A System for Wearable Audio Navigation (SWAN)

Integrating Advanced Localization and Auditory Display

Frank Dellaert and Bruce Walker
College of Computing, School of Psychology
Georgia Institute of Technology

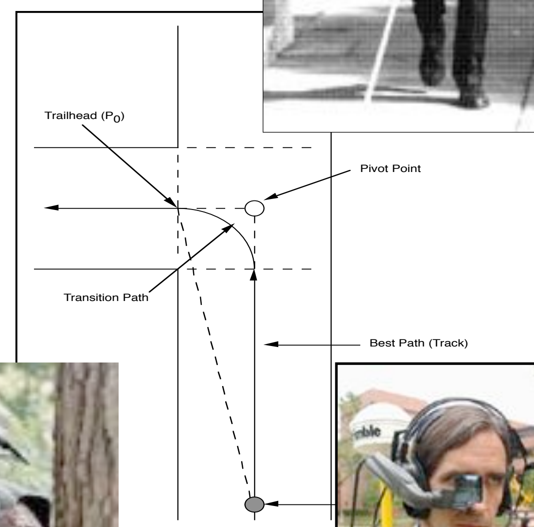
NSF IIS-0534332

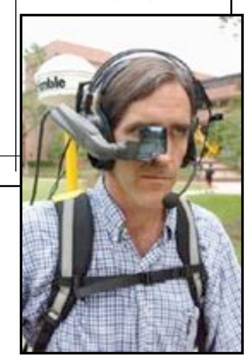
Outline

- Motivation
- Localization
  - GPS
  - Vision-based
- Auditory Display & Sonification
- Evaluation/Results
- Future Directions

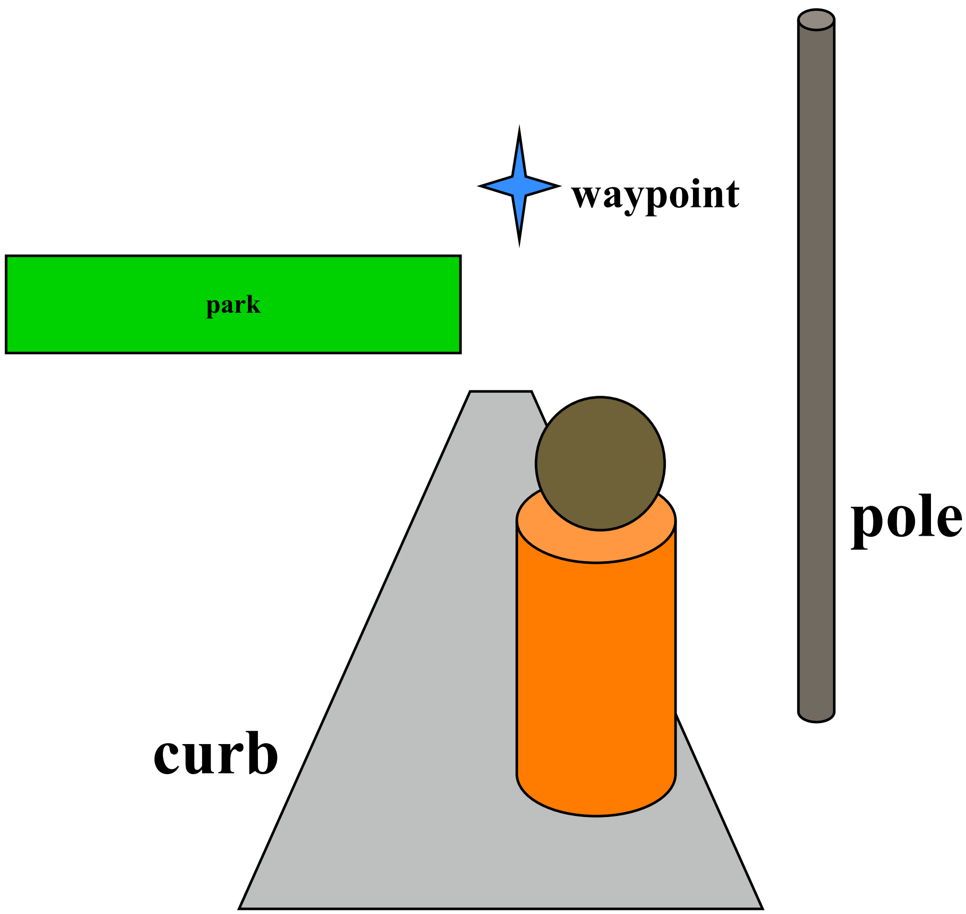
Motivation for SWAN

- System for Wearable Audio Navigation
- Wayfinding tool for those who cannot ***look*** or cannot ***see***
- Accessibility applications (blind)
- Tactical applications



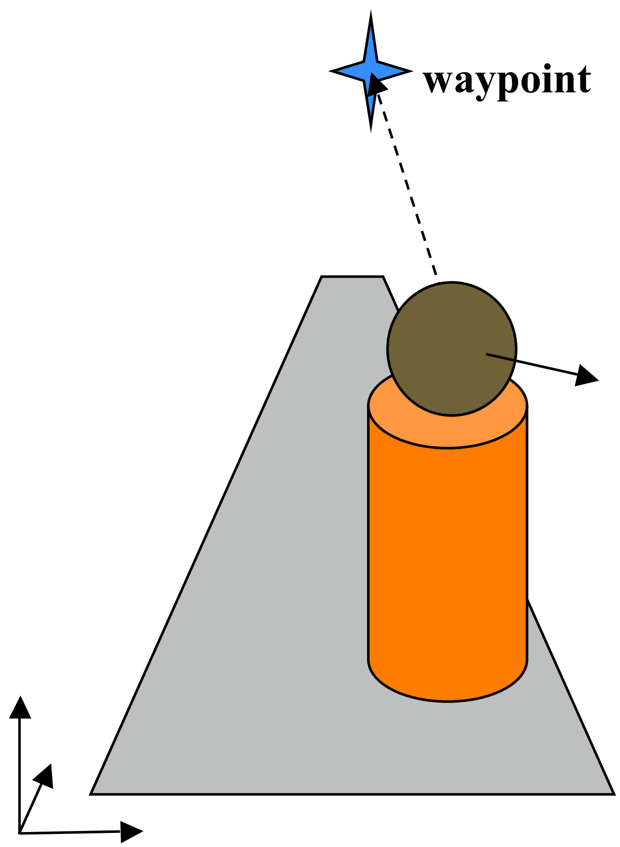


Wayfinding via Auditory Display



- Determine user's location
- Figure out what's around them (parks, curbs, poles, buildings, benches, etc.)
- Represent each object with unique sounds
- Listener learns what a location "sounds like"
- Also add audio waypoints along a path to destination

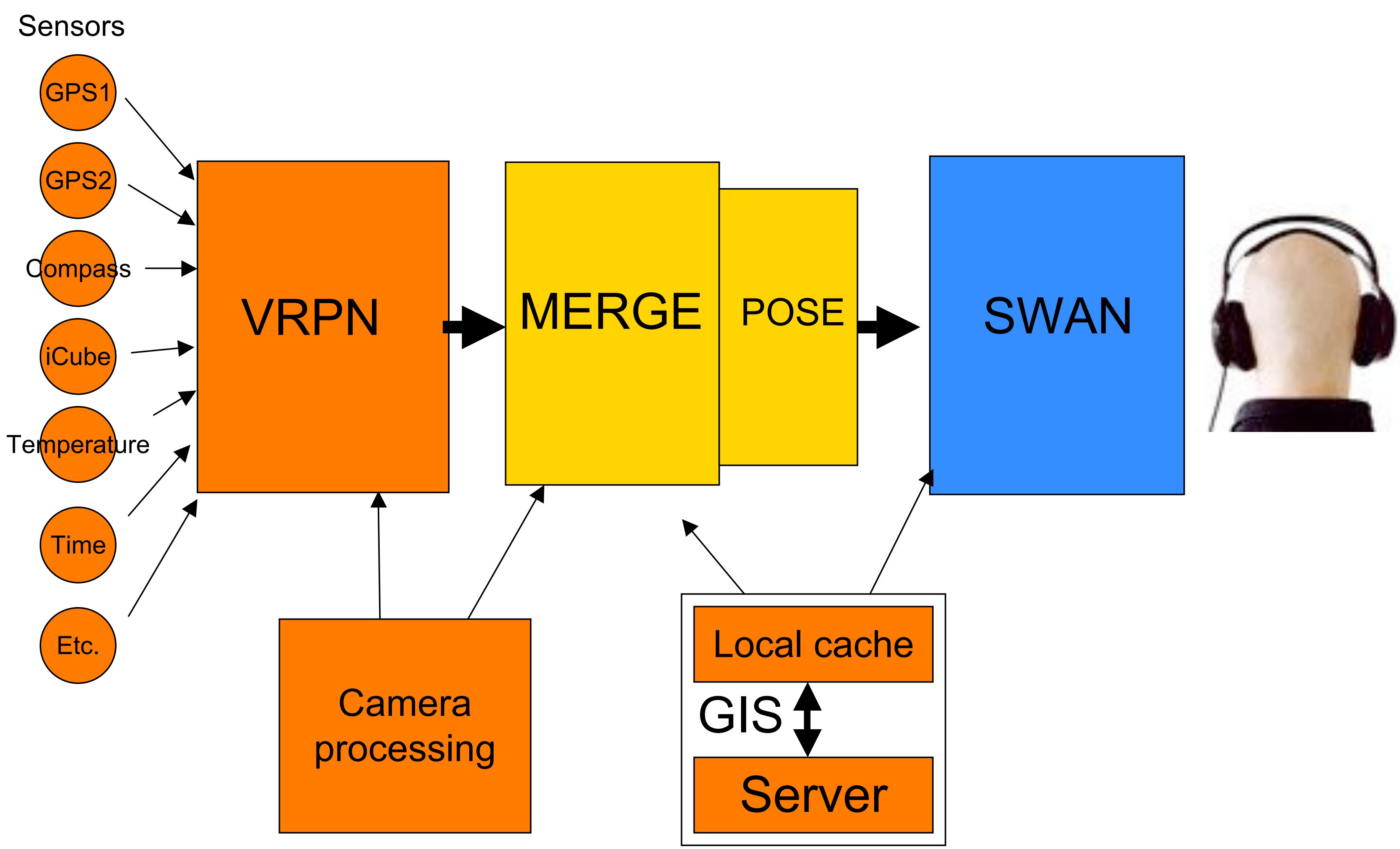
Attach Sounds to Objects: How?



