VISION-BASED REACQUISITION FOR TASK-LEVEL CONTROL

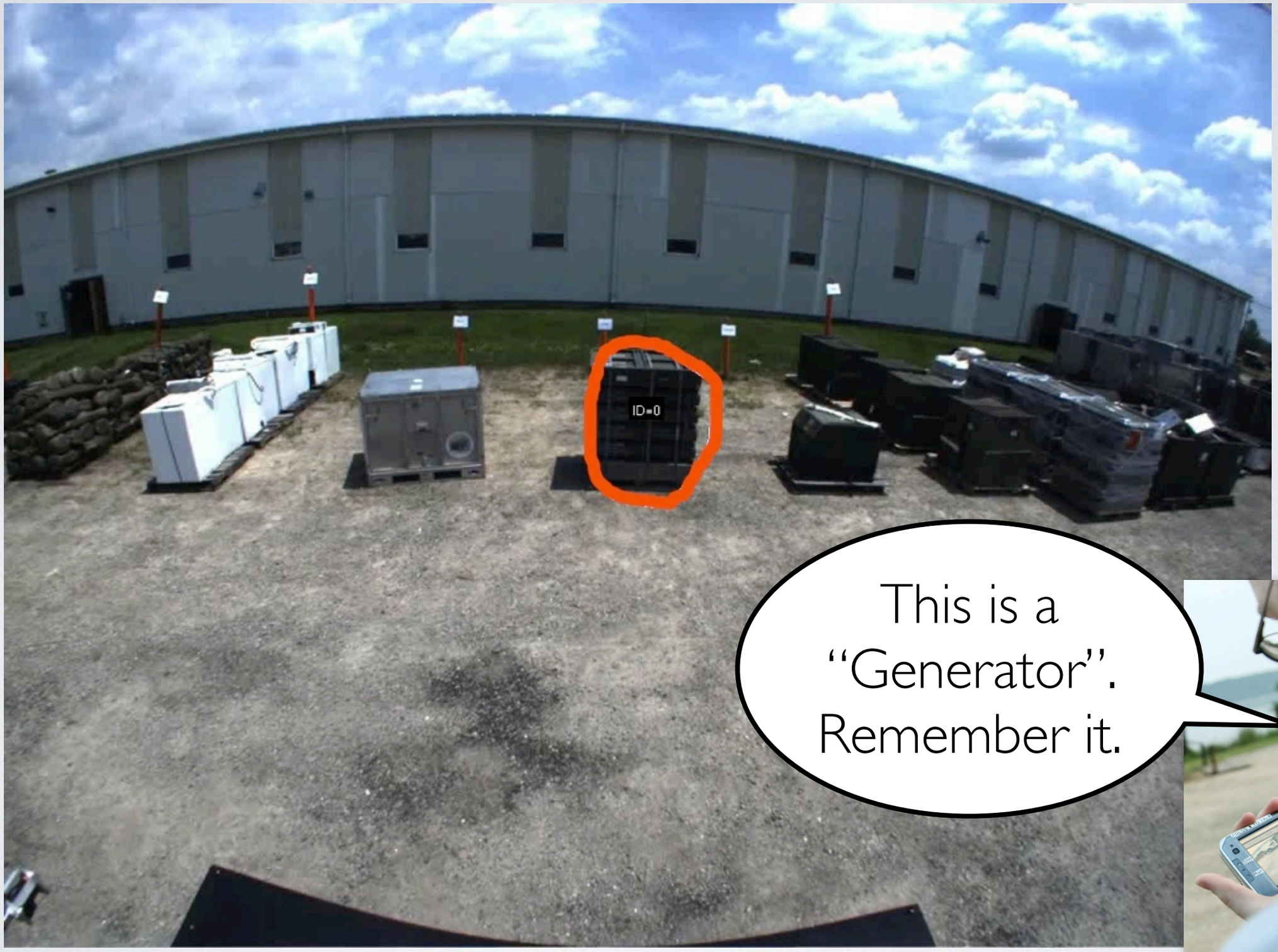
[video: 2010_waverly_reacquisition.mp4]

Matthew Walter[1], Yuli Friedman[2], Matthew Antone[2], & Seth Teller[1]

