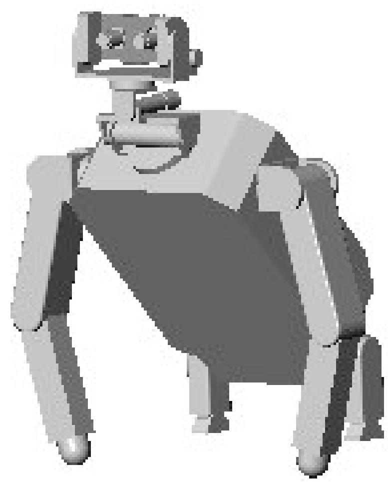
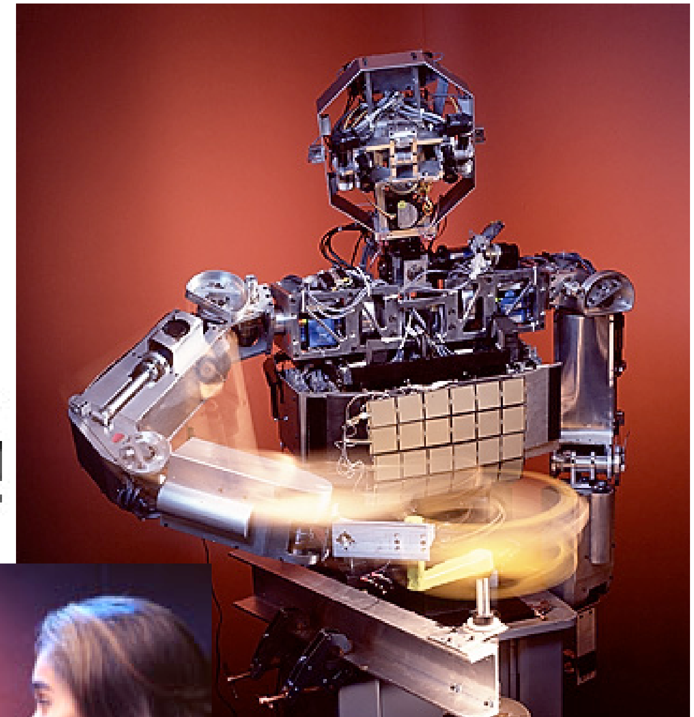
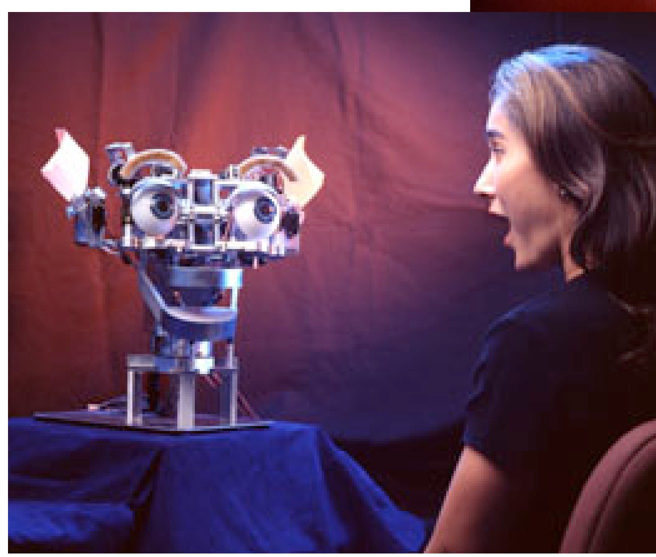
Natural Tasking of Robots Based on Human Cues -- MIT AI Lab

- **Challenge: Make commanding robots as intuitive and natural as commanding soldiers**
- **We will build software systems (on 4 robot platforms) that**
  - **interact with people using intuitive social cues**
  - **are self-adapting through developmental learning**
  - **exploit their physical embodiment to simplify computation**
  - **use multi-modal sensory integration to maximize effectiveness**


Research Issues: Learning by Imitating Humans

- **Knowing what to imitate**
  - what sensory signals are relevant to the current task?
- **Mapping between bodies**
  - how to convert sensory view of one agent to self-mapping?
- **Correcting failures and recognizing success**
  - how can the robot automatically develop success measures?
- **Chaining actions together**
  - how can the learned pieces be chained together into complete task sequences?
- **Generalizing to more complex tasks**
  - how to use invariants and common-sense knowledge to adapt the learned sequences to new situations
- **Intuitive interactions for the human**
  - how to use natural social signals to make it intuitive for humans and robots to communicate with each other

Our Approaches Lead to Specific Progress on Each Research Issue

Research Issues

Our Approaches

	Knowing what to imitate	Mapping between bodies	Correcting failures and recognizing success	Chaining actions together	Generalizing to more complex tasks	Intuitive Interactions
Social Interaction	Uses attentional cues to recognize task relevant events		Identifies and displays emotional states for reinforcement			Allows humans and robots to share similar social cues
Development	Limits search space by incremental refinement of perception	Simplifies mapping by incremental refinement of motor skills		Provides natural decomposition of complex tasks	Exploits incremental learning to build new skills	Provides methods for building social skills
Embodiment	Constrains movement which assists directed perception	Simplifies body mapping between human and robot	Enables simple but robust low-level behaviors	Places physical limits on successive actions		Allows the human to observe natural social feedback cues
Integration	Allows robot to recognize cues in multiple modalities		Increases robustness through redundancy		Assists in transfer to new modalities	Allows human to use natural modalities (voice, gesture, facial expressions)

The Research Team

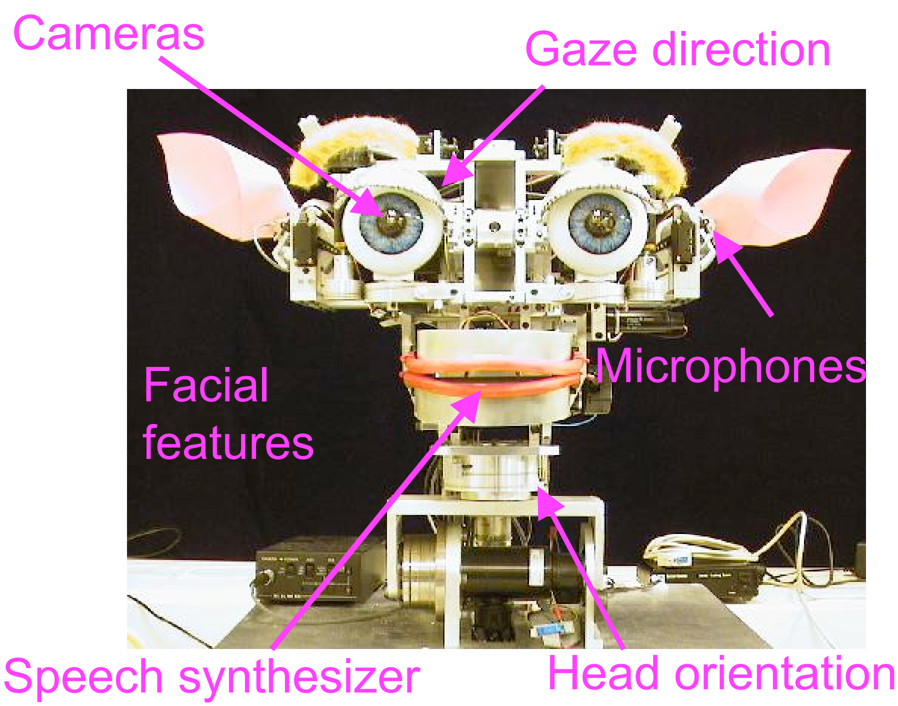
Our Approaches Lead to Specific Progress on Each Research Issue

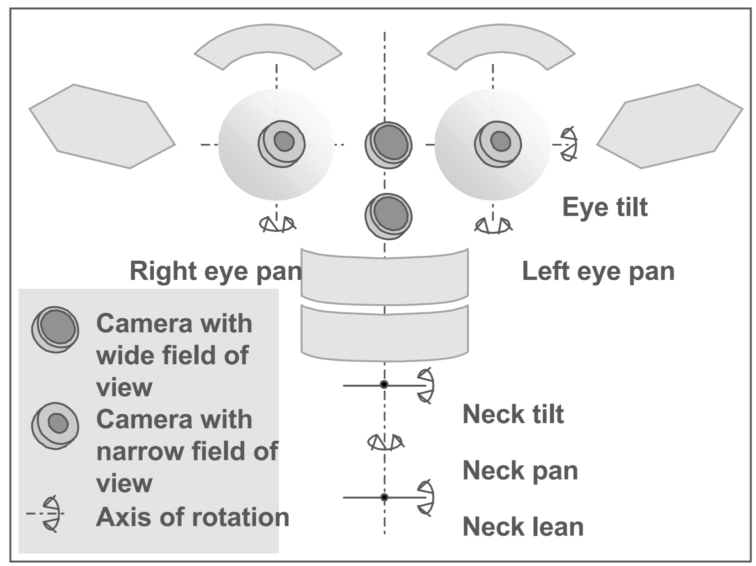
Research Issues

Our Approaches

	Knowing what to imitate	Mapping between bodies	Correcting failures and recognizing success	Chaining actions together	Generalizing to more complex tasks	Intuitive Interactions
Social Interaction	Uses attentional cues to recognize task relevant events		Identifies and displays emotional states for reinforcement			Allows humans and robots to share similar social cues
Development	Limits search space by incremental refinement of perception	Simplifies mapping by incremental refinement of motor skills		Provides natural decomposition of complex tasks	Exploits incremental learning to build new skills	Provides methods for building social skills
Embodiment	Constrains movement which assists directed perception	Simplifies body mapping between human and robot	Enables simple but robust low-level behaviors	Places physical limits on successive actions		Allows the human to observe natural social feedback cues
Integration	Allows robot to recognize cues in multiple modalities		Increases robustness through redundancy		Assists in transfer to new modalities	Allows human to use natural modalities (voice, gesture, facial expressions)

