the sequence

Keyframes
Consistency

- Examples from rendered sequences



Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Keyframes
Consistency

MPI Face Database

Car Video Database

Recognition Results

- Tübingen face database
- 60 faces, 2 **different** illuminations, 2 **different** poses
- Recognition rate **98%**

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

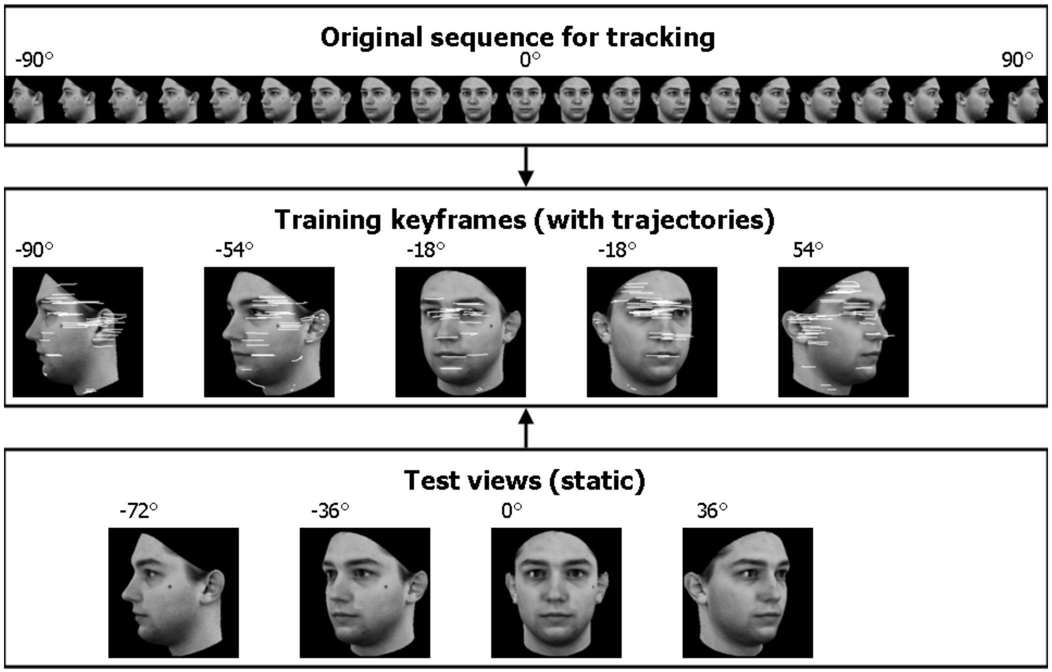
Experiments

Binding Problem

Temporal Association

Computer Vision

Recognition Results



Training: 100 face sequences
→ 5 keyframes

Testing: with 4 intermediate views which are part of the representation

Task: face identification

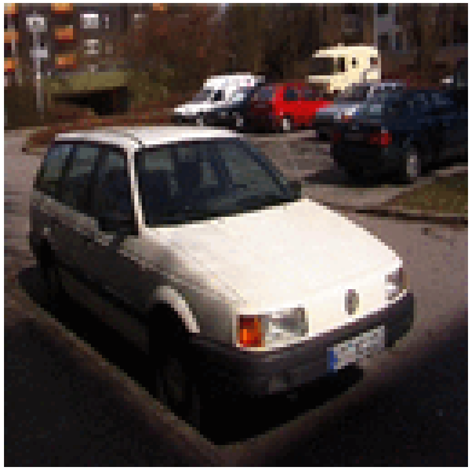
Spatio-temporal representation wins!!!