[1]MIT / CSAIL [2]BAE Systems

21 December 2010

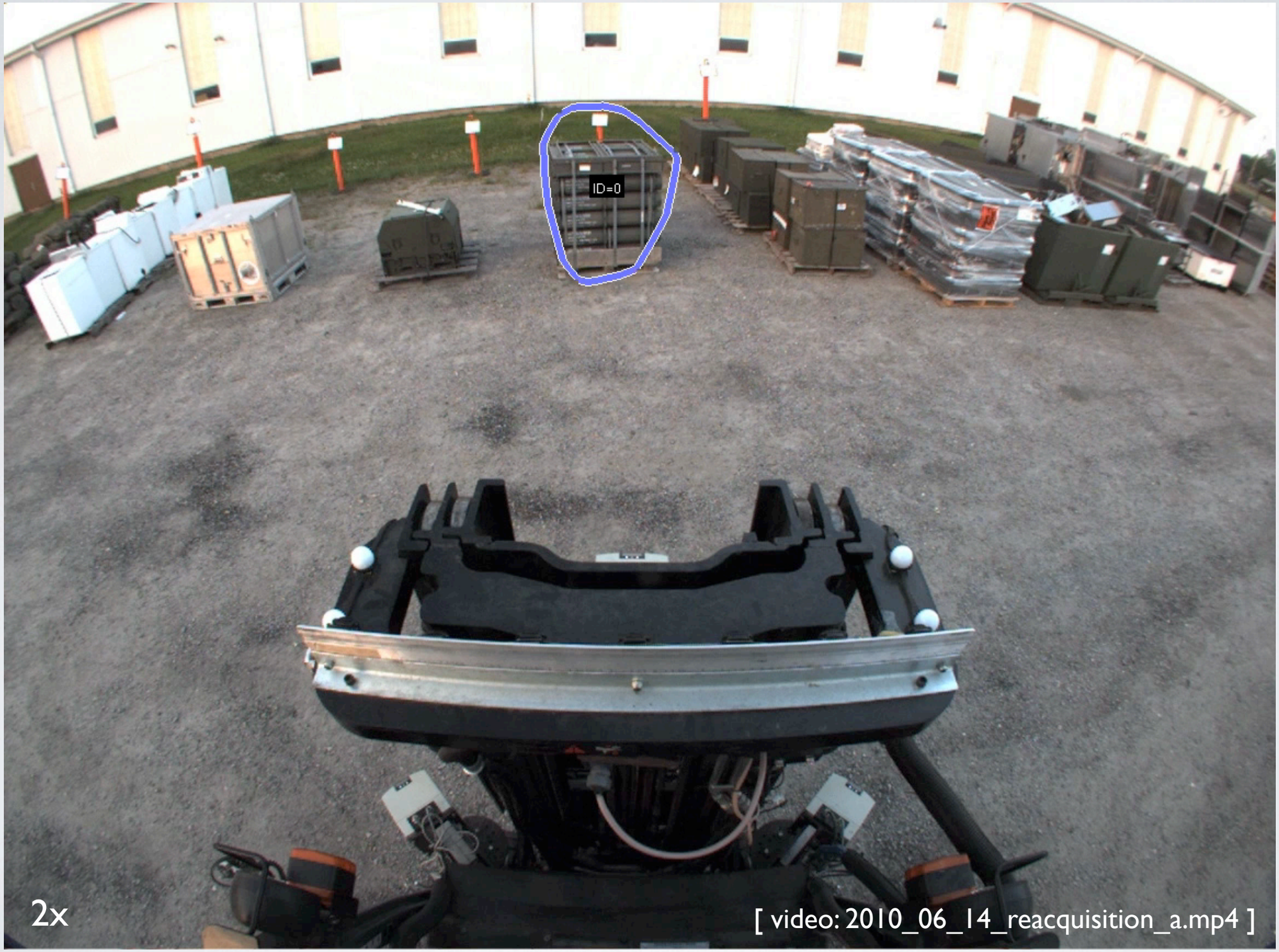
VISION-BASED OBJECT REACQUISITION



VISION-BASED OBJECT REACQUISITION

VISION-BASED OBJECT REACQUISITION

VISION-BASED OBJECT REACQUISITION



VISION-BASED OBJECT REACQUISITION

SHARED SITUATIONAL AWARENESS

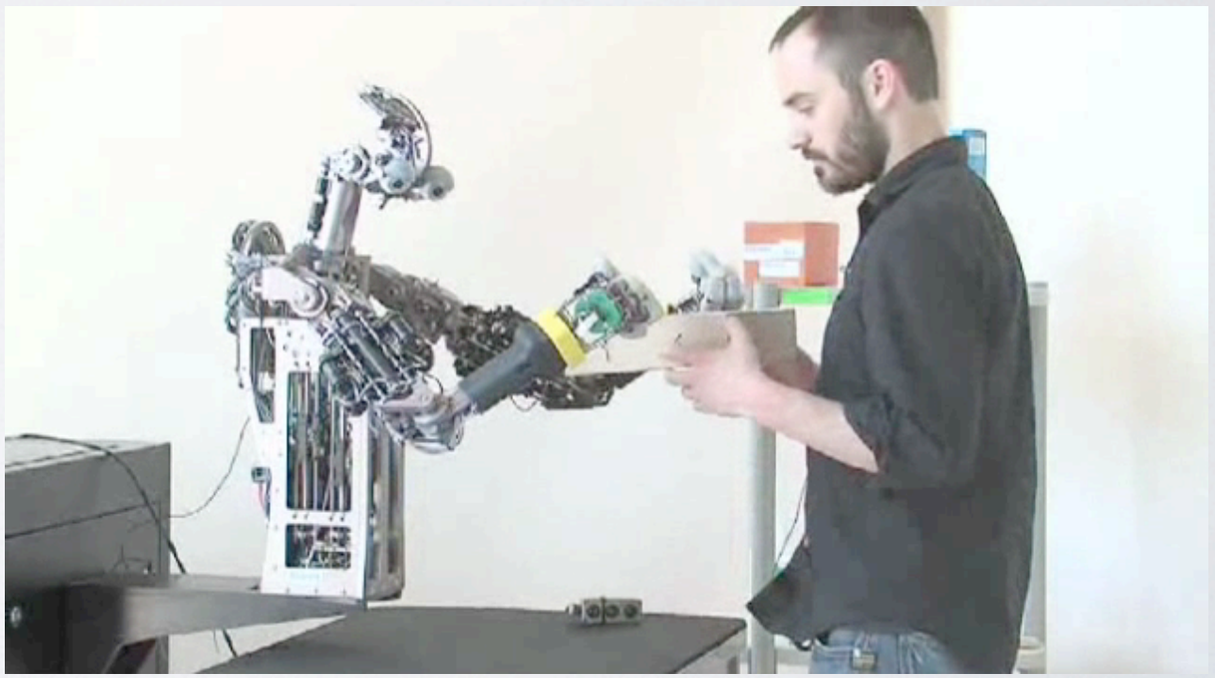
Enables robots that accommodate people

- Robots that operate in *our* human-populated, unstructured environments
- Safe, human-centered operation is critical
- Effective command and control mechanisms needed

Acquiring Local Knowledge: Narrated Guided Tour

[video: 2010_tbh_mit_wheelchair.mp4]

TBH Wheelchair (MIT/CSAIL)



Domo (MIT/CSAIL)

OUR PROBLEM: HUMAN-COMMANDABLE AUTONOMOUS FORKLIFT

- Operate in existing dynamic, unstructured facilities
- Usable by existing personnel
- Autonomously manipulate cargo
- Natural, effective command and control

Shared situational awareness

TASK-LEVEL MOBILE MANIPULATION

Multimodal command interface

- Stylus gestures: "Drawing on the world"
- Spoken language interface

Autonomous tasks

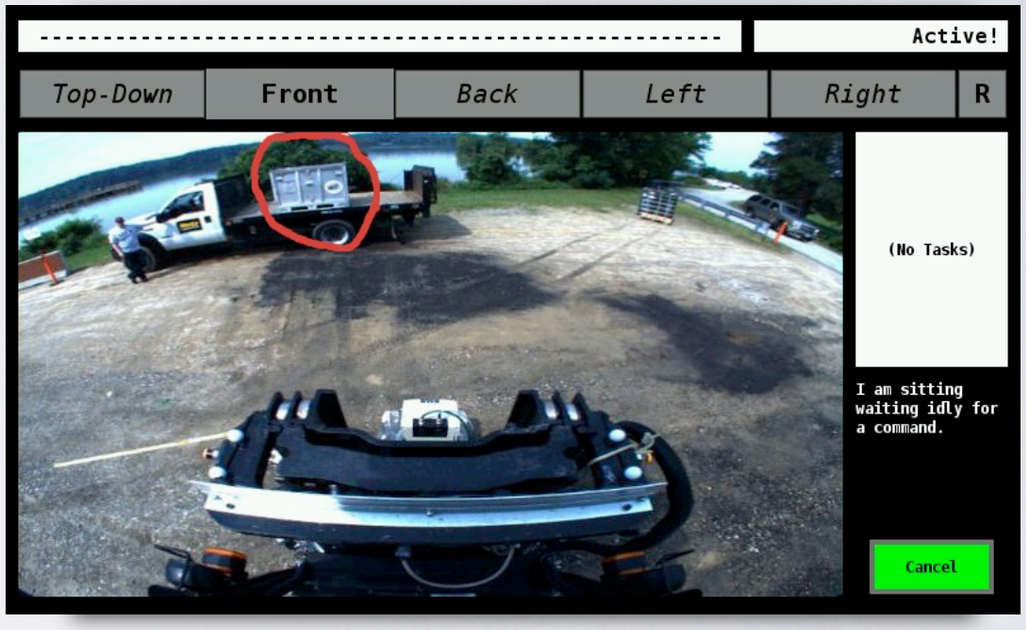
- Manipulation to/from ground and truck
- Transport to desired location

[Correa, et al., HRI 2010;
Teller et al., ICRA 2010; Walter et al., IROS 2010]

TASK-LEVEL MOBILE MANIPULATION

Multimodal command interface

- Stylus gestures: "Drawing on the world"
- Spoken language interface



Autonomous tasks

- Manipulation to/from ground and truck
- Transport to desired location

[Correa, et al., HRI 2010;
Teller et al., ICRA 2010; Walter et al., IROS 2010]

EXTENDING SITUATIONAL AWARENESS

- Utilize sensing to extend situational awareness
- Richer interactions via shared world model
- Increased autonomy
  - Understand higher-level commands
  - Extended-duration tasks

Guided tour of named objects

[video: 2010_06_guided_tour.mp4]

EXTENDING SITUATIONAL AWARENESS

- Utilize sensing to extend situational awareness
- Richer interactions via shared world model
- Increased autonomy
  - Understand higher-level commands
  - Extended-duration tasks

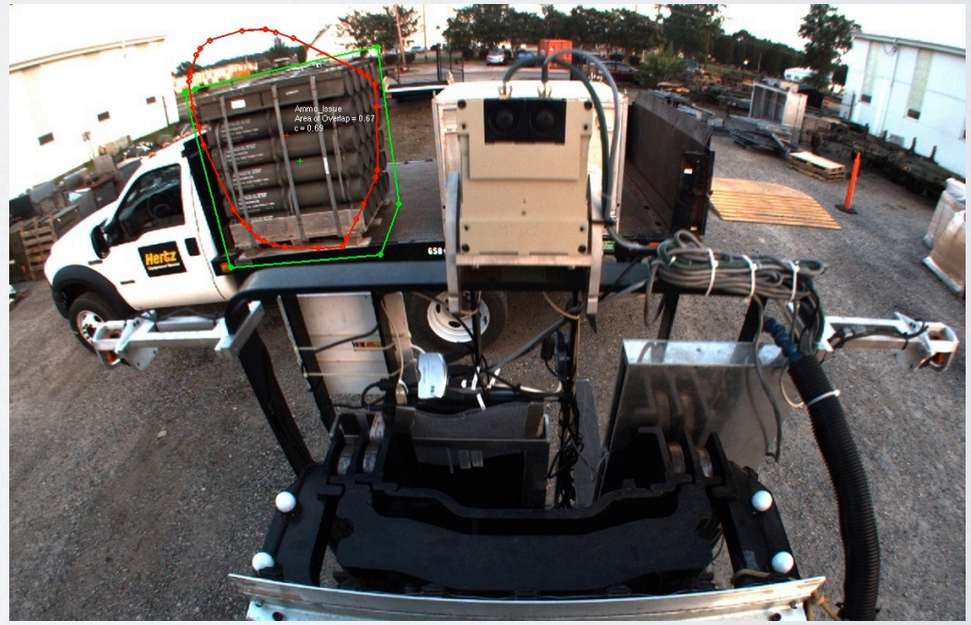
Retrieve objects simply by name

[video: 2010_06_reacquire_pickup.mp4]

VISION-BASED OBJECT REACQUISITION

- Goal: Recognize specific objects hours/days later
  - Reconstitute user gesture (segmentation)
- Challenges
  - Outdoors: Varying lighting
  - Viewpoint changes over time
  - Object relocation
  - Coarse localization