- Accurate Head Pose
- Transform Object into head-centered coordinates
- 3D Sonification

- 6DOF Needed !
- GPS can't do it alone

SWAN System Overview

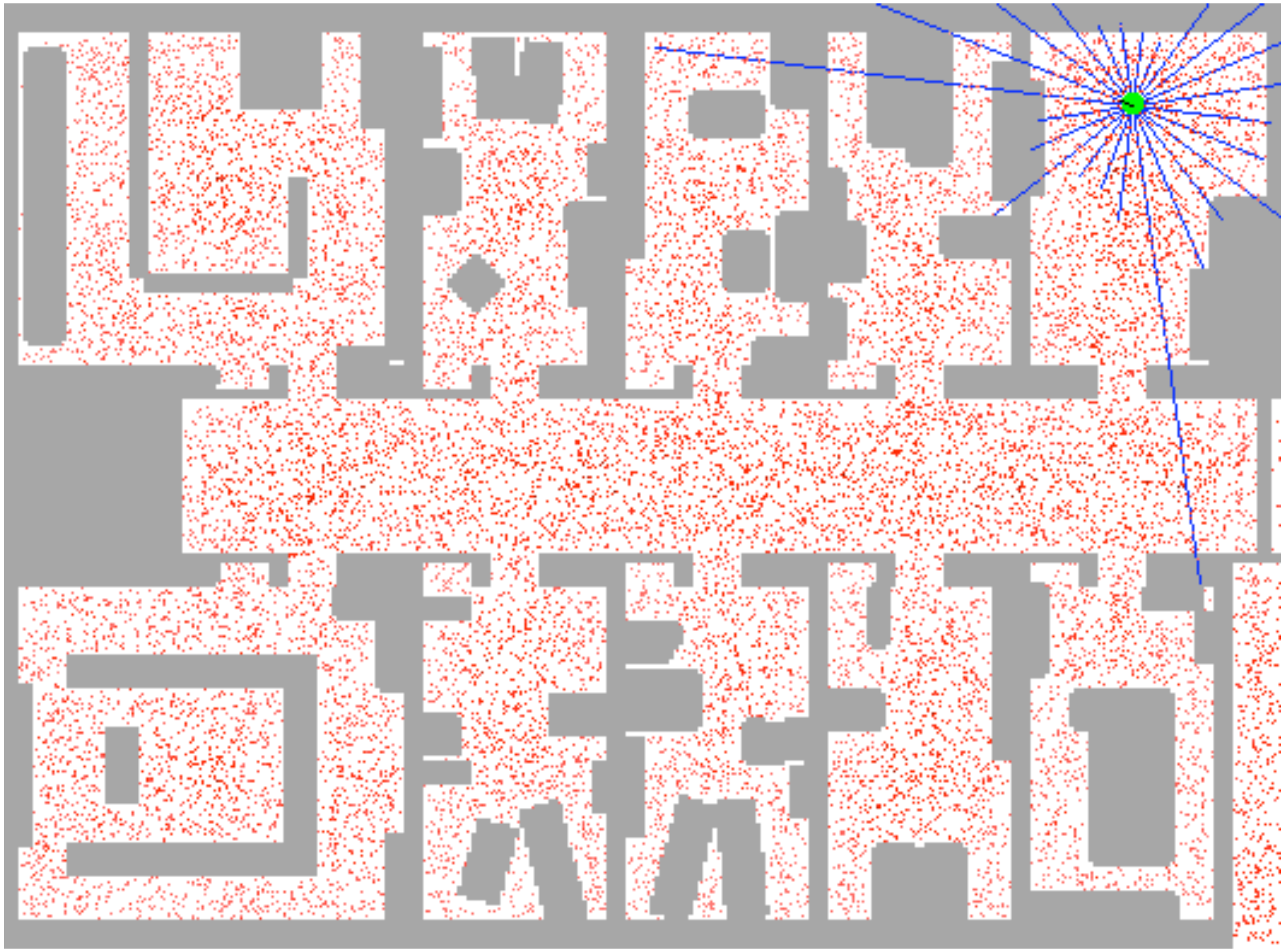


Localization

- Multiple Sensors, Sensor Fusion Required
  - Cameras
  - Maps
  - GPS
  - Compass
  - Head tracker
  - Thermometer
  - Light meter
  - Clock, calendar
  - etc.

Particle Filters

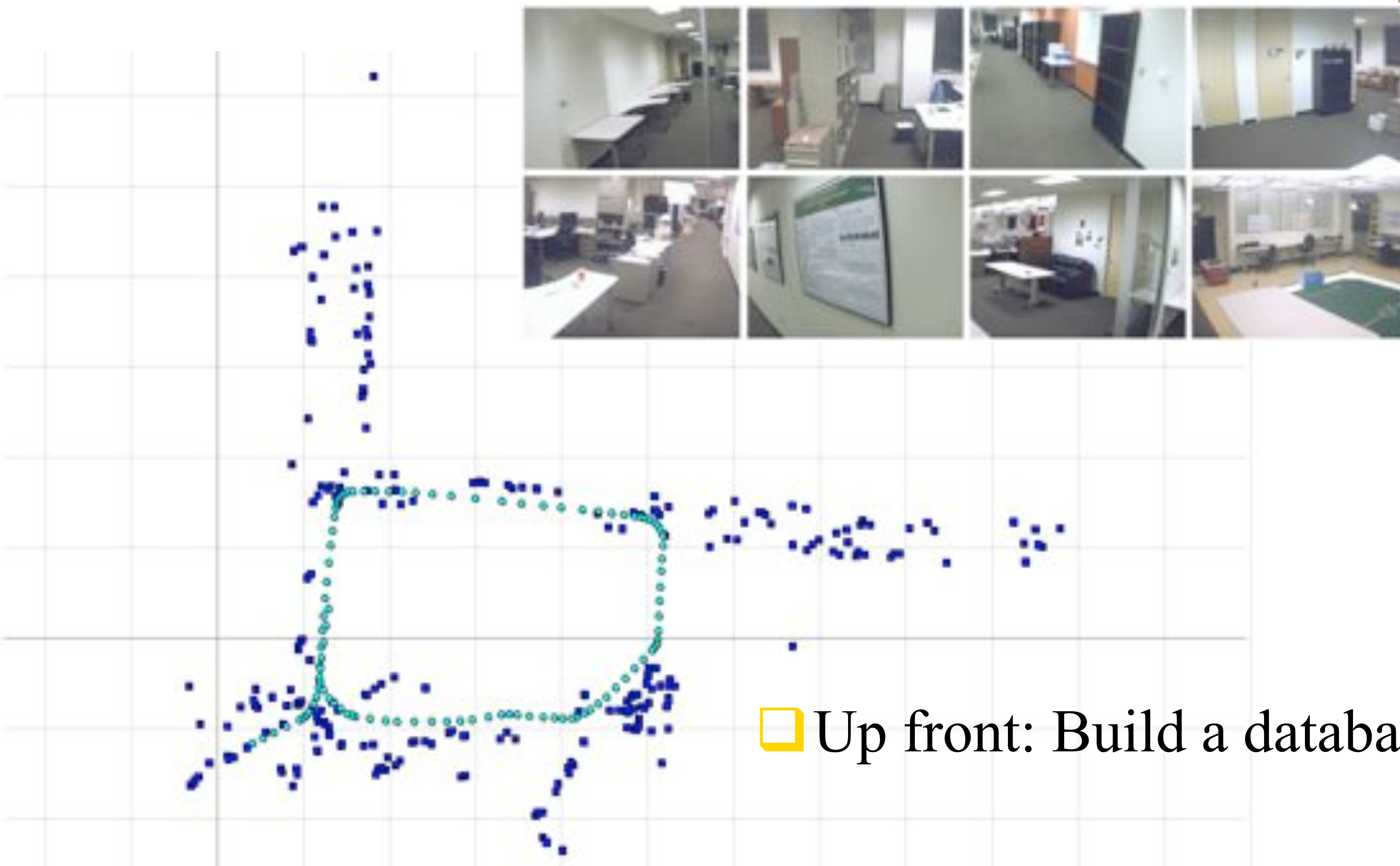
- Samples approximate 2D pose probability density



Map-based Priors

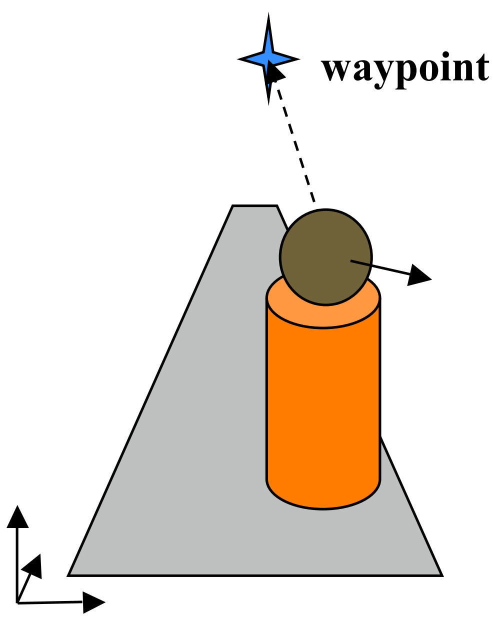
- Maps fetched from GIS
- Biases particle filter to stay on course

Vision-Based Localization

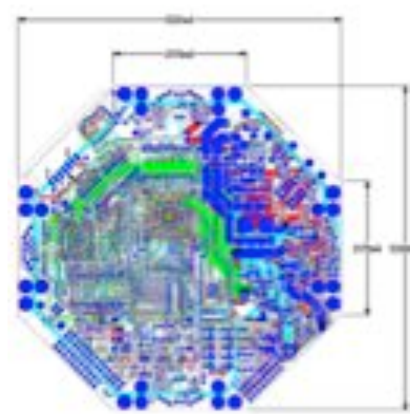


- Up front: Build a database

Vision-based Localization (run-time)

