perception, action and the information knot that ties them

stefano soatto
ucla

http://vision.ucla.edu

icra workshop on mobile manipulation

- "ask *not* what perception can do for manipulation - ask what manipulation can do for perception"

- (vision in particular)

i manipulate, therefore i am

- "signal-to-symbol barrier" problem
- what is an "object"? what "information" does an image contain about the object?
- "information theory" in the context of decision/control tasks

gibson's information

- task ⇨ data = "information" & (structured) "nuisance"
- information = complexity of the data after the effects of nuisances has been discounted
  - nuisances in vision:
    - viewpoint
    - illumination
    - visibility (occlusion, cast shadows)
    - quantization/noise

gibson: *"my notion is that information consists of invariants underlying change [...] of illumination, point of observation, overlapping samples [...] and disturbance of structure"*

is a "gibsonian information theory" viable? (take I)

- general-case viewpoint invariants do not exist [burns et al., '92]
- non-trivial illumination invariants do not exist [chen et al., '00]

is a "gibsonian information theory" viable? (take II)

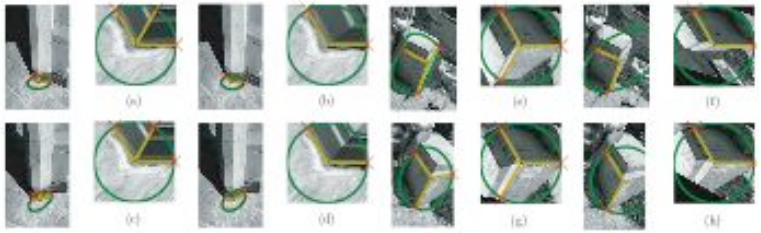
- ☑ general-case viewpoint invariants do exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto '05-'06]
- ☑ non-trivial contrast invariants do exist, and are sufficient statistics [morel & c., '93-'05]
- ☐ what is invariant to contrast (geometry of the level lines) is not invariant to viewpoint
- ☐ what is invariant to viewpoint (image range in a canonized domain) is not invariant to contrast

is a "gibsonian information theory" viable? (take III)

- ☑ general-case viewpoint invariants exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto '05-'06]
- ☑ non-trivial contrast invariants exist, and are sufficient statistics [morel & c., '93-'05]
- ☑ viewpoint-illumination invariants exist (ambient-lambert)
- ☑ they are "discrete" structures (attributed reeb tree, ART), supported on a thin set
- ☑ they are sufficient statistics! (equivalent to the image up to changes of viewpoint and contrast) [sundaramoorthi et al., '09]

"the set of images modulo viewpoint and contrast changes"

[sundaramoorthi-petersen-varadarajan-soatto '09]



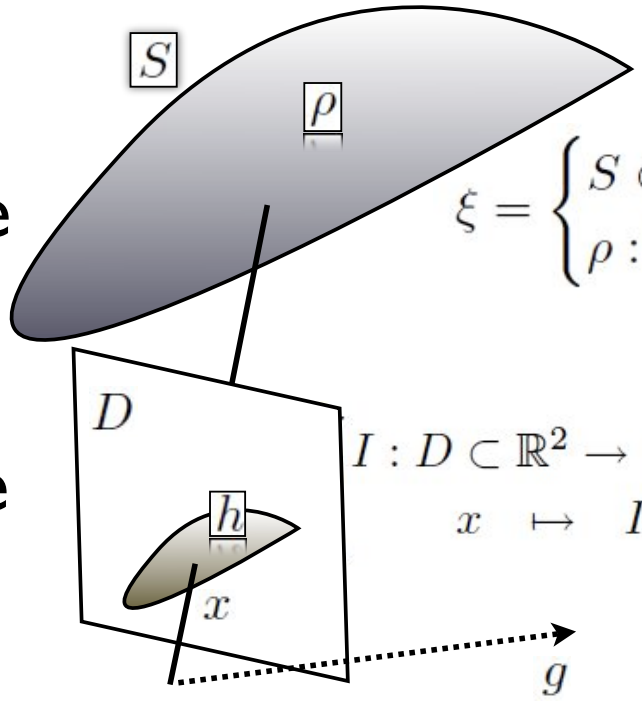
- viewpoint changes induce (epipolar-homeomorphic) deformations of the image domain; diffeomorphic closure (general non-planar surfaces)
- viewpoint-contrast invariants exists
- they are (supported on) a zero-measure subset of the image domain (attributed reeb tree)
- they are sufficient statistics! (equivalent to the image up to contrast and viewpoint transformations)

is a "gibsonian information theory" viable? (take III)

- ☑ general-case viewpoint invariants exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto '05-'06]
- ☑ non-trivial contrast invariants exist, and are sufficient statistics [morel & c., '93-'05]
- ☑ viewpoint-illumination invariants exist (ambient-lambert)
- ☑ they are "discrete" structures (attributed reeb tree, ART), supported on a thin set
- ☑ they are sufficient statistics! (equivalent to the image up to changes of viewpoint and contrast) [sundaramoorthi et al., '09]
- ☐ occlusions and quantization admit no invariants!

some notation

scene



$$\xi = \begin{cases} S \subset \mathbb{R}^3 & \text{shape (geometry)} \\ \rho : S \to \mathbb{R}^k & \text{radiance (reflectance)} \end{cases}$$

image

$$I : D \subset \mathbb{R}^2 \to \mathbb{R}^k_+ \qquad x \mapsto I(x)$$

$\mathbb{L}^2$ integrable $\overset{\text{dense}}{\to}$ morse

$k = 1, 3, N$ spectral bands

nuisance

$$\nu = \begin{cases} h : \mathbb{R}^k \to \mathbb{R}^k & \text{contrast } (\simeq \text{illumination}) \\ g : \mathcal{Z}_+ \to SE(3) & \text{motion (viewpoint)} \end{cases}$$

lambert-ambient

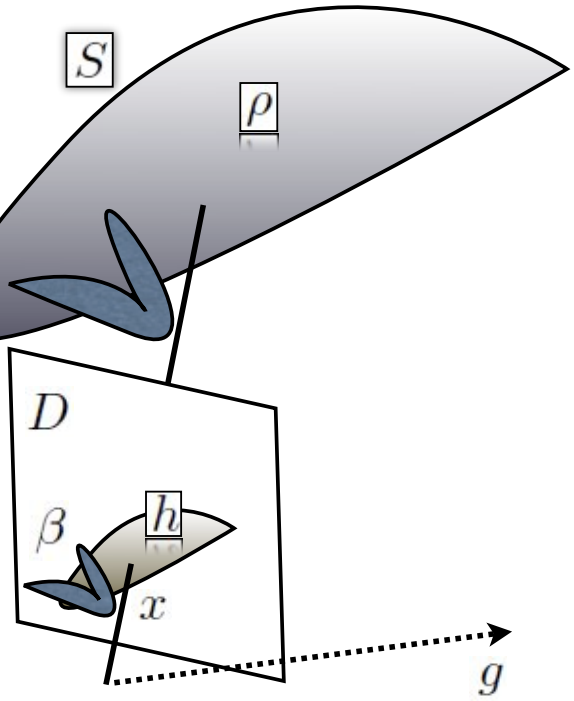
$$\begin{cases} I(x,t) = h(t) \circ \rho(p) + n(x,t) \\ x = \pi(g(t)p) \end{cases}$$

$$I(x,t) = f(\underbrace{\rho, S}_{\xi}; \underbrace{g, h, n}_{\nu})$$

some notation

scene

image



occlusions

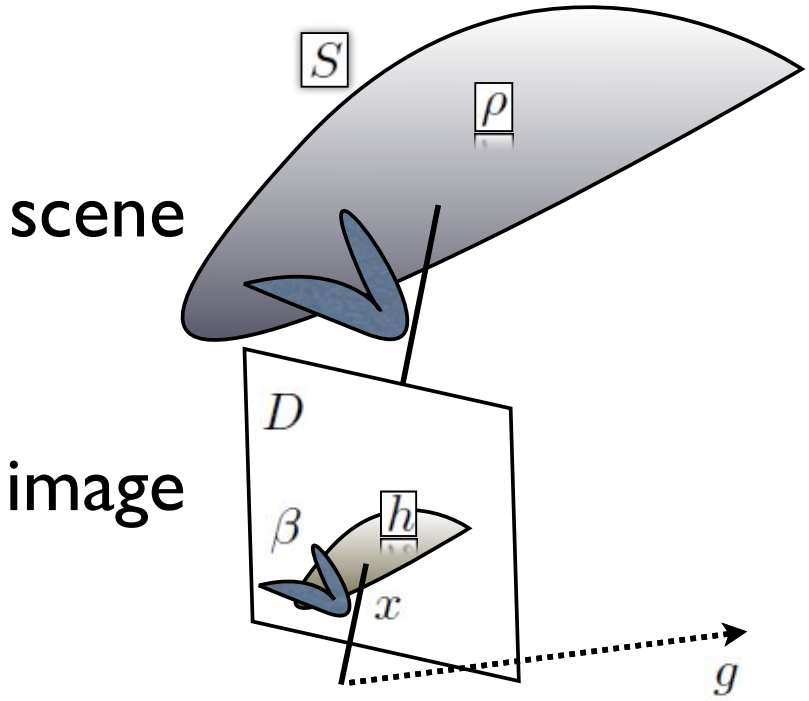
$$I(x) = \begin{cases} f(\rho, S; g, h, n) & x \in D \backslash \Omega \\ \beta(x) & x \in \Omega \end{cases}$$

lambert-ambient

$$\begin{cases} I(x,t) = h(t) \circ \rho(p) + n(x,t) \\ x = \pi(g(t)p) \end{cases}$$

$$I(x,t) = f(\underbrace{\rho, S}_{\xi}; \underbrace{g, h, n}_{\nu})$$

some notation



nuisance $\nu = g \quad h \quad \beta \quad n$

image formation model
(formal notation)

$$\boxed{I = f(\xi, \nu)}$$

$$I = f(g\xi, \nu) + n$$

some definitions

feature $\phi : \{I(x), x \in D\} \rightarrow \mathbb{R}^K$

$I \mapsto \phi(I)$

invariant $\phi \circ f(\xi, \nu) = \phi \circ f(\xi, \bar{\nu}) \quad \forall\, \nu, \bar{\nu};\ \forall\, \xi$

maximal invariant $\phi^{\wedge}(I)$

sufficient statistic $\phi \mid R(u|I) = R(u|\phi(I))$

conditional risk $R(u|I) \doteq \int L(u, \bar{u}) dP(\bar{u}|I)$

loss function L decision/control policy u

minimal sufficient statistic $\phi^{\vee}_{\xi}(I)$

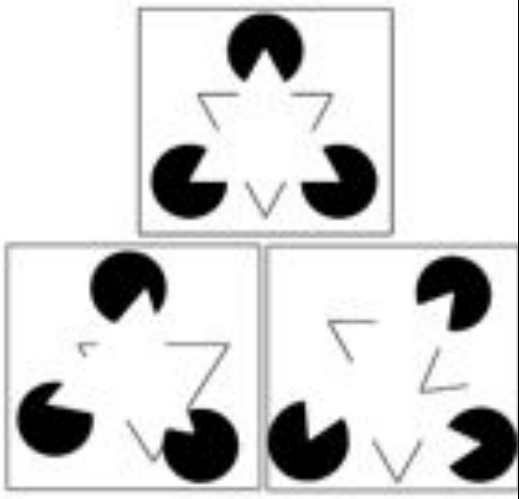
representation

given one or more images $\{I\}$ a representation

$\hat{\xi}$ is a statistic $\hat{\xi} = \phi(\{I\})$ such that

$$\boxed{\{I\} \in \{f(g\hat{\xi}, \nu), \quad g \in G, \nu \in \mathcal{V}\} \doteq \mathcal{L}(\hat{\xi})}$$

i.e., it is a statistic from which
the images can be hallucinated



$$\mathcal{L}(\hat{\xi}) = \mathcal{L}(\xi)$$

complete representation
minimal complete representation
(note it is invariant to G)

information gap

- actionable information: coding length of a maximal invariant statistic; can be computed from an image.

