Perceptive Context

Trevor Darrell
Vision Interface Group
MIT CSAIL

Perceptive Context

Awareness of the User -- Visual Conversation Cues:

*Interfaces (kiosks, agents, robots…) are currently **blind** to users…machines should be aware of presence, pose, expression, and non-verbal dialog cues…*

Awareness of the Environment -- Perceptive Devices:

*Mobile devices (cellphones, PDAs, laptops) bring computing and communications with us wherever we go, but they are **blind** to their environment…they should be able to see things of interest in the environment just as we do…*

Today

- Visually aware conversational interfaces ("*read my body language!*")
  - head modeling and pose estimation
  - articulated body tracking

- Mobile devices that can see their environment ("*what's that thing there?*")
  - mobile location specification
  - image-based mobile web browsing

Head modeling and pose tracking

3D Head Pose Tracker

Stereo camera

Current frame

intensity

range

rigid stereo motion estimation
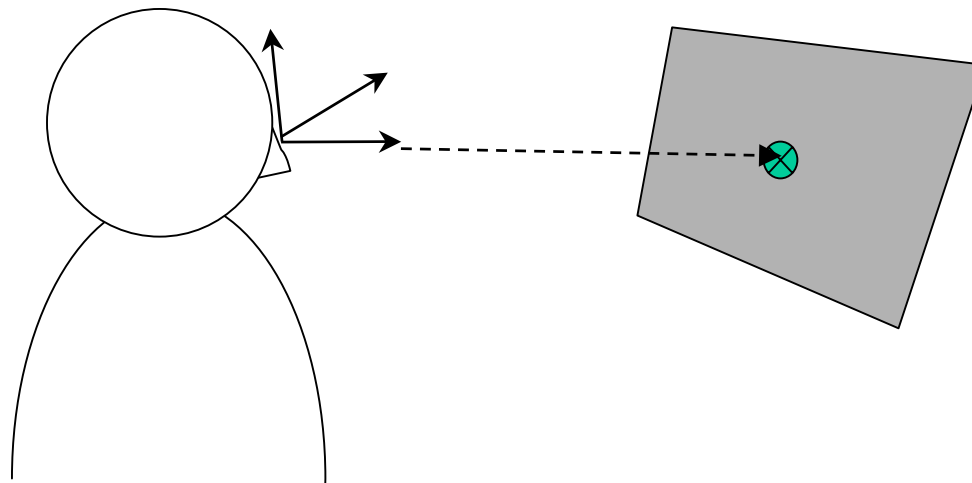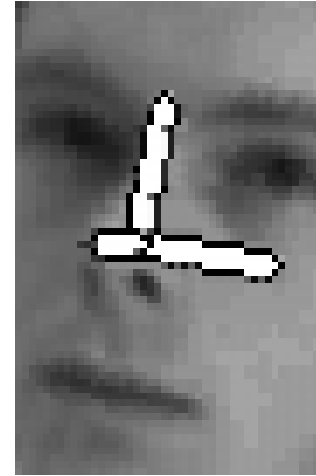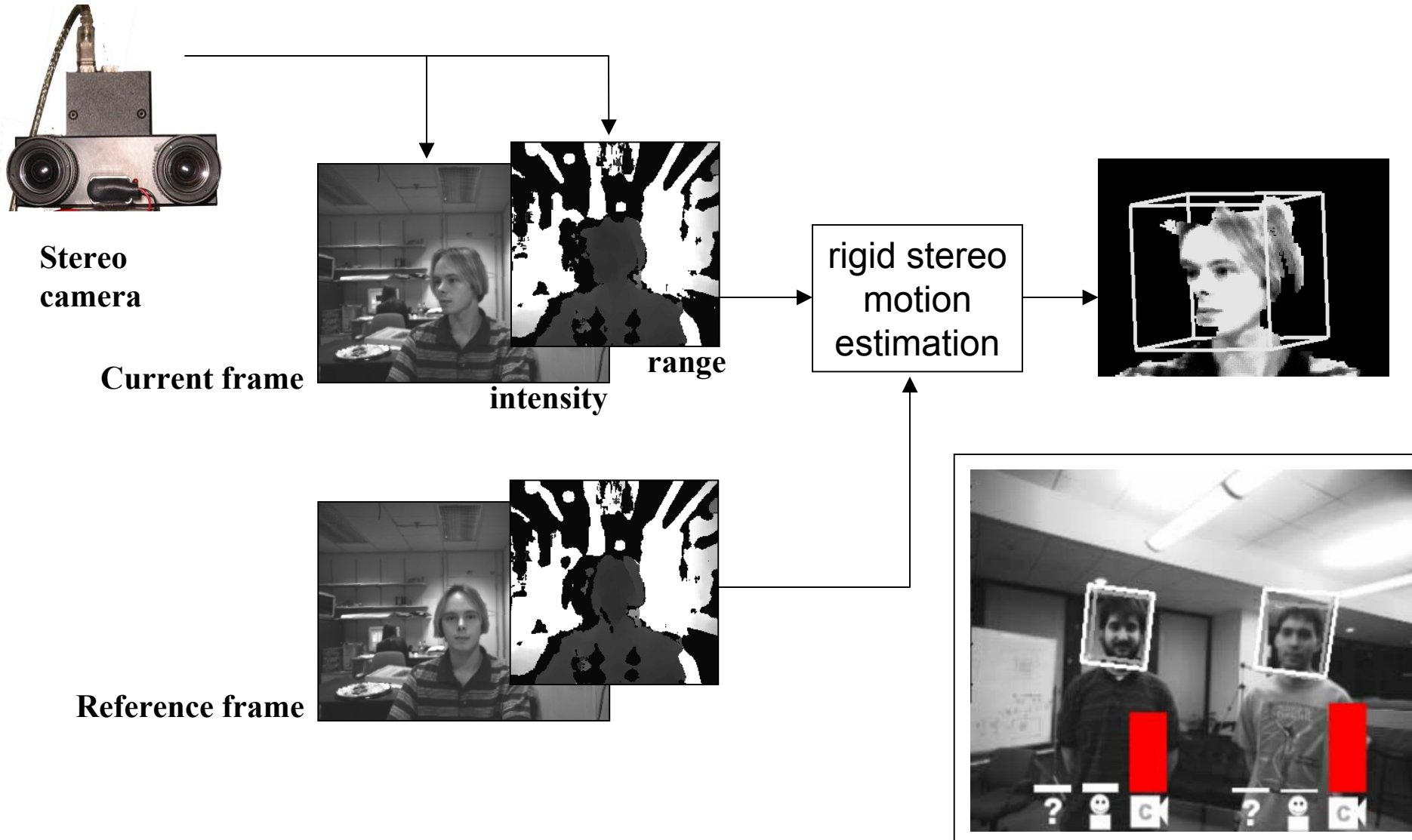
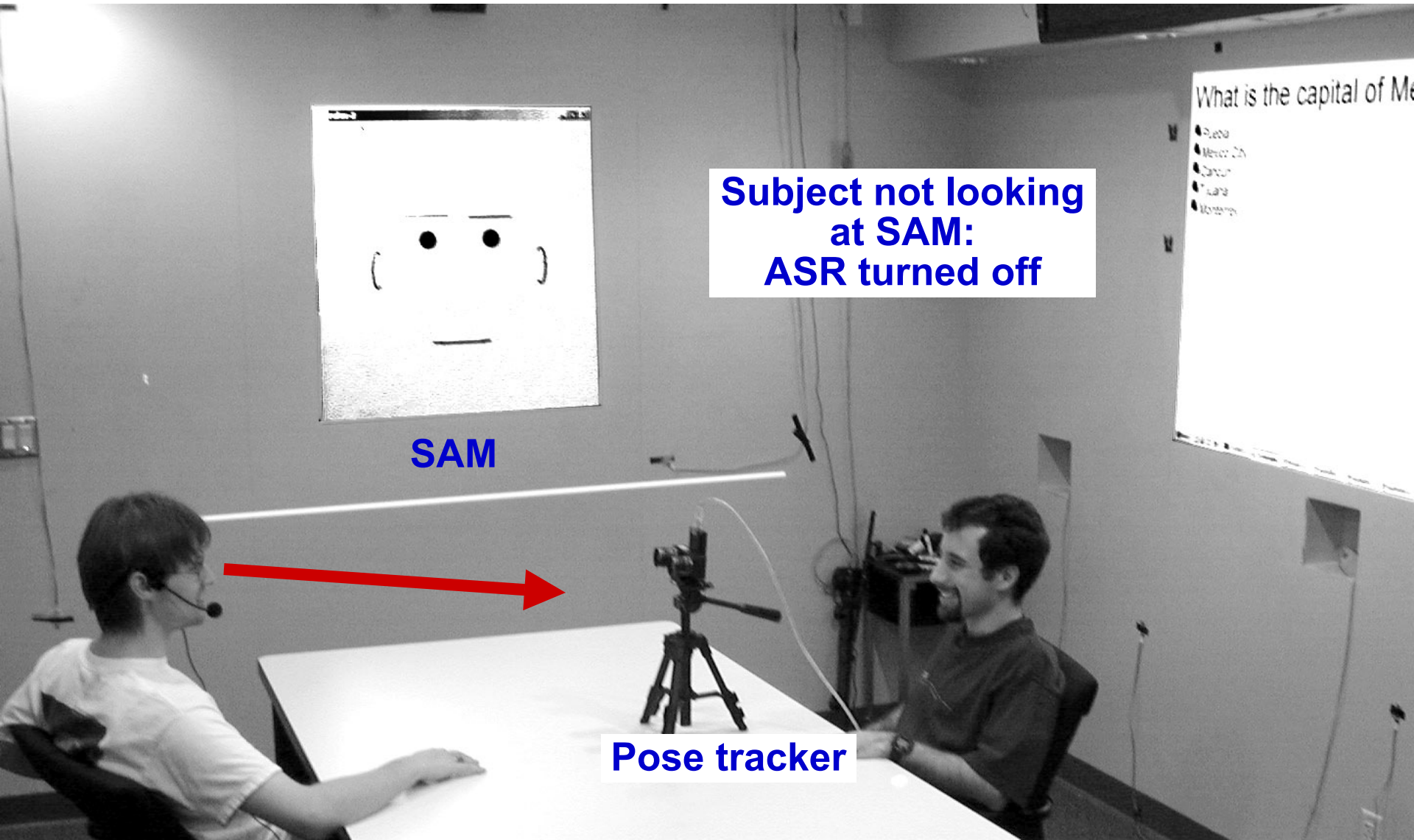Reference frame