Method	Rec. error
Images only	37.5%
Features only	14.5%
Features + Trajectories	2.0%

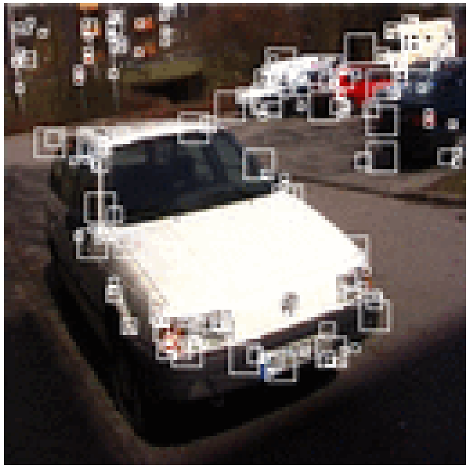
Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Keyframes
Compression

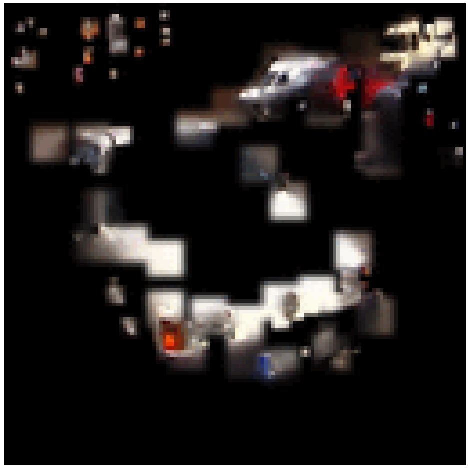
- Examples from video sequences
  - Small database of 20 car videos
  - Videos taken under un-controlled conditions
  - Compression rate over 99%



720 frames



21 frames



21 frames
only features

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision

Recognition Results

- Car Database
  - Variations in lighting, size, occlusion in real-world
  - 20 car videos, 50 test pictures with a digital camera
  - Recognition rate **88%**

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Keyframes

- provide a way of automatically acquiring scene representations suitable for recognition
- main criticism against image-based recognition is the storage requirement
  - key-frame technique provides a low dimensional representation of scenes
  - compression rate of over 99%

Learning multi-modal Object Representations through Active Manipulation

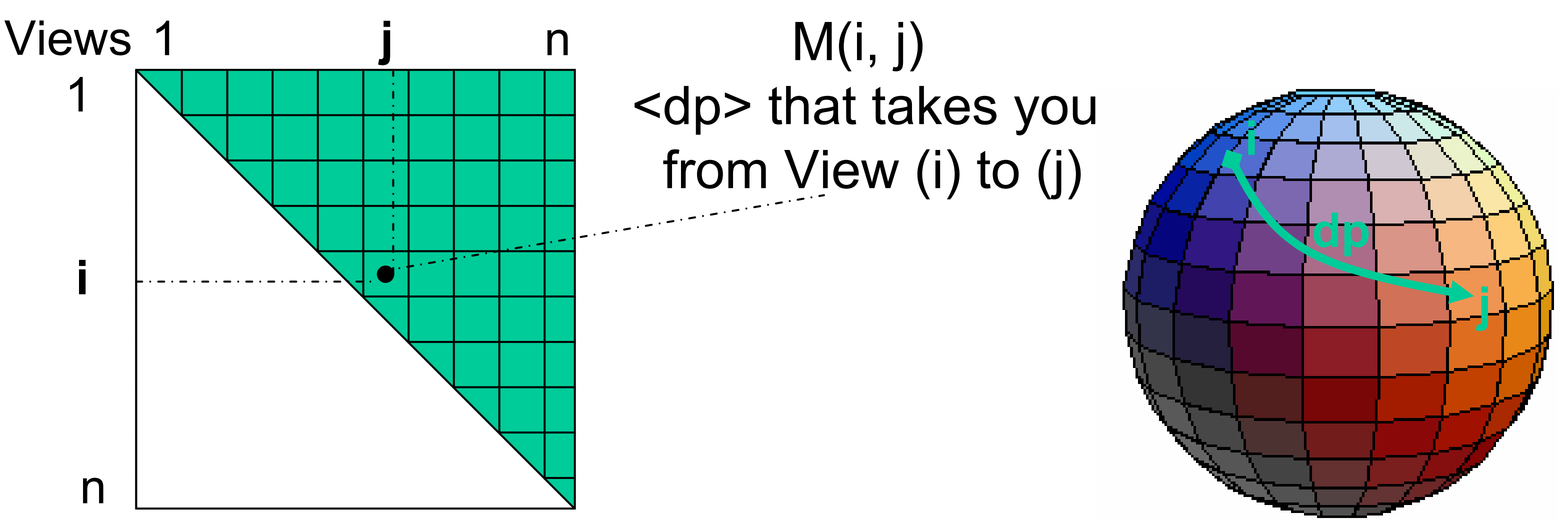
Christian Wallraven (MPI), Sajit Rao (DIST)

"How can **Proprioception, Vision and Active Control** make object recognition more robust?"