$$\mathcal{H}(I) \doteq H(\phi^{\wedge}(I))$$

- complete information: coding length of a minimal sufficient statistic of a (complete) representation

$$\mathcal{I} = H(\phi^{\vee}(\hat{\xi}))$$

- actionable information gap (AIG)

$$\mathcal{G}(I) \doteq \mathcal{I} - \mathcal{H}(I)$$

invertible nuisances

invertible nuisance $f(\xi, \emptyset) \mapsto f(\xi, \nu)$ injective $\boxed{\mathcal{G} = 0}$

contrast $\nu = h$ $\boxed{\phi^{\wedge}(I) = \dfrac{\nabla I(x)}{\|\nabla I(x)\|} \quad (\equiv \text{geom. level curves})}$

viewpoint $\nu = \begin{cases} w : D \subset \mathbb{R}^2 \to \mathbb{R}^2 \\ \qquad x \mapsto w(x) = \pi \circ g^{-1} \circ \pi^{-1}(x) \end{cases}$

$\boxed{\phi^{\wedge}(I) = ART}$ away from occlusions

(non)invertible nuisances

- visibility (occlusions, cast shadows); quantization
- invertibility depends on the sensing process: control authority
- j. j. gibson: "*the occluded becomes unoccluded*" in the process of "information pickup"

is a "gibsonian information theory" viable? (take IV)

- ☑ general-case viewpoint invariants exist, and are non-trivial, for lambertian scenes in ambient light [vedaldi-soatto '05-'06]
- ☑ non-trivial contrast invariants exist, and are sufficient statistics [morel & c., '93-'05]
- ☑ viewpoint-illumination invariants exist (ambient-lambert)
- ☑ they are "discrete" structures (attributed reeb tree, ART), supported on a thin set
- ☑ they are sufficient statistics! (equivalent to the image up to changes of viewpoint and contrast) [sundaramoorthi et al., '09]
- ☑ occlusions and quantization are invertible! [gibson '50s]

how to build representations?

1. canonizability (sparse yet lossless)
2. commutativity (beyond existing local descriptors)
3. structural stability (BIBO vs. structural stability)
4. proper sampling (beyond nyquist)
5. exploration (gibson)

how to build representations? feature optimality by design

co-variant detector: a functional $\psi : \mathcal{I} \times G \to \mathbb{R}^{dim(G)}; (I, g) \mapsto \psi(I, g)$

I. the zero-level set $\psi(I, g) = 0$ uniquely determines $\hat{g} = \hat{g}(I)$

II. if $\psi(I, \hat{g}) = 0$ then $\psi(I \circ g, \hat{g} \circ g) = 0 \quad \forall\ g \in G$

canonizable: an image region is canonizable if it admits at least one co-variant detector

canonized descriptor: $\phi(I) \doteq I \circ \hat{g}^{-1}(I) \ \mid \ \psi(I, \hat{g}(I)) = 0$

what is the "best" descriptor? when is it optimal? 1. canonizability

- Thm 1: canonized descriptors are complete invariant statistics (wrt canonized group)
- Thm 2: if a complete invariant descriptor can be constructed, an equi-variant classifier can be designed that attains the Bayes' risk
- the best descriptor can be derived analytically (BTD)
- What about non-group nuisances?

2. commutativity

- commutative nuisance: $I \circ g \circ \nu = I \circ \nu \circ g$
- Thm 3: the only nuisances that are invertible and commutative are the isometric group of the plane and contrast range transformations
- Corollary: do not canonize scale (nor affine/ projective transformations)

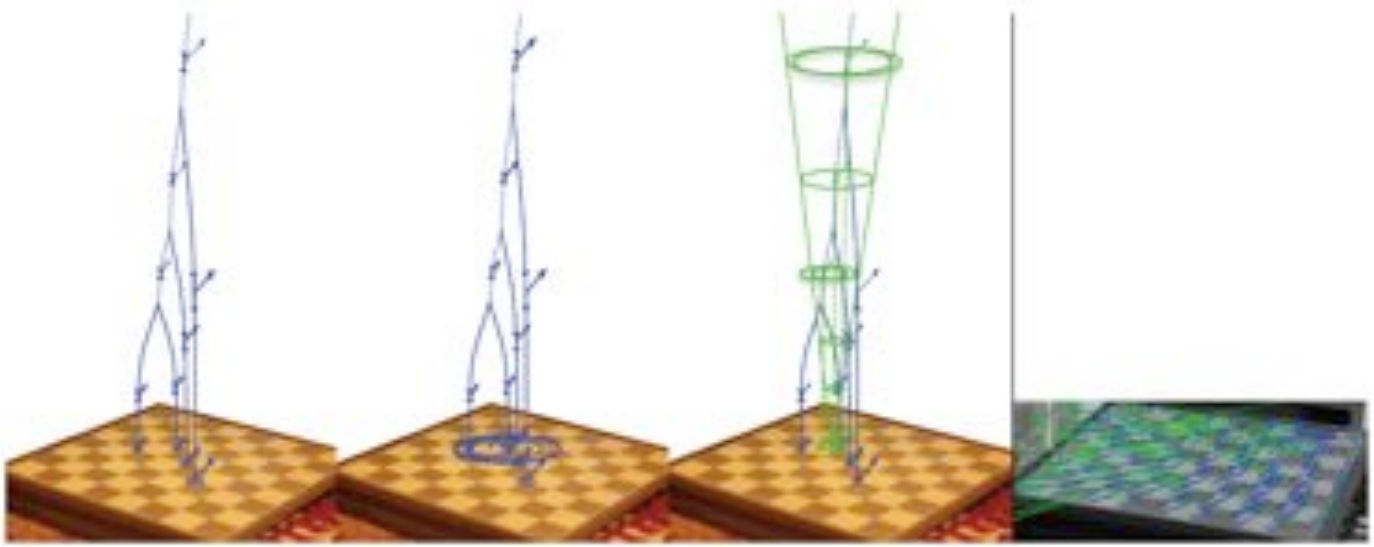
- (Thm 5: an image region is a texture if and only if it is not canonizable)

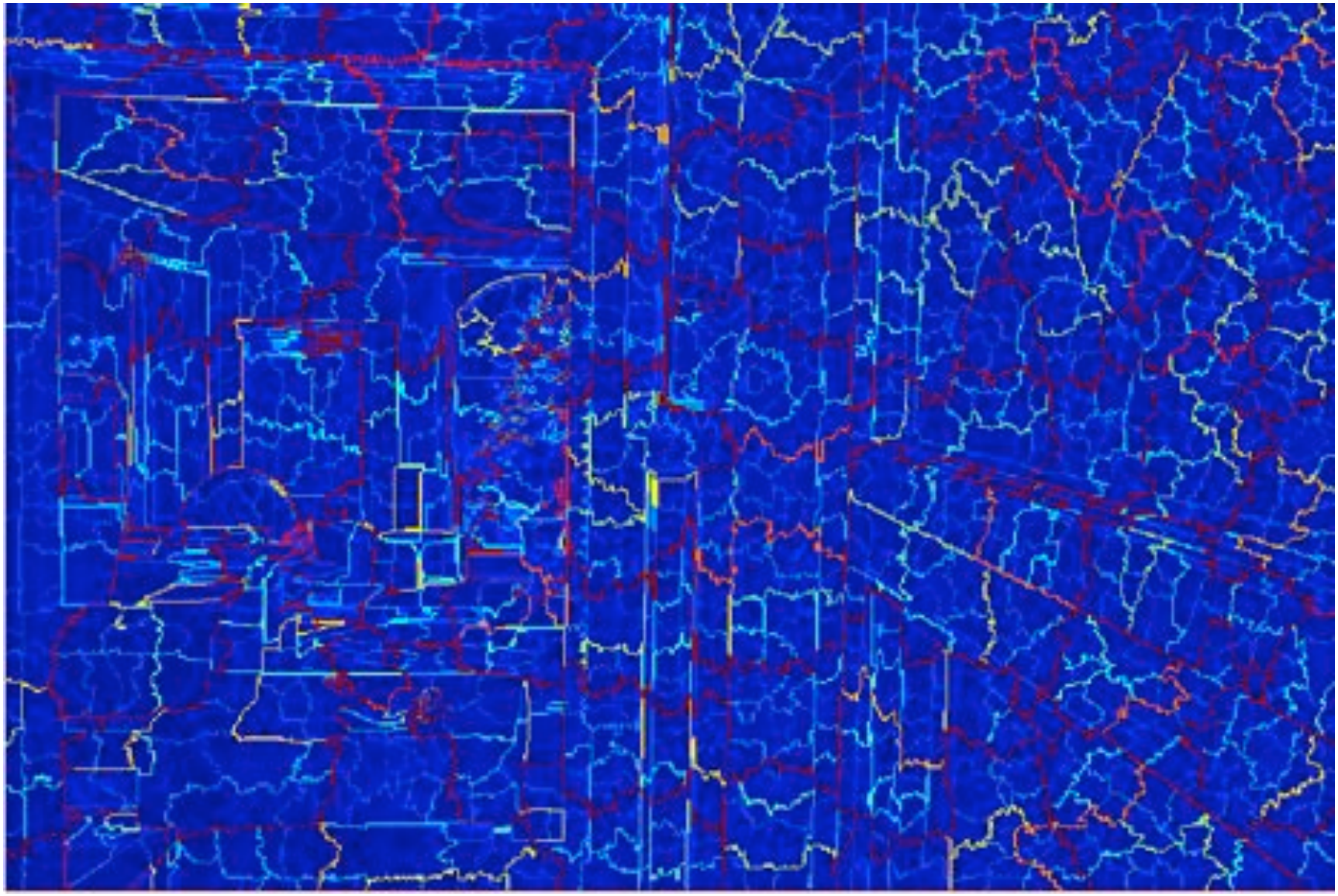
3. BIBO stability (sensitivity)

- BIBO sensitivity: a detector is BIBO insensitive (stable) if small nuisance variations cause small changes in the canonical element.
- Thm 6: any co-variant detector is BIBO stable
- BIBO stability is irrelevant for visual decisions!