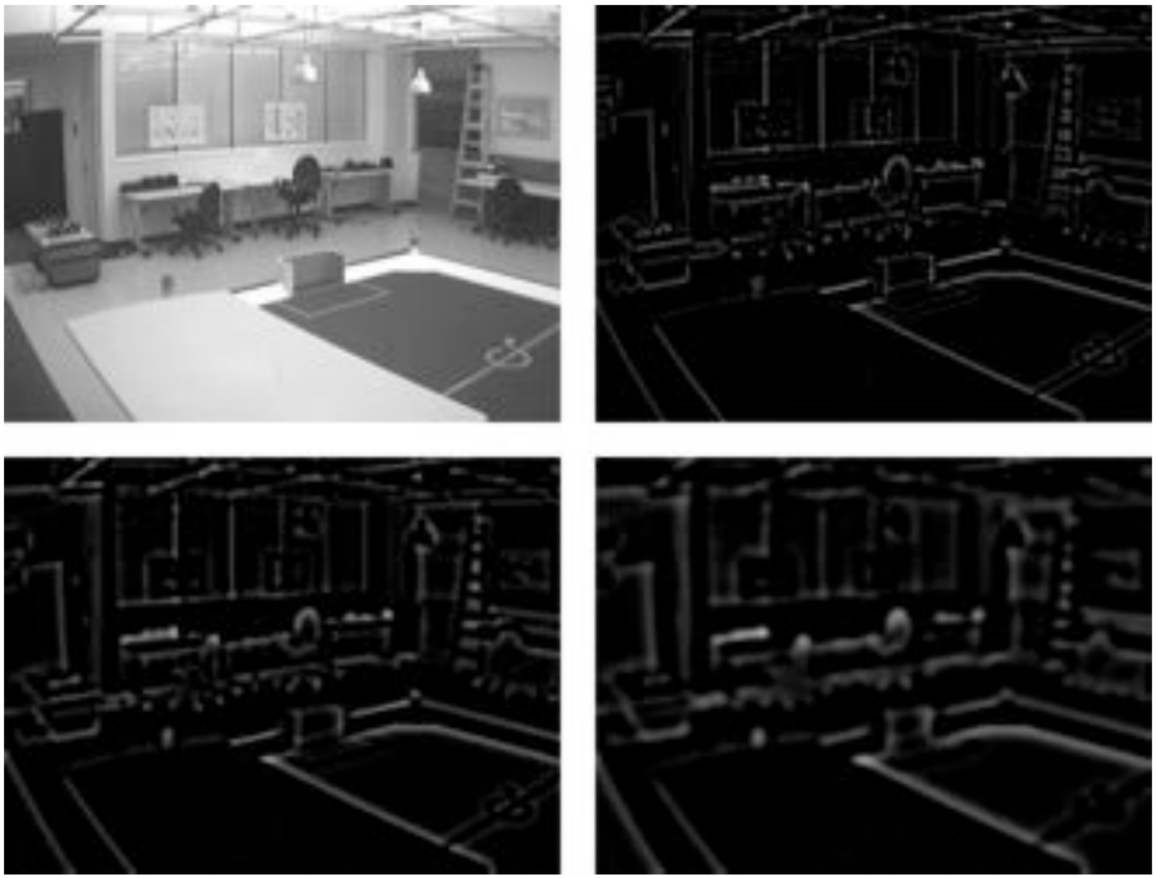


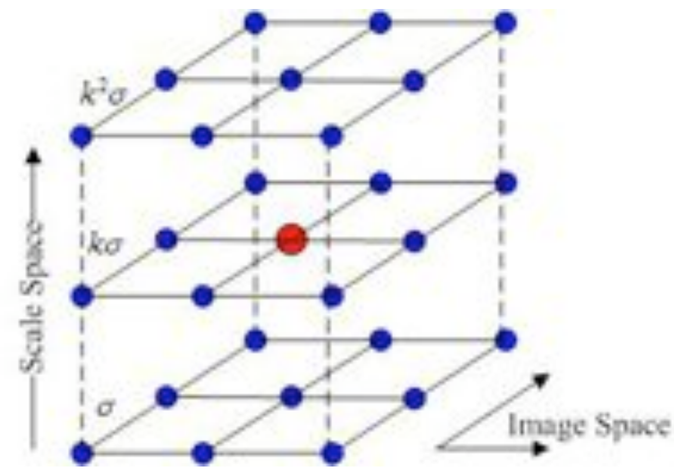
- At run-time: Accurate 6DOF localization using a multi-camera rig
- FPGA hardware for real-time performance



DOG Feature Detection

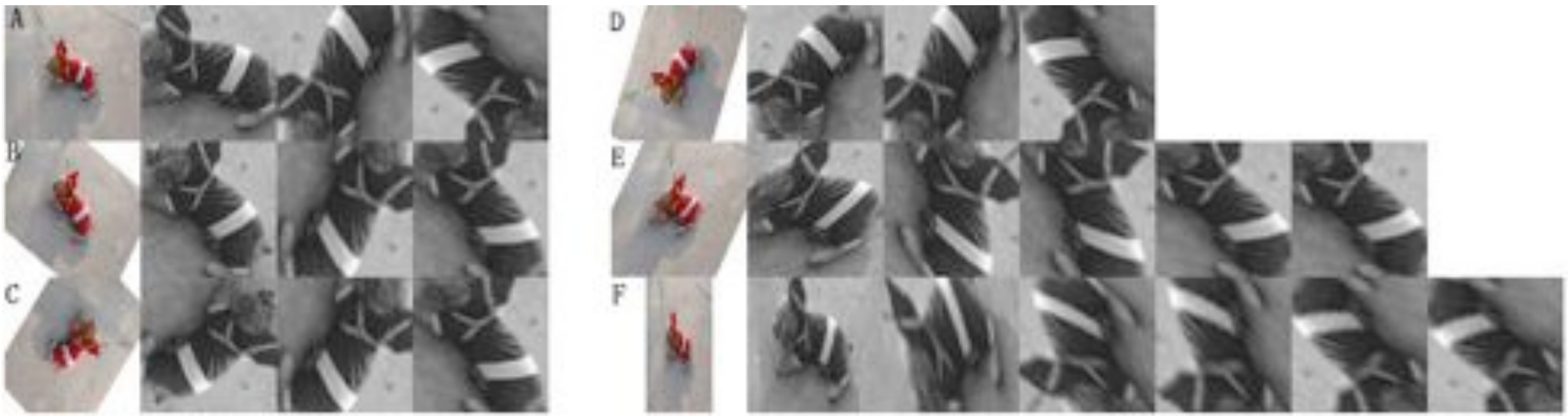
With Kai Ni





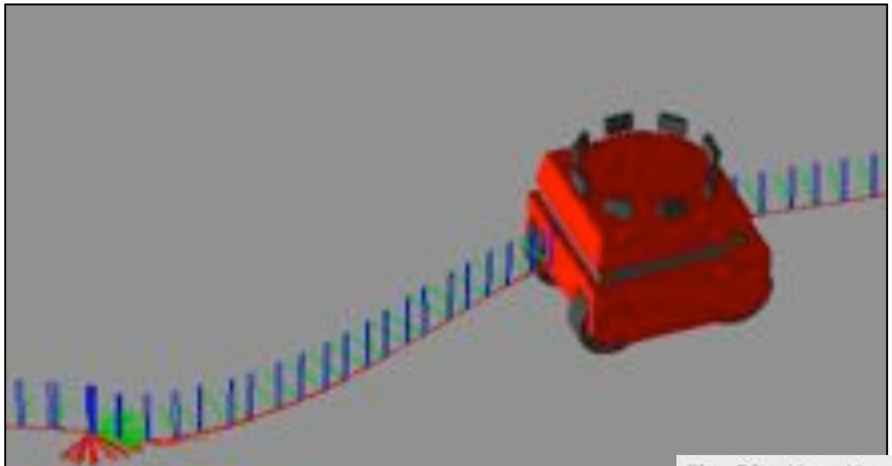
- **D**ifference **O**f **G**aussian Filters at Different Scales, Lowe, IJCV 2004

Affine-invariant Feature Descriptors

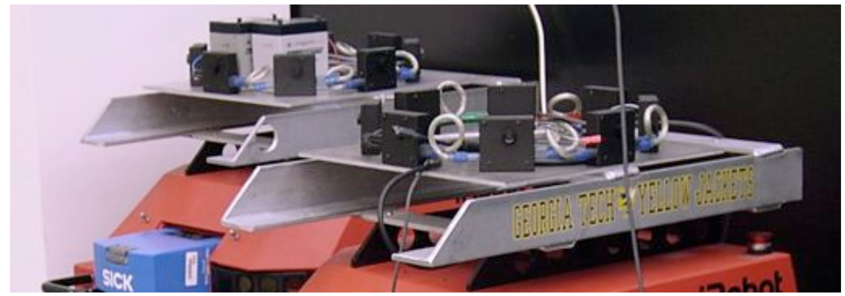


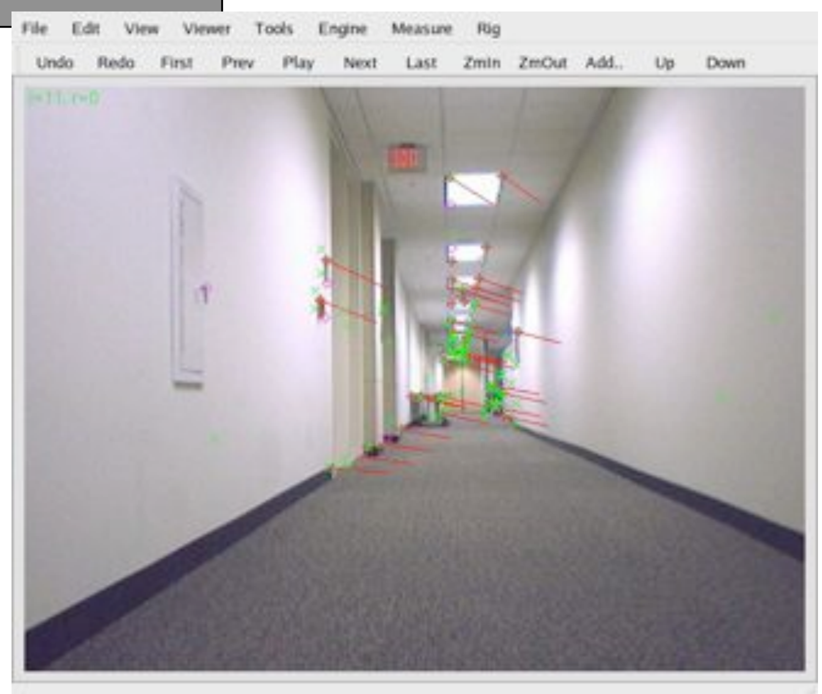
- Invariant to affine deformations
- Mikolajczyk & Schmid, ECCV 2002
- Appearance is then compressed using PCA (40 dimensions, 160 bytes per landmark)