Designing for the Human In the Loop



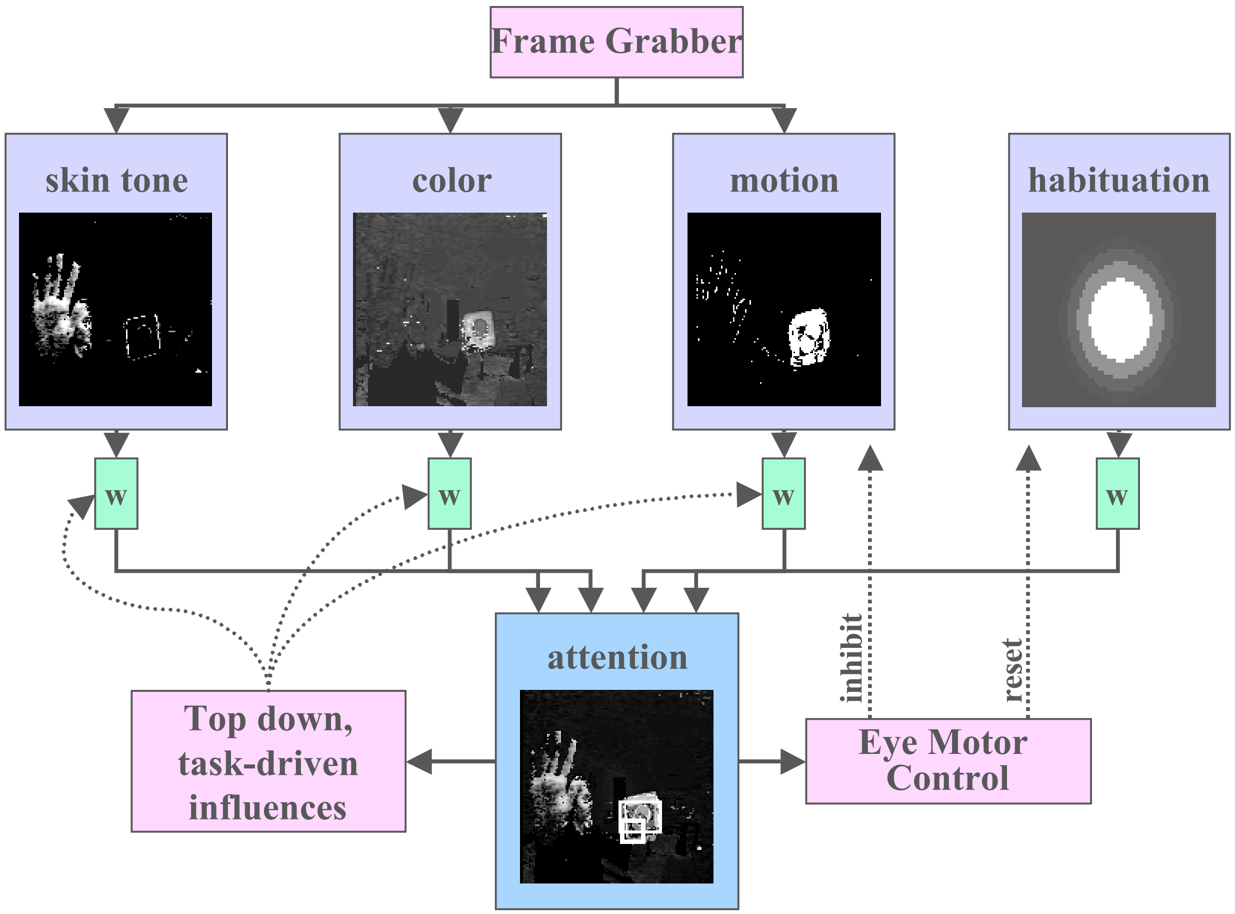


Evaluation Metrics

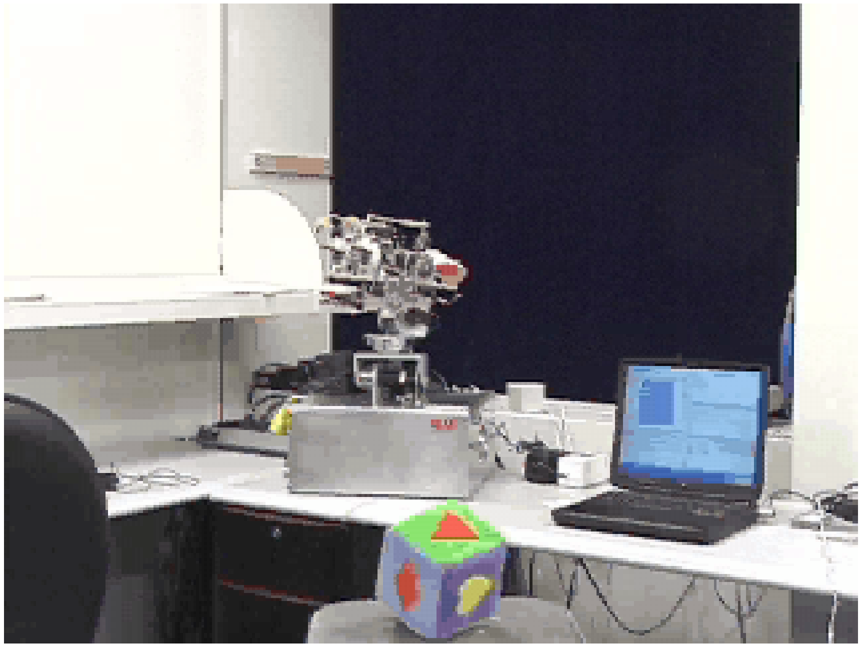
- **Must consider human response as well as robot's behavior**
  - **Subjective but quantifiable**

- **Four Criteria used to evaluate Kismet:**
  - **Can people intuitively read and do they naturally respond to Kismet's social cues?**
  - **Does Kismet readily elicit social cues (scaffolding) from the human that could be used to benefit learning?**
  - **Can Kismet perceive and appropriately respond to these naturally offered social cues?**
  - **Does the human adapt to the robot, and robot adapt to the human, in a way that benefits the interaction? (entrainment)**

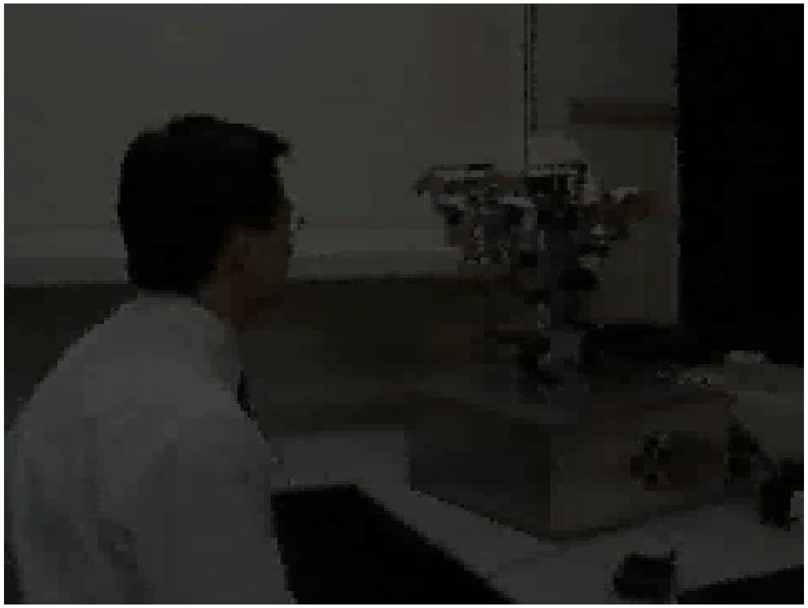
Visual Attention-Source of Coherence



Video of Visual Search for Toy



Turn-Taking and Proto-Dialog



Our Approaches Lead to Specific Progress on Each Research Issue

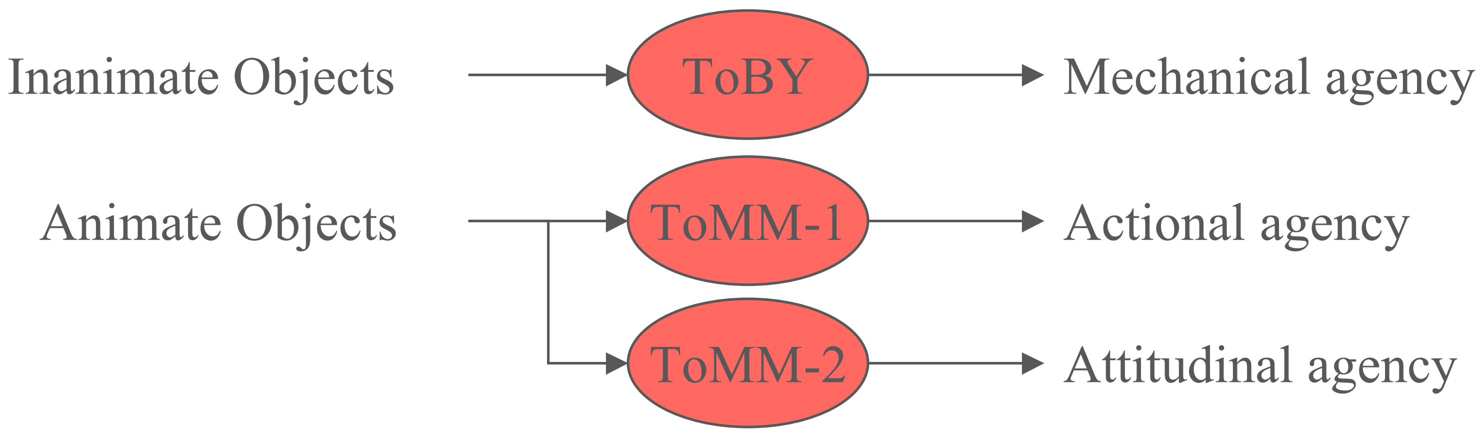
Research Issues

Our Approaches	Knowing what to imitate	Mapping between bodies	Correcting failures and recognizing success	Chaining actions together	Generalizing to more complex tasks	Intuitive Interactions
Social Interaction	Uses attentional cues to recognize task relevant events		Identifies and displays emotional states for reinforcement			Allows humans and robots to share similar social cues
Development	Limits search space by incremental refinement of perception	Simplifies mapping by incremental refinement of motor skills		Provides natural decomposition of complex tasks	Exploits incremental learning to build new skills	Provides methods for building social skills
Embodiment	Constrains movement which assists directed perception	Simplifies body mapping between human and robot	Enables simple but robust low-level behaviors	Places physical limits on successive actions		Allows the human to observe natural social feedback cues
Integration	Allows robot to recognize cues in multiple modalities		Increases robustness through redundancy		Assists in transfer to new modalities	Allows human to use natural modalities (voice, gesture, facial expressions)

Theory of Mind

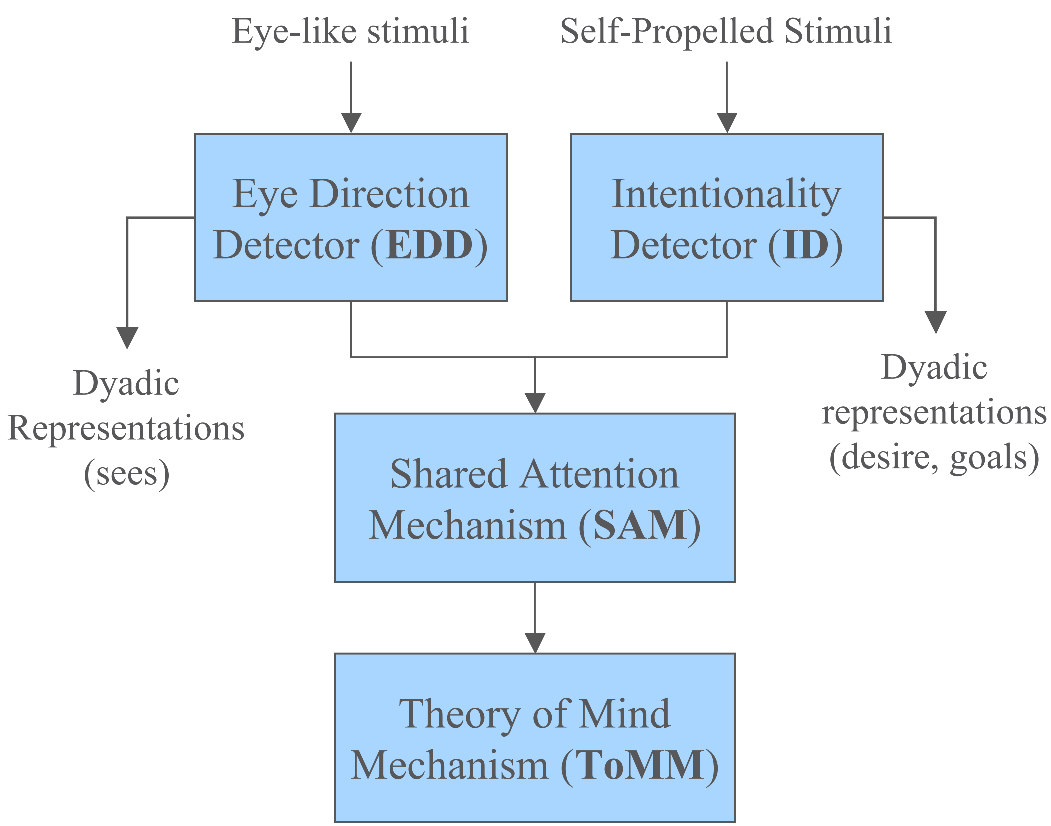
- For complex learning tasks the robot needs to
  - understand what it is the human knows
    - about the world
    - about the robot's current state
  - understand what the human is doing and why
  - understand what is the human