3. structural stability

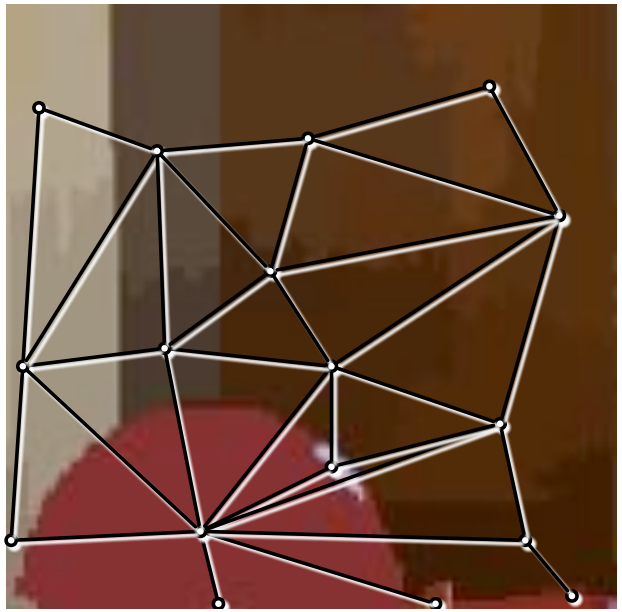
- structural stability: small changes in the nuisance do not cause catastrophic (singular) perturbations in the detector
- design detectors by maximizing structural stability margins: the selection tree

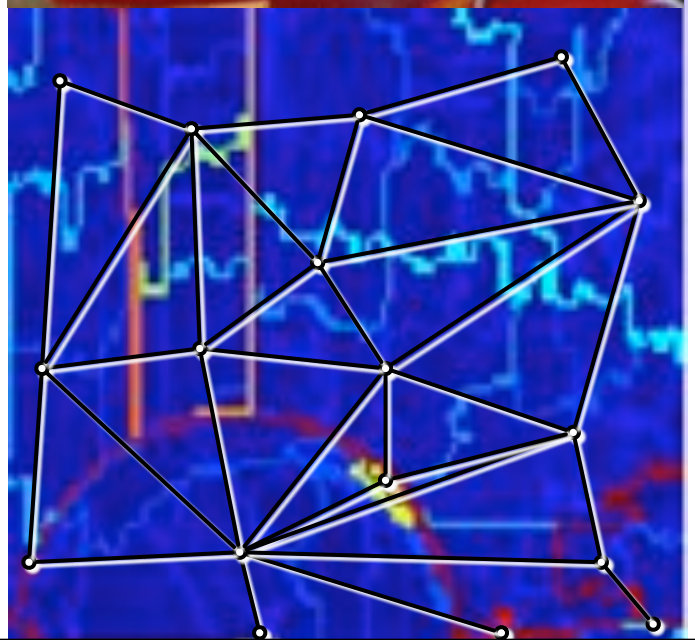


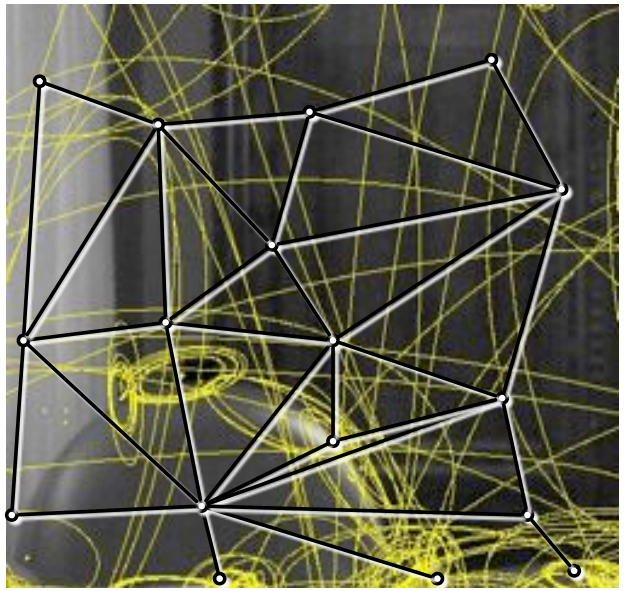


quickshift [vedaldi-soatto '08]
(non-iterative, constant-time, returns entire segmentation tree)

representational (hyper)graph





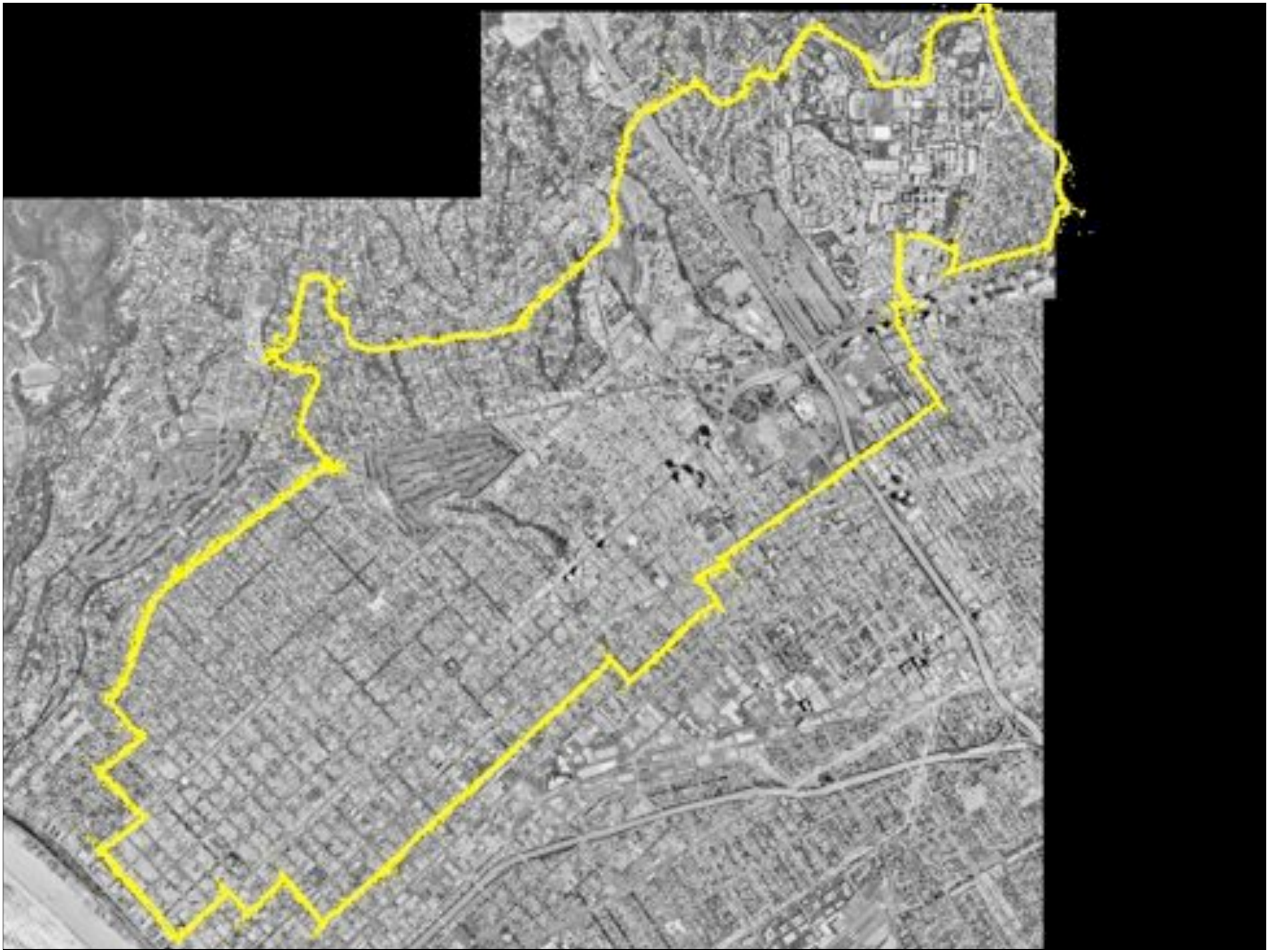


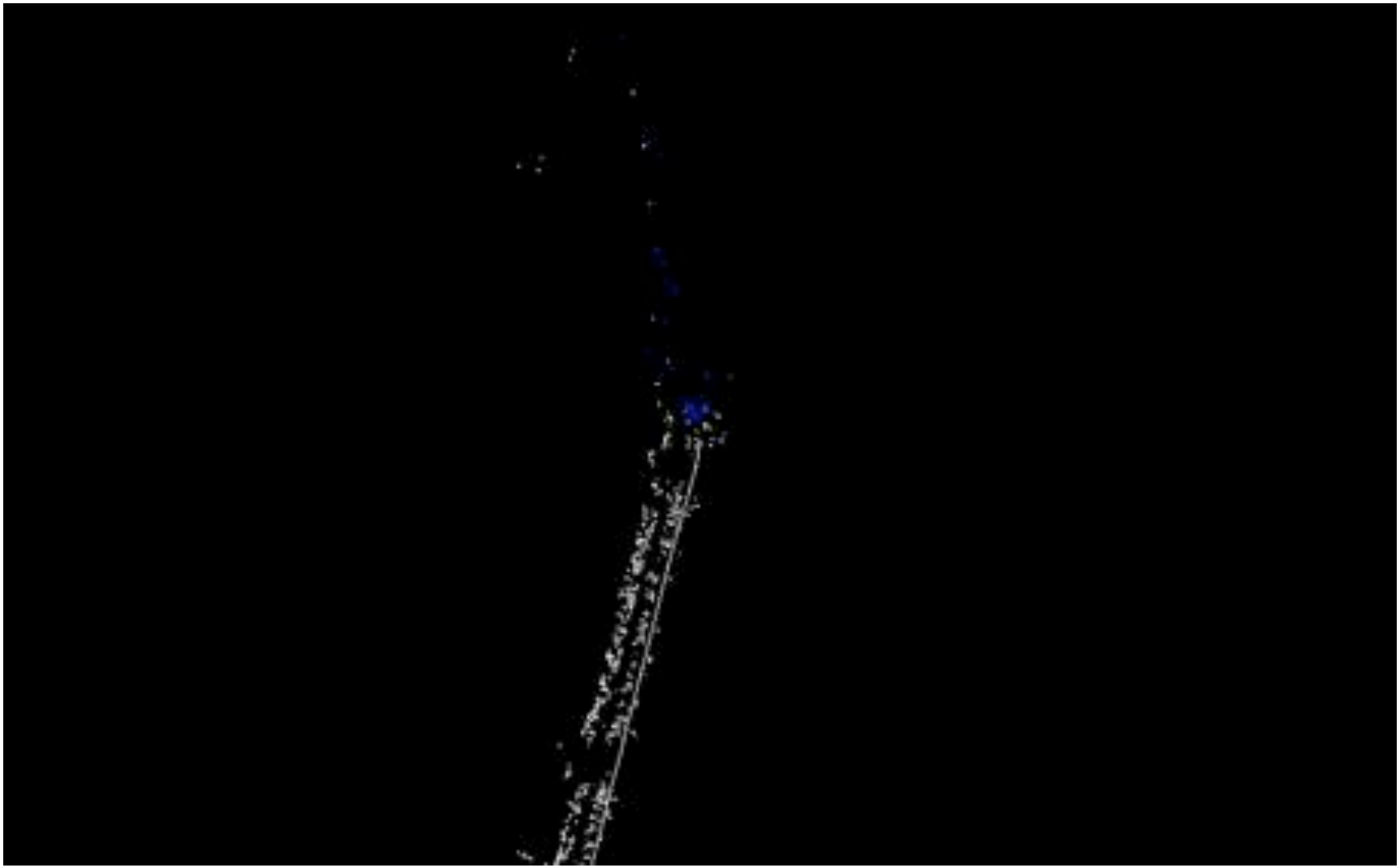
iphone demo

- http://www.youtube.com/watch?v=cMv-McHw660

5. visual exploration

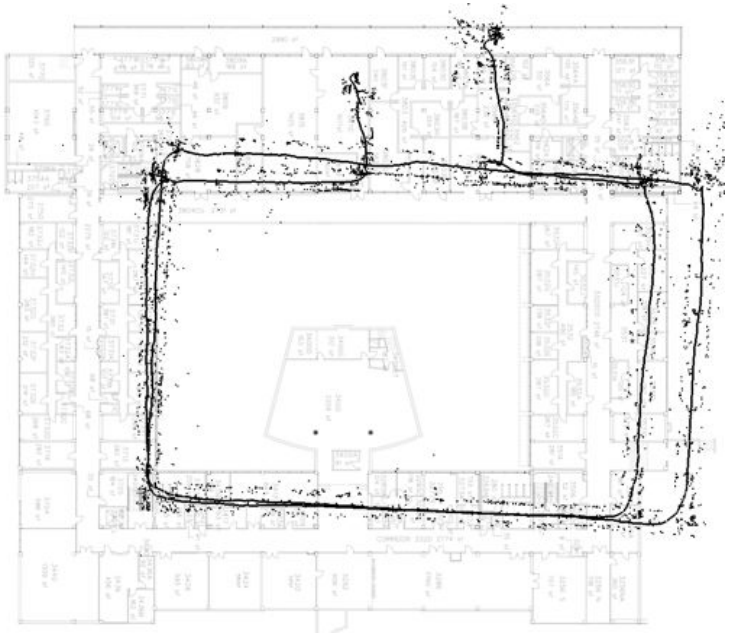
- Exploit gravity (but don't assume you know it!)
- Visual-Inertial navigation + Community Map Building [E. Jones and S. Soatto, IJRR 2011]