Database: Structure from Motion

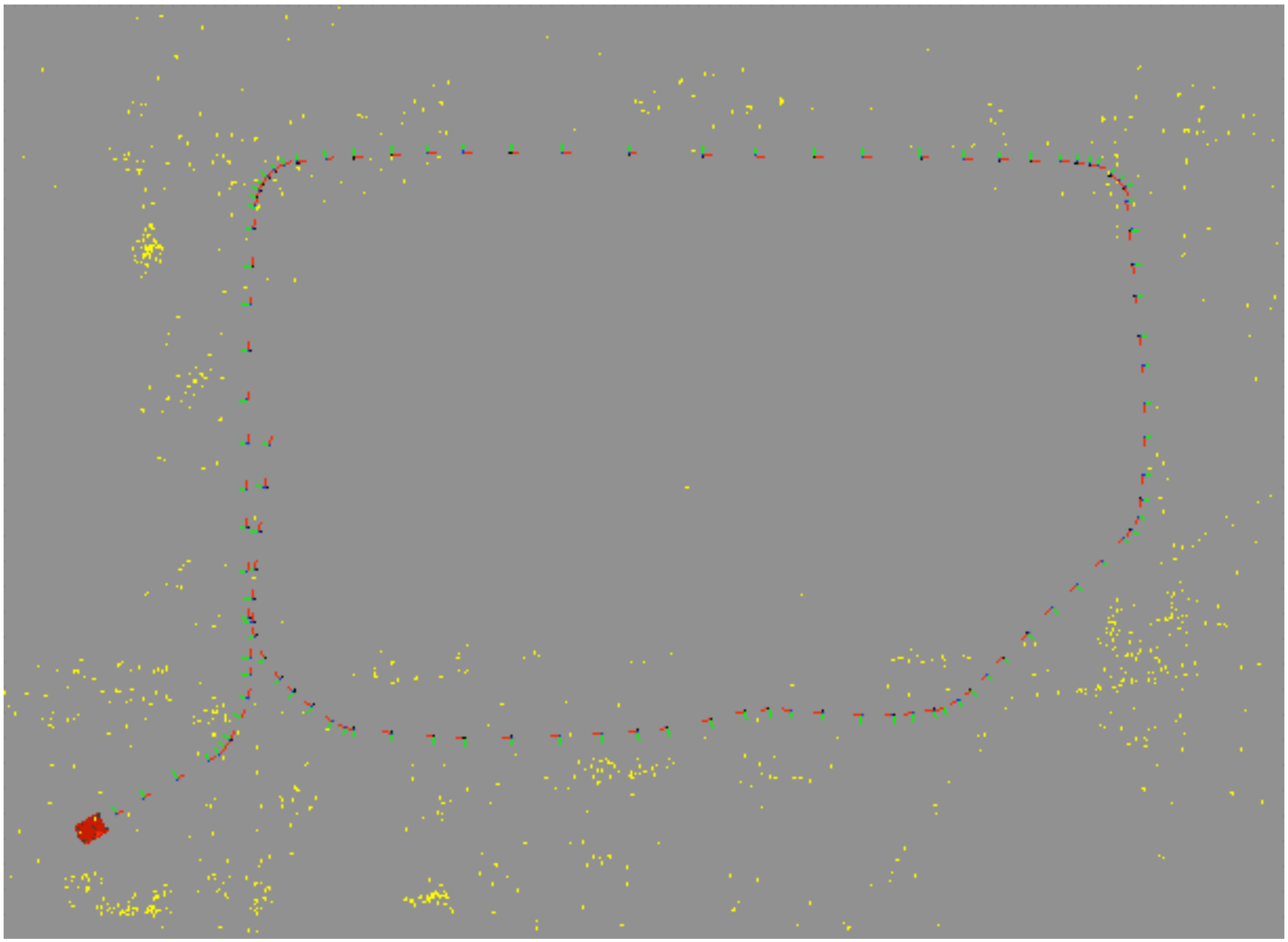


with Michael Kaess



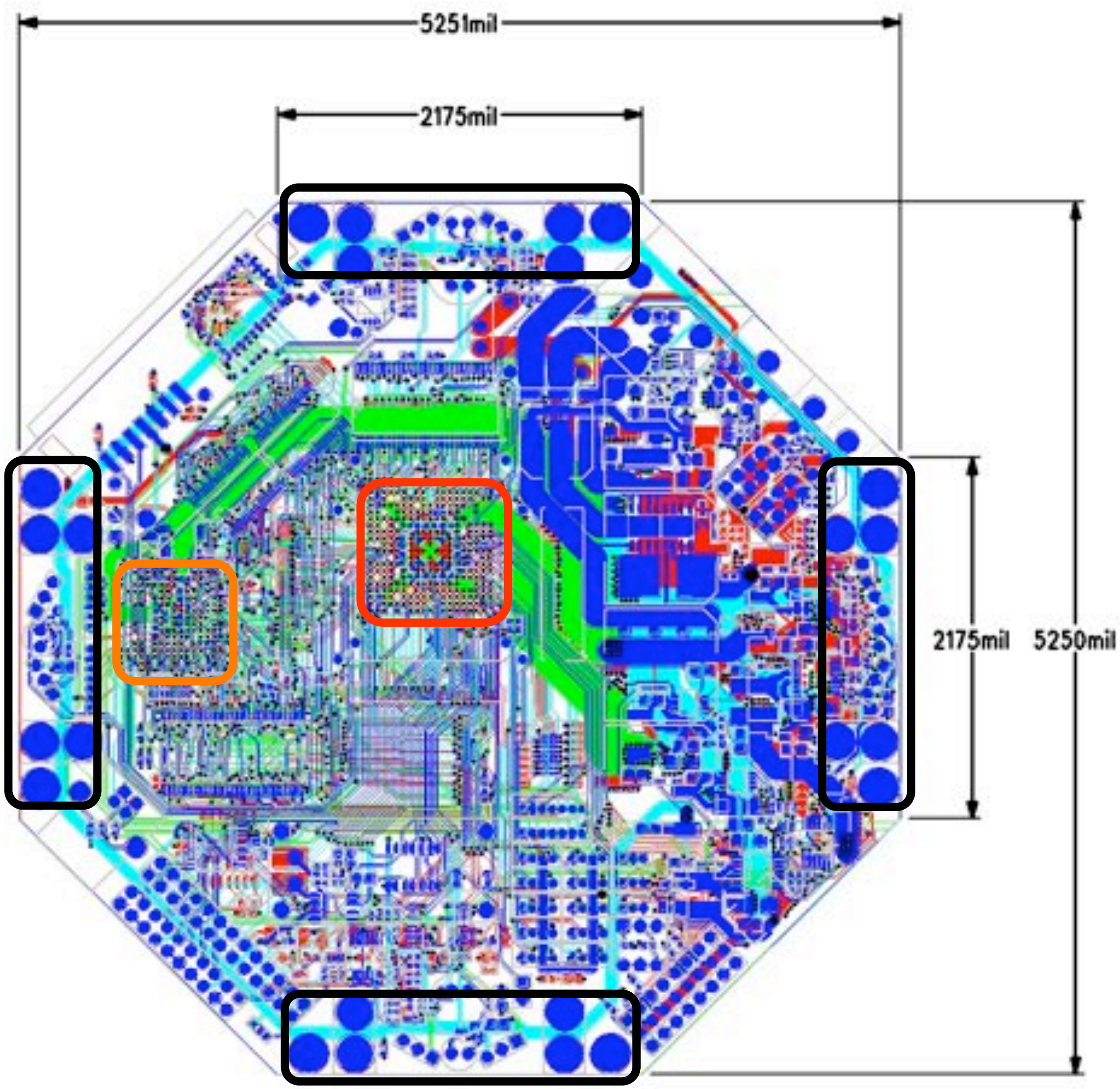


Animation



Run-time: Custom Hardware

with Daniel Walker and Tucker Balch



4 Cameras covering the viewing circle

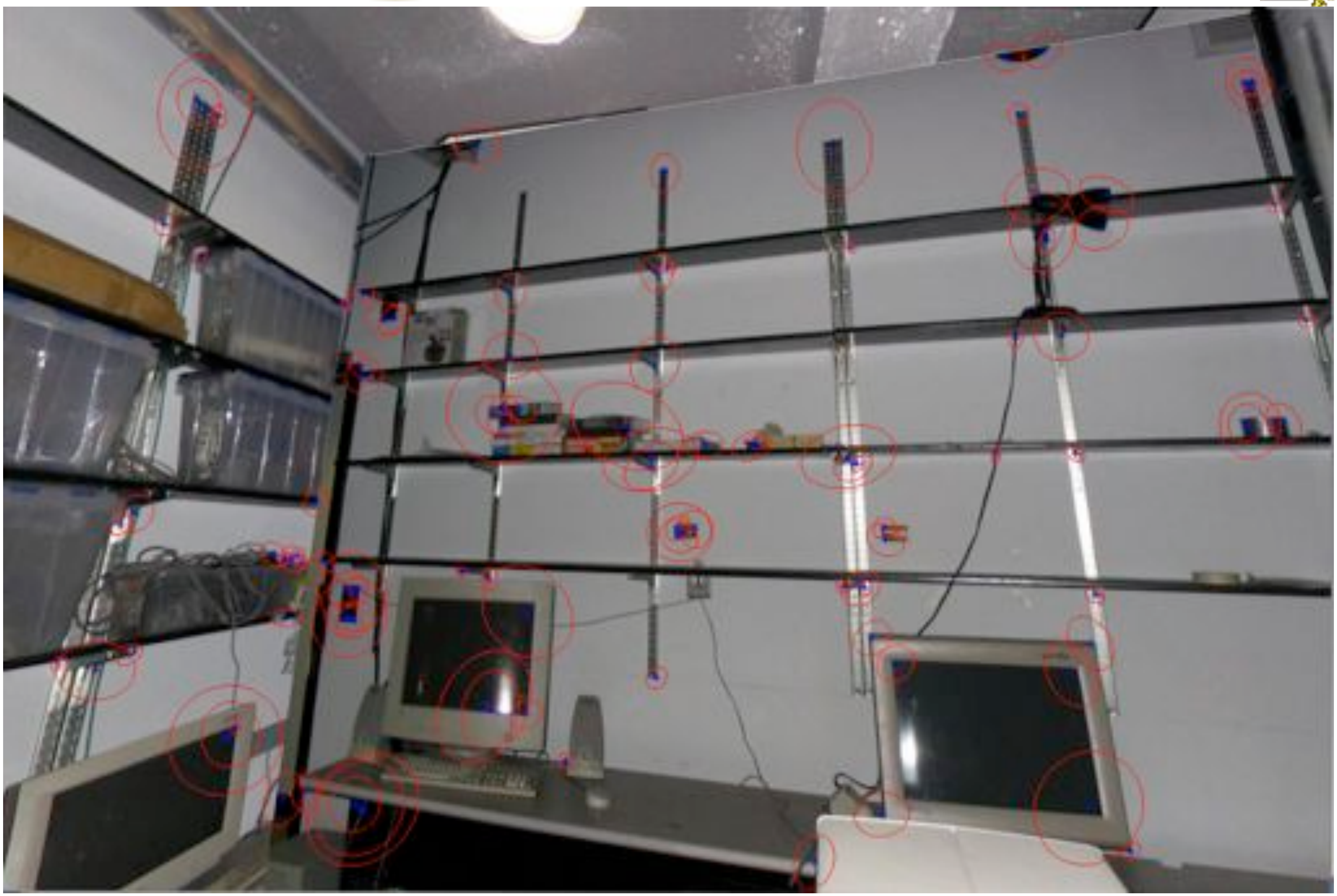
FPGA for Real-time Feature Detection

Xscale Processor for Real-Time Localization

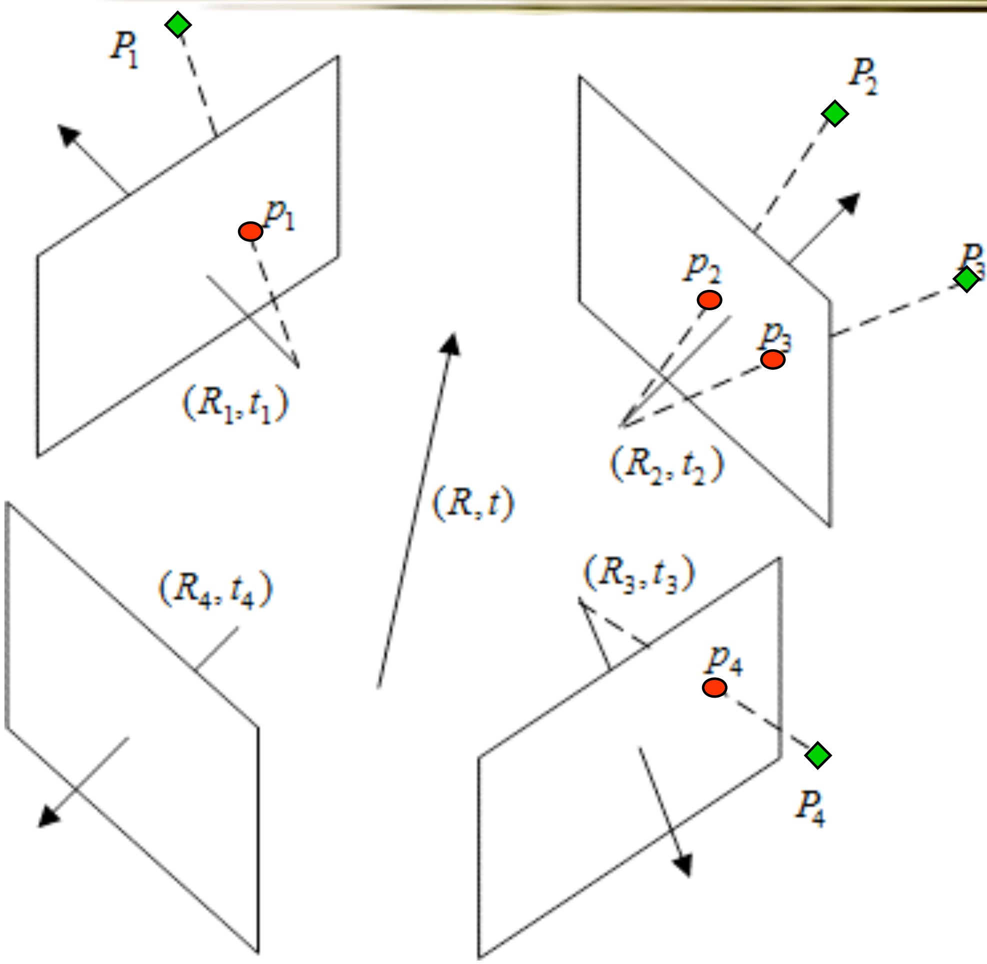
Recent Prototype :-)