Leslie's Model



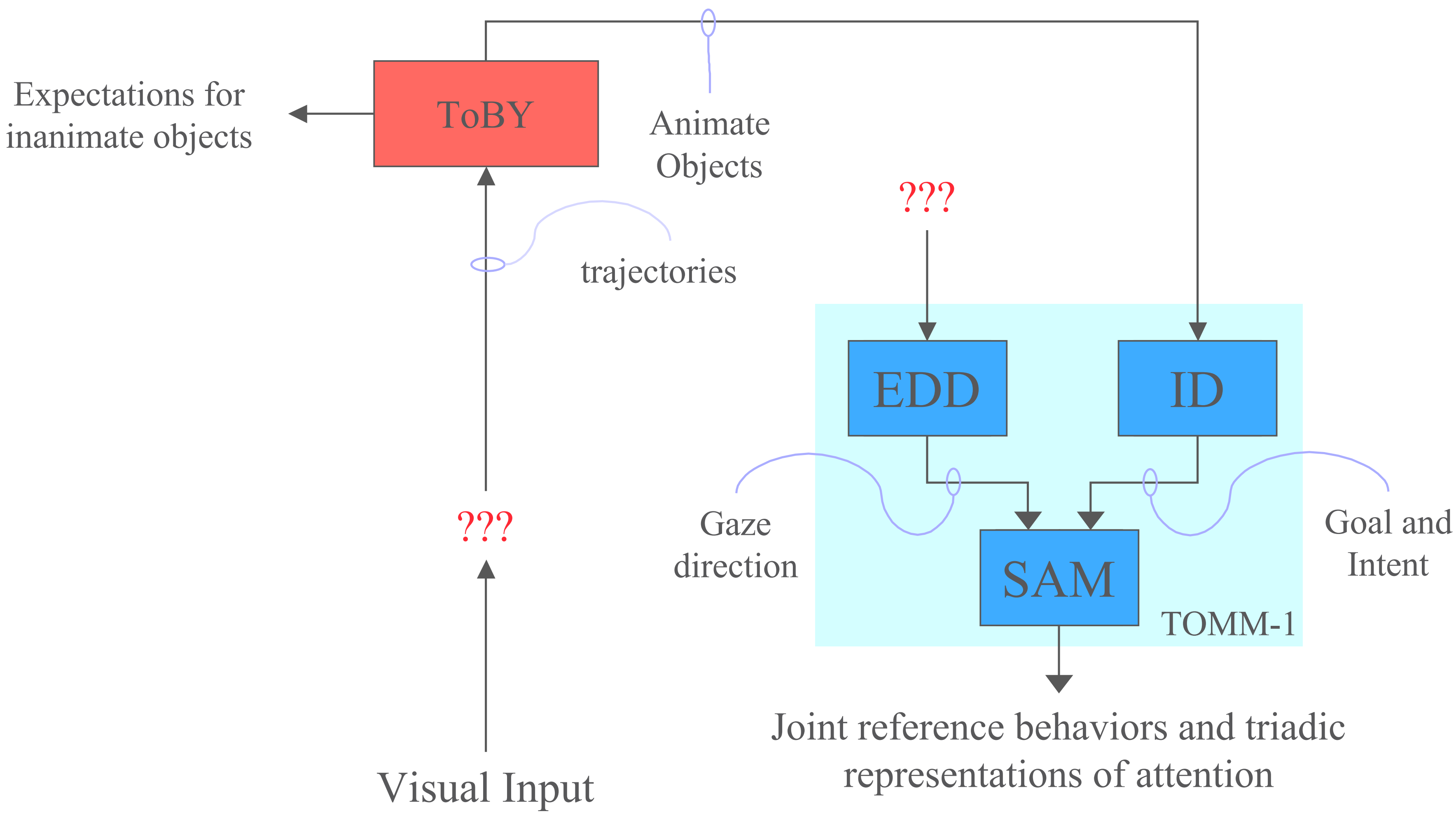
- Divides the world into three spheres of causation, each of which is processed by a separate module:
  - Theory of Body (ToBY) – applies naïve physics to objects
  - Theory of Mind Mechanism (ToMM) – applies folk psychology to agents
    - » System 1 applies rules of goals and desires
    - » System 2 applies rules of belief and knowledge

Baron-Cohen's Model

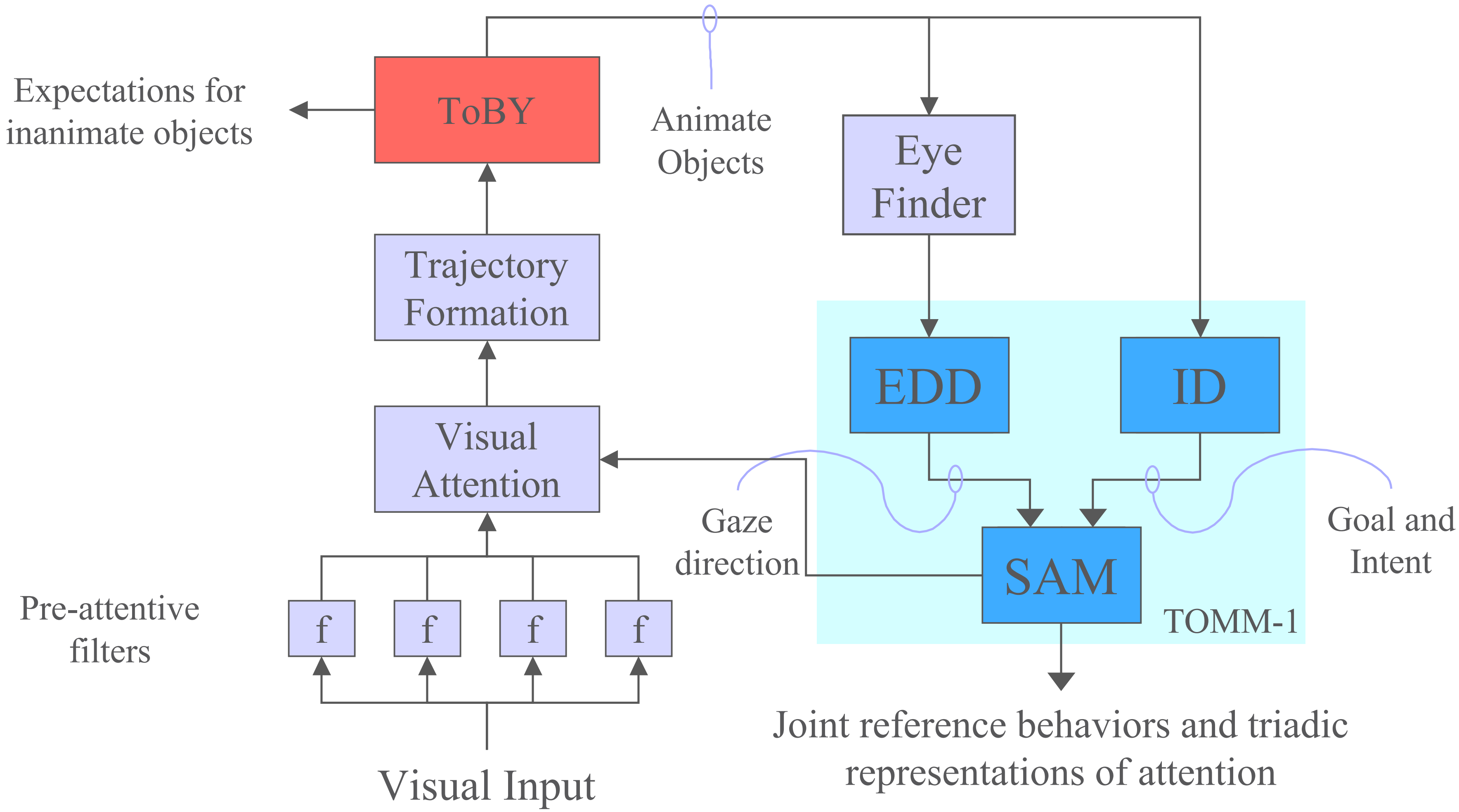


- **Requires two types of input stimuli:**
  - **Eye-like stimuli**
  - **Self-propelled (animate) stimuli**
- **Provides greater detail on the inner workings of Leslie's ToMM-1 module**
- **Proposes that autism is an impairment of either SAM (subgroup A) or ToMM (subgroup B)**