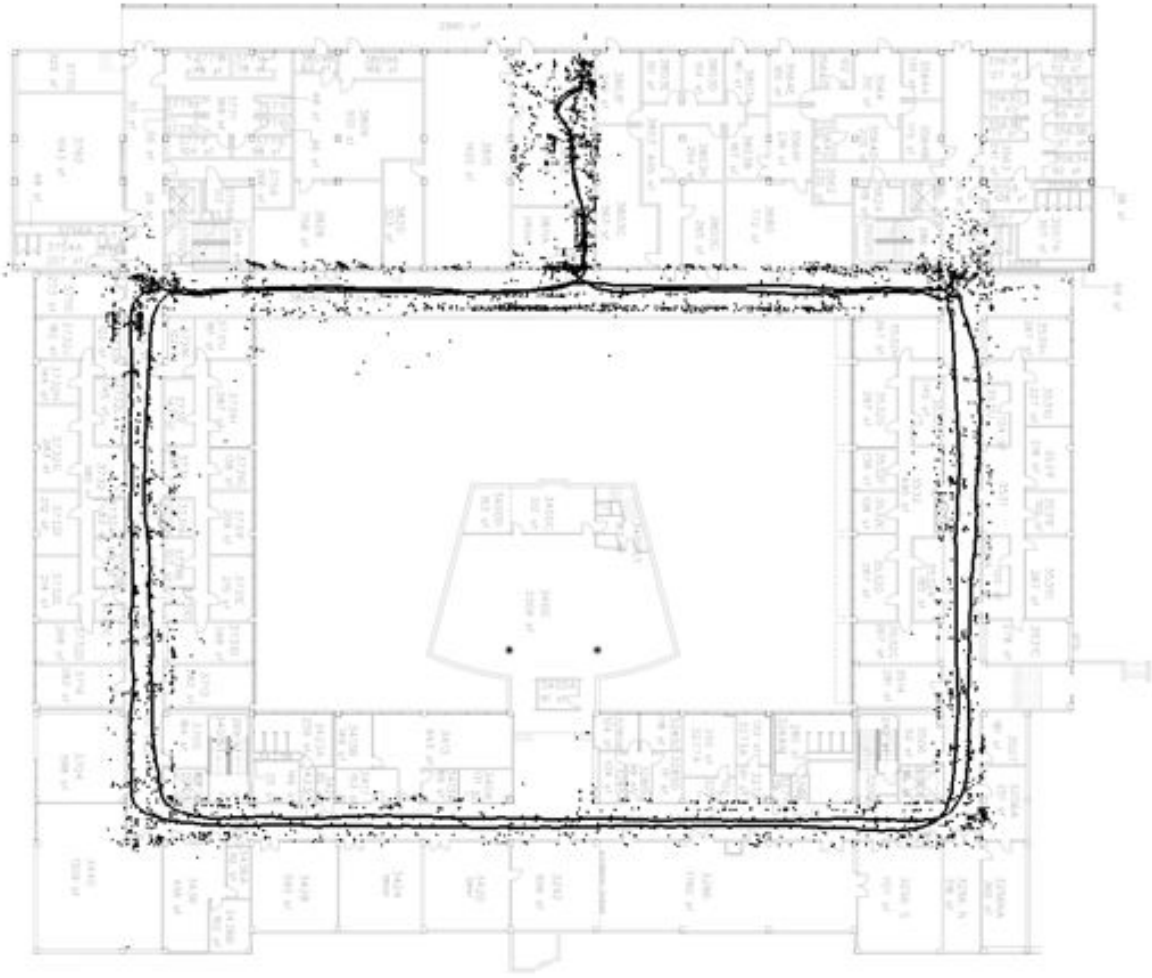


Inertial Only

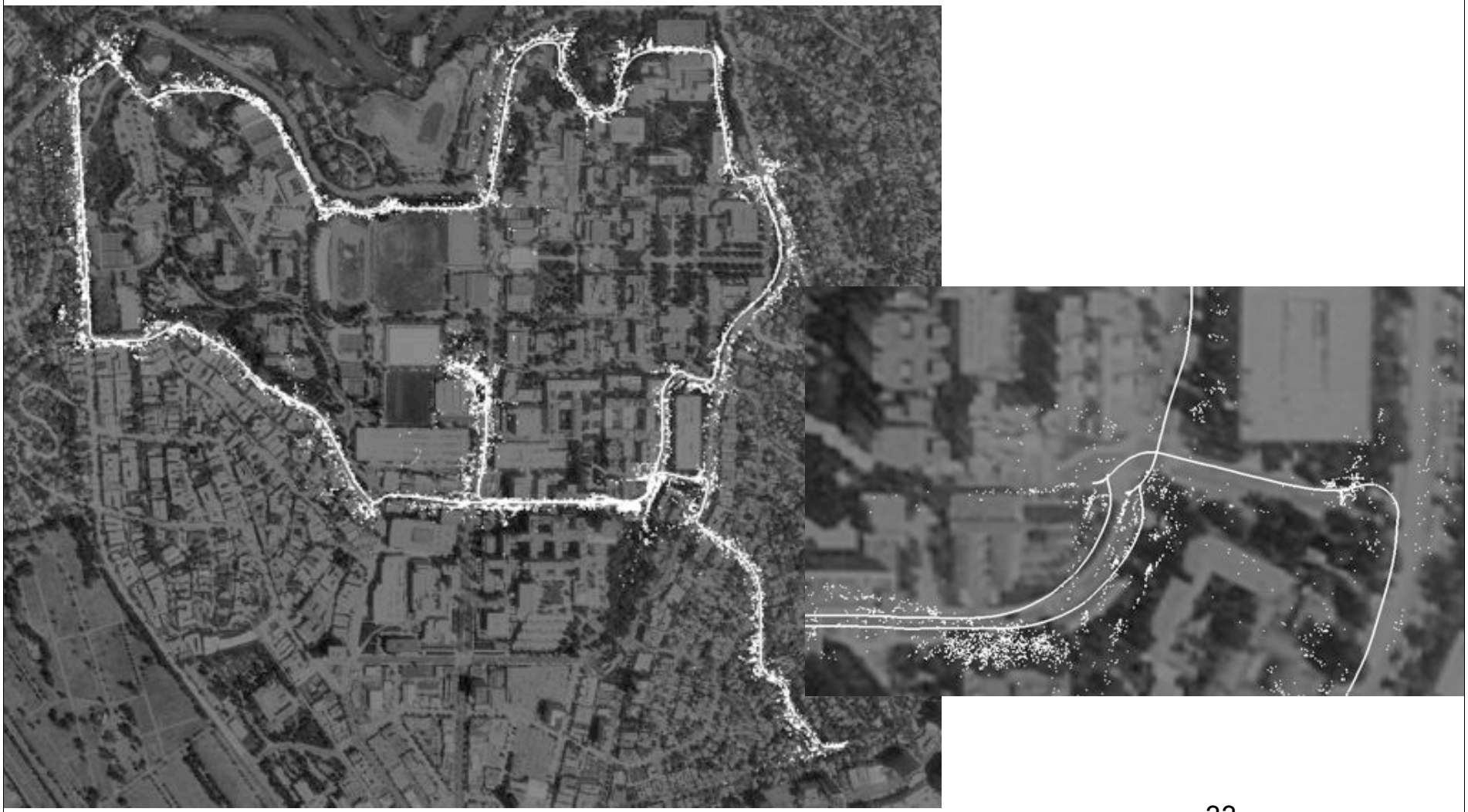
Vision Only



Drift: 0.19% (500 m)



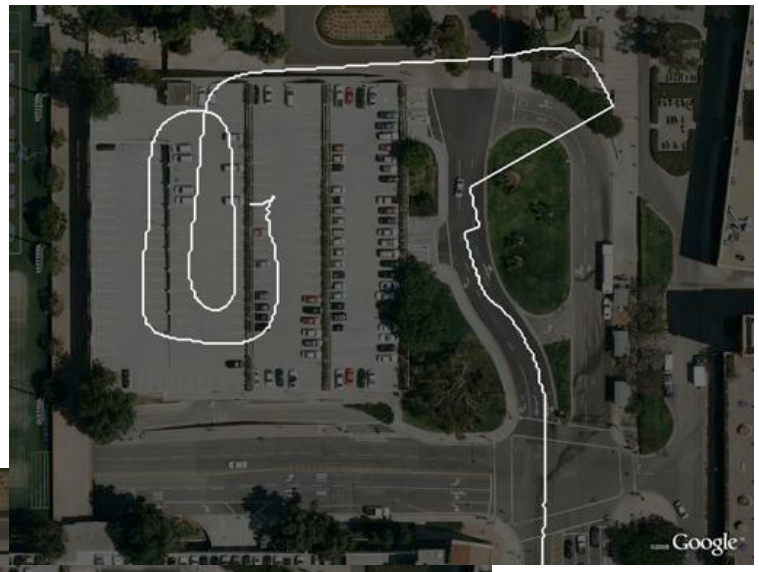
Drift: 0.27% (8 km)



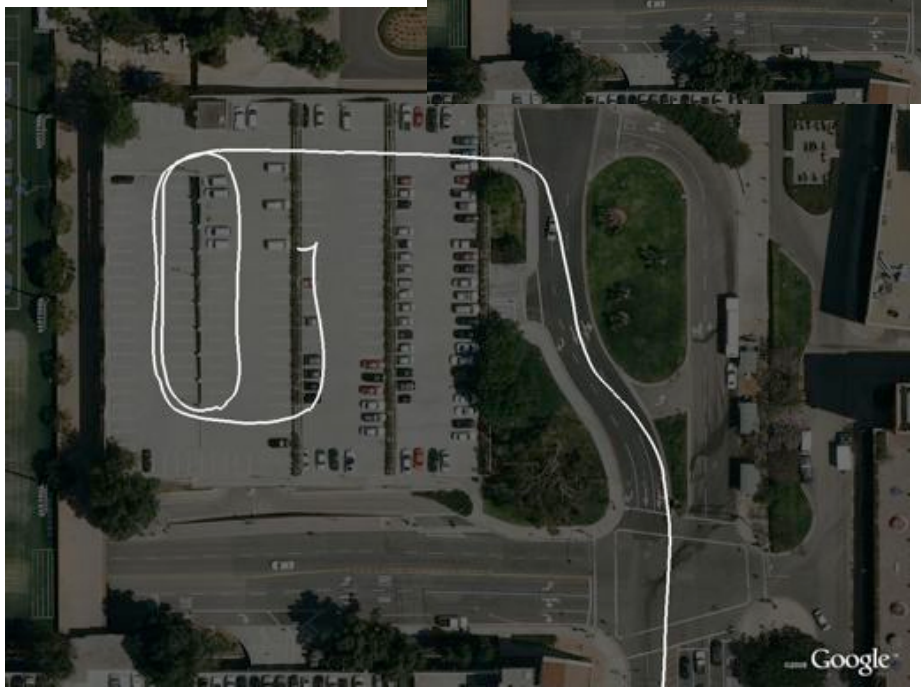
Drift: 0.5% (30km)