- 4 Firewire cameras on a single bus
- Connected to a laptop

Typical Output of Feature Detector Stage



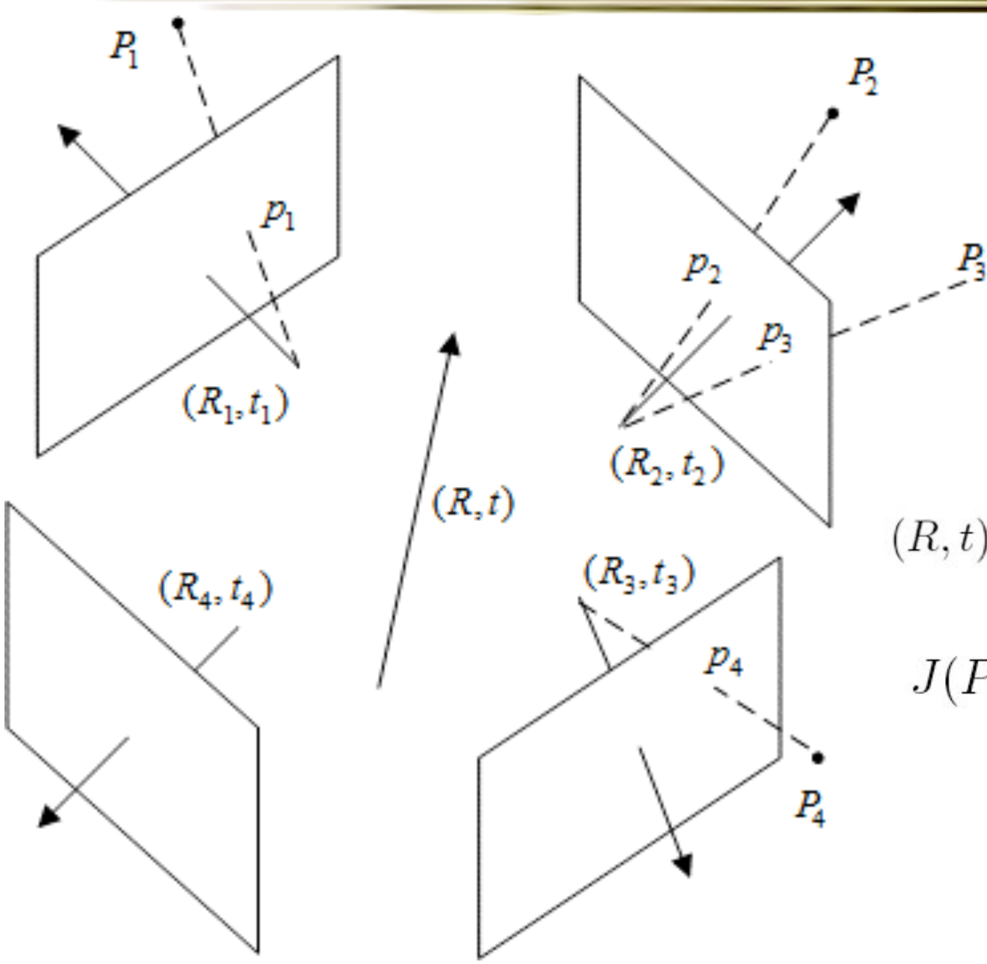
Multi-camera Localization



- Maximum a-posteriori estimation:
- Pose R,t given >3 feature measurements

$$(R,t)^* = \operatorname*{argmax}_{R,t} \left\{ P(R,t) \prod_{j=1}^{n} P\left(\boxed{P_j}, i_j, \boxed{p_j} \middle| R,t\right) \right\}$$

Multi-camera Localization (cont'd)



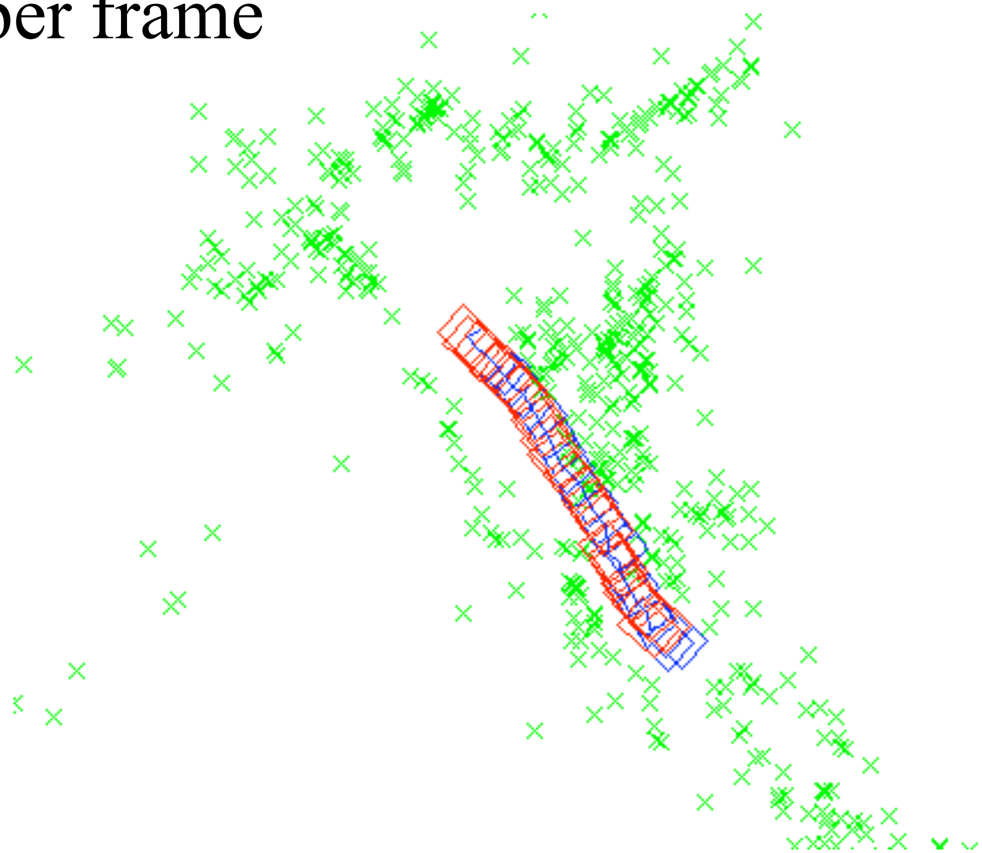
- MAP = Least-squares
- RANSAC w 3 points

$$(R,t)^* = \underset{R,t}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{j=1}^{n} J(P_j, i_j, p_j) - \log P(R,t) \right\}$$

$$J(P,i,p) \triangleq \|p - \Pi_i(K_i, R_i(R(P-t) - t_i)\|^2_{\Sigma_i}$$

Preliminary "Real-time" Results

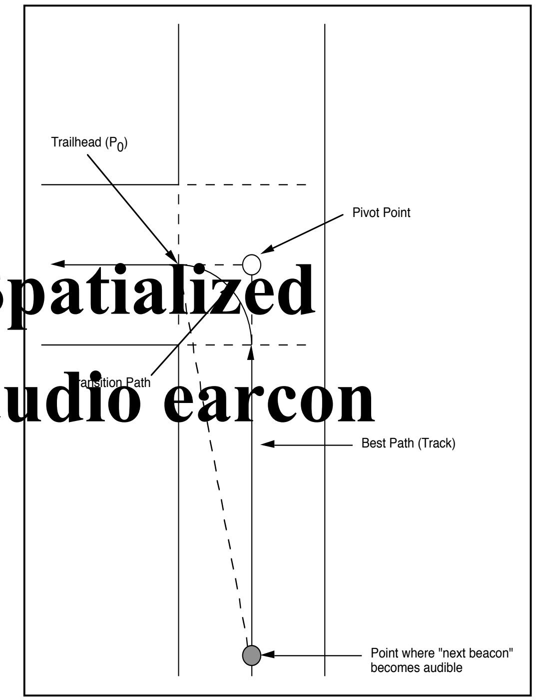
- 1.6GHz Pentium II Laptop
- Approx. 5 secs per frame



SWAN Auditory Display

- Navigation Beacons
  - Spatialized audio <u>beacons</u> form a path which can be followed
- Objects & obstacles
  - e.g., a desk in the hall; phone booth
- Surface Transitions
  - e.g., sidewalk to grass; start of stairway
- Location
  - e.g., lecture hall; intersection; office

- Annotations
  - e.g., "Puddle here whenever it rains"
  - e.g., "Ramp on left side of entrance"