Face aware interfaces

- Agent should know when it's being attended to
- Turn-taking discourse cues: who is talking to whom?
- Model attention of user
- Agreement: head nod and shake gestures
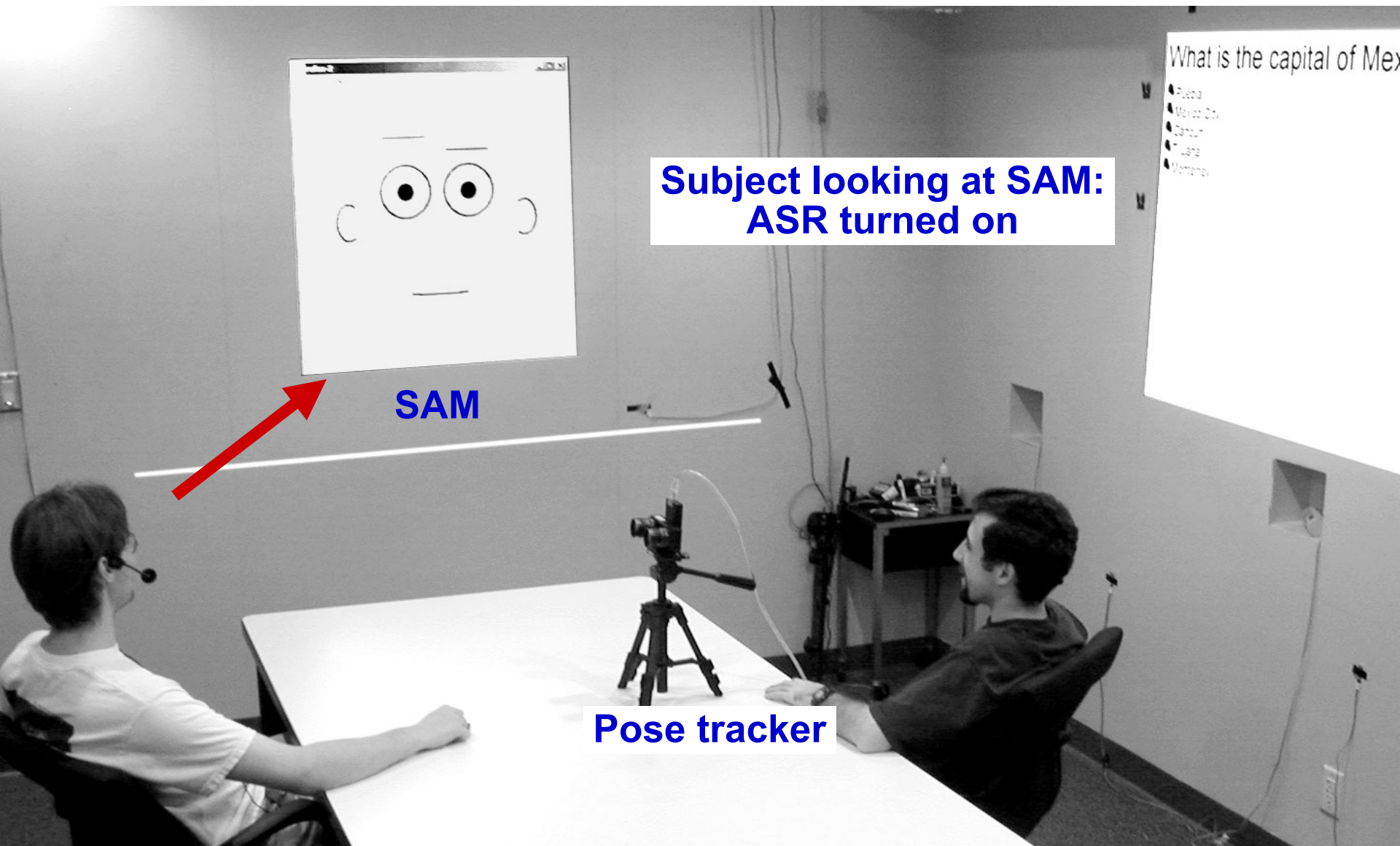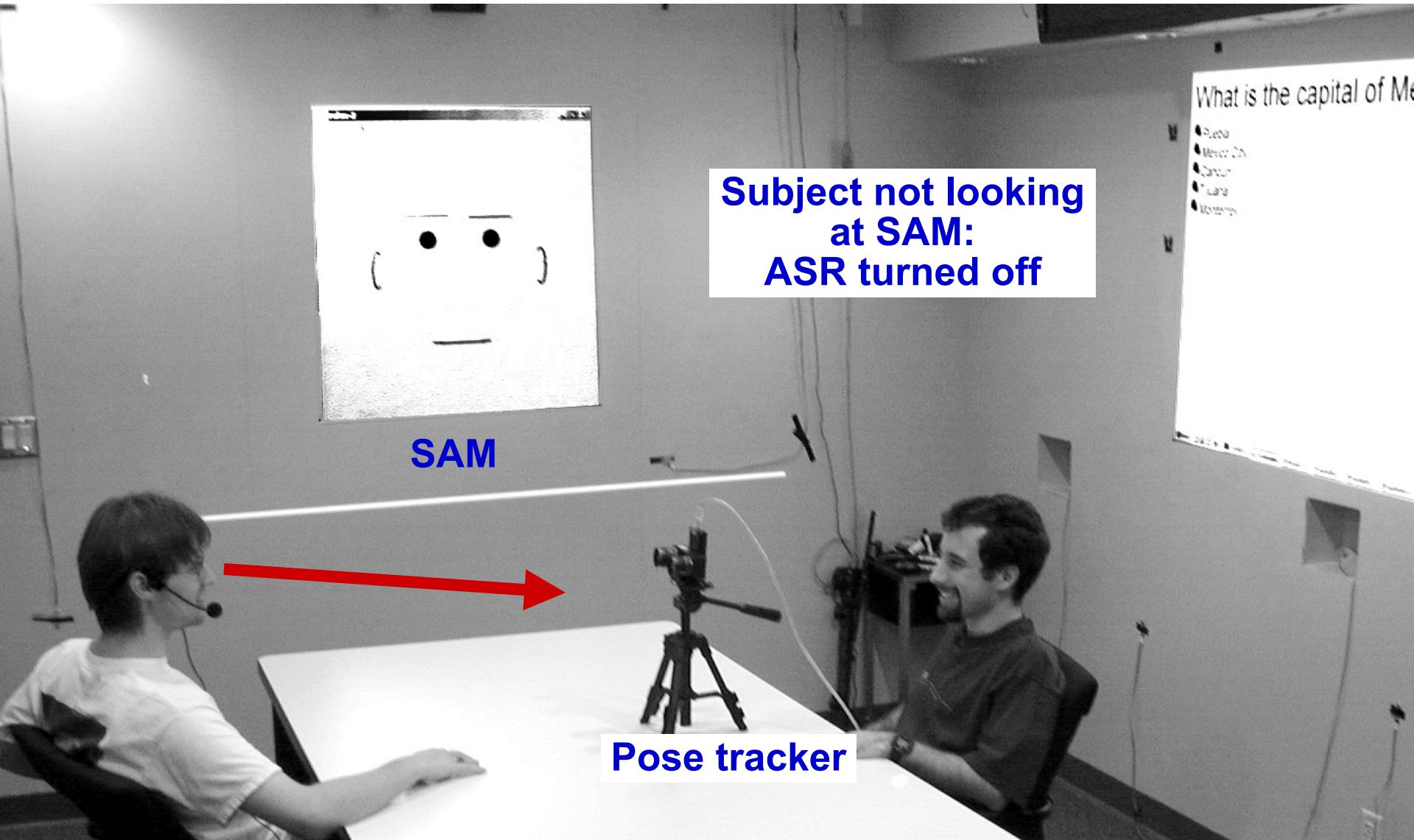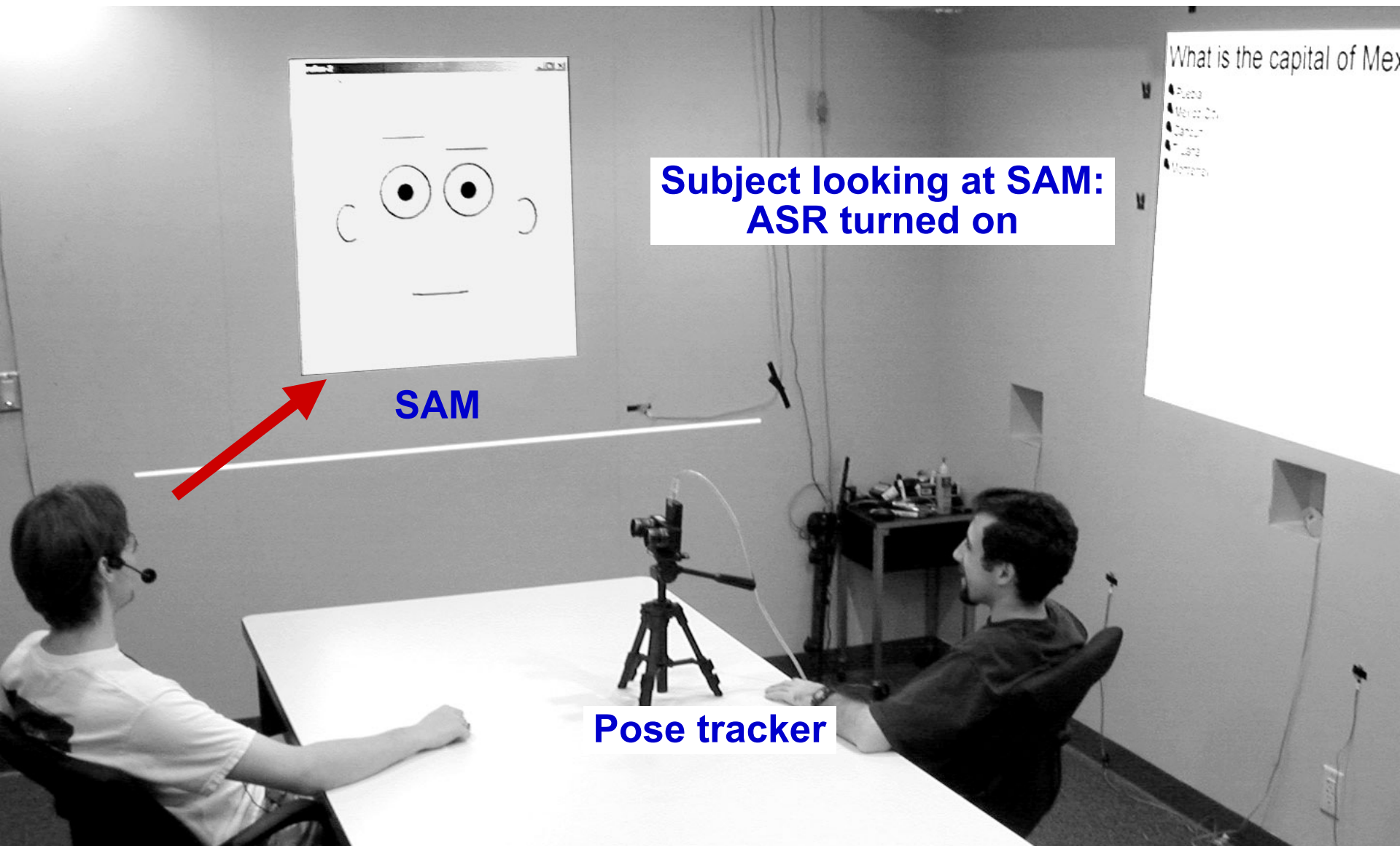- Grounding: shared physical reference