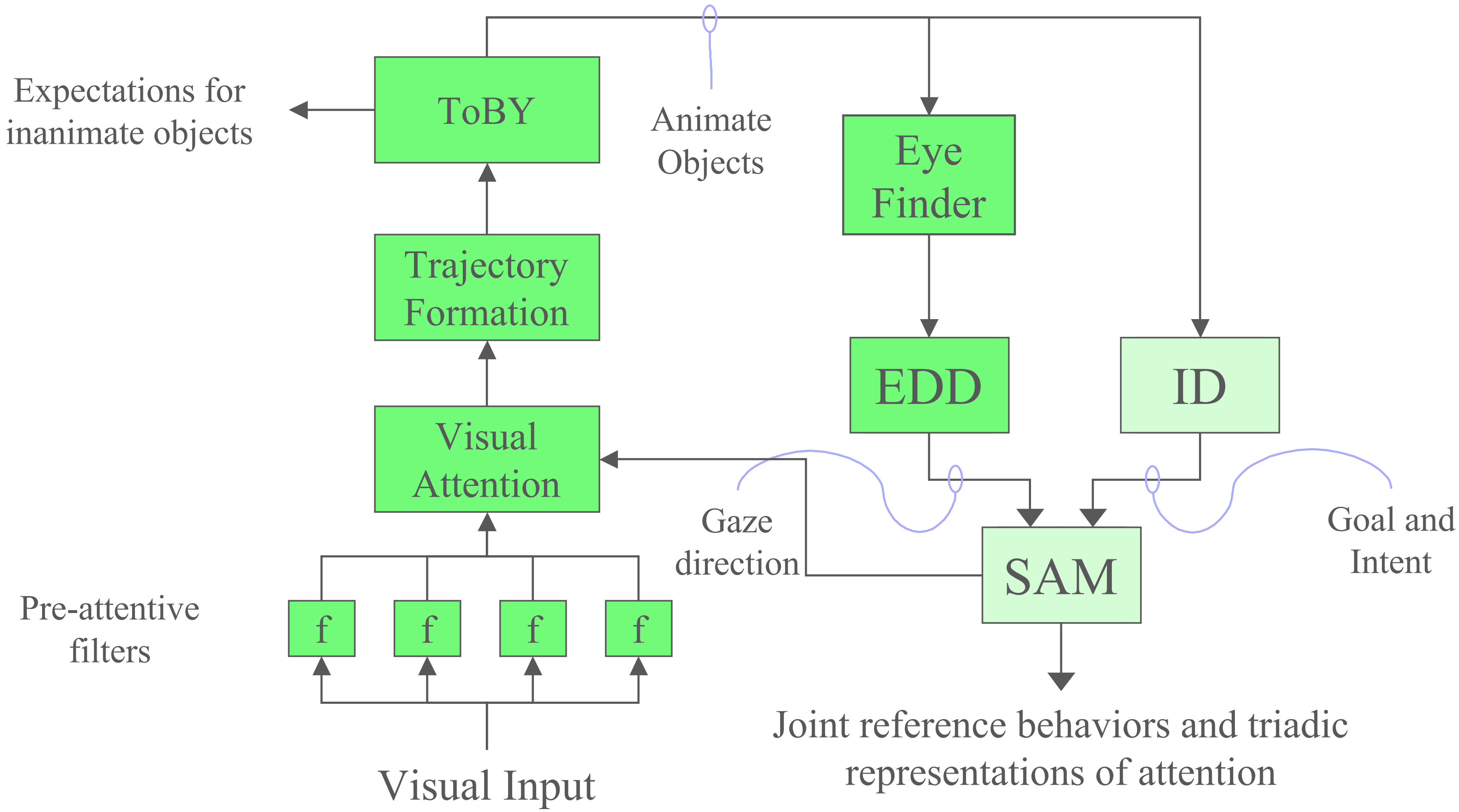
Scassellati's Model



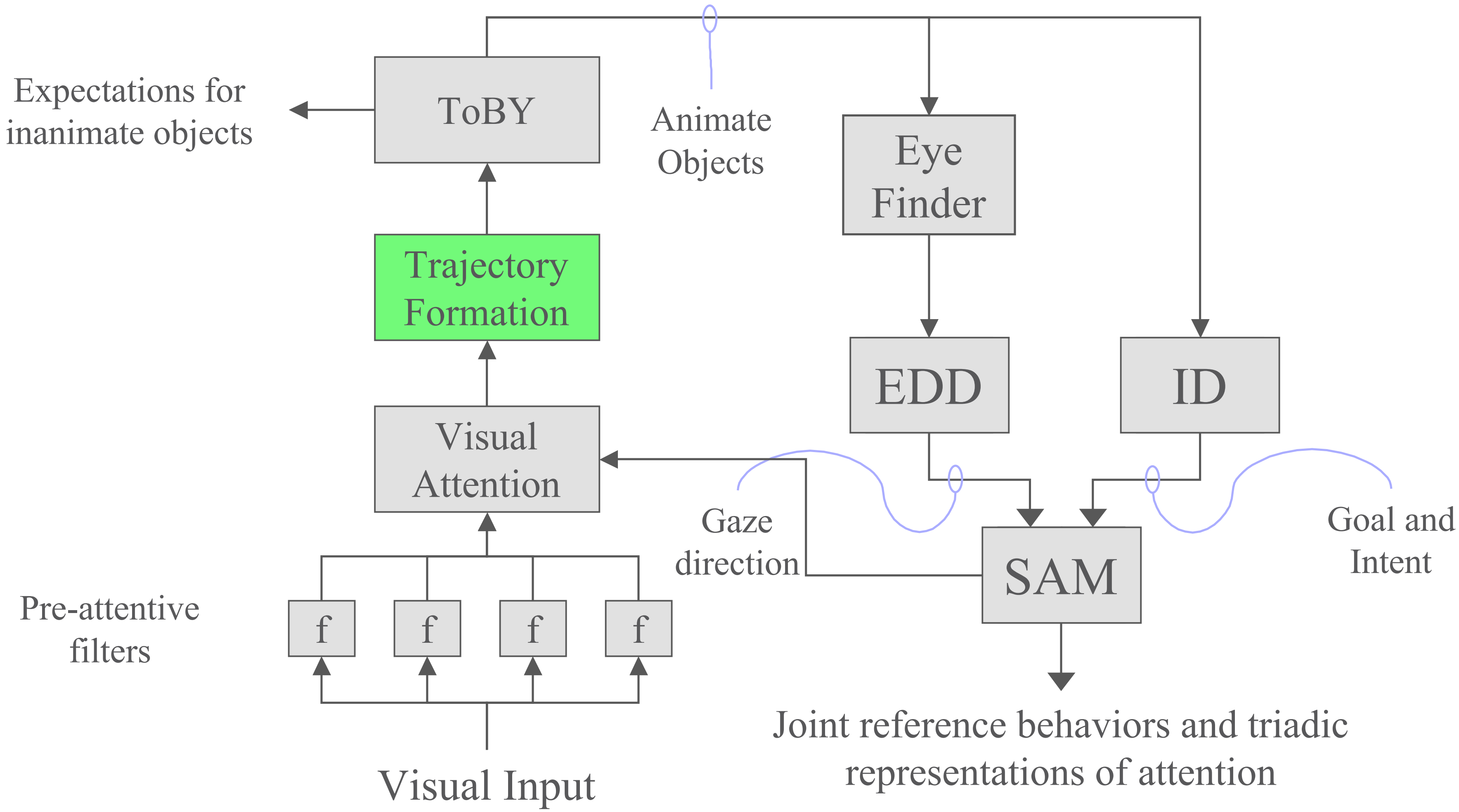
Scassellati's Model



Where We Are Now

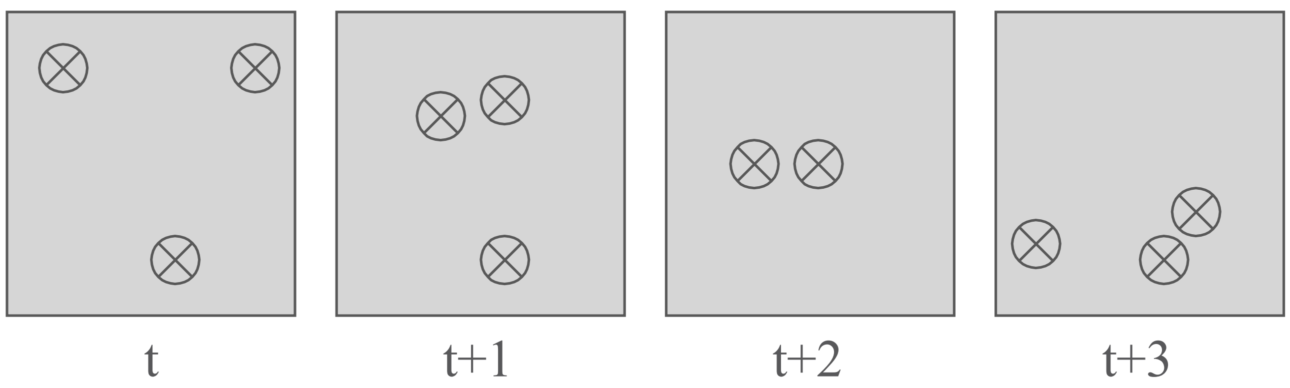


Roadmap



Expectations for inanimate objects
ToBY
Animate Objects
Eye Finder
Trajectory Formation
EDD
ID
Visual Attention
Gaze direction
SAM
Goal and Intent
Pre-attentive filters
f
f
f
f
Joint reference behaviors and triadic representations of attention
Visual Input

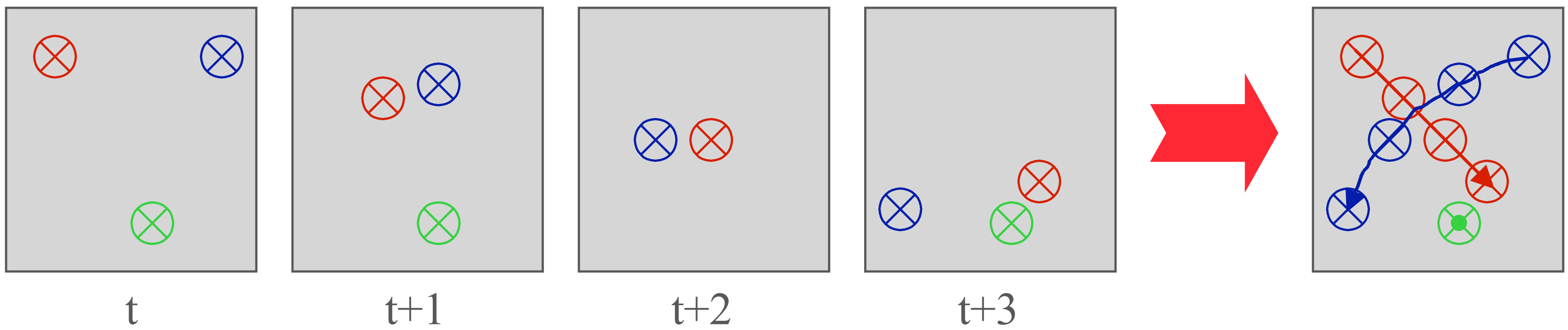
Trajectory Formation



- **Each frame produces a set of target points**

$$\{P_t^1, P_t^2 \ldots P_t^n\}, \{P_{t+1}^1, P_{t+1}^2 \ldots P_{t+1}^n\}, \ldots \{P_{t+k}^1, P_{t+k}^2 \ldots P_{t+k}^n\}$$

Motion Correspondence



- **Each frame produces a set of target points**

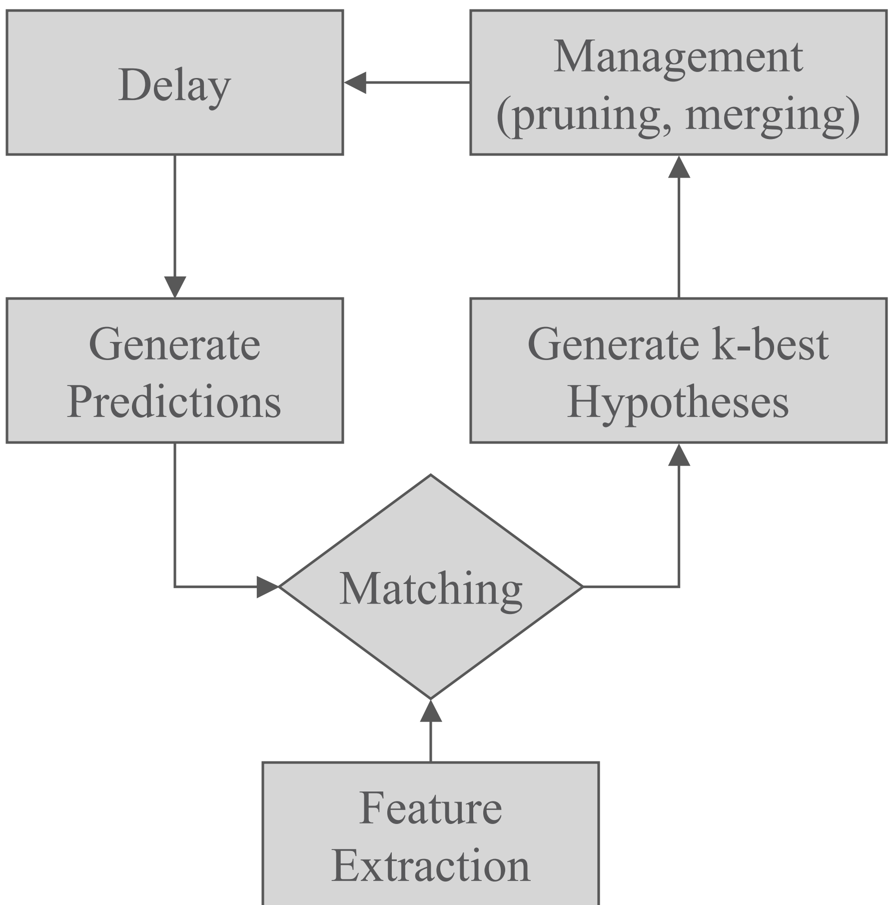
$$\{P_t^1, P_t^2 \ldots P_t^n\}, \{P_{t+1}^1, P_{t+1}^2 \ldots P_{t+1}^n\}, \ldots \{P_{t+k}^1, P_{t+k}^2 \ldots P_{t+k}^n\}$$

- **Objective is to identify sequences through a subset of the frames**

$$\{P_t^{\alpha_{11}}, P_{t+1}^{\alpha_{12}} \ldots P_{t+k}^{\alpha_{1k}}\}, \{P_{t+1}^{\alpha_{21}}, P_{t+1}^{\alpha_{22}} \ldots P_{t+1}^{\alpha_{2k}}\}, \ldots \{P_{t+k}^{\alpha_{k1}}, P_{t+k}^{\alpha_{k2}} \ldots P_{t+k}^{\alpha_{kk}}\}$$