Self-terminating Learning ← **Proprioceptive View-Transition Map** → Object Recognition

The Proprioceptive View-Transition Map (PVTM)

- Object representation that links model views in proprioceptive space, hand-centered (= proprioceptive viewing-sphere)

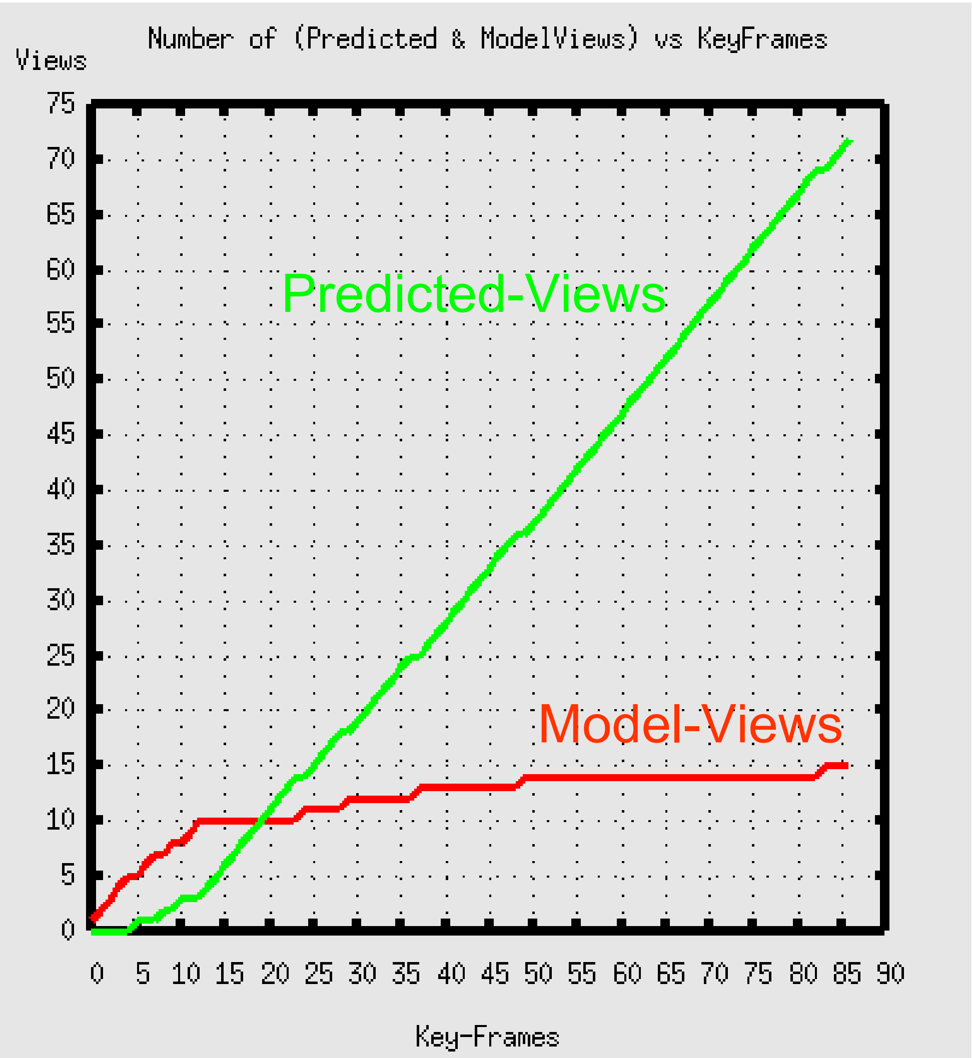


Learning the PVTM with keyframes

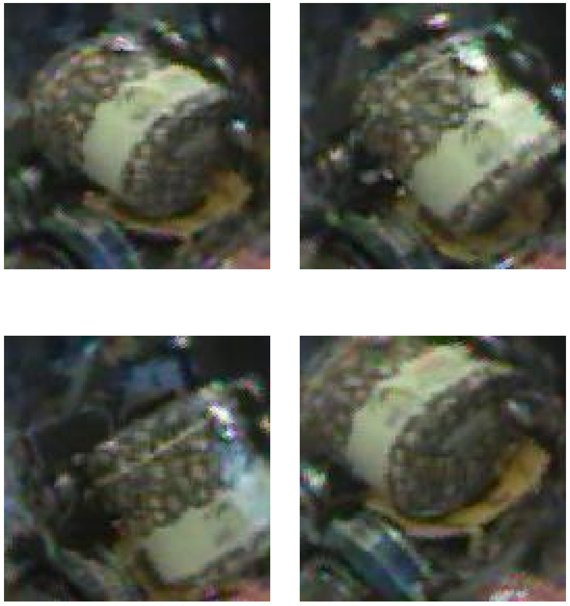
- Robot performs explorative motor-program for any given object to learn the PVTM
- Each view of the PVTM is given by a **keyframe**

External View Keyframes Tracking

Self-Terminating Learning
Results: Box Example



- 30 sec exploration,
- repeated yaw, roll 4 times
- ~ 750 views
- 90 Keyframes + proprioceptive state vectors
- 15 Model Views are sufficient to predict all keyframes



Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

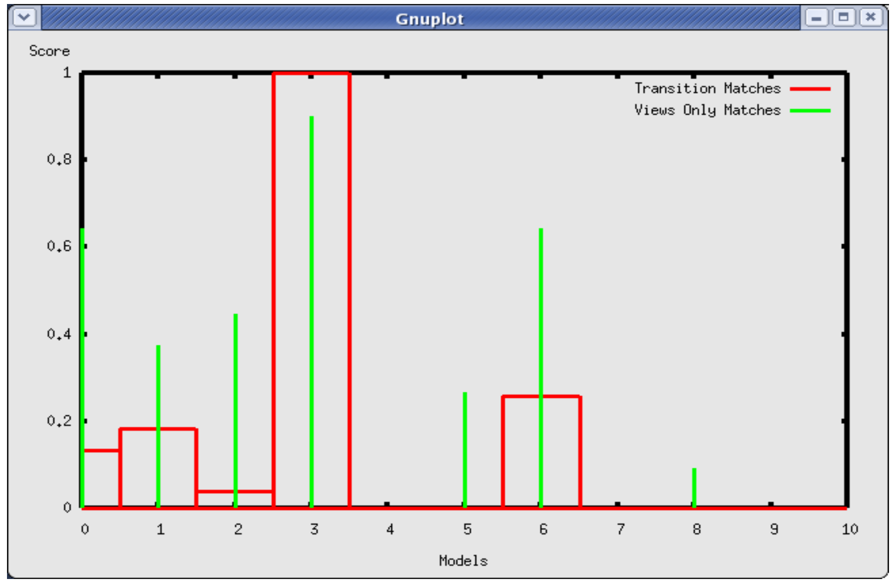
Temporal Association

Computer Vision

Object Recognition

Comparison of Multi-modal vs Visual-only

Bricks



- Visual matching is sufficient to predict the best model
  - but **not** very discriminatory
- Multi-modal matching profile is more **"sharply tuned"**
  – higher on the best match, lower on the distractors
- Transition map match appears to be more discriminative by bringing metric 3D information to bear

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Possible uses of PVTM

- 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
  - Generalizability to other sensory channels

View-based Computer Vision

- Objects can be represented as collections of associated views
- A view-based computer-vision recognition framework motivated by this research was successfully implemented and tested
  - Feature representation enables full control over matching/learning process
  - Easily expandable
  - Modeling of various psychophysical experiments
  - Successful implementation on robotic setup
  - Extended to multi-modal representations

Conclusion

- Human and monkey experiments suggest that objects are represented as **collections of views** linked by **temporal association**.
- Computers can be taught to see if we use appearance-based strategies.
- The information for recognition is in the 2D image.
- Artists have known this already for quite some time.