Face cursor

Face-responsive agent

SAM

Subject not looking at SAM: ASR turned off

What is the capital of Me

Pose tracker

Face-responsive agent

Face-responsive agent

SAM
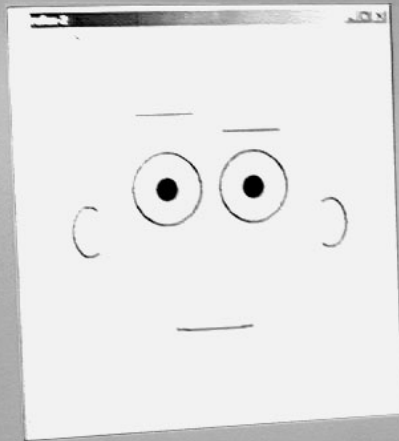
Subject not looking
at SAM:
ASR turned off

Pose tracker

What is the capital of Me

Face-responsive agent

What is the capital of Mex

Subject looking at SAM: ASR turned on

SAM

Pose tracker

Face-responsive agent

Subject looking at SAM:
ASR turned on

SAM

- General conversational turn-taking
- Agreement (Nod/Shake)
- Grounding / Object reference…

Room tracking for Location Context

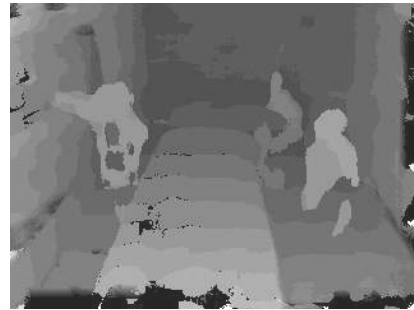Location is an important cue for pervasive computing applications…

- Location context should provide a finer scale cue than room-ID, but more abstract than 3-space position and orientation.

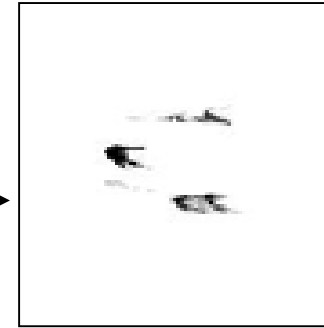- Regions ("zones") should be learned from observing actual user behavior.