vs GPS+IMU

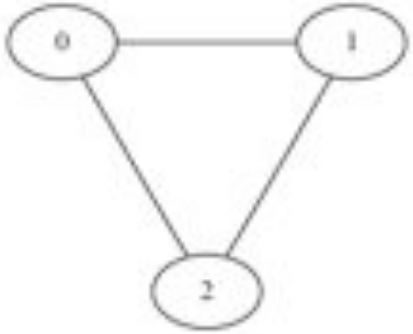
GPS+Inertial

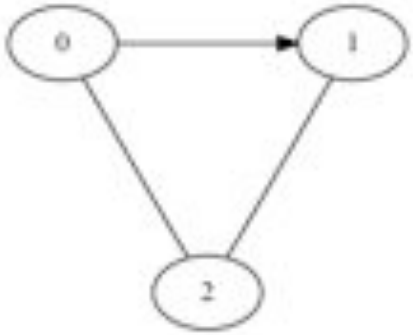


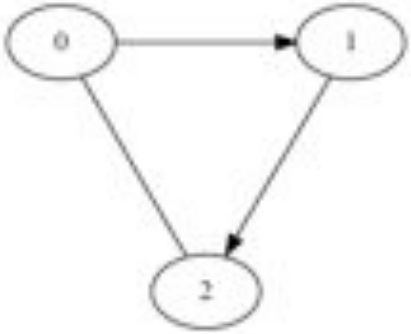
Vision+Inertial

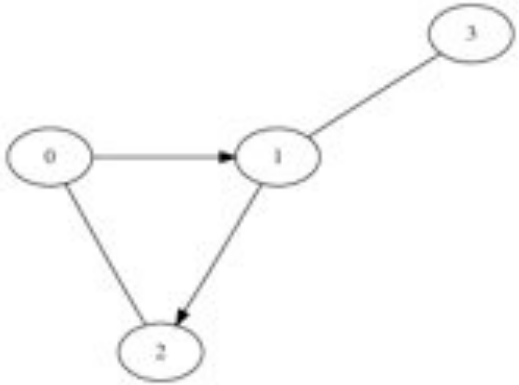


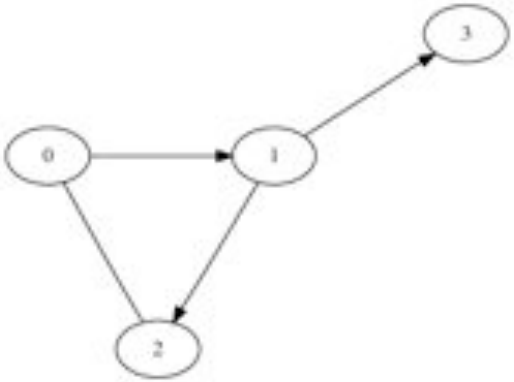
“location”, topology and co-visibility



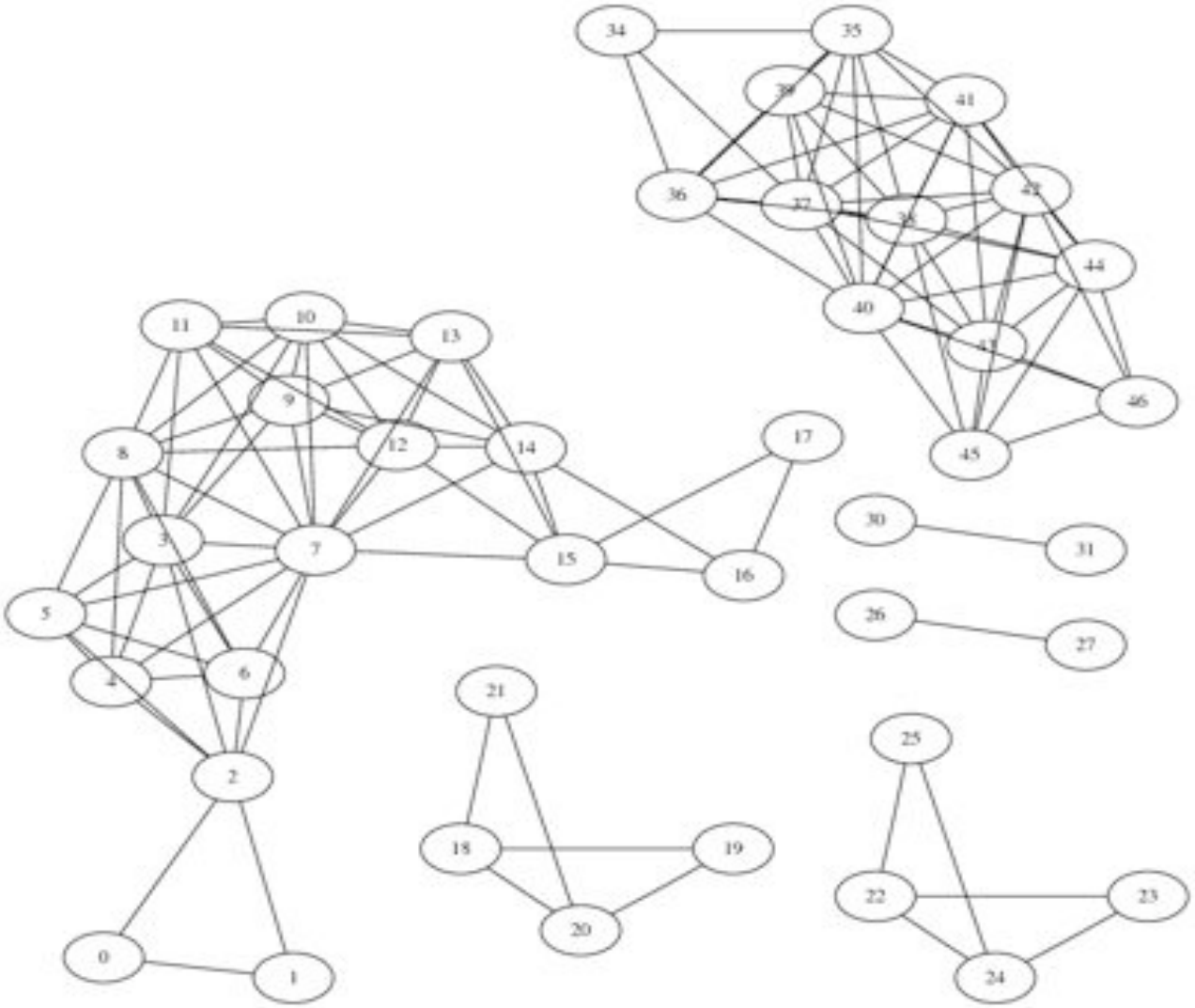




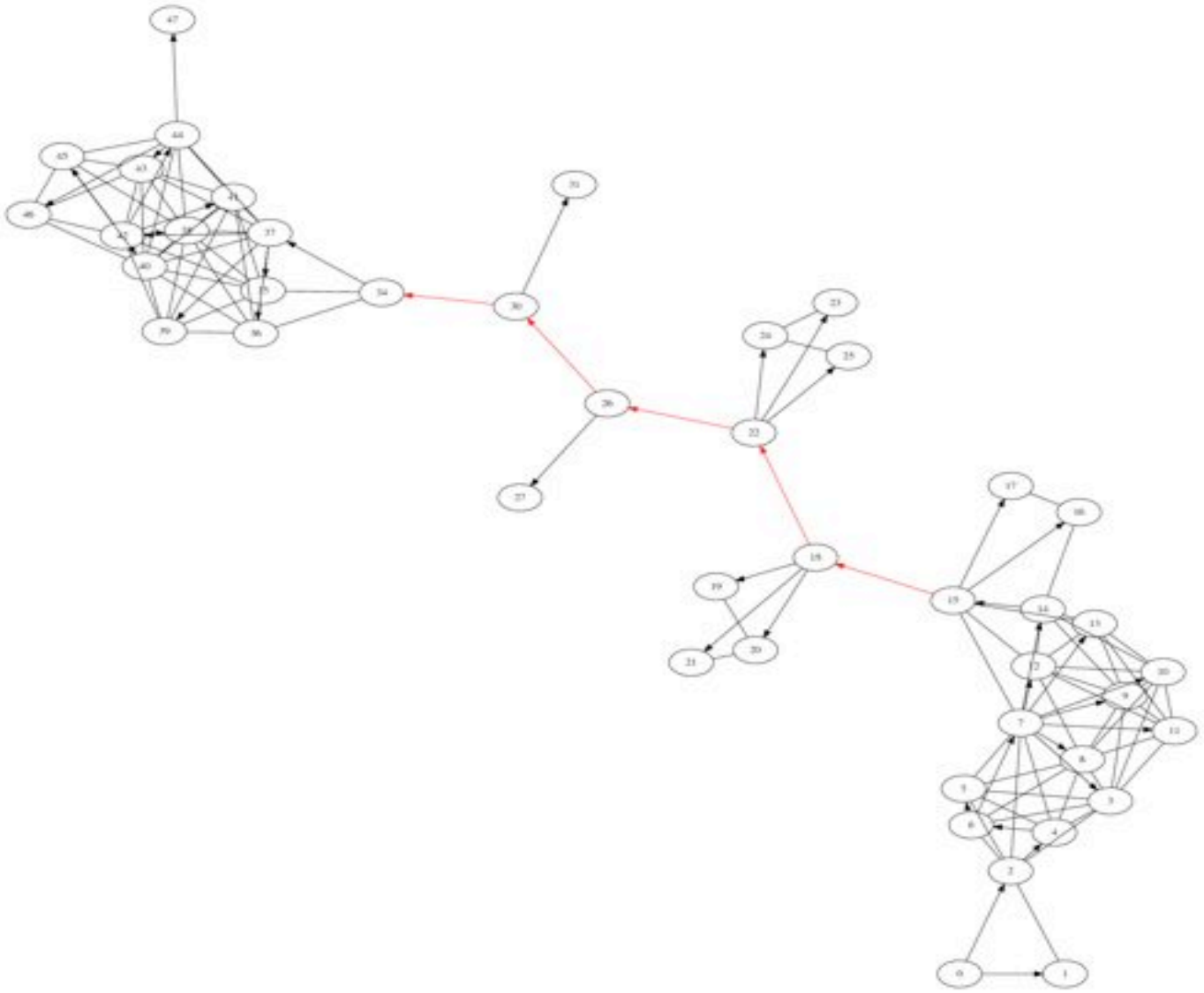




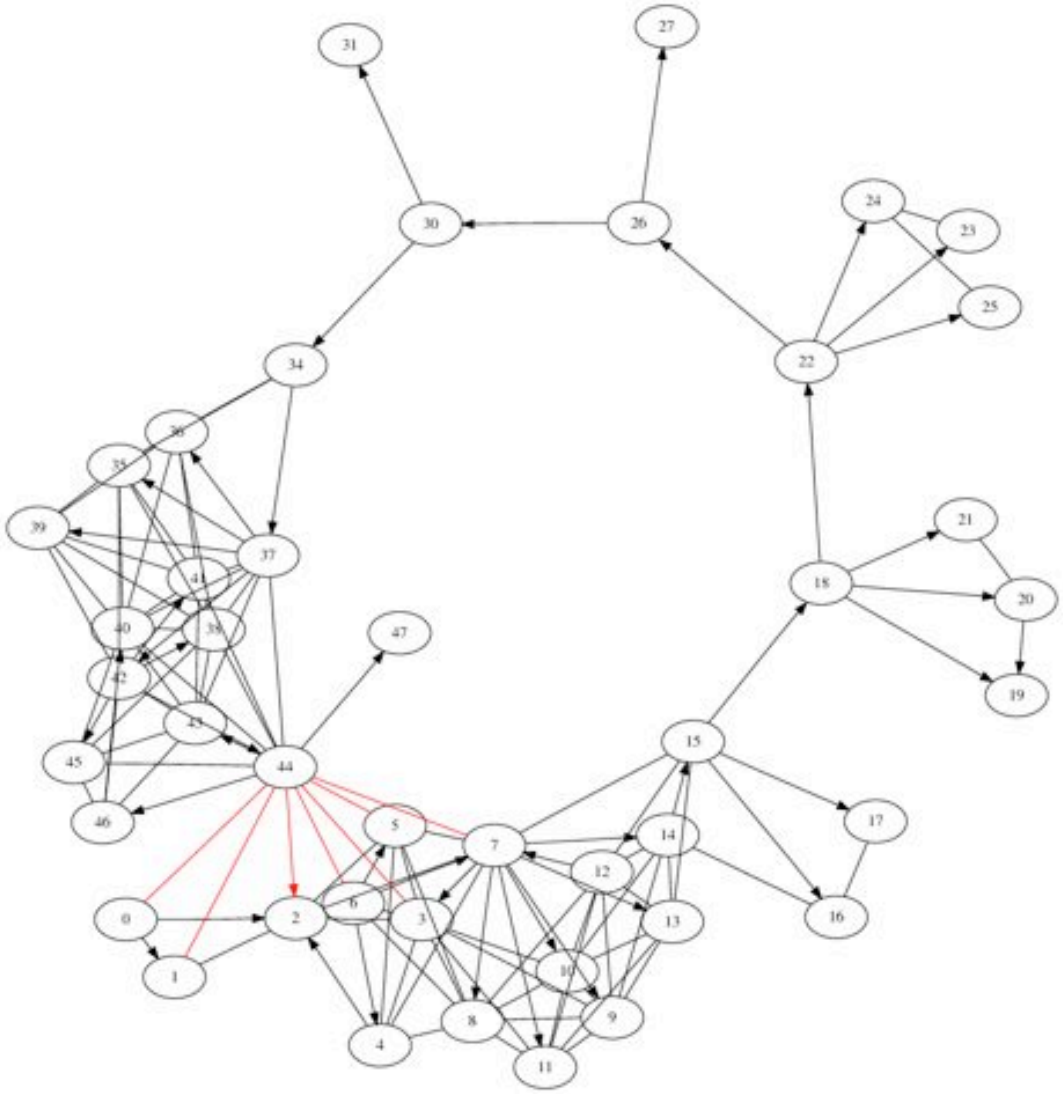