Spatialized audio earcon

Recorded speech or TTS

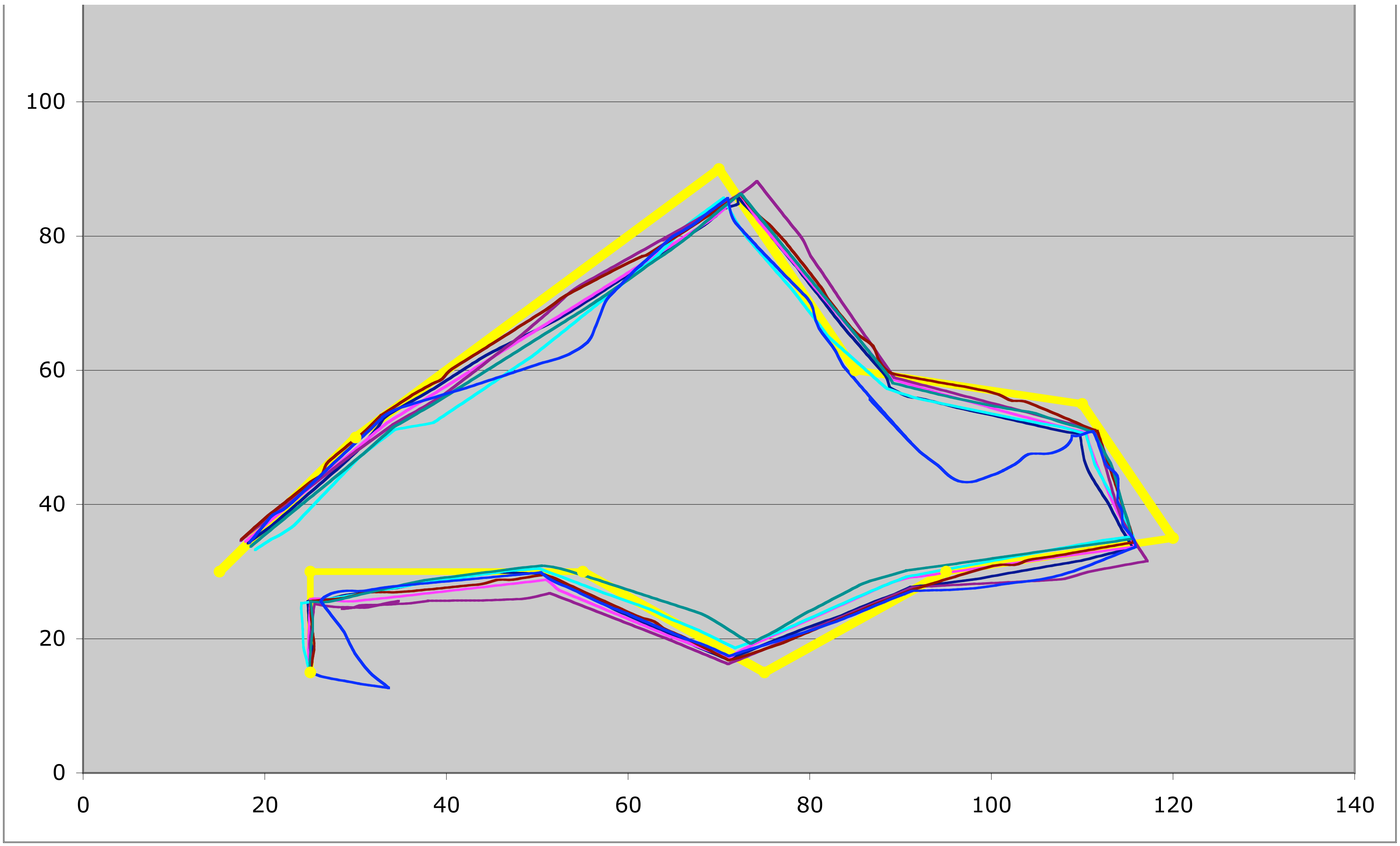
System Evaluation

- Does SWAN help the user safely accomplish specific tasks?
  - Navigation effectiveness
  - Situational awareness
  - Movement speed, efficiency
  - Exploration of novel environment
  - Comfort, satisfaction
  - Safety

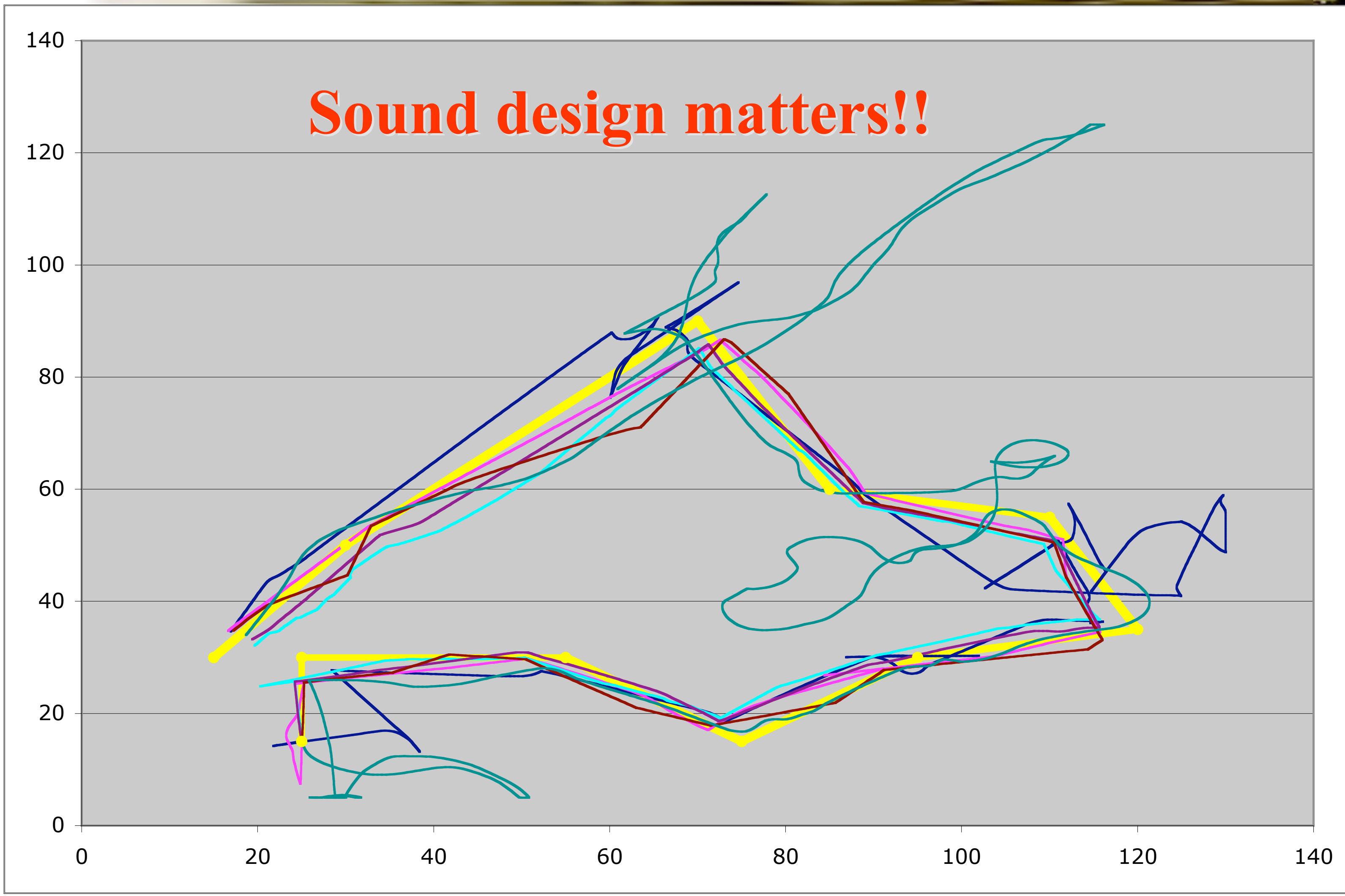
Auditory Display: Some Factors to Evaluate

- Beacon Sound
- Capture Radius
- Sound device
  - headphones vs. bonephones

Sound Design: “Good” Beacons

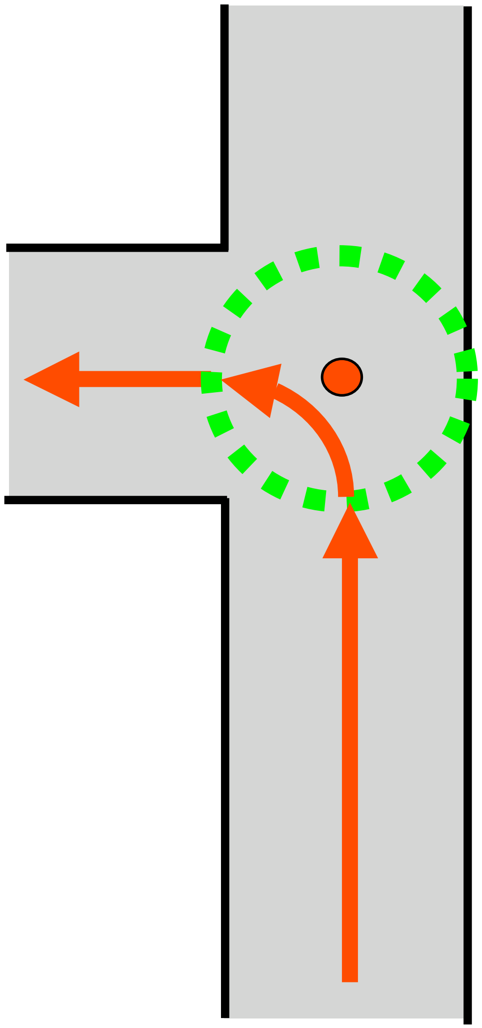


Sound Design: Bad Beacons (!)