Room   **Range**   **Foreground**   **Plan view**
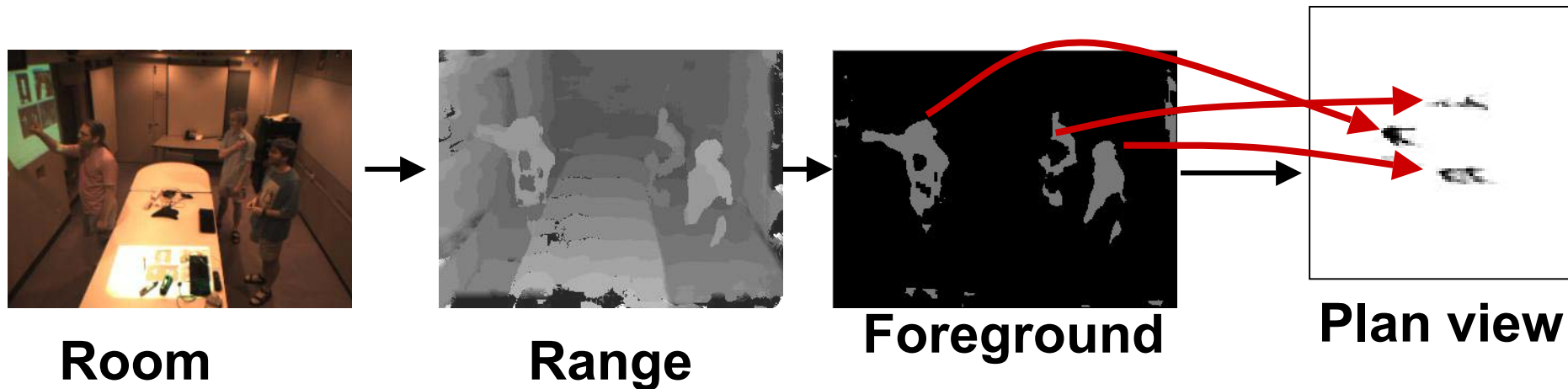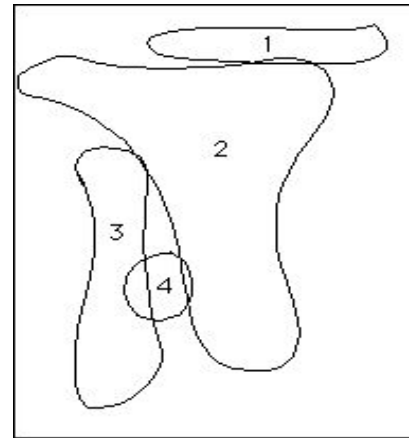
Learning activity zones

Room

Range

Foreground

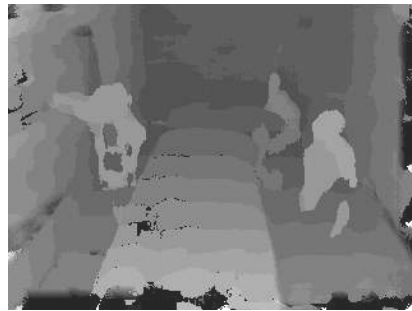**Plan view**

Activity zones

Motion Clustering

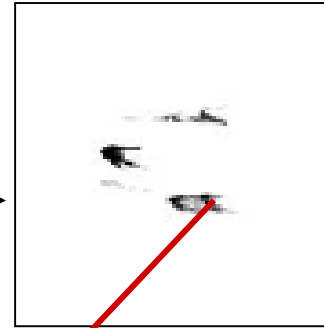*Zone map formed from observing user behavior*

Using activity zones

Room

Range

Foreground

Plan view

Activity zones
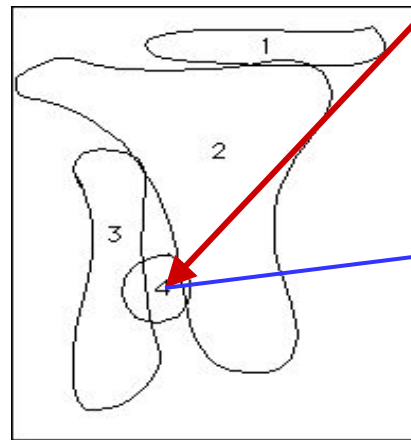
zone 4 prefs

Current zone determines application context

[Koile, Darrell, et. al, UBICOMP 03]
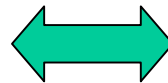
Articulated pose sensing
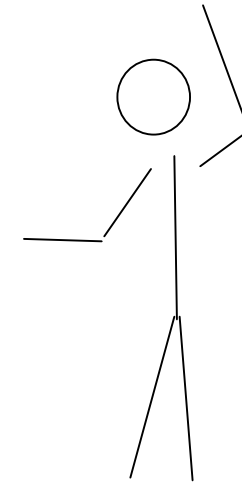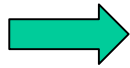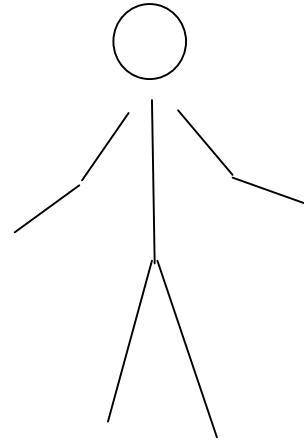
Model-based Approach

depth image

ICP with
articulation
constraint

model
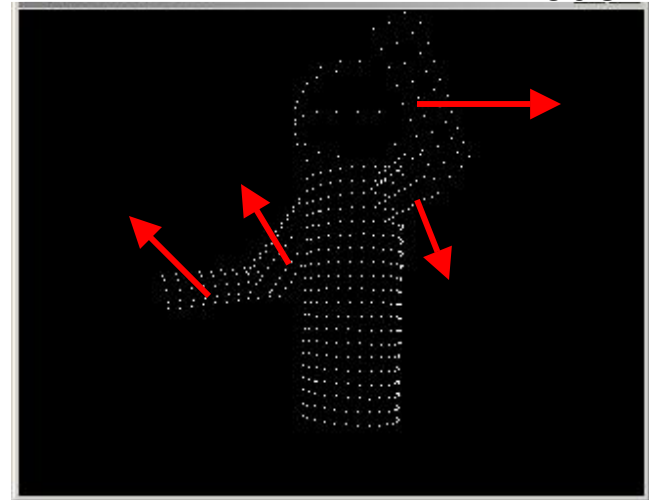
1. Find closest points
2. Update poses
3. Constrain…

slow

Interactive Wall

Multimodal studio

multimodal_studio.mpeg

Articulated Pose from a single image?