Covisibility Graph

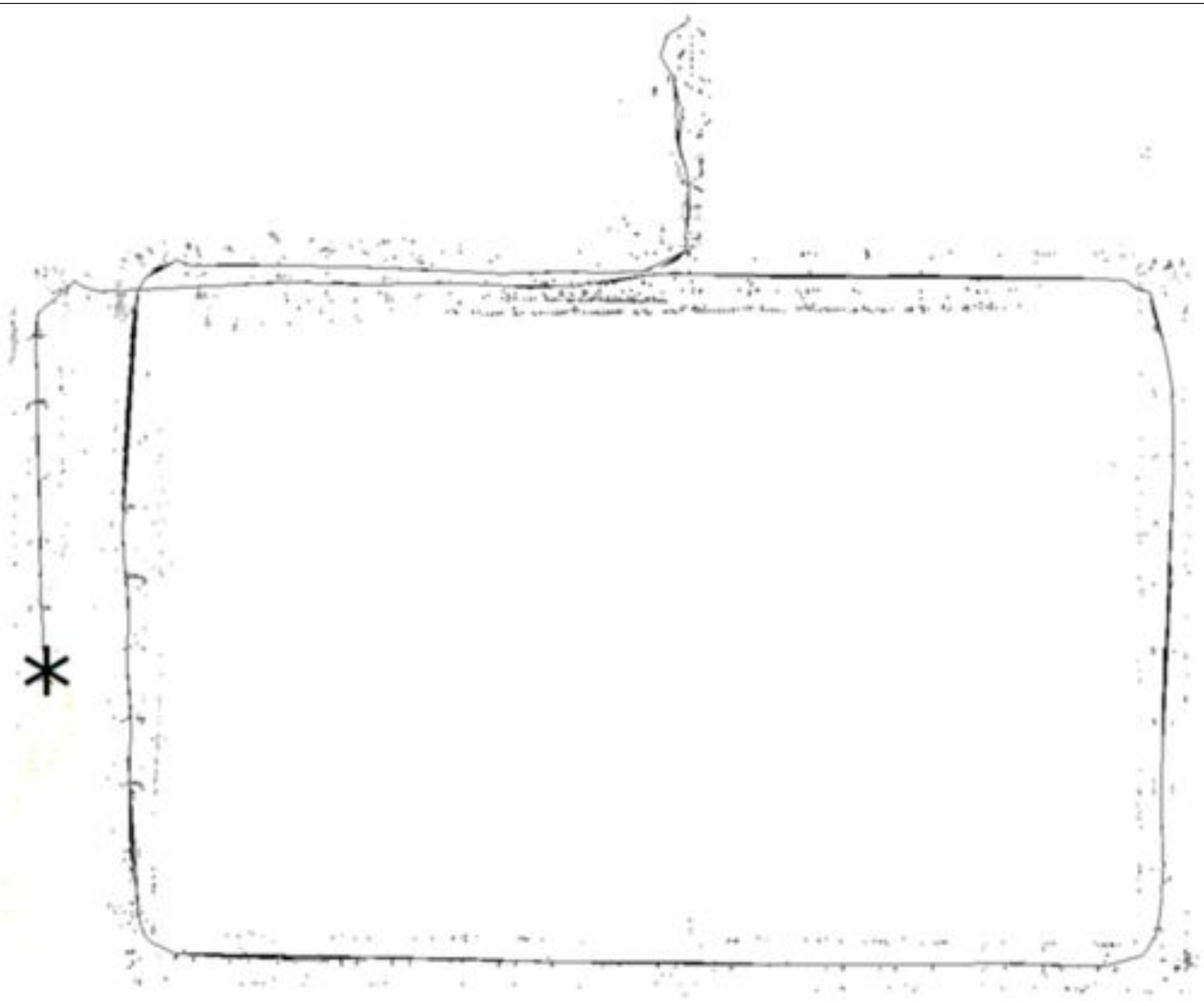


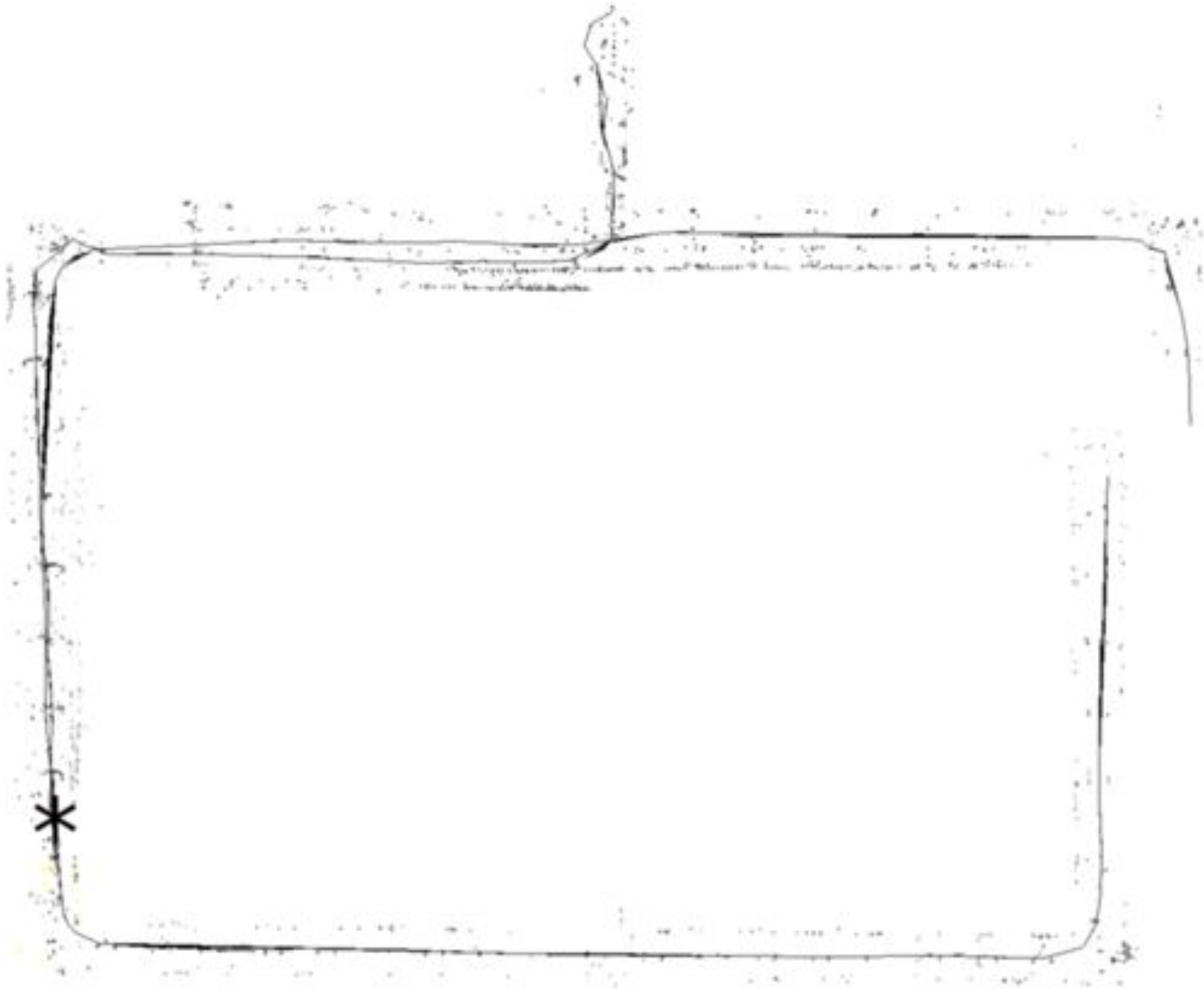
Adding Geometry

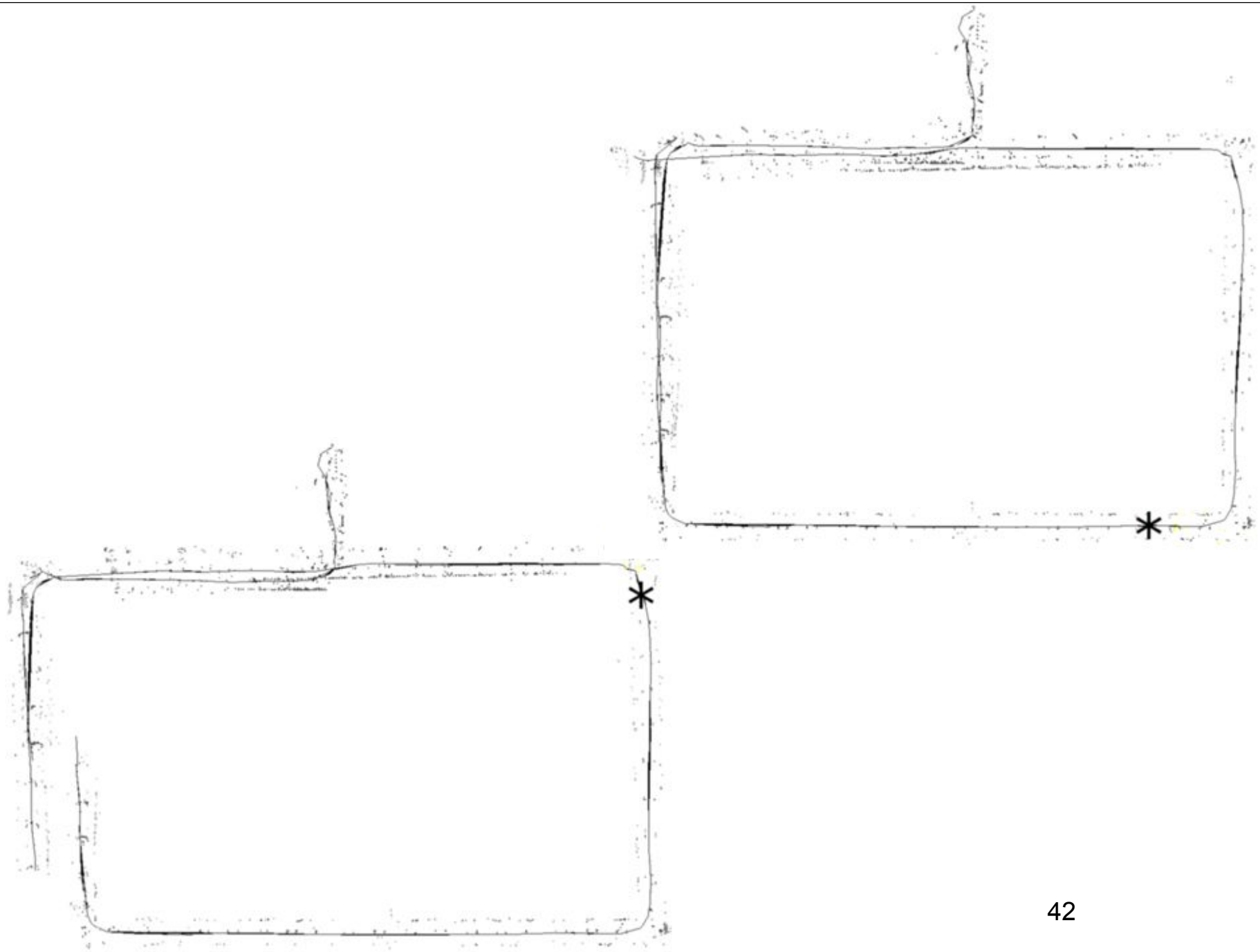


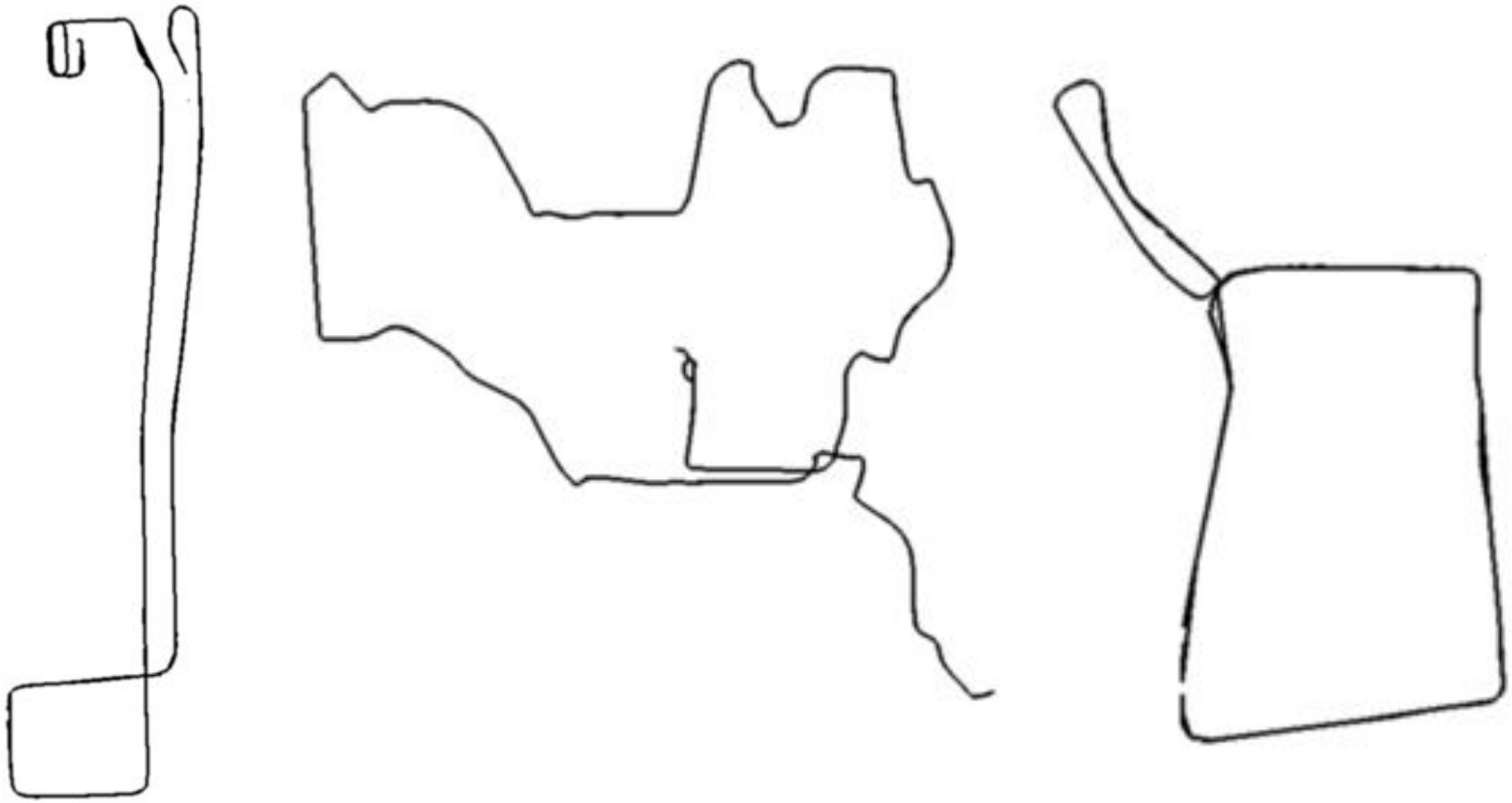
Loop Closing

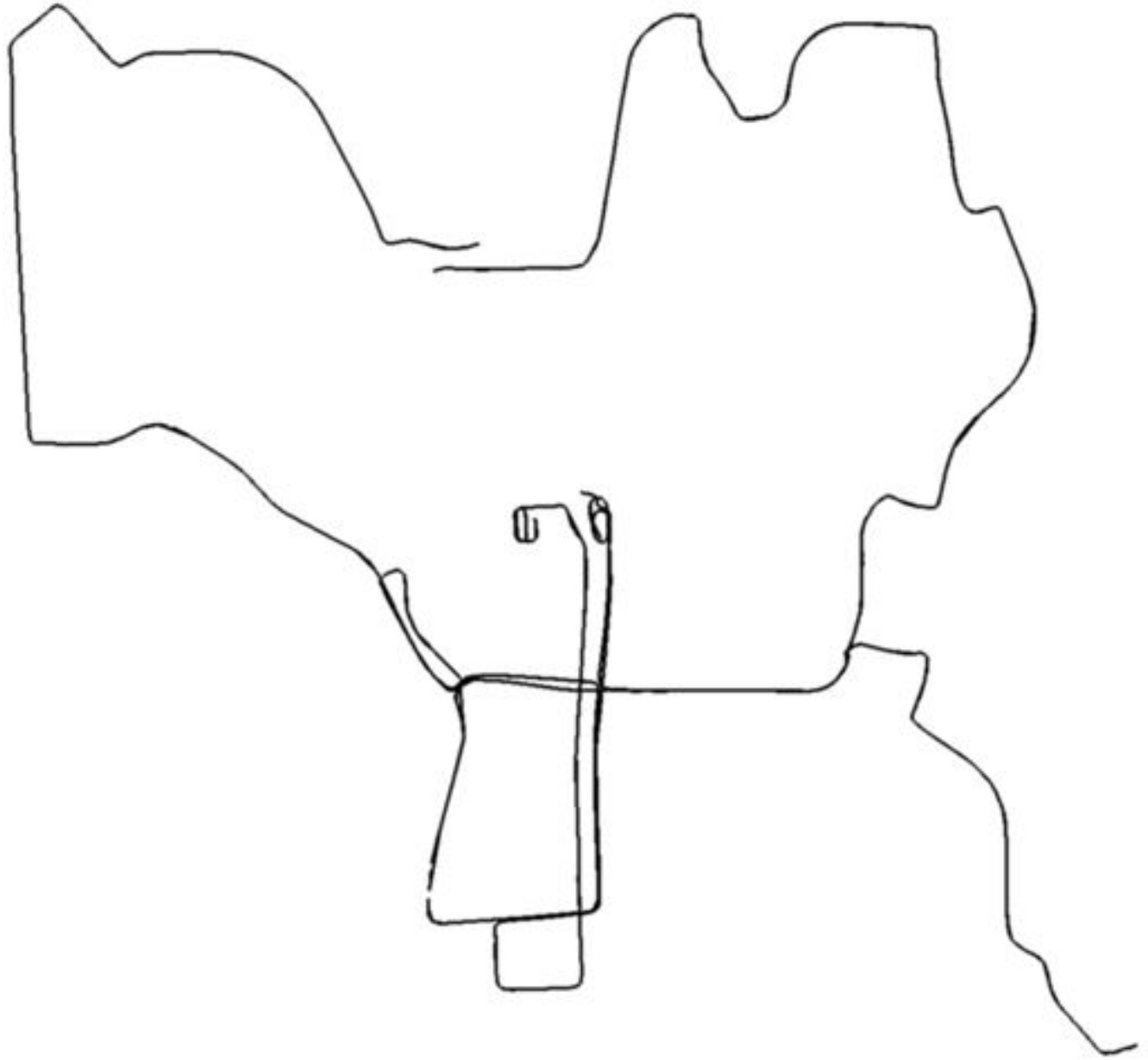












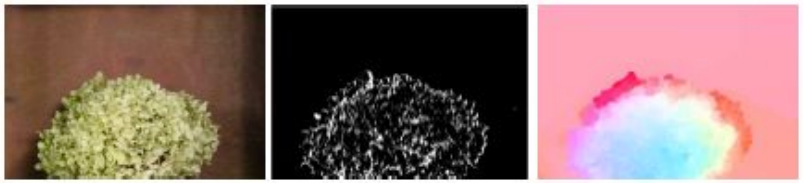
“The Black Box”

- Sensor Platform
  - Battery
  - Computation
  - D-GPS
  - Stereo, Omni Cameras
  - LADAR
  - IMU
- Portable
  - Wheels
  - Vehicle
  - Human

visual exploration

- occlusion detection
- myopic exploration
- memory and representation
- min(inference)-max(control) entropy

occlusion detection



- (i) lambertian reflection (ii) constant illumination, (iii) co-visibility:

small dense

$$I(x,t+dt) = \begin{cases} I(w(x,t),t) + n(x,t) & x \in D \backslash \Omega(d,dt) \\ \nu(x,t), & x \in \Omega(t,dt) \end{cases}$$