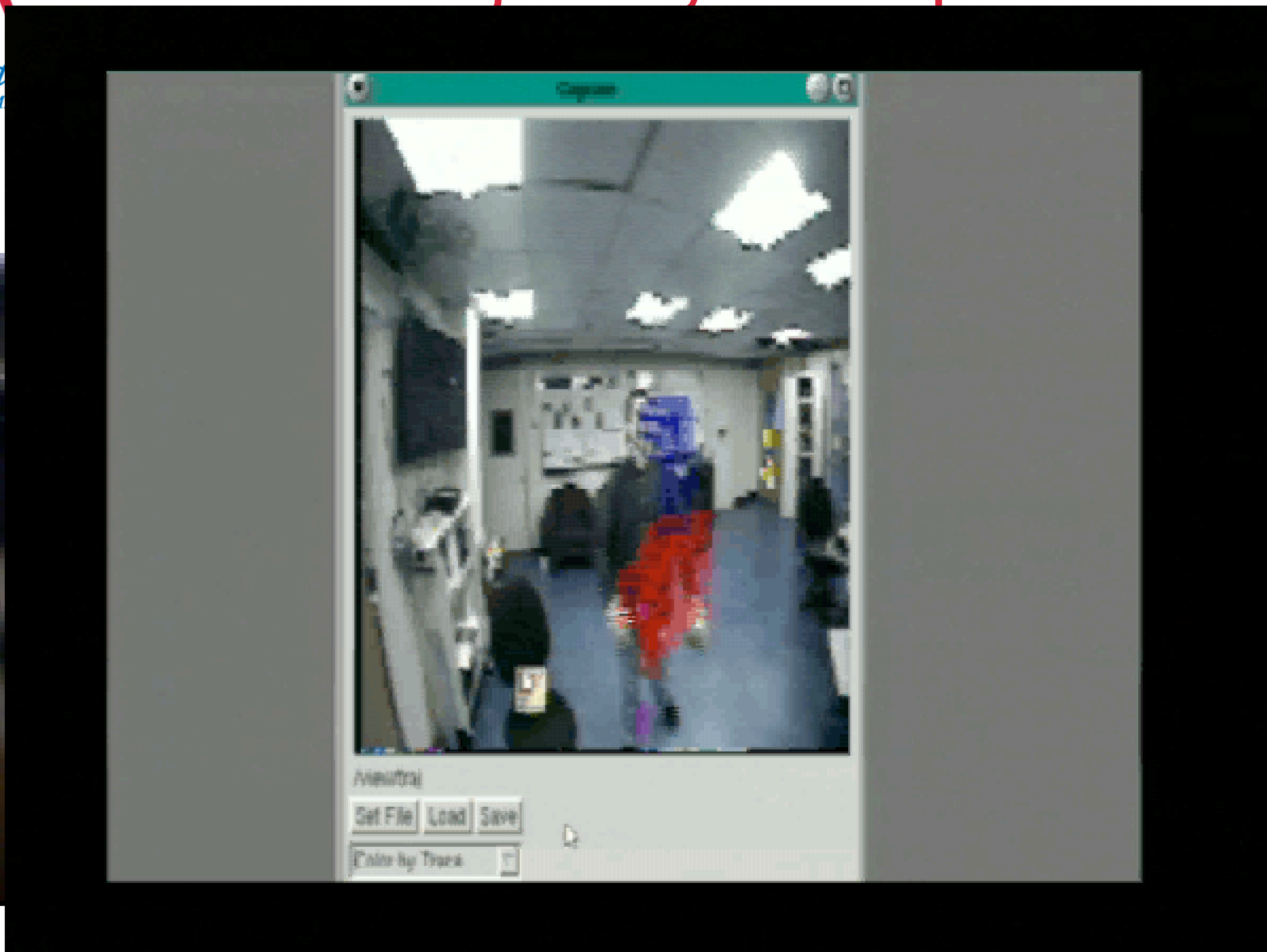
Multiple Hypothesis Tracking

(Reid, 1979)(Cox and Hingorani, 1996)

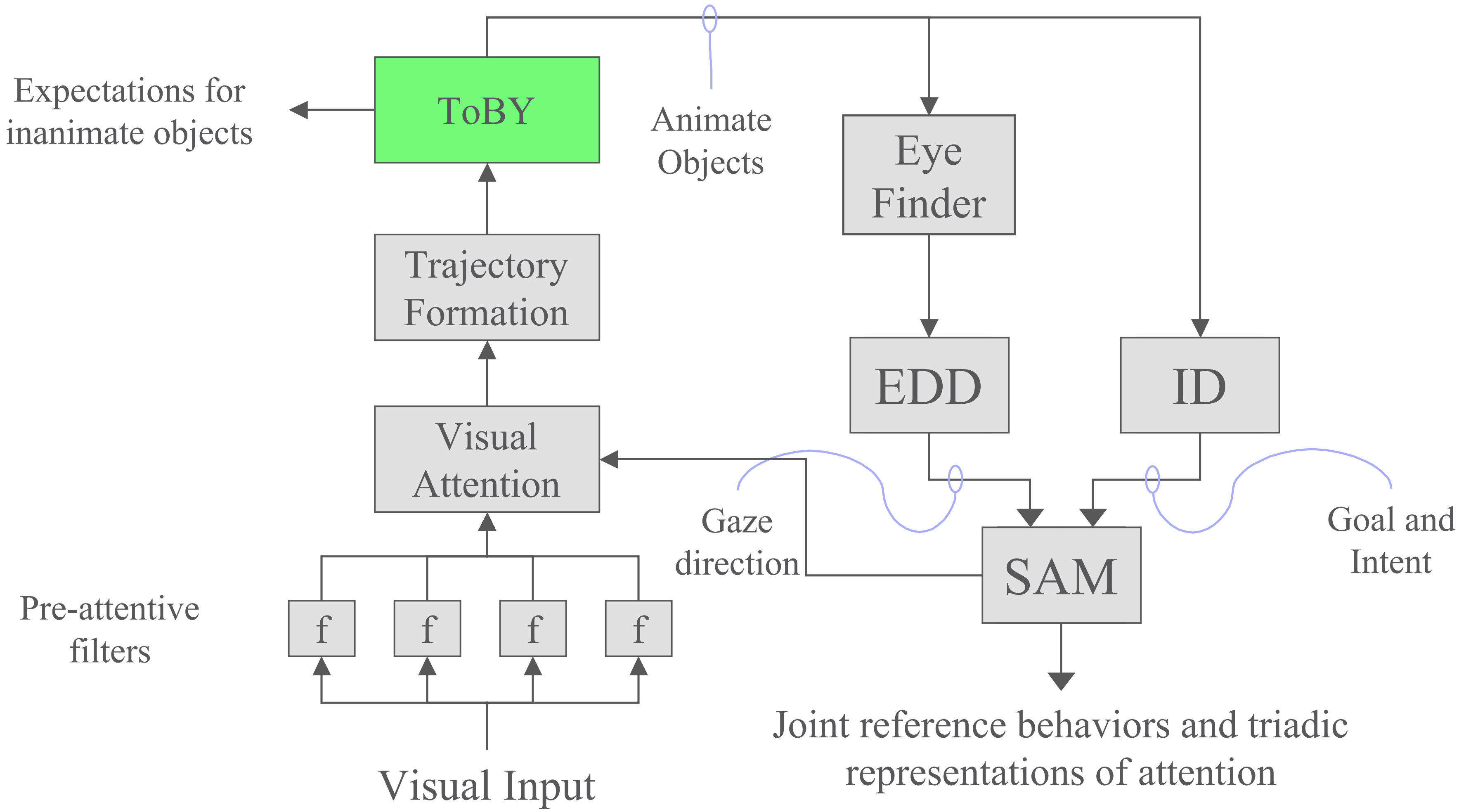


- **Allows for**
  - **Trajectory Initiation**
  - **Trajectory Termination**
  - **Minor occlusion**
- **Matching based on**
  - **Mahalanobis distance**
  - **Cross-correlation of features including**
    - **Area**
    - **Overall Saliency**
    - **Saliency among the individual feature channels**

Trajectory Example

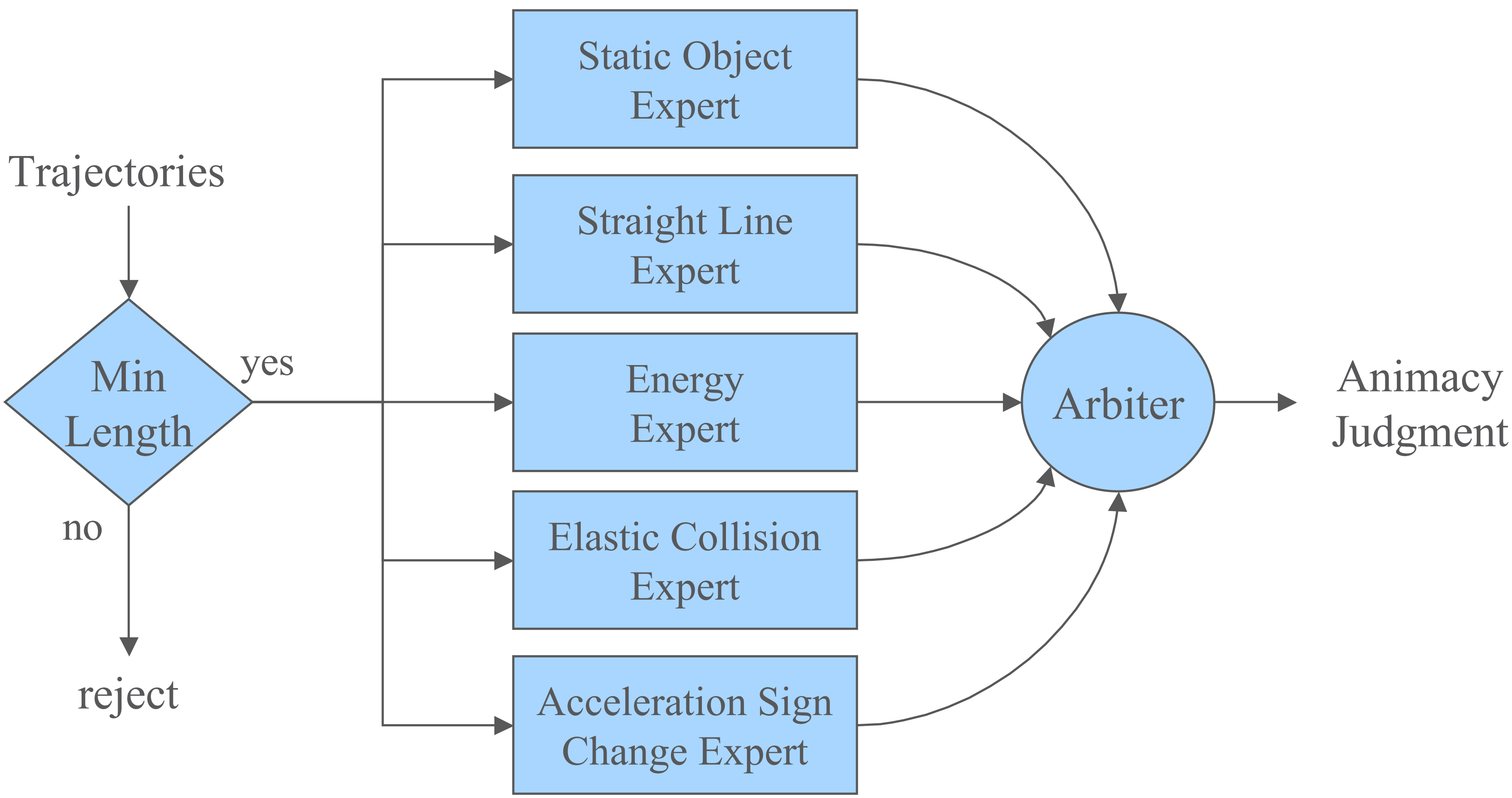


Roadmap



Expectations for inanimate objects
ToBY
Animate Objects
Eye Finder
Trajectory Formation
EDD
ID
Visual Attention
Gaze direction
SAM
Goal and Intent
Pre-attentive filters
f
f
f
f
Joint reference behaviors and triadic representations of attention
Visual Input