Model based approach difficult with more impoverished observations:

- contours
- edge features
- texture
- (noisy stereo…)

hard to fit a single image reliably!

➢ *Example-based learning paradigm*

Example-based matching

- Match 2-D features against large corpus of 2-D to 3-D example mappings

- Fast hashing for approximate nearest neighbor search

- Feature selection using paired classification problem

- Data collection: use motion capture data, or exploit synthetic (but realistic) models

Parameter sensitive hashing
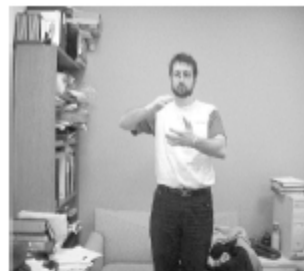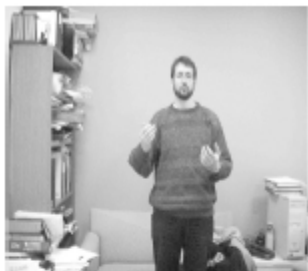

Input image   Compute features   Find similar examples

Robust LWR

2D->3D with Parameter sensitive hashing

INPUT

TOP MATCH
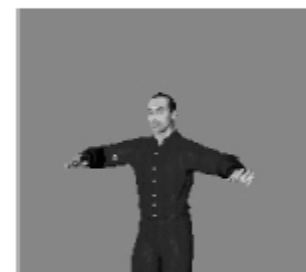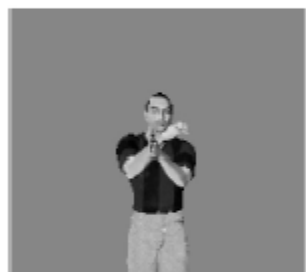
RLWR

Today

- Visually aware conversational interfaces -- *read my body language!*

  - head modeling and pose estimation

  - articulated body tracking

- Mobile devices that can see their environment -- *what's that thing there?*

  - mobile location specification

  - image-based mobile web browsing

Physical awareness

How can device be aware of what user is looking at?

User

Physical awareness

Asking a friend, "What's this?"

User

Human Expert

What is this?

MIT Dome

IDeixis
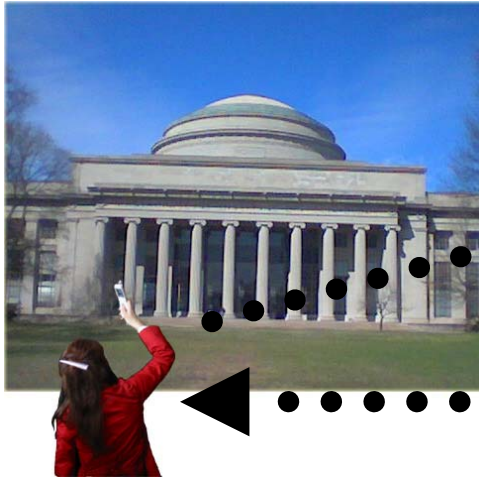
Instead, use CBIR (Content-based Image Retrieval) system:

User　　　　　　　　　　　　　　　　　　　　　　　　**CBIR System**
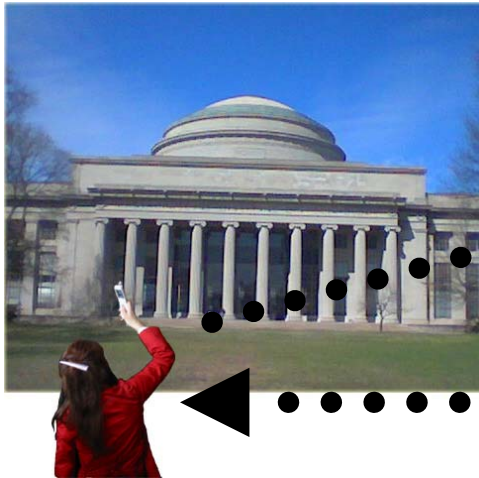
What is this?

http://mit.edu/..

CBIR: Content-based Image Retrieval

- Use image (or video) query to database.
- For place recognition, many current matching methods can be successful
  - PCA
  - Gobal orientation histograms [Torralba et al.]
  - Local features (Affine-invariant detectors/descriptors [Schmid], SIFT [Lowe], etc.)

...where to get the database?