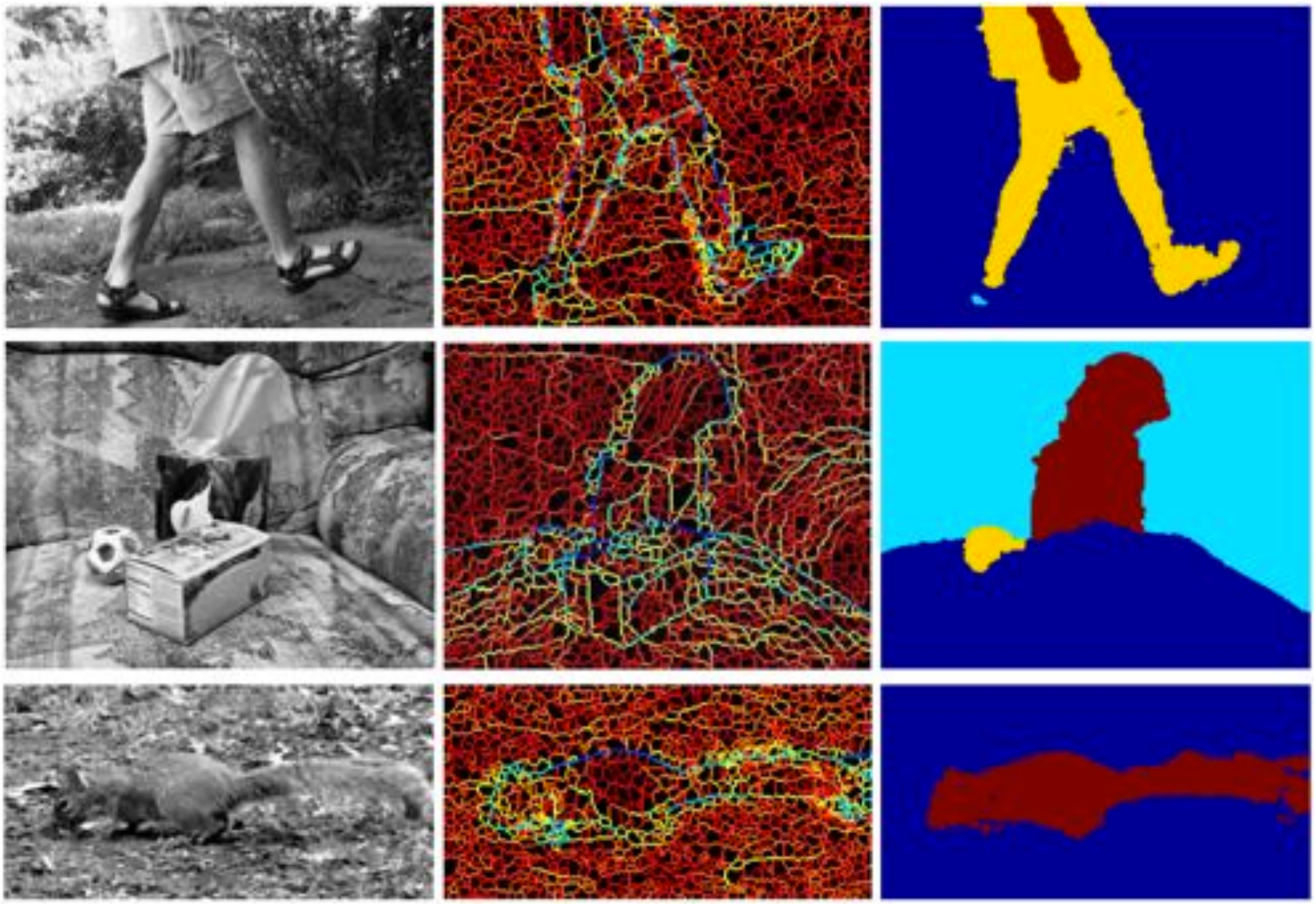
large sparse $\Omega \overset{dt \to 0}{\to} \emptyset$

$$\hat{\Omega}, \hat{w} = \arg\min \|\nu\|_0 + \lambda \|n\|_1$$

re-weighted ℓ^1

nesterov vs. split-bregman w/ isotropic reg. TV for w

[ayvaci-raptis-soatto nips '10]

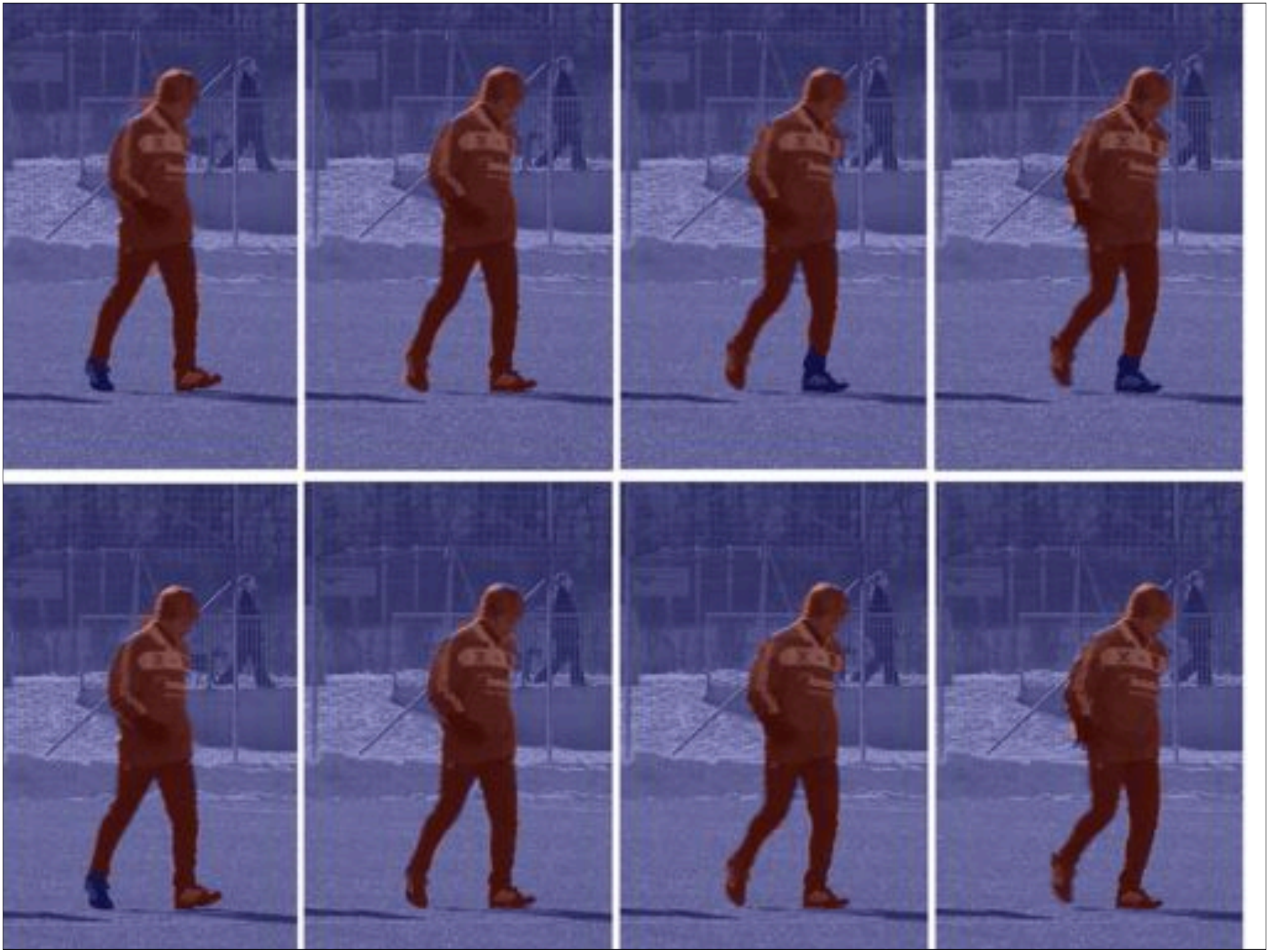


occlusion detection

- code at http://vision.ucla.edu/~ayvaci/occlusion-detection

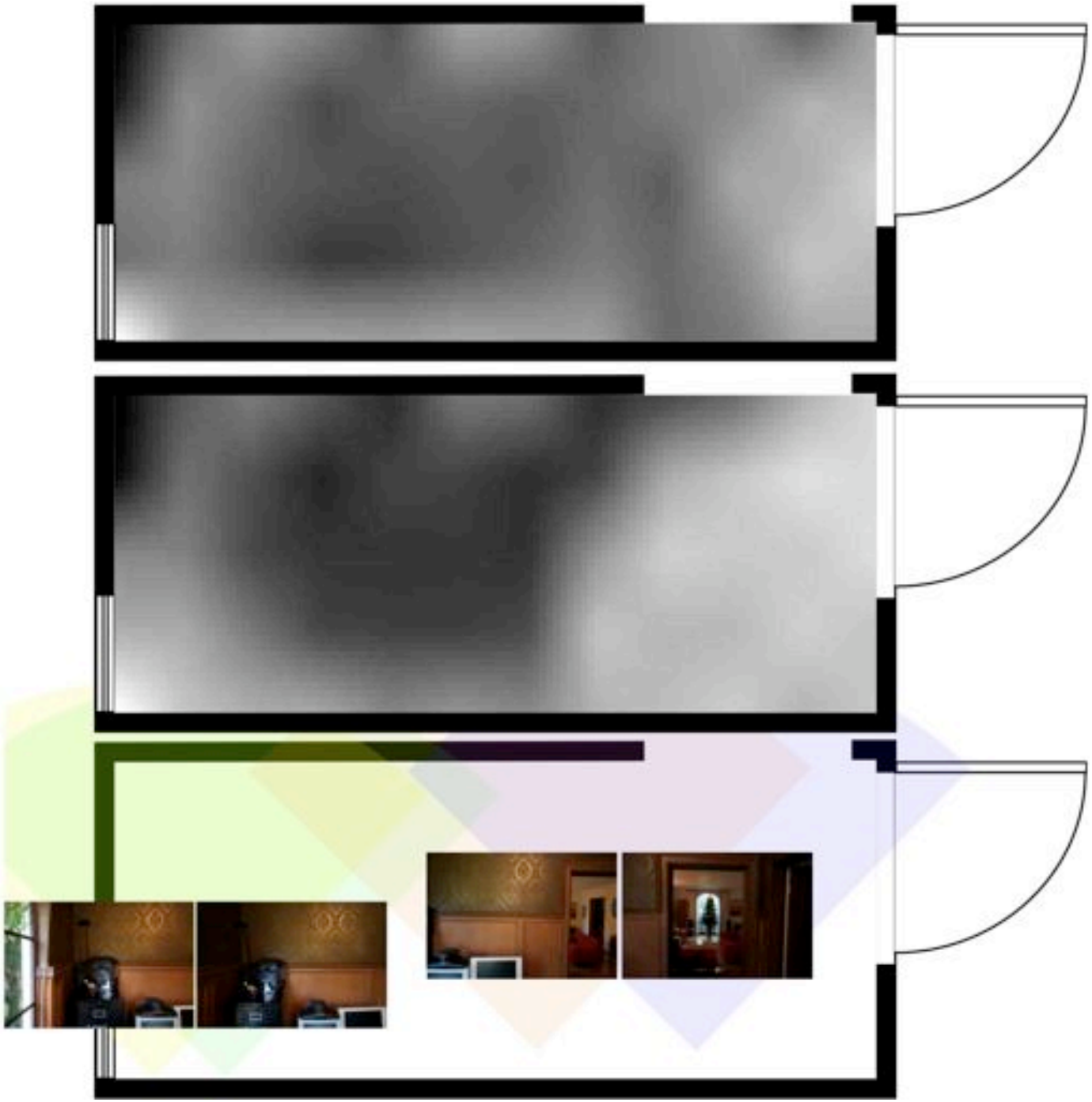
detachable object detection

- "efficient model selection for detachable object detection", proc. of EMMCVPR, July 2011



building a representation: perceptual explorers

$$\begin{cases} \hat{\xi}_{t+dt} = \hat{\xi}_t \oplus \epsilon(I_{t+dt}, t+dt; \hat{u}_t, \hat{\xi}_t) \\ \hat{u}_t = \arg\max_u H(\epsilon(I_t, t; u, \hat{\xi}_t)) \\ \hat{\xi}_0 = h^{-1}(I_0) \end{cases}$$



what's peculiar about vision?

- *scaling* makes continuous limit relevant
- *occlusions* make mobility/control relevant
- phenomena critical in any *remote sensing* modality (EO, IR, MS, radar, laser, lidar, TOF, ...)

references

- S. Soatto, "Actionable Information in Vision", in Machine Learning for Computer Vision, R. Cipolla et al. (eds), Springer Verlag 2011 (short version in the Proc. of the Intl. Conf. on Comp. Vision, ICCV 2009).
- E. Jones and S. Soatto, "Visual-inertial navigation, localization and mapping", IJRR 2011
- A. Ayvaci et al. "Optical Flow and Occlusion Detection with Convex Optimization", NIPS 2010