ToBY Architecture



ToBY Agents

Straight Line Expert

$$\min \Box \sqrt{(v-\bar{v})^2}$$

Minimize the sum of the deviations from the mean velocity

Energy Expert

$$\tfrac{1}{2} m v_y^2 + mgy$$

Constant mass and the inertial system provides the gravity vector

Elastic Collision Expert

$$v_{t-1}^1 + v_{t-1}^2 = v_{t+1}^1 + v_{t+1}^2$$

Look for transfer in velocities before and after collision

Acceleration Sign Change Expert

$$\mathrm{sgn}(v_t) \ ? \ \mathrm{sgn}(v_{t+1})$$

Look for multiple sign changes in the acceleration

Animacy Results

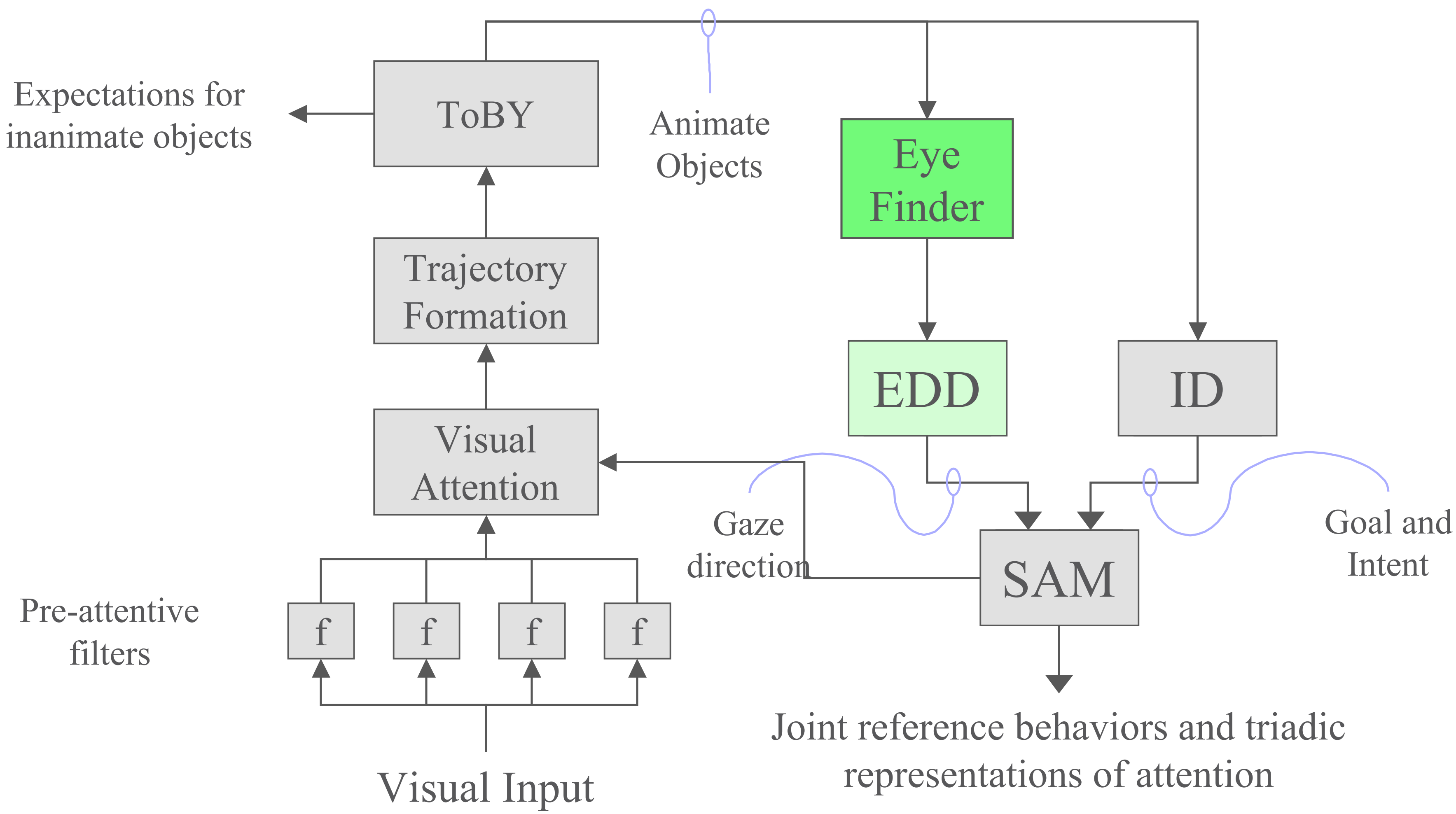
Animate

Inanimate

Animate

- Arbitration methods
  - Weighted sum
  - Winner-take-all

Roadmap



Expectations for inanimate objects
ToBY
Animate Objects
Eye Finder
Trajectory Formation
EDD
ID
Visual Attention
Gaze direction
SAM
Goal and Intent
Pre-attentive filters
f
f
f
f
Joint reference behaviors and triadic representations of attention
Visual Input

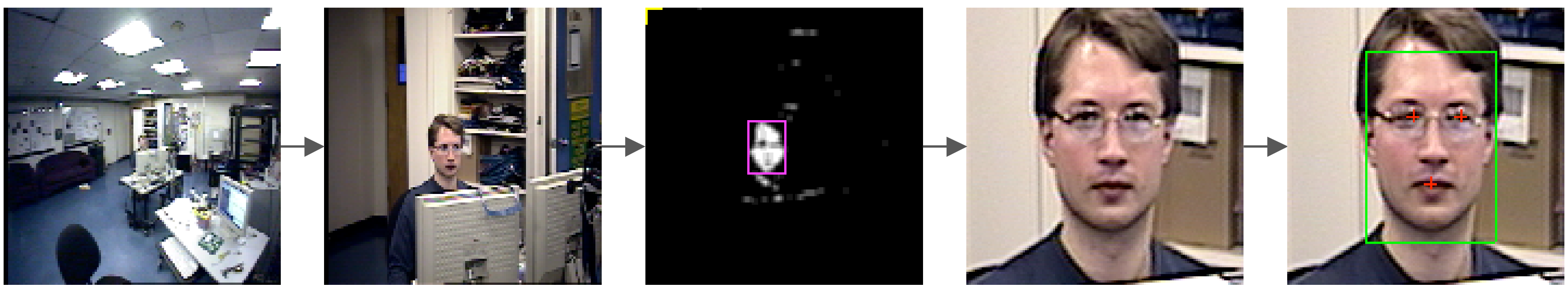
Post-Attentive Visual Processing: Finding Faces and Eyes

Can you find the face in this image?

Can you tell where Scaz is looking?

Post-Attentive Visual Processing: Finding Faces and Eyes

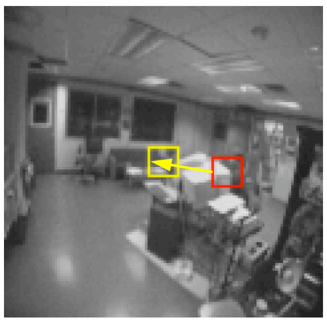
Locate target In wide field → Foveate Target → Apply Face Filter → Software Zoom → Feature Extraction



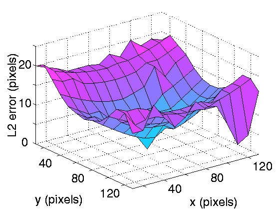
300 msec → 66 msec

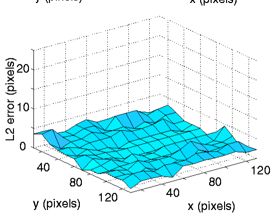
Two Sensory-Motor Mappings

Saccade Map



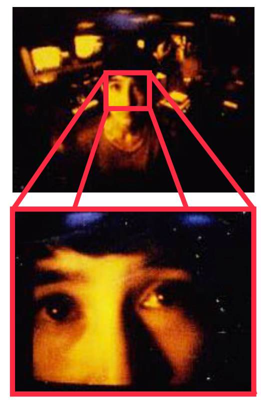
Maps image positions to motor commands necessary to center that location in the visual image.



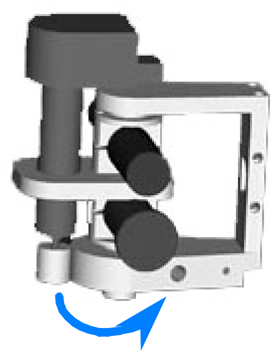


Learned using standard self-supervised techniques (lookup tables, neural nets, etc.)

Peripheral-Foveal Map

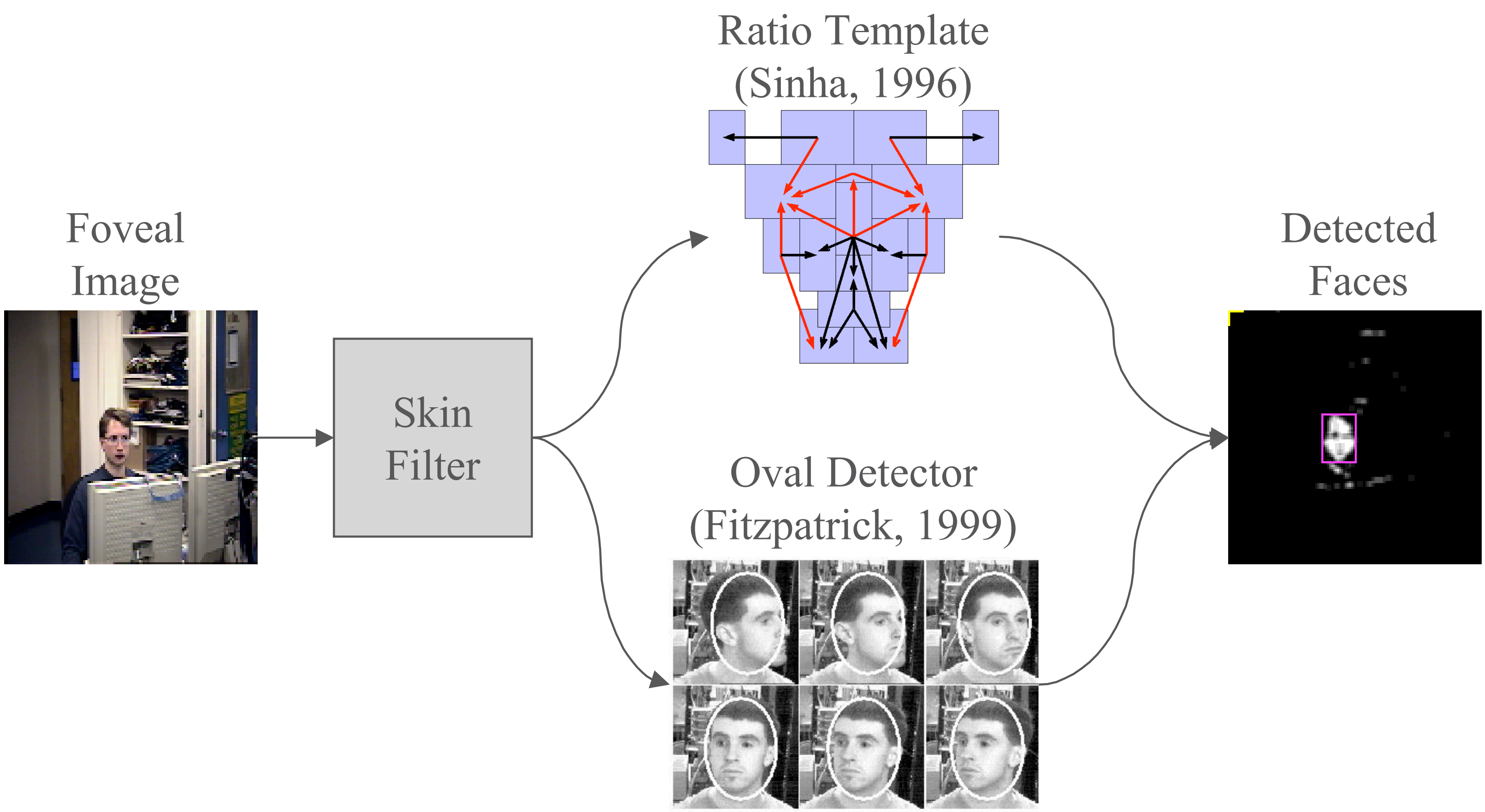


Maps pixels in the peripheral image to pixels in the foveal image.