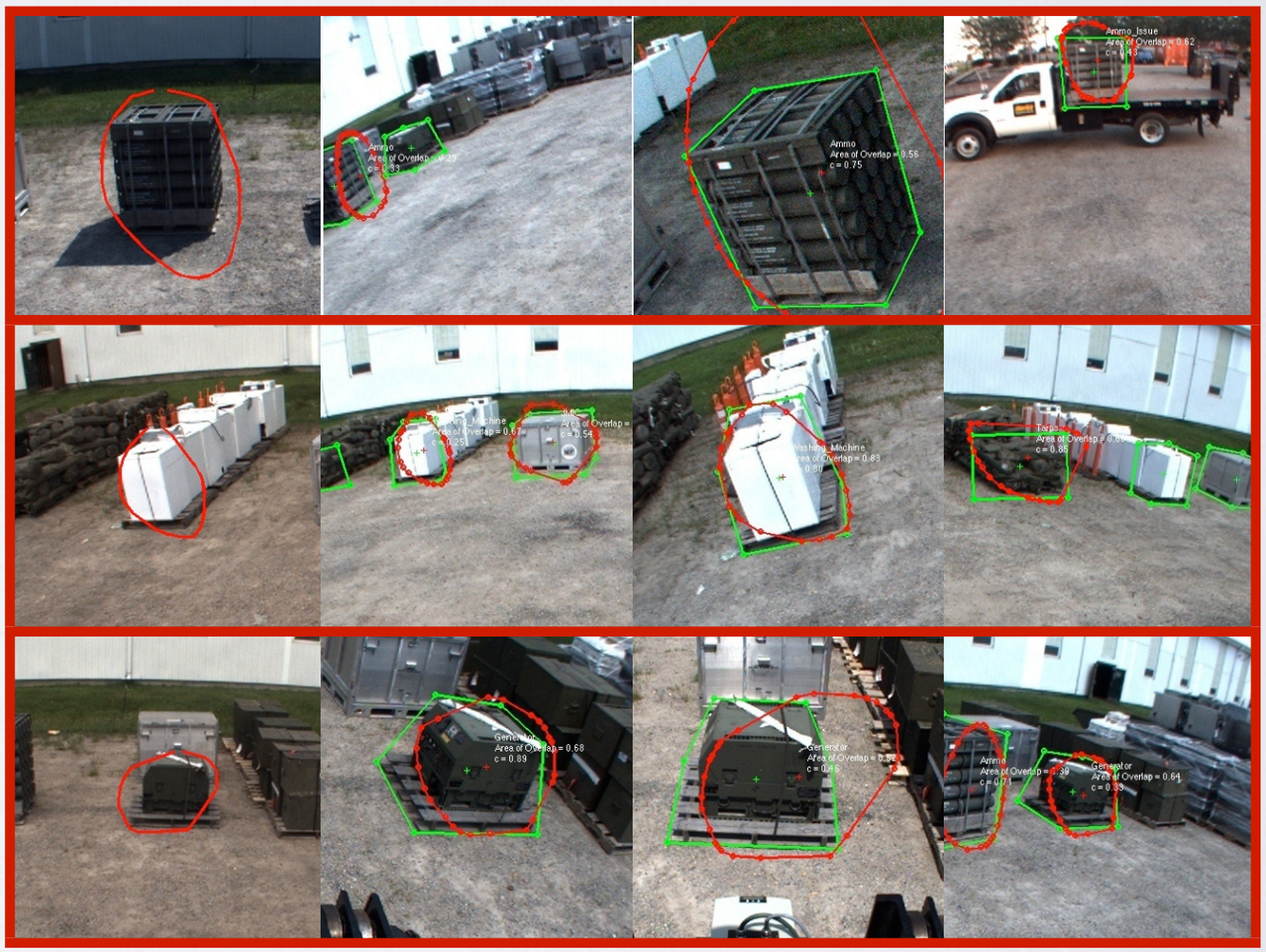
Original segmentation



Reacquisition with virtual gesture

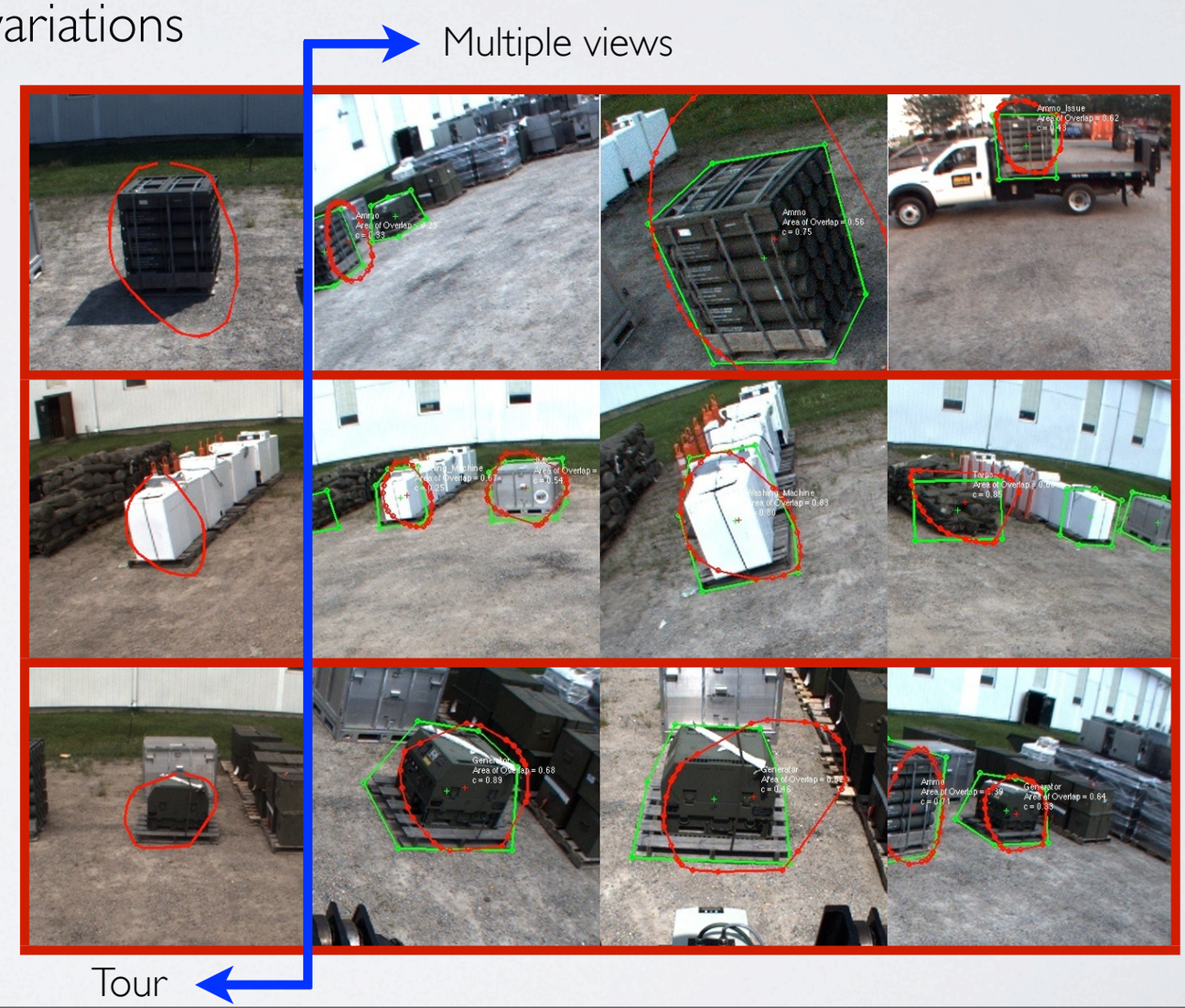
VISION-BASED OBJECT REACQUISITION

- Goal: Recognize specific objects hours/days later
  - Reconstitute user gesture (segmentation)
- Challenges
  - Outdoors: Varying lighting
  - Viewpoint changes over time
  - Object relocation
  - Coarse localization



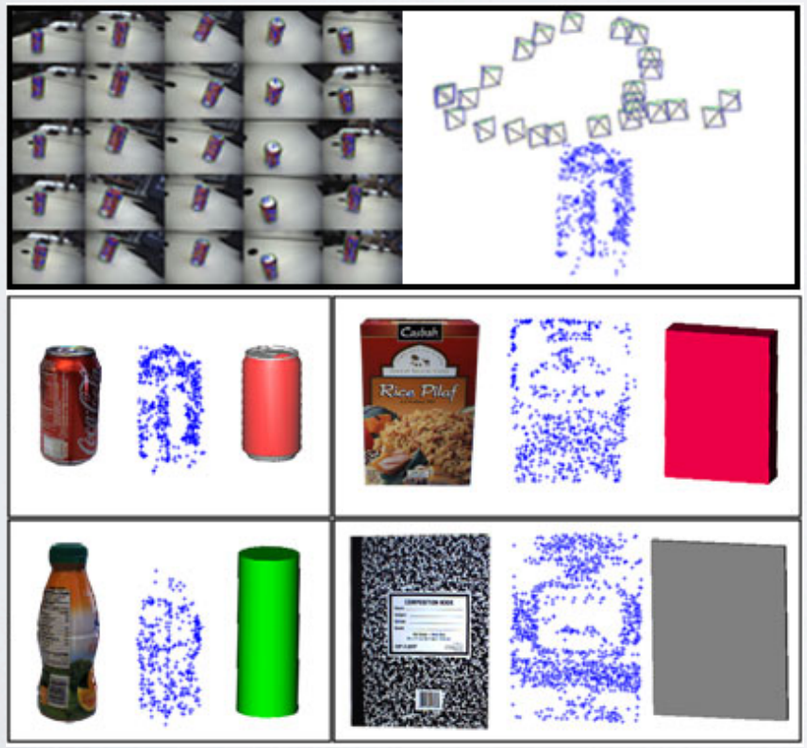
VISION-BASED OBJECT REACQUISITION

- Maintain visual descriptions of objects
- Opportunistically capture appearance of each object online
  - Maintain multiple-view model
  - Set of views capture appearance variations