The Web

- Many location images can be found on the web
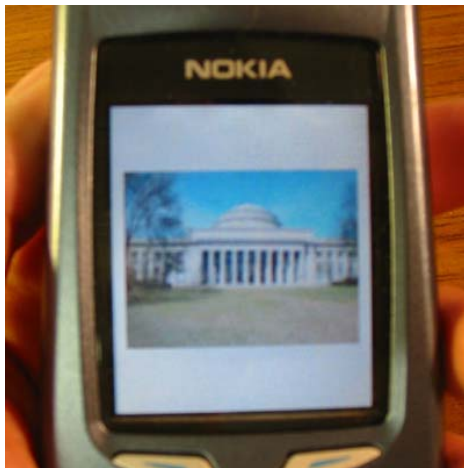
First Prototype

1. Take an Image

2. Send image via MMS

3. View search result (matching location images)

4. Browse a relevant webpage

Images -> keywords (-> images)

- Hard to compile an image database of entire web!
- But given matches in subset of web:
    - Extract salient keywords
    - Keyword-based image search
    - Apply content-based filter to keyword-matched pages
- And/or allow direct keyword search
- Weighted term/bigram frequency sufficient for early experiments…

Bootstrap image web search
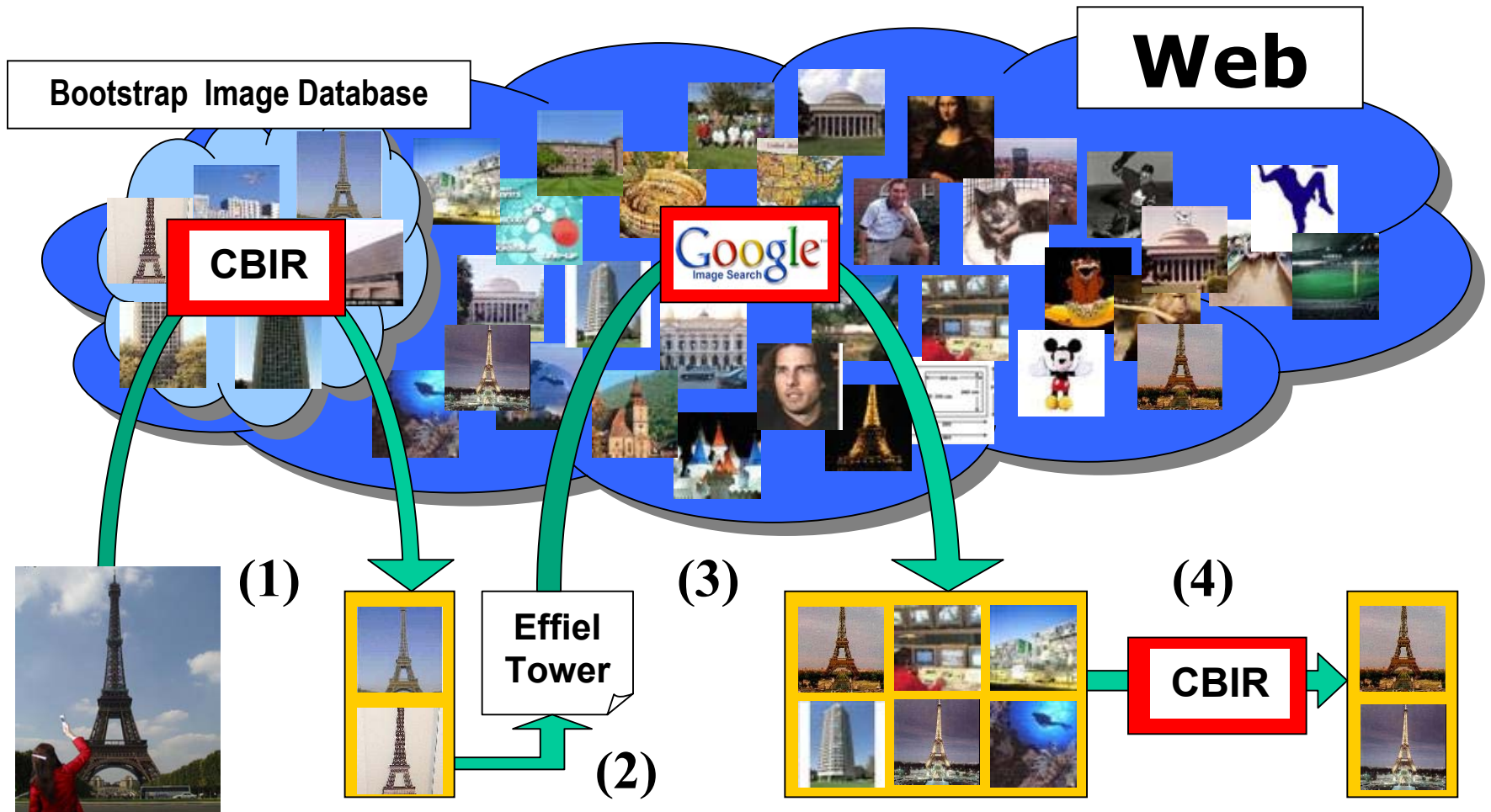
Advantages

- Recognizing distant location (by taking photo)

- Infrastructure free (by using the web)

- Large-scale image-based web search (by bootstrapping keywords)

- With advances in segmentation, can apply to many other object recognition problems
  - mobile signs
  - appliance
  - product packaging

Visual Interfaces and Devices

Interfaces (kiosks, agents, robots…) are currently blind to users…machines should be aware of presence, pose, expression, and non-verbal dialog cues…

Mobile devices (cellphones, PDAs, laptops) bring computing and communications with us wherever we go, but they are blind to their environment…they should be able to see things of interest in the environment just as we do…

Acknowledgements

David Demirdjian

Kimberlie Koile

Louis Morency

Greg Shakhnarovich

Mike Siracusa

Konrad Tollmar

Tom Yeh

& many others…

END