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

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

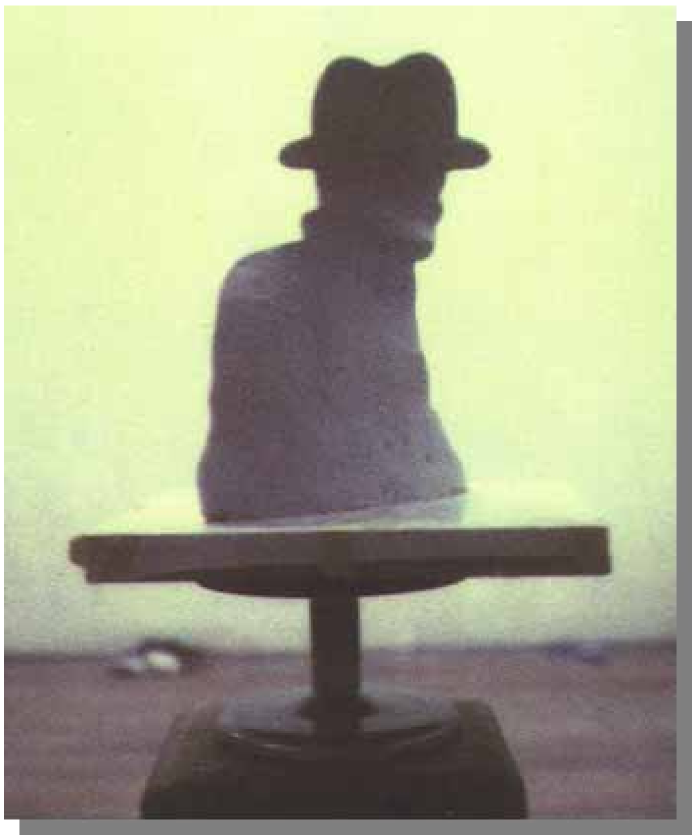
Experiments

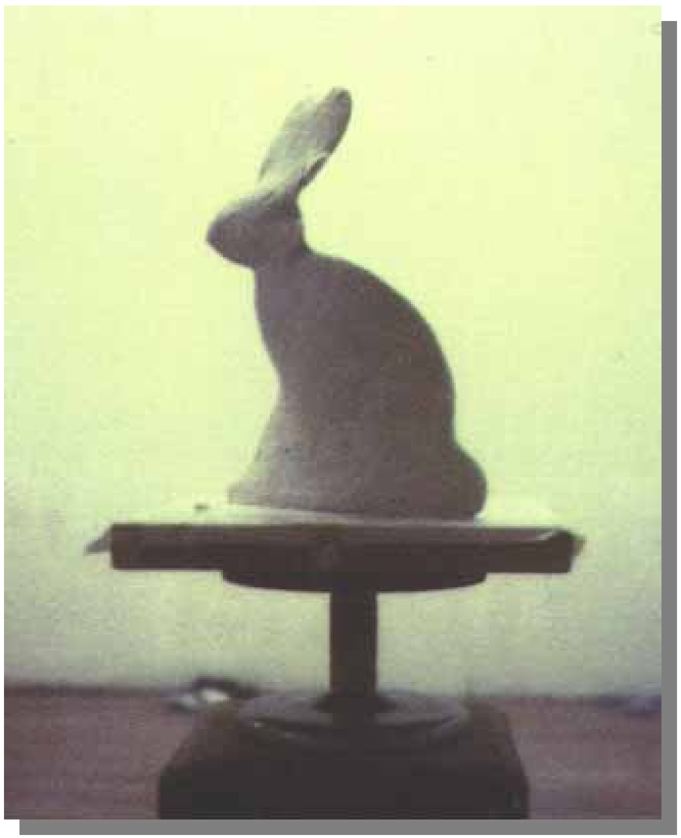
Binding Problem

Temporal Association

Computer Vision

One Object –Two Views
Man or Hare ?





Markus Raetz

Introduction

Visual Recog.

Models

Demonstration

Haptic Recog.

Experiments

Binding Problem

Temporal Association

Computer Vision

Isabelle Bülthoff

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

Introduction
Visual Recog.
Models
Demonstration
Haptic Recog.
Experiments
Binding Problem
Temporal Association
Computer Vision


Isabelle Bülthoff
MPI f. biol. Cybernetics
Tübingen, Germany