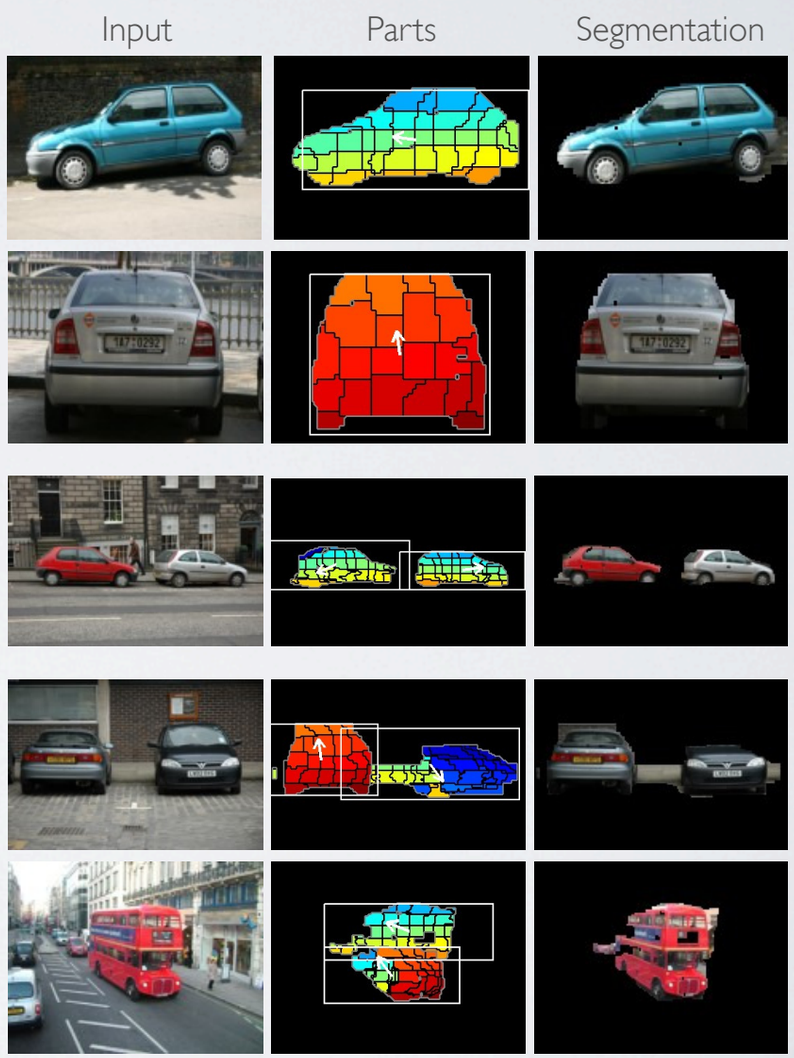


RELATED WORK

- Object category classification [Nister06, Hoiem07, Savaerese07]
  - Invariant descriptors and constituent parts
  - Recognize object *categories* rather than *instances*
  - Train offline with many samples
- Multiple-view matching [Lowe01, Gordon06, Collet09]
  - Learn model offline from controlled viewpoints

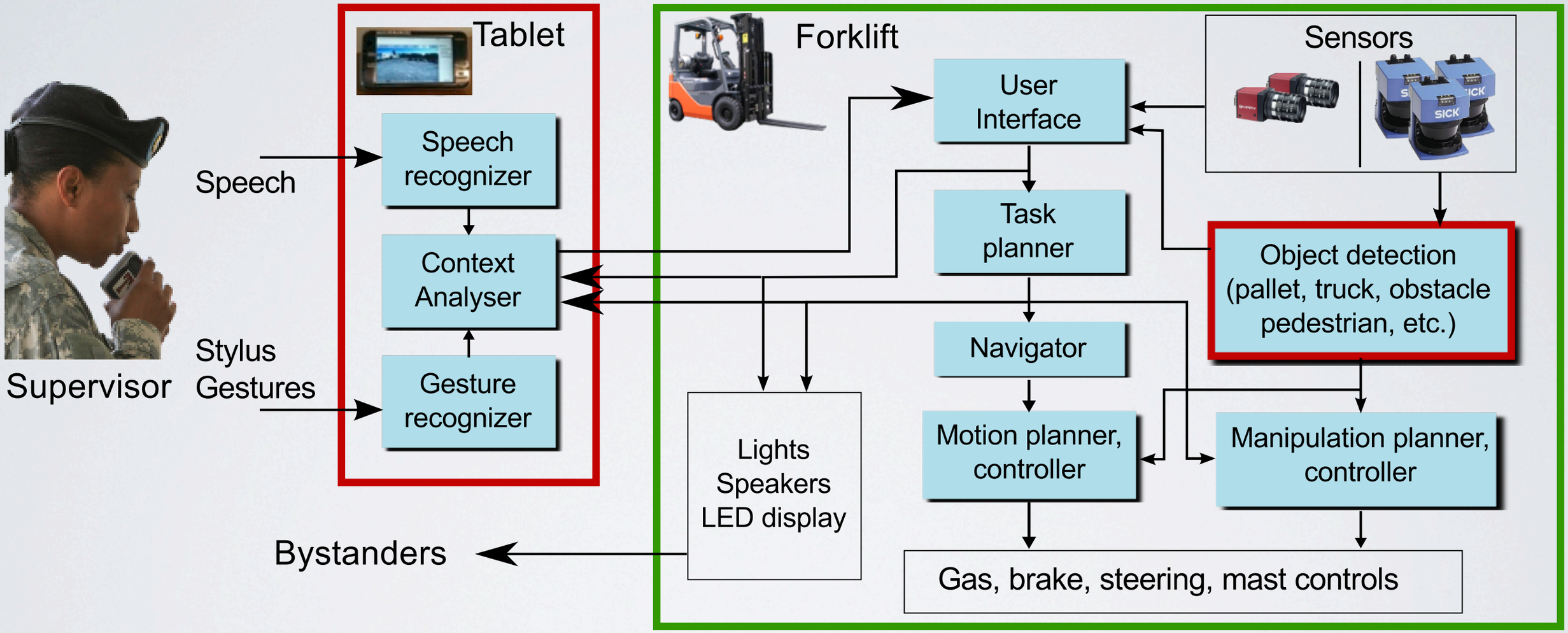


[Collet et al., ICRA 2009]

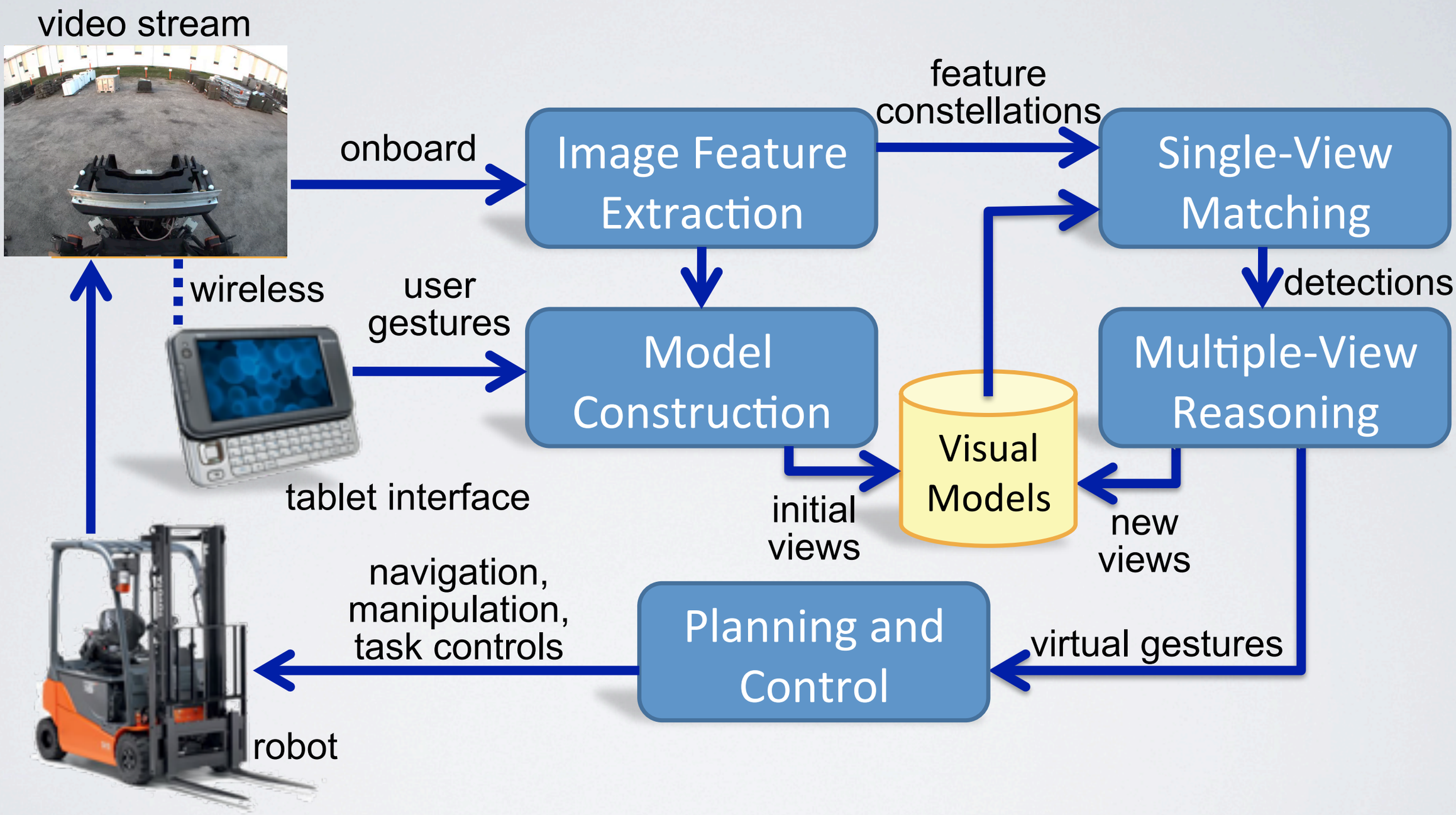


[Hoiem et al., CVPR 2007]

