– Face to Face –

Robot vision in social settings
Robot vision in social settings

- Humans (and robots) recover information about objects from the light they reflect
- Human head and eye movements give clues to attention and motivation
- In a social context, people constantly read each other’s actions for these clues
- Anthropomorphic robots can partake in this implicit communication, giving smooth and intuitive interaction
Humanoid robotics at MIT
Kismet
Kismet

- Built by Cynthia Breazeal to explore expressive social exchange between humans and robots
  - Facial and vocal expression
  - Vision-mediated interaction (collaboration with Brian Scassellati, Paul Fitzpatrick)
  - Auditory-mediated interaction (collaboration with Lijin Aryananda)
Vision-mediated interaction

- **Visual attention**
  - Driven by need for high-resolution view of a particular object, for example, to find eyes on a face
  - Marks the object around which behavior is organized
  - Manipulating attention is a powerful way to influence behavior

- **Pattern of eye/head movement**
  - Gives insight into level of engagement
Expressing visual attention

Attention can be deduced from behavior
Or can be expressed more directly
Building an attention system

Current input → Pre-attentive filters → Saliency map → Tracker

Primary information flow
Modulatory influence

Behavior system
Eye movement

Core ideas:
- Pre-attentive filters
- Saliency map
- Tracker
- Modulatory influence
- Primary information flow
Initiating visual attention

Current input

Skin tone

Color

Motion

Habituation

Saliency map

Weighted by behavioral relevance

Pre-attentive filters
Example filter – skin tone
Image pixel \( p(r,g,b) \) is NOT considered skin tone if:

- \( r < 1.1 \times g \) (red component fails to dominate green sufficiently)
- \( r < 0.9 \times b \) (red component is excessively dominated by blue)
- \( r > 2.0 \times \max(g,b) \) (red component completely dominates)
- \( r < 20 \) (red component too low to give good estimate of ratios)
- \( r > 250 \) (too saturated to give good estimate of ratios)

Lots of things that are not skin pass these tests

But lots and lots of things that are not skin fail
Modulating visual attention

low skin gain, high color saliency gain  high skin gain, low color saliency gain

Looking time –
28% face, 72% block

Looking time –
86% face, 14% block
Manipulating visual attention
Maintaining visual attention
Persistence of attention

- Want attention to be responsive to changing environment
- Want attention to be persistent enough to permit coherent behavior
- Trade-off between persistence and responsiveness needs to be dynamic
Influences on attention

Current input → Pre-attentive filters → Saliency map → Tracker

Primary information flow

Modulatory influence

Behavior system

Eye movement
Eye movement

Ballistic saccade to new target

Smooth pursuit and vergence co-operate to track object

Right eye

Left eye

Vergence angle
Eye/neck motor control

- Neck movements combine :-
  - Attention-driven orientation shifts
  - Affect-driven postural shifts
  - Fixed action patterns

- Eye movements combine :-
  - Attention-driven orientation shifts
  - Turn-taking cues
Active vision involves choosing a robot’s pose to facilitate visual perception.

Focus has been on immediate physical consequences of pose.

For anthropomorphic head, active vision strategies can be “read” by a human, assigned an intent which may then be completed beyond the robot’s immediate physical capabilities.

Robot’s pose has communicative value, to which human responds.
Comfortable interaction distance

Too close – withdrawal response

Comfortable interaction distance

Person draws closer

Too far – calling behavior

Beyond sensor range

Comfortable interaction speed

Too fast, Too close – threat response

Too fast – irritation response
Video: Withdrawal response
Kismet’s cameras

- Right eye pan
- Left eye pan
- Eye tilt

- Camera with wide field of view
- Camera with narrow field of view
Simplest camera configuration

- Single camera
  - Multiple camera systems require careful calibration for cross-camera correspondence
- Wide field of view
  - Don’t know where to look beforehand
- Moving infrequently relative to the rate of visual processing
  - Ego-motion complicates visual processing
Missing components

- High acuity vision – for example, to find eyes within a face
  - Need cameras that sample a narrow field of view at high resolution
- Binocular view, for stereoscopic vision
  - Need paired cameras
  - May need wide or narrow fields of view, depending on application
Typical visual tasks require both high acuity and a wide field of view.

High acuity is needed for recognition tasks and for controlling precise visually guided motor movements.

A wide field of view is needed for search tasks, for tracking multiple objects, compensating for involuntary ego-motion, etc.
A common trade-off found in biological systems is to sample part of the visual field at a high enough resolution to support the first set of tasks, and to sample the rest of the field at an adequate level to support the second set.

This is seen in animals with foveate vision, such as humans, where the density of photoreceptors is highest at the center and falls off dramatically towards the periphery.
Simulated example

- Compare size of eyes and ears in transformed image – eyes are closer to center, and so are better represented
(From C. Graham, “Vision and Visual Perception”)
Mechanical approximations

- Imaging surface with varying sensor density (Sandini et al)
- Distorting lens projecting onto conventional imaging surface (Kuniyoshi et al)
- Multi-camera arrangements (Scassellati et al)
- Cameras with zoom control directly trade-off acuity with field of view (but can’t have both)
- Or do something completely different!
Multi-camera arrangement

Wide view camera gives context used to select region at which to point narrow view camera

If target is close and moving, must respond quickly and accurately, or won’t gain any information at all

Wide field of view, low acuity

Narrow field of view, high acuity
Mixing fields of view

- Small distance between cameras with wide and narrow field of views, simplifying mapping between the two
- Central location of wide camera allows head to be oriented accurately independently of the distance to an object
- Allows coarse open-loop control of eye direction from wide camera – improves gaze stability
Tip-toeing around 3D

Field of view

Object of interest

Wide View camera

Narrow view camera

Rotate camera

New field of view
Using 3D

- Right eye pan
- Left eye pan
- Eye tilt

Fixed with respect to head
Kismet’s little secret
speech synthesis
affect recognition
Linux
Speech
recognition

Cameras
Eye, neck, jaw motors

Face
Control
Percept & Motor
Emotion
Drives & Behavior

Ear, eyebrow, eyelid, lip motors

QNX
Motor ctrl
Tracker
Eye finder
Attent. system
Dist. to target
Motion filter
Skin filter
Color filter
audio speech comms
dual-port RAM

NT
speech synthesis
affect recognition

sockets, CORBA
 Speakers

CORBA

Linux
Speech recognition

Microphone

CORBA

Cameras
Eye, neck, jaw motors

Ear, eyebrow, eyelid, lip motors
- Responsiveness to the human face is vital for a robot to partake in natural social exchange
- Need to locate and track facial features, and recover their semantic content
Modeling the face

- Match oriented regions on face against vertical model to isolate eye/brow region
- Match eye/brow region against horizontal model to find eyes, bridge
- Each model scans one spatial dimension, so can formulate as HMM, allowing fast optimization of match
It is useful to link the robot’s representation of its own face with that of humans.

Conclusion

- Vision community working on improving machine perception of human
- But equally important to consider human perception of machine
- Robot’s point of view must be clear to human, so they can communicate effectively – and quickly!
Video: Turn taking
Video: Affective intent