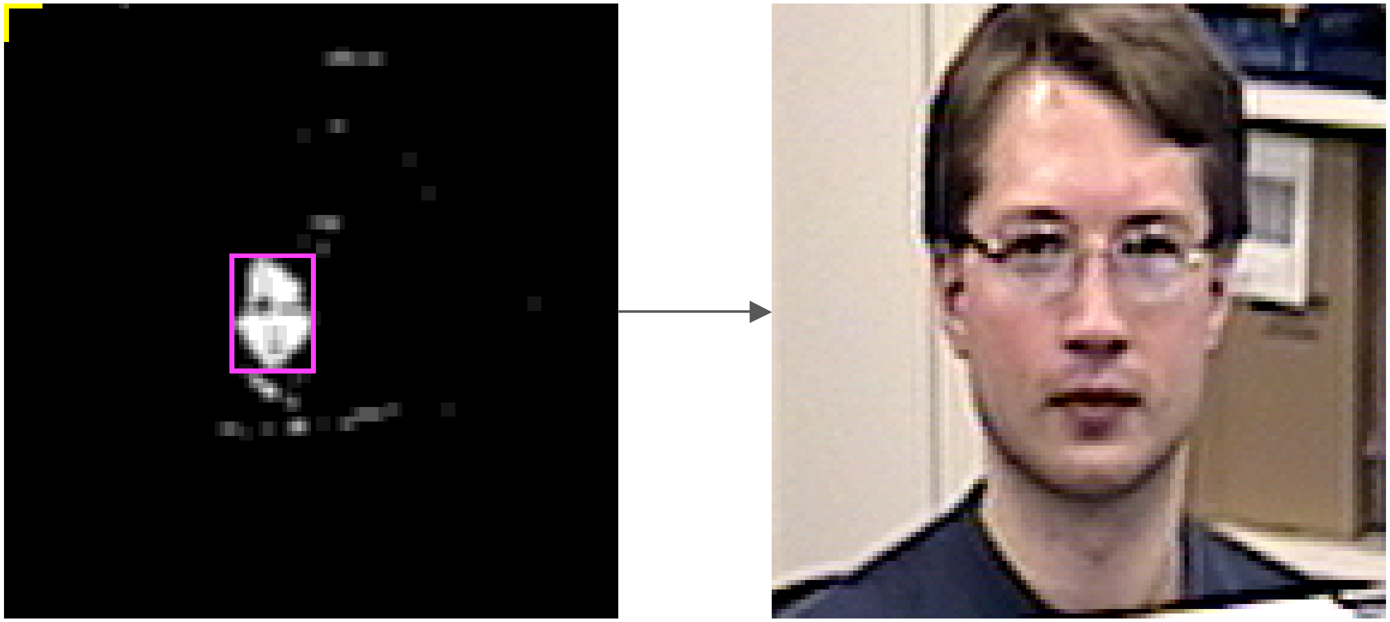


Scale is learned using optic flow rates obtained while the cameras are moving. Position is learned using correlation.

Face Finding

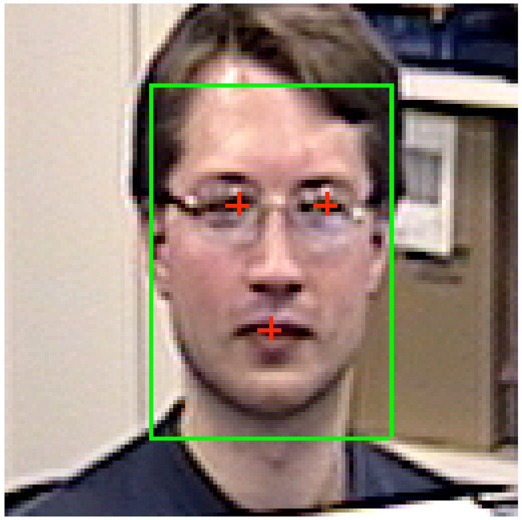


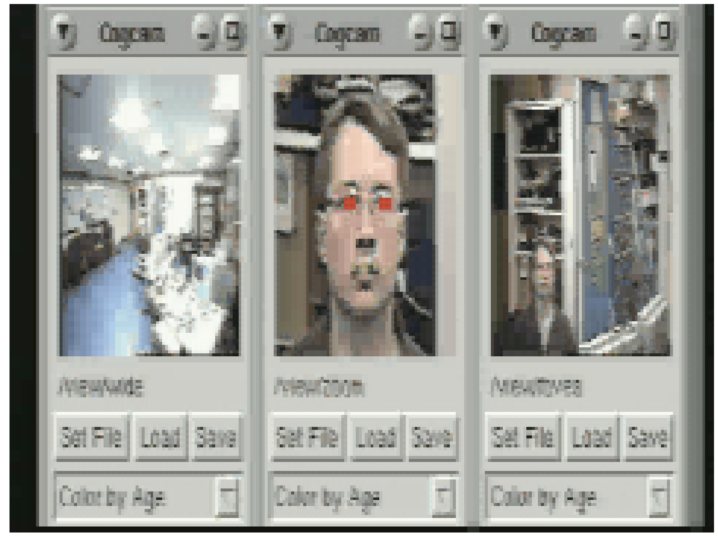
Software Zoom



- From full 640x480 image, extract the relevant 128x128 sub-image
- Introduces the majority of the system delay

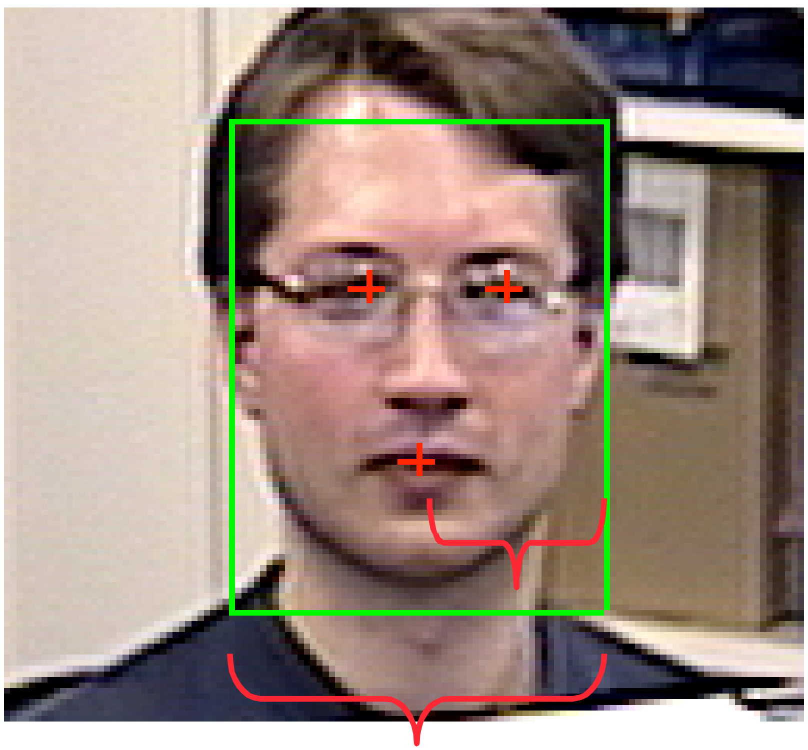
Feature Finding





- **Locate eyes and mouth by looking for centroid of luminance minima**
- **Mouth:**
  - **Iterative algorithm with adaptive regions provides performance similar to simulated annealing**
- **Eyes:**
  - **Add symmetry requirement**

Head Pose Derivation

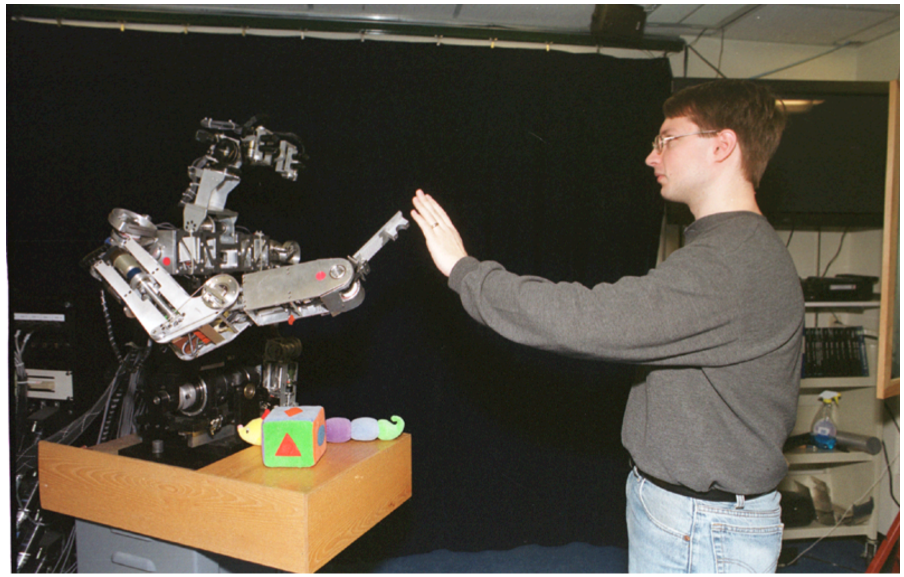


- Obtain estimate of head pose from the positions of the mouth, nostrils, and eyes

$$\frac{Mouth - to - Left\ Edge}{Face\ Width}$$

- Failure modes
  - Match to nostril
  - Vertical Lengthening
- Accuracy of +/- 5 degrees at a distance of 6 meters

An Application to Mimicry



- **Imitation of arm trajectories**
  - **No body model**
  - **Imitation based around detected faces**
  - **Imitation only of animate motion trajectories**

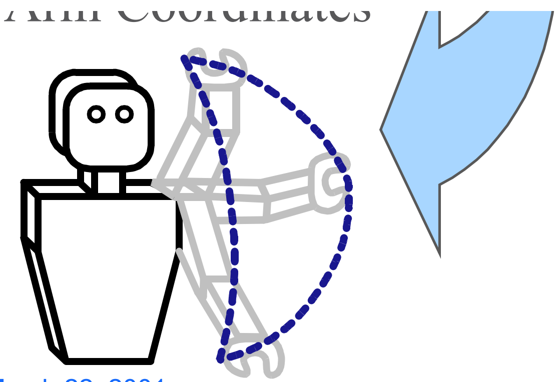
Mapping Visual Trajectories to Arm Movements