SYSTEM ARCHITECTURE

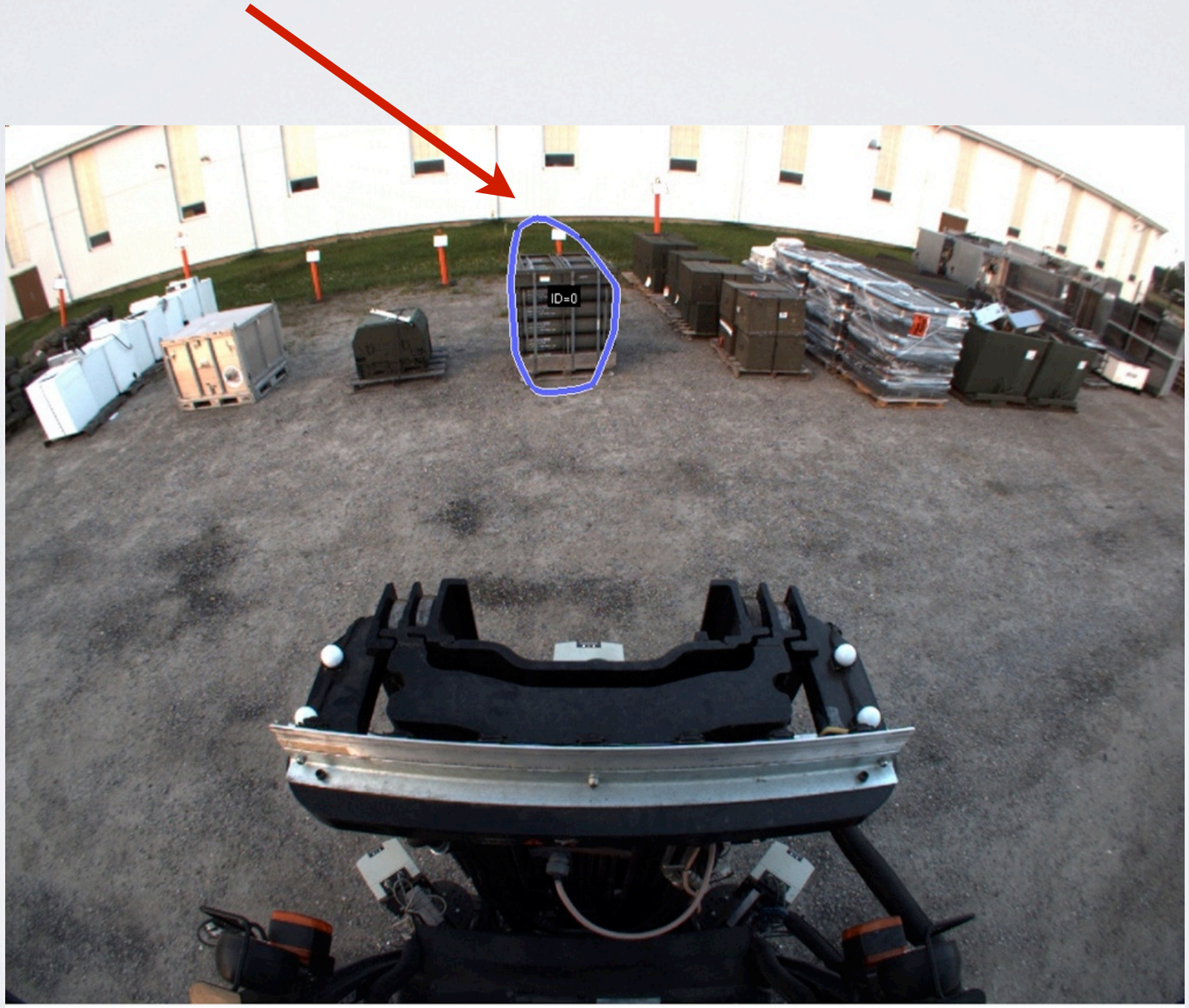


SYSTEM ARCHITECTURE



USER INITIATES VISUAL APPEARANCE MODEL

User circles object in tablet image

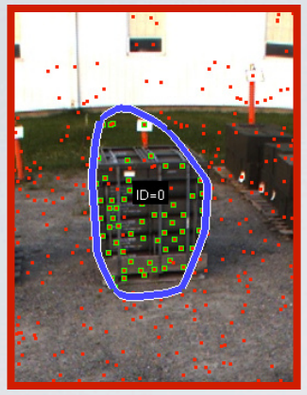


Robot's forward-facing camera image

USER INITIATES VISUAL APPEARANCE MODEL

SIFT features extracted from initial image

USER INITIATES VISUAL APPEARANCE MODEL

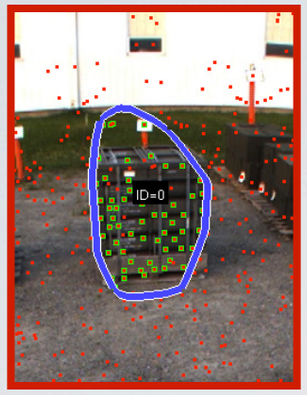


View 0 (user gesture)

SIFT features extracted from initial image

Initialize model ***M*** to contain single view ***V***

SINGLE-VIEW MATCHING

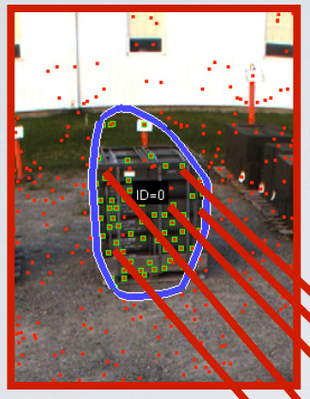


View 0 (user gesture)

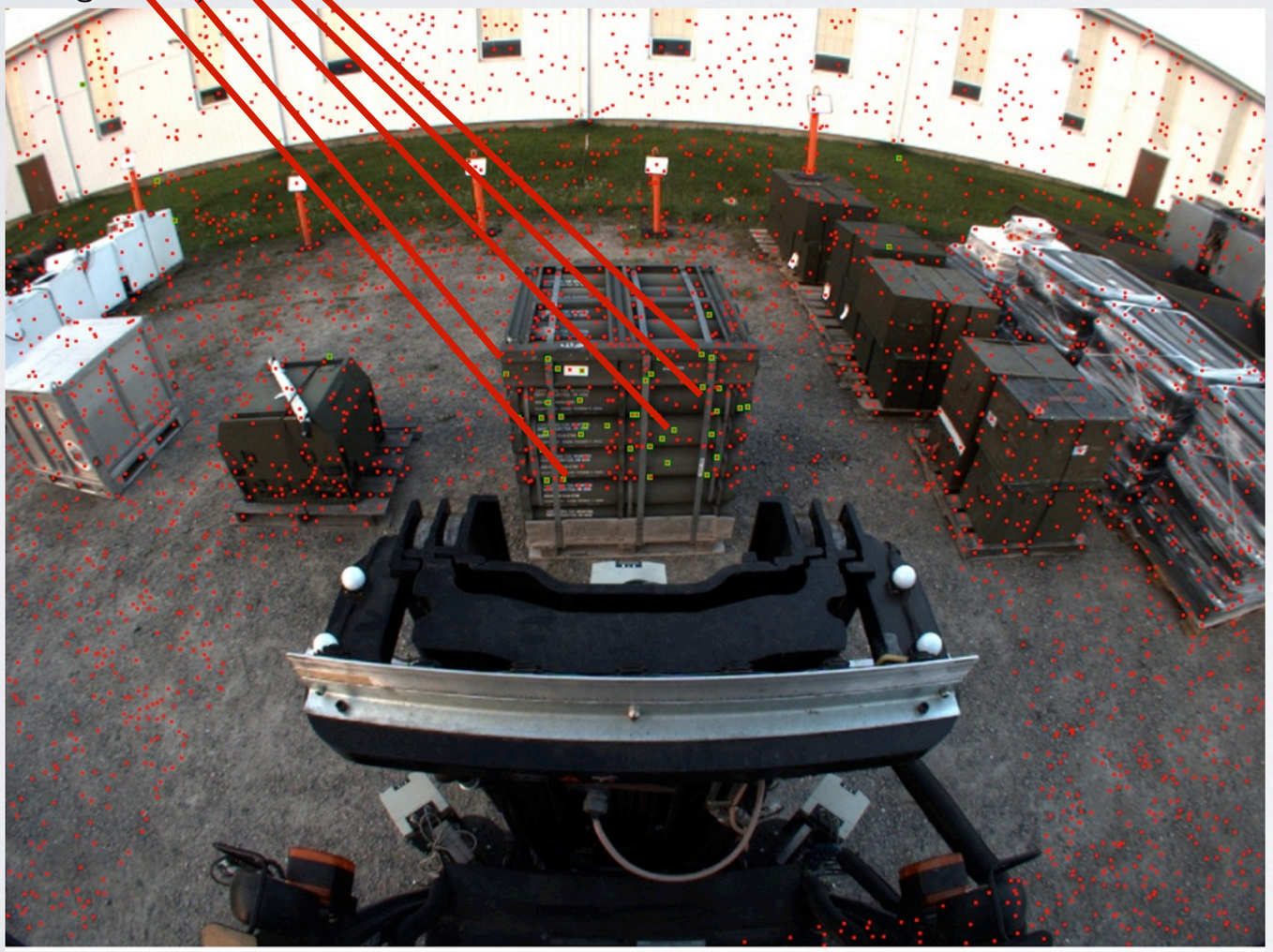
SIFT features extracted from new image

Extract features and match against all views

SINGLE-VIEW MATCHING



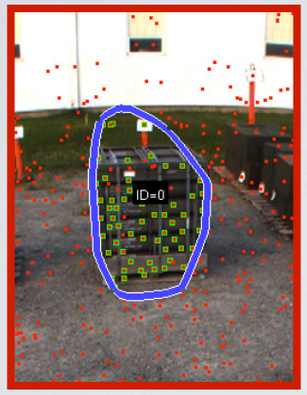
View 0 (user gesture)