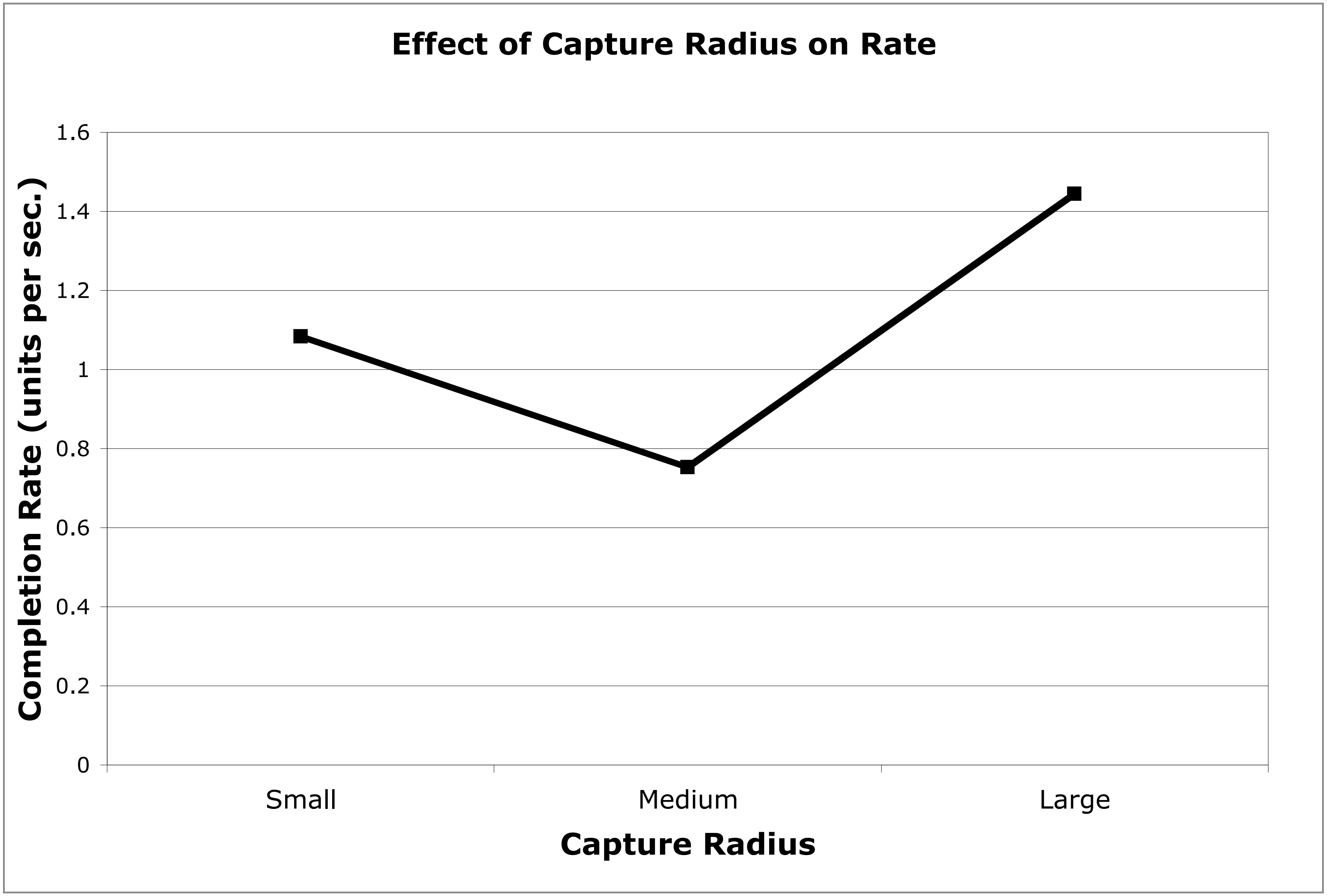


Capture Radius: Real-World Interaction

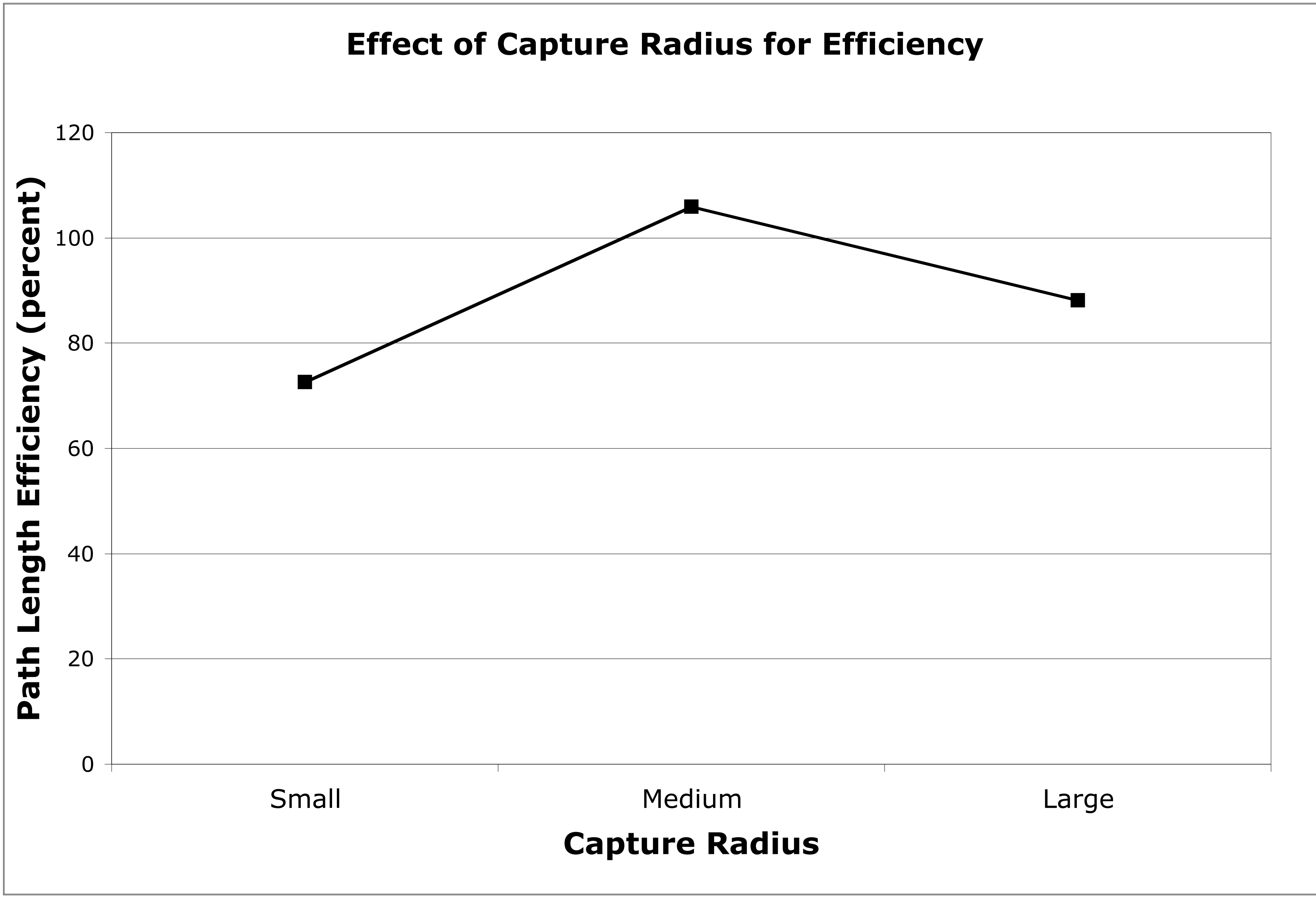
- You never *exactly* reach waypoint
- Capture radius = distance from the waypoint that is "close enough"
  - The "next beacon" sound begins
  - Intended to allow for more natural walking around corners and turns
- Participants in the study "bounced" off edge of capture radius
  - Artifact of movement with flight stick in VR (not real walking)



Effect of CR on Rate of Travel



Effect of CR on Path Efficiency (accuracy)



Capture Radius: Findings

- Speed-accuracy tradeoff based on the size of the capture radius
  - Medium capture radius had the slowest rate, but also had the greatest efficiency (accuracy)
- Capture radius must be considered in design of navigation interfaces
- Depends on goals
  - Stay on path or move quickly?

Sound Hardware

- Headphones
  - Benefits
  - Problems

- Bone conduction headphones (bone phones)
  - Benefits
  - Issues

Bone Phones

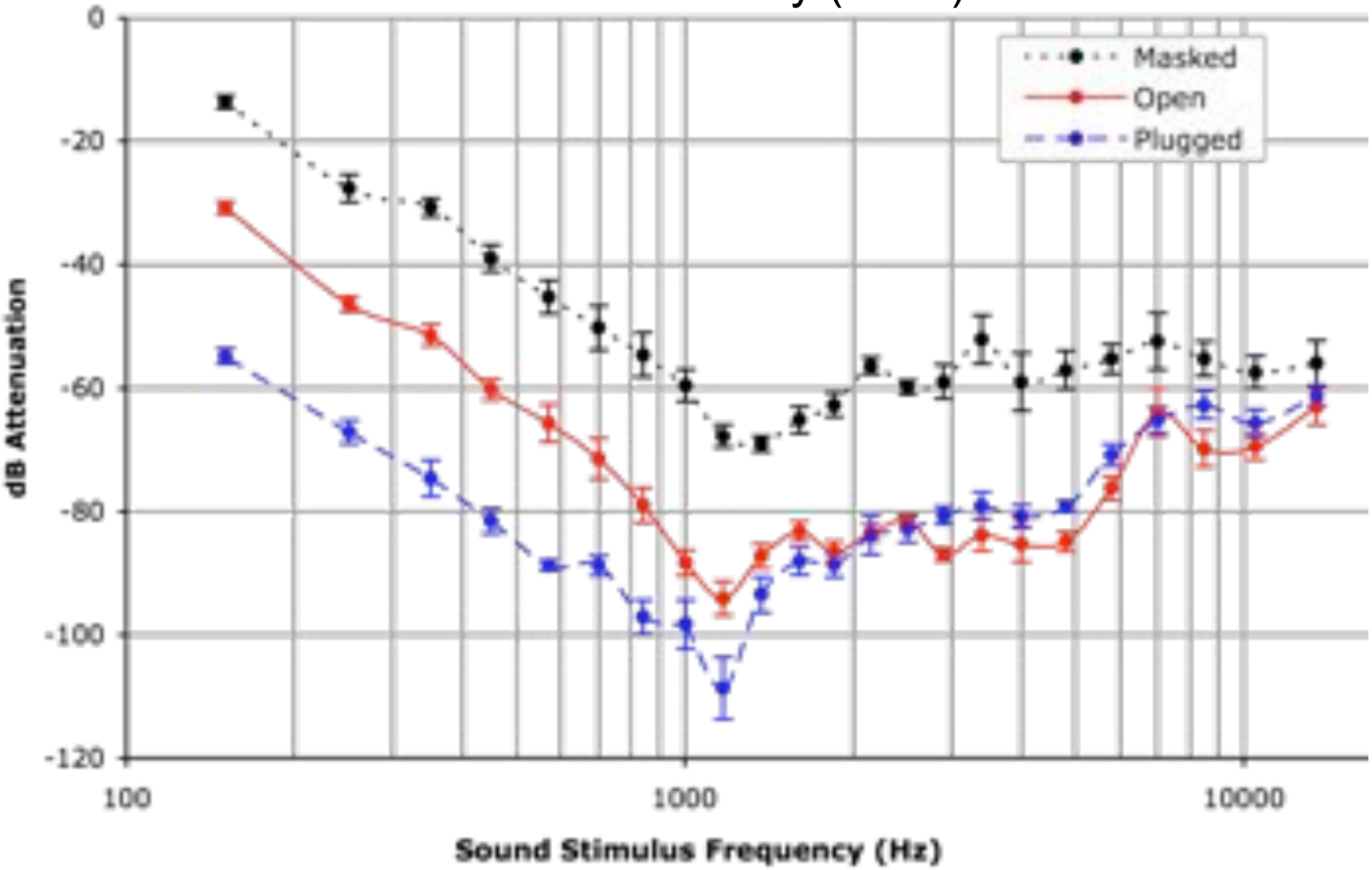
- Bone conduction
- Discrete
- Ears open (or plugged)
- Stereo separation (?)

Bonephone Research

- Psychophysics
  - Hearing thresholds
  - Frequency response
- Practical Applications
  - Lateralized sound: speech separation
    - “Ready Charlie” task (with Brungart, Simpson, et al.)
  - Spatialized sound: SWAN
    - Need “BRTFs” (bone related transfer function)

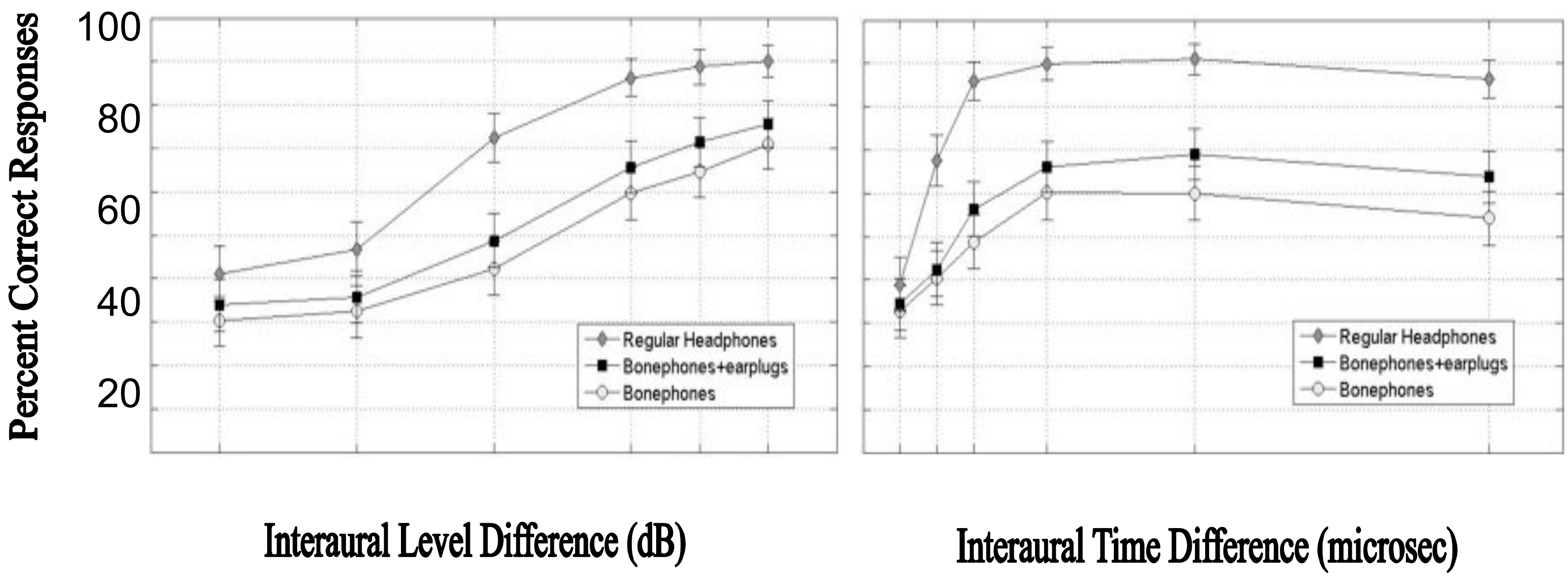
Bonephones Threshold

Walker & Stanley (2005)