Visual Coordinates

- Postural primitives define a sub-

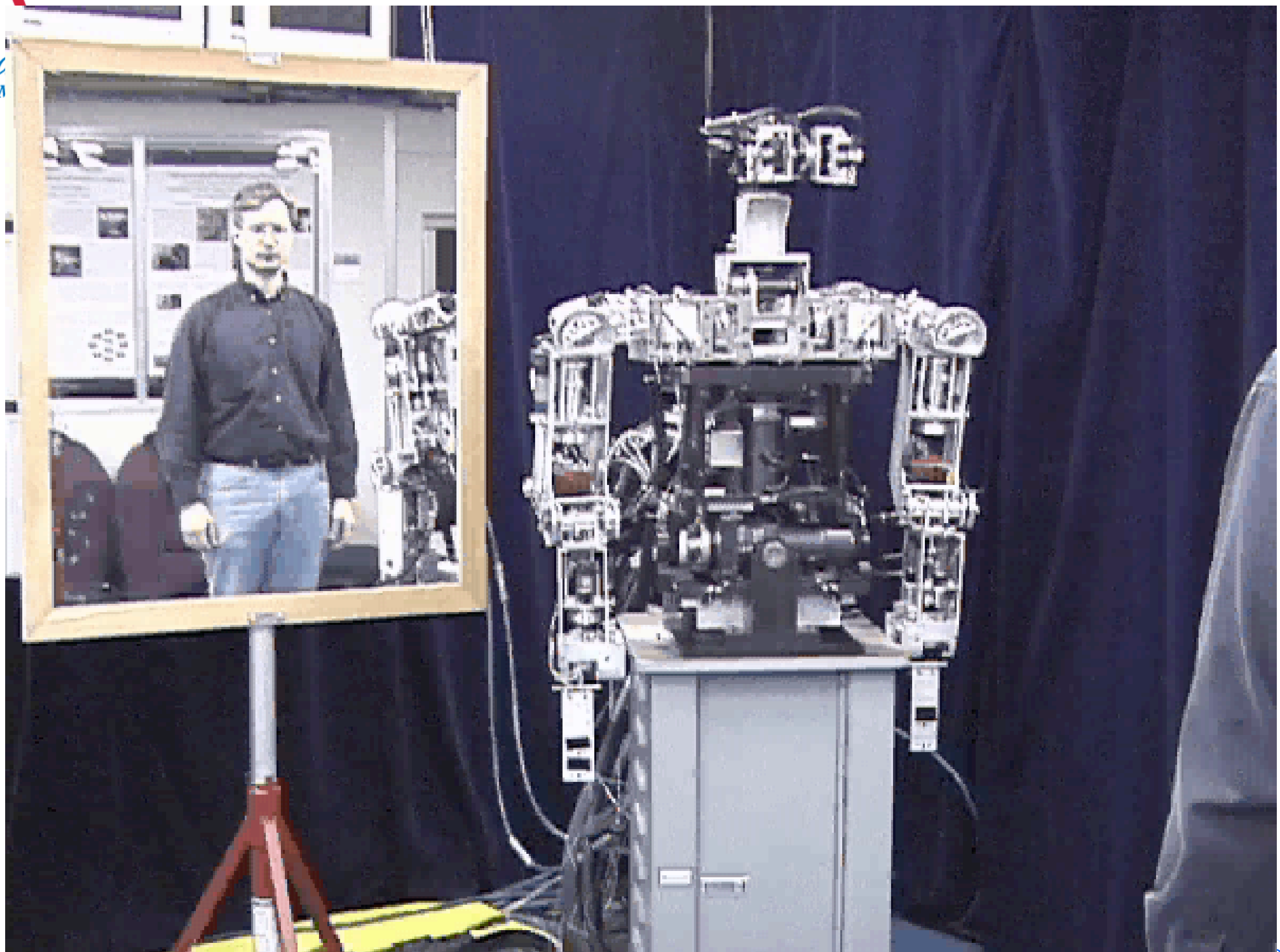
Current demos suffer from low stiffness as we rebuild our arm control system

Arm Coordinates

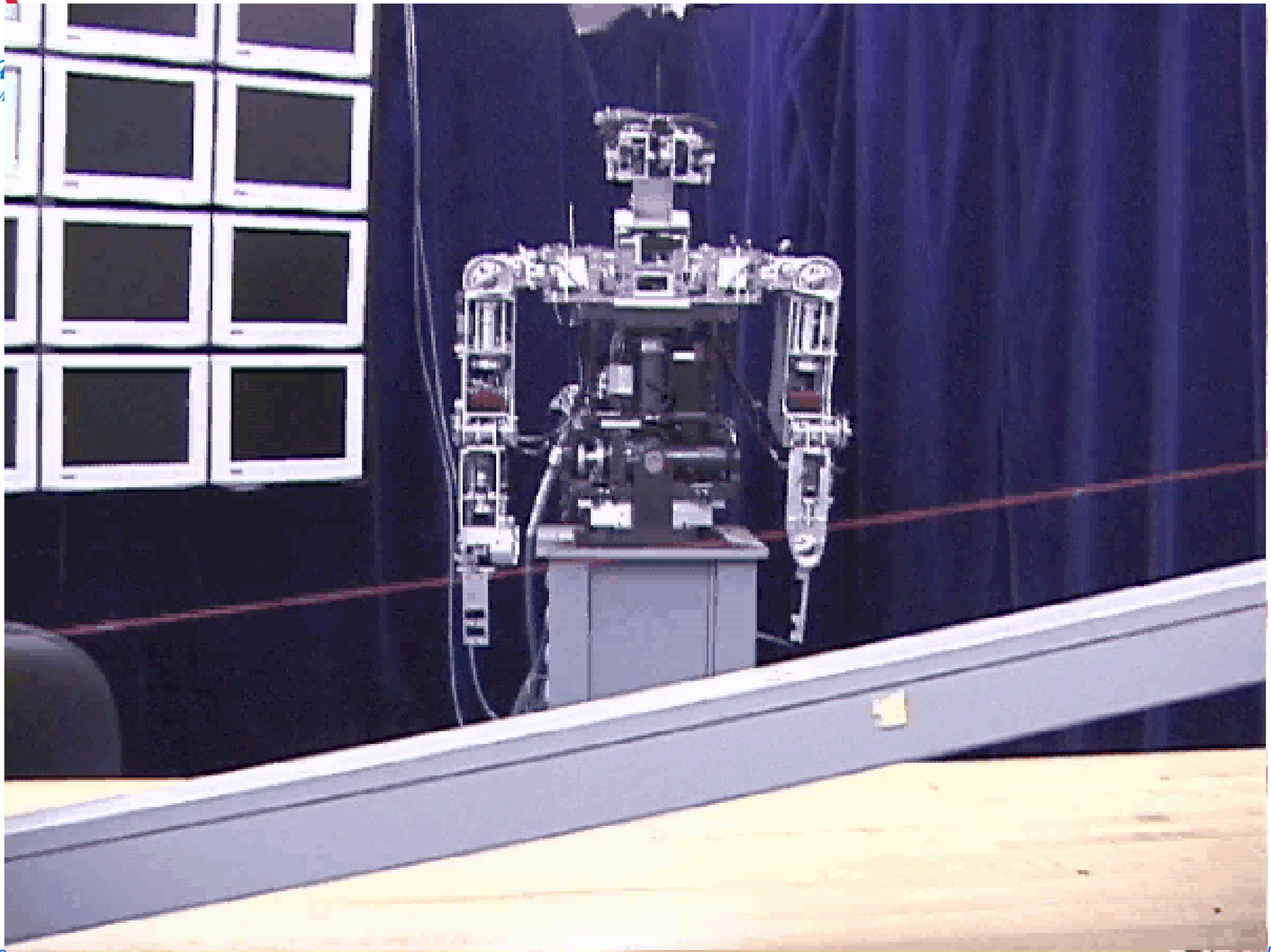


- **Mapping is based on perceived head position and robot's own symmetry axis**

Imitation Based on Social Cues



Imitating an Object



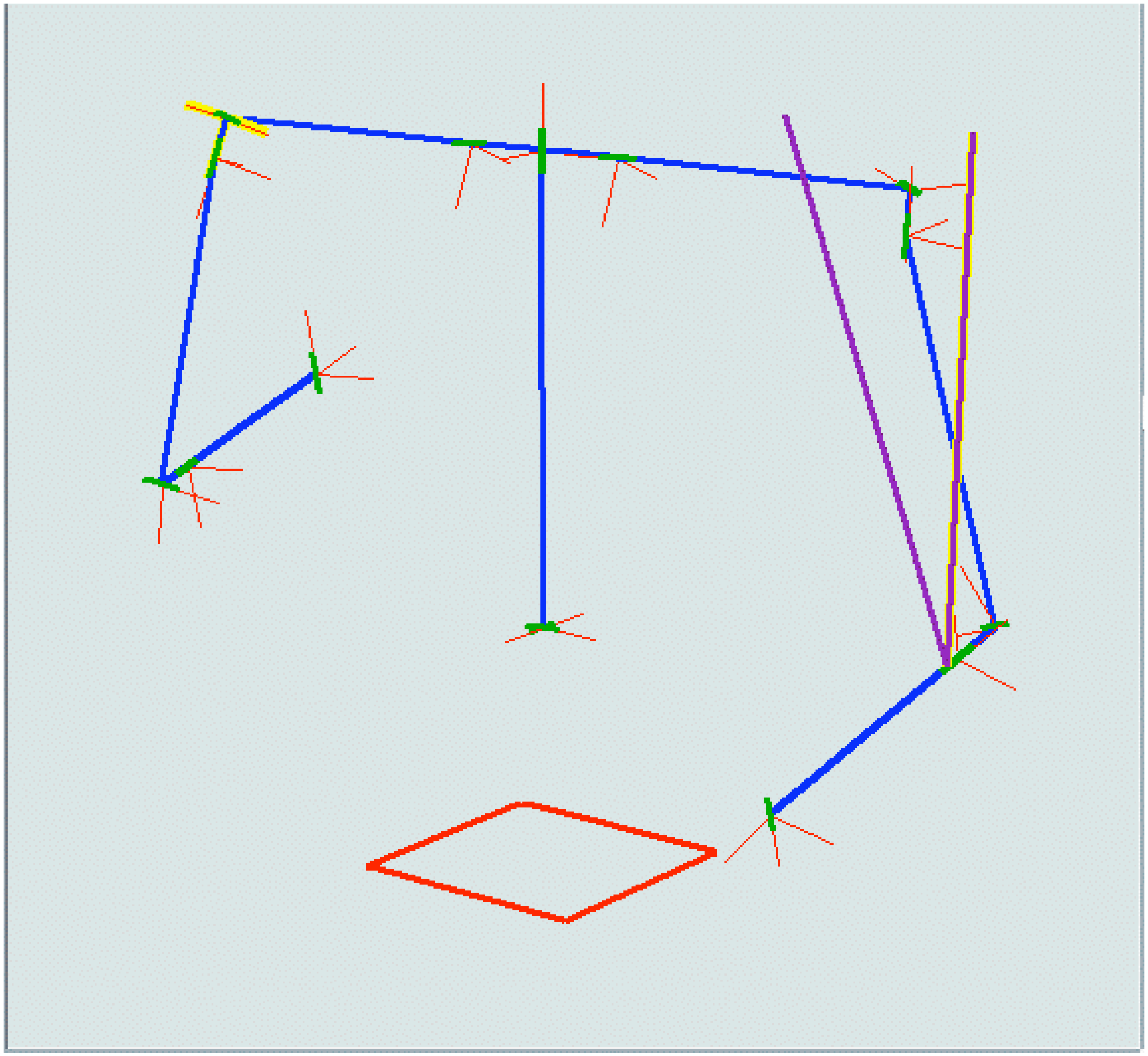
Our Approaches Lead to Specific Progress on Each Research Issue

Research Issues

Our Approaches

	Knowing what to imitate	Mapping between bodies	Correcting failures and recognizing success	Chaining actions together	Generalizing to more complex tasks	Intuitive Interactions
Social Interaction	Uses attentional cues to recognize task relevant events		Identifies and displays emotional states for reinforcement			Allows humans and robots to share similar social cues
Development	Limits search space by incremental refinement of perception	Simplifies mapping by incremental refinement of motor skills		Provides natural decomposition of complex tasks	Exploits incremental learning to build new skills	Provides methods for building social skills
Embodiment	Constrains movement which assists directed perception	Simplifies body mapping between human and robot	Enables simple but robust low-level behaviors	Places physical limits on successive actions		Allows the human to observe natural social feedback cues
Integration	Allows robot to recognize cues in multiple modalities		Increases robustness through redundancy		Assists in transfer to new modalities	Allows human to use natural modalities (voice, gesture, facial expressions)

Work in Progress: Muscle Groups



Work in Progress: Sequence Learning