SIFT features extracted from new image

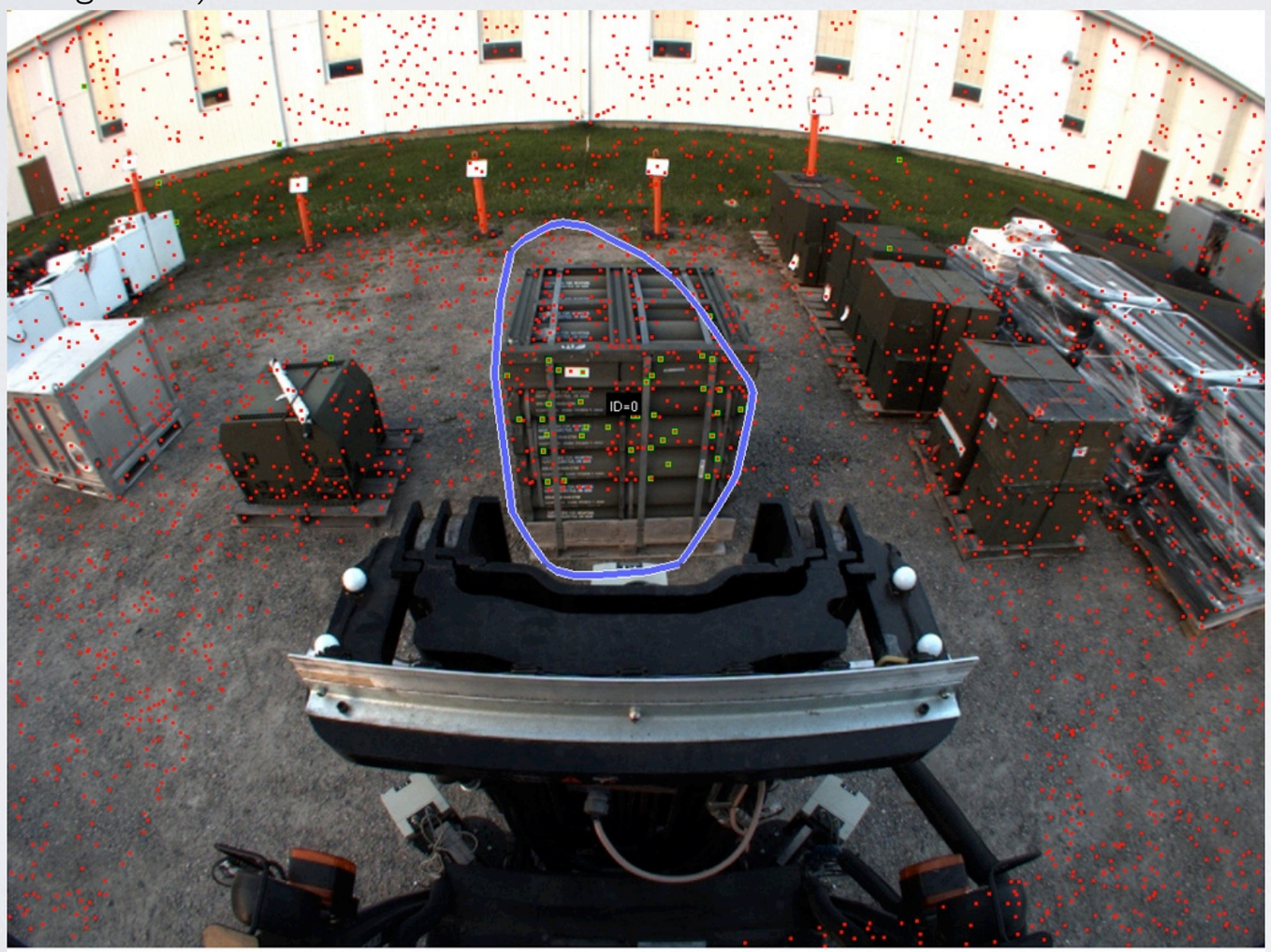
Estimate plane-projective homography for view candidate via RANSAC

MODEL AUGMENTATION



View 0 (user gesture)

View 1



SIFT features extracted from new image

Generate segmentation and add new view

MODEL AUGMENTATION

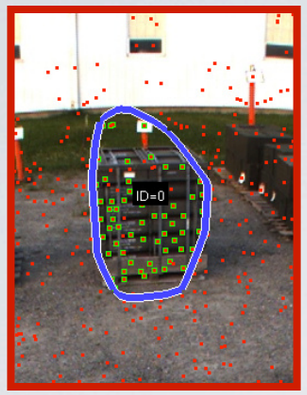
View 0 (user gesture)

View 1

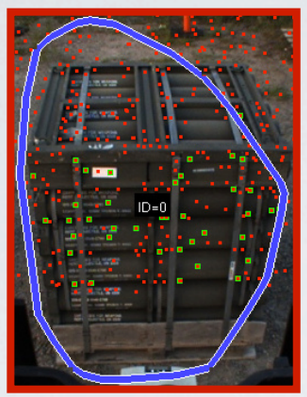
SIFT features extracted from new image

Repeat as object appearance changes

MODEL AUGMENTATION



View 0 (user gesture)



View 1

View 2

SIFT features extracted from new image

Repeat as object appearance changes

MULTIPLE-VIEW REASONING

- Multiple views form object's appearance model
  - Capture aspect and scale variability
  - Improve robustness to pose and lighting variation

View 0 (user gesture)

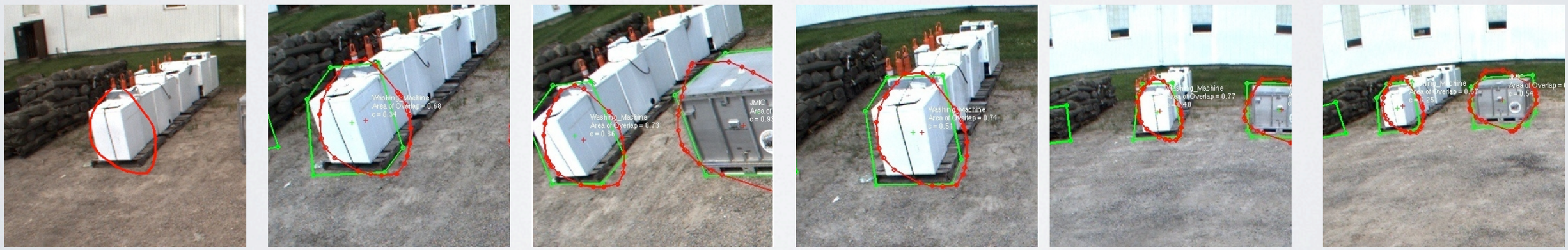
MULTIPLE-VIEW REASONING

- Multiple views form object's appearance model
  - Capture aspect and scale variability
  - Improve robustness to pose and lighting variation

Model 1

View 0 (user gesture) | View 1 | View 2 | View 3 | View 4 | View 5

Model 2

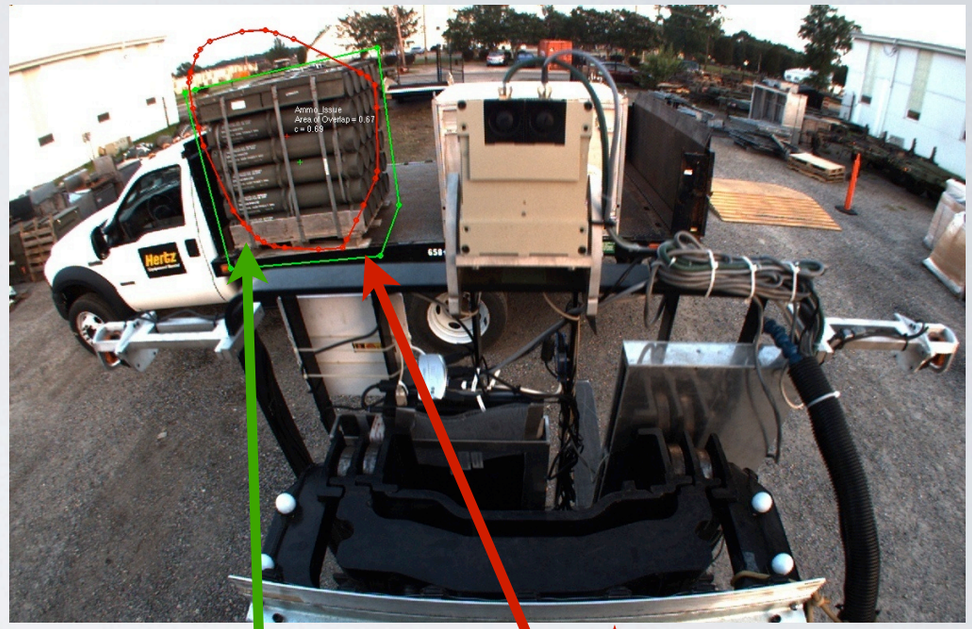


View 0 (user gesture) | View 1 | View 2 | View 3 | View 4 | View 5

MULTIPLE-VIEW REASONING

- Reasoning component resolves ambiguity
  - Score different hypotheses
  - Identify one detection per model