Bonephones CRM Data: ILDs & ITDs

Walker, Stanley, Iyer, Simpson, & Brungart (2005)

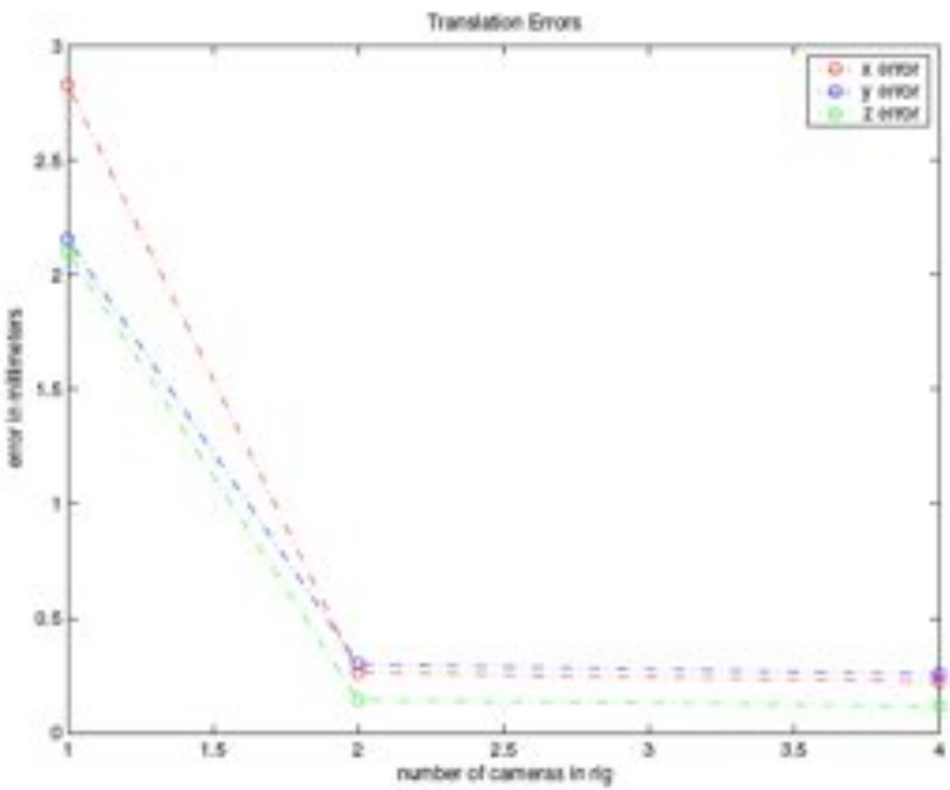


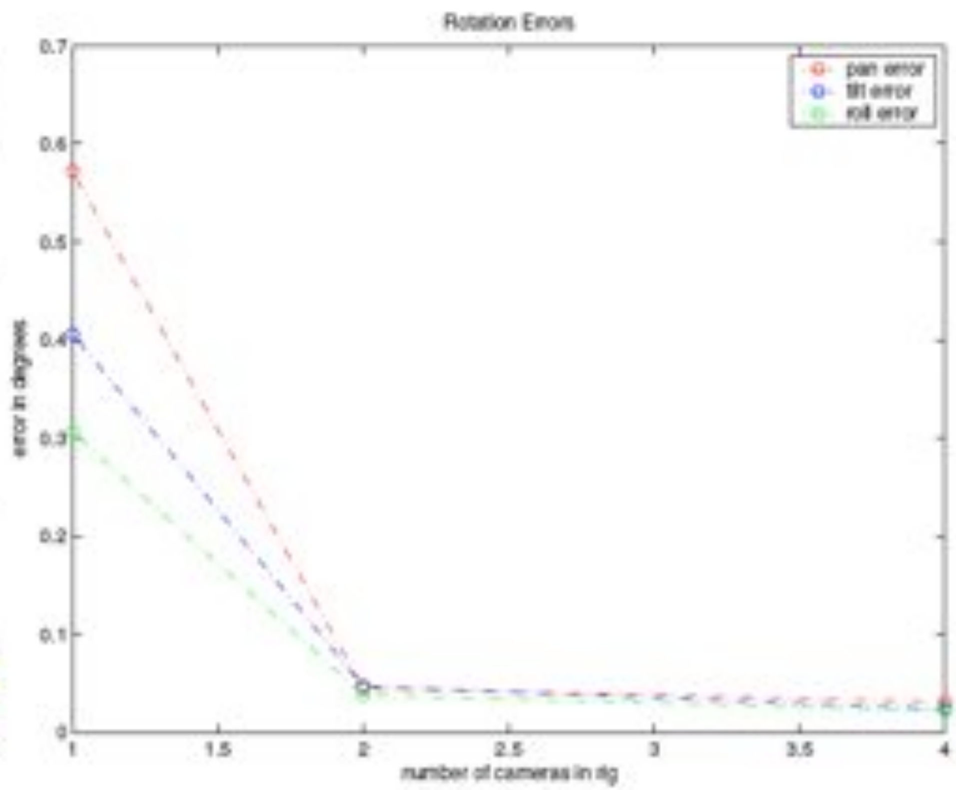
Future Directions

- Lots of "live" tests (in addition to in the VR)
  - sighted and visually impaired participants
  - indoors, outdoors, and mixed (the *real* test!)
- Integrating more pedestrian-level GIS data
  - including accessibility information
- Expanding to more "discover and explore" tasks, in addition to simple wayfinding
- Using the cameras for more tasks
  - face recognition, object identification, text/OCR
- Expand to different user populations
  - fire fighters, police, military

The End

Synthetic Results to Support Design
Increasing Nr. Of Cameras

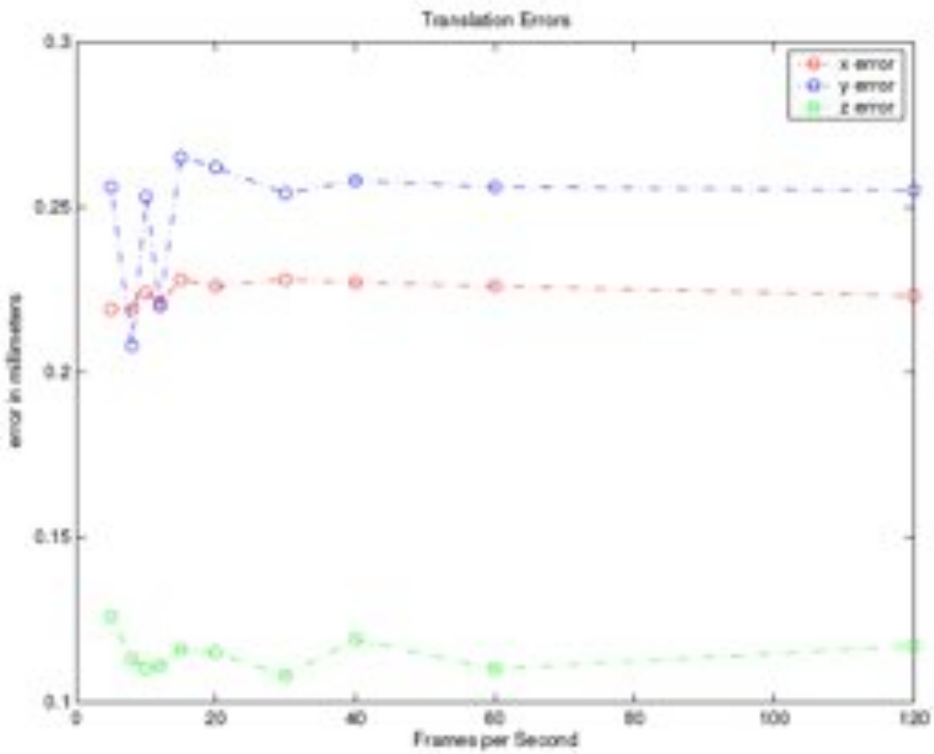


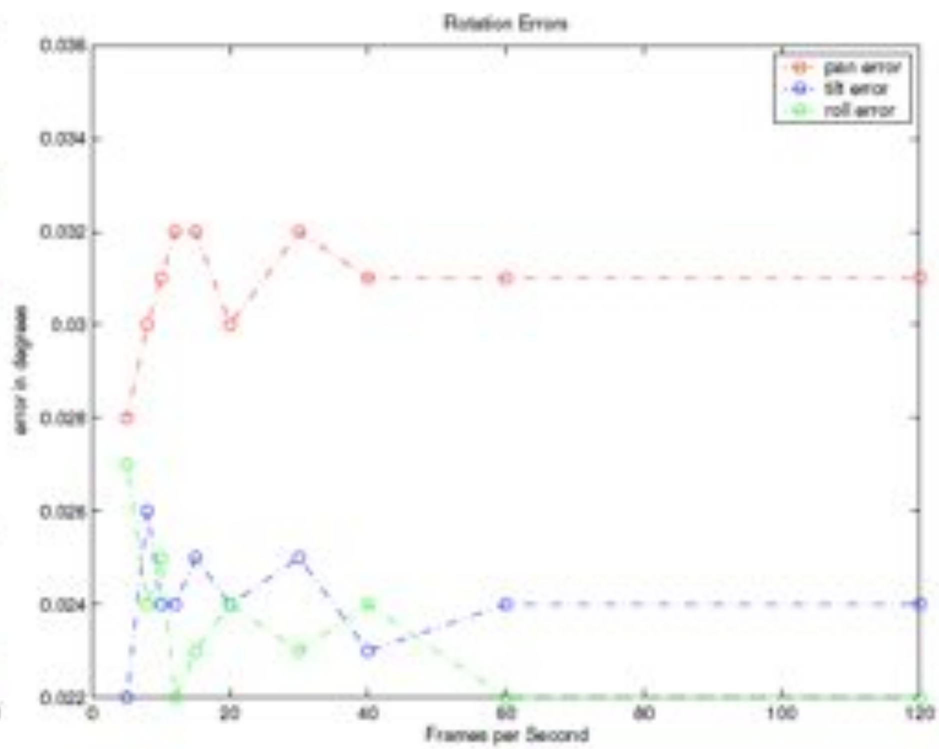


- Textured Cube
- Motion-capture data for realism at 120 fps
- 1172 frames of a subject looking around

Synthetic Results to Support Design

Increasing Frame Rate





- 3386 frames, subject walking and looking in various directions
- Frame rates were 5,8,10,12, 15, 20, 30, 40, 60, and 120 fps