Confidence $\propto$ # feature matches and # RANSAC inliers

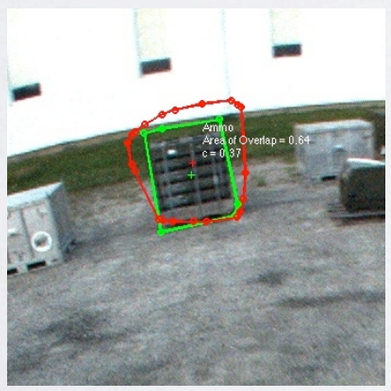


Hypothesis 1: Homography + confidence

Hypothesis 2: Homography + confidence

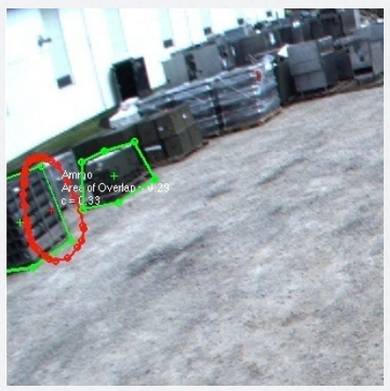
Model 1


View 0 (user gesture)

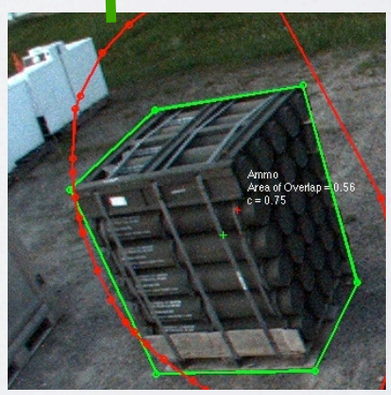


View 1

View 2



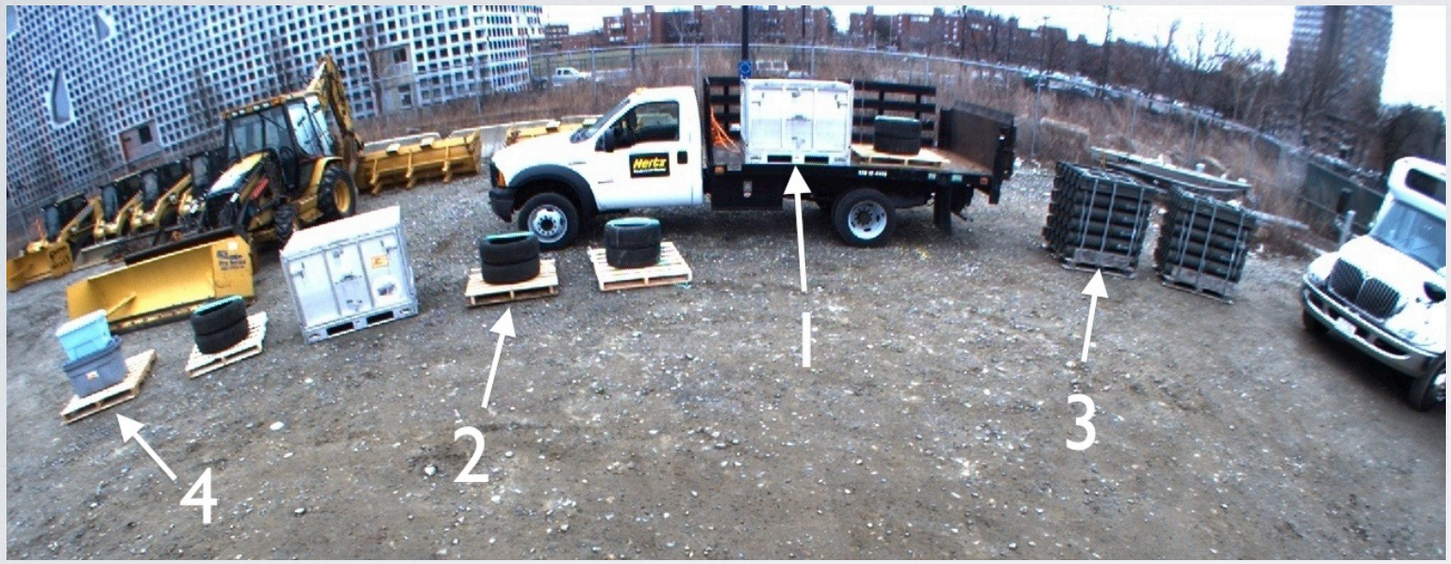
View 3

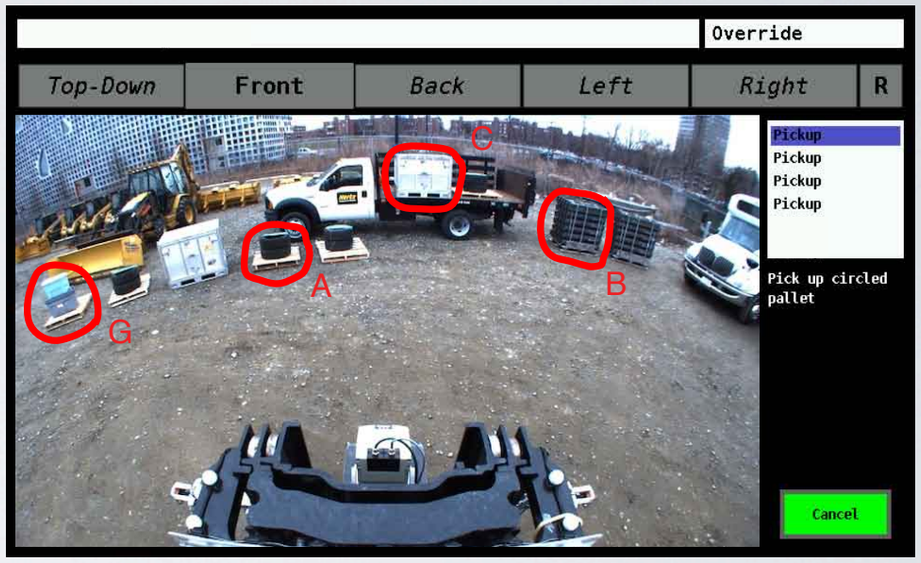


View 4

View 5

SINGLE-VIEW VS. MULTIPLE-VIEW





- Environment: Active, outdoor, gravel lot
- User gestured four objects of nine objects
- Objects transported 50m away
- Reacquisition of objects upon returning to the scene

SINGLE-VIEW VS. MULTIPLE-VIEW

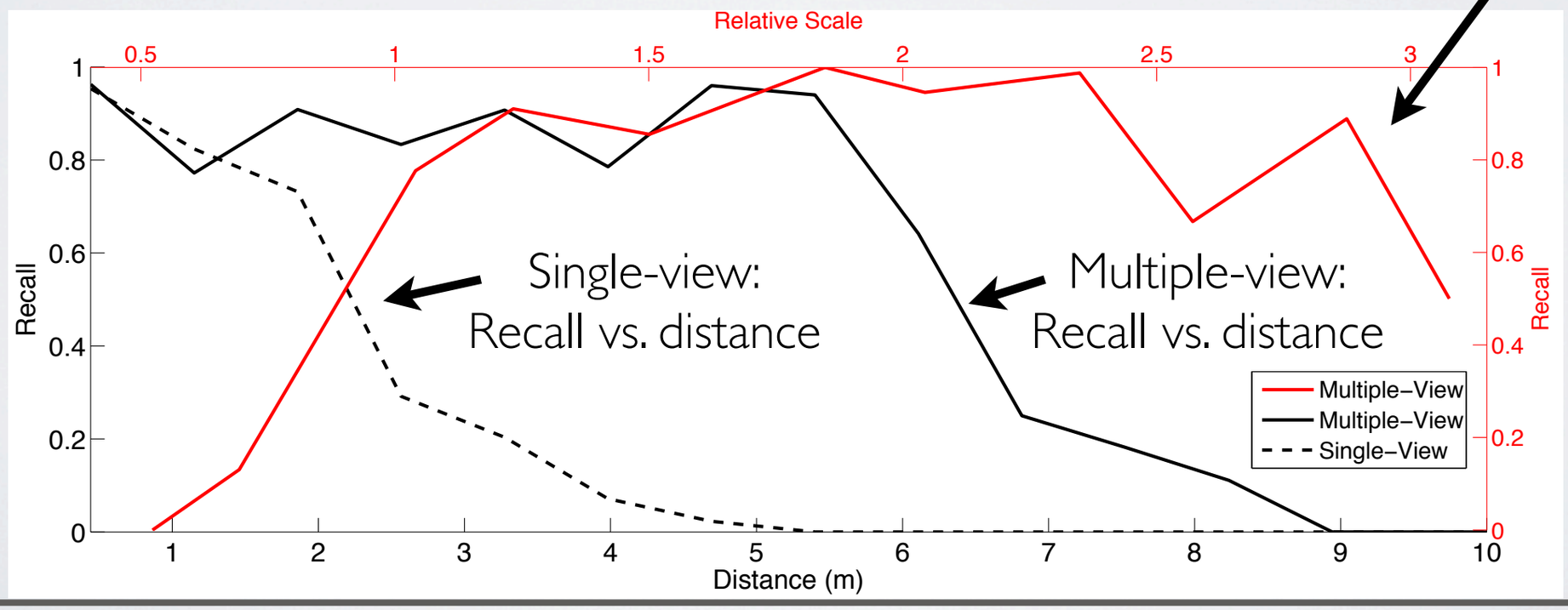
- Environment: Active, outdoor, gravel lot
- User gestured four objects of nine objects
- Objects transported 50m away
- Reacquisition of objects upon returning to the scene

SINGLE-VIEW VS. MULTIPLE-VIEW

$$\text{scale} = \sqrt{\frac{\text{current truth area}}{\text{tour truth area}}}$$

$$\text{recall} = \frac{\text{TP}}{\text{truth}}$$

Multiple-view:
Recall vs. scale



VARYING VIEWING CONDITIONS

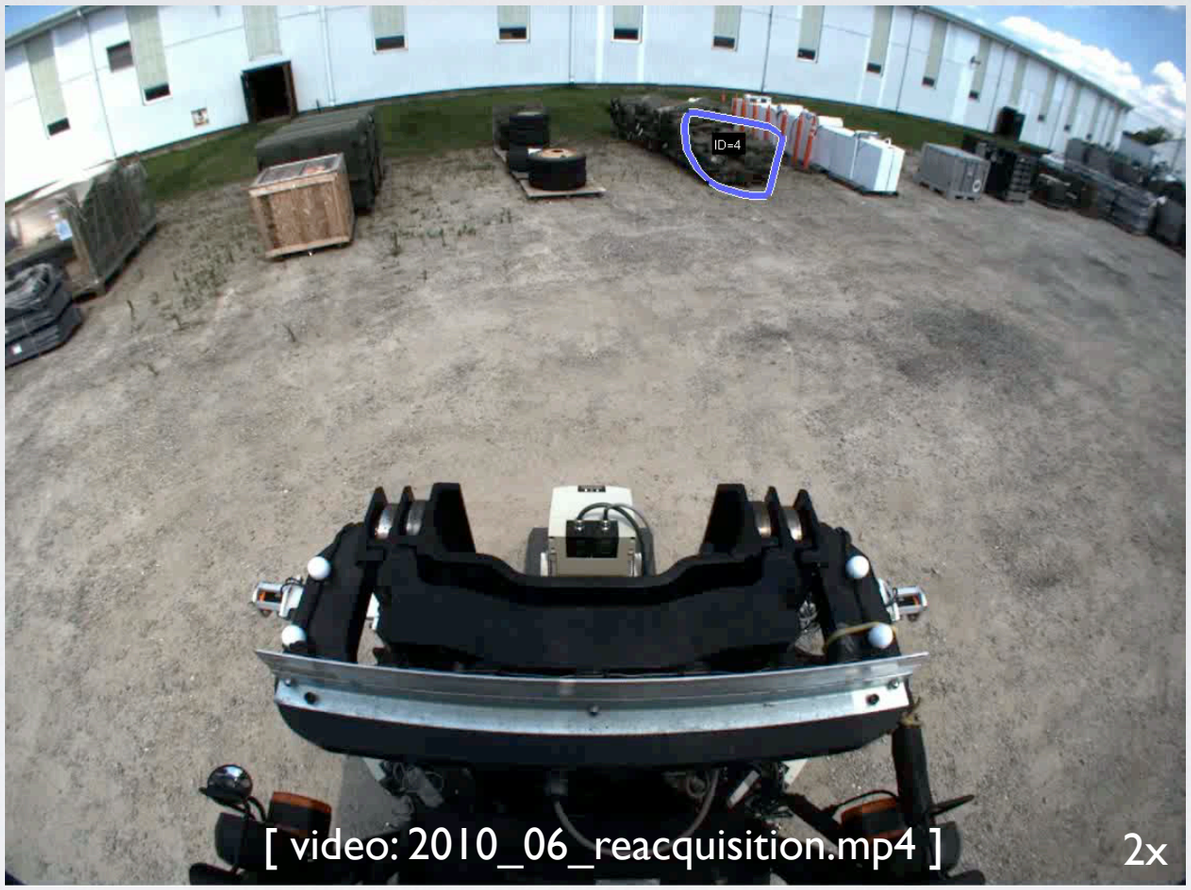
- Guided-tour, reacquisition hours/days later
- Active outdoor military warehouse
- Training and detection with different cameras
- Evaluate performance under different conditions
  - Lighting
  - Viewpoint
  - Context
  - Ambiguity

VARYING VIEWING CONDITIONS

VARYING VIEWING CONDITIONS

$$\text{recall} = \frac{\text{TP}}{\text{truth}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$



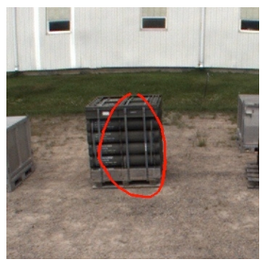
Scenario 1

Scenario	Train	Test	Delta T	Truth	Precision	Recall
1	Afternoon	Afternoon	5 min	1781	94.23%	54.13%
2	Evening	Evening	5 min	167	100.00%	94.61%
3	Morning	Evening	14 hours	165	100.00%	93.33%
4	Morning	Evening	10 hours	260	100.00%	94.53%
5	Noon	Evening	7 hours	377	100.00%	94.55%

VARYING VIEWING CONDITIONS

$$\text{recall} = \frac{\text{TP}}{\text{truth}}$$

$$\text{scale} = \sqrt{\frac{\text{current truth area}}{\text{tour truth area}}}$$



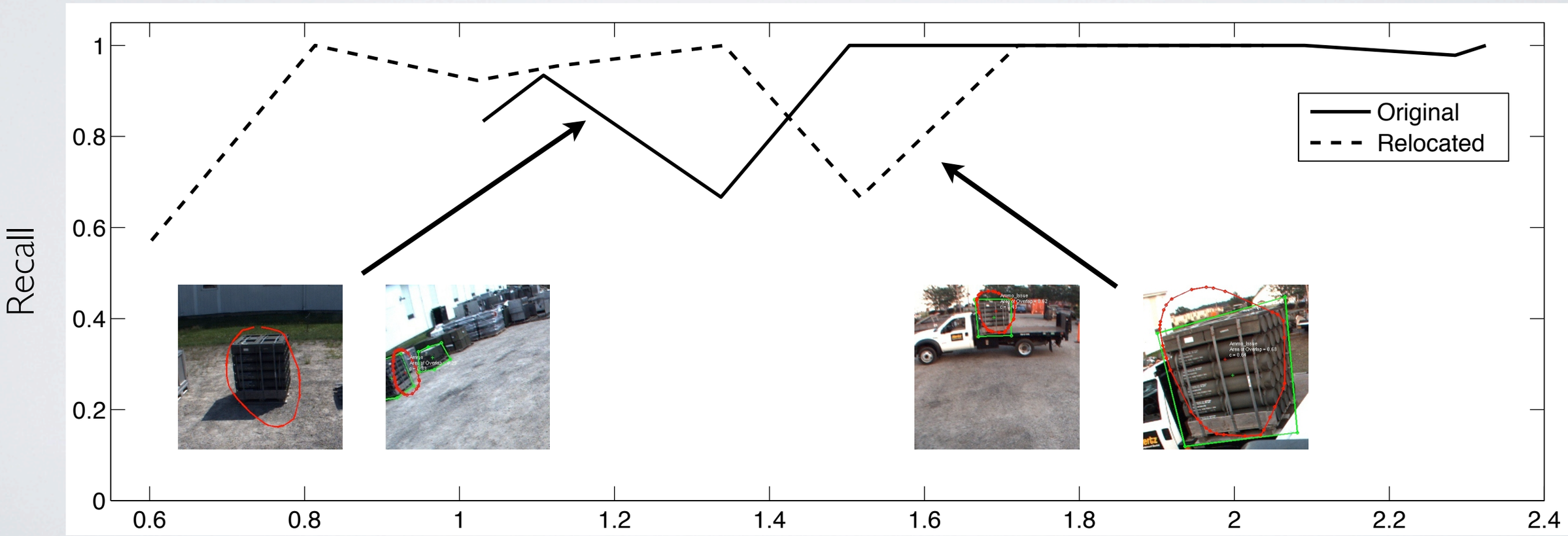
Per object, averaged over scenarios

Low contrast: Few features

VARYING VIEWING CONDITIONS

Performance under object relocation

Effect of context: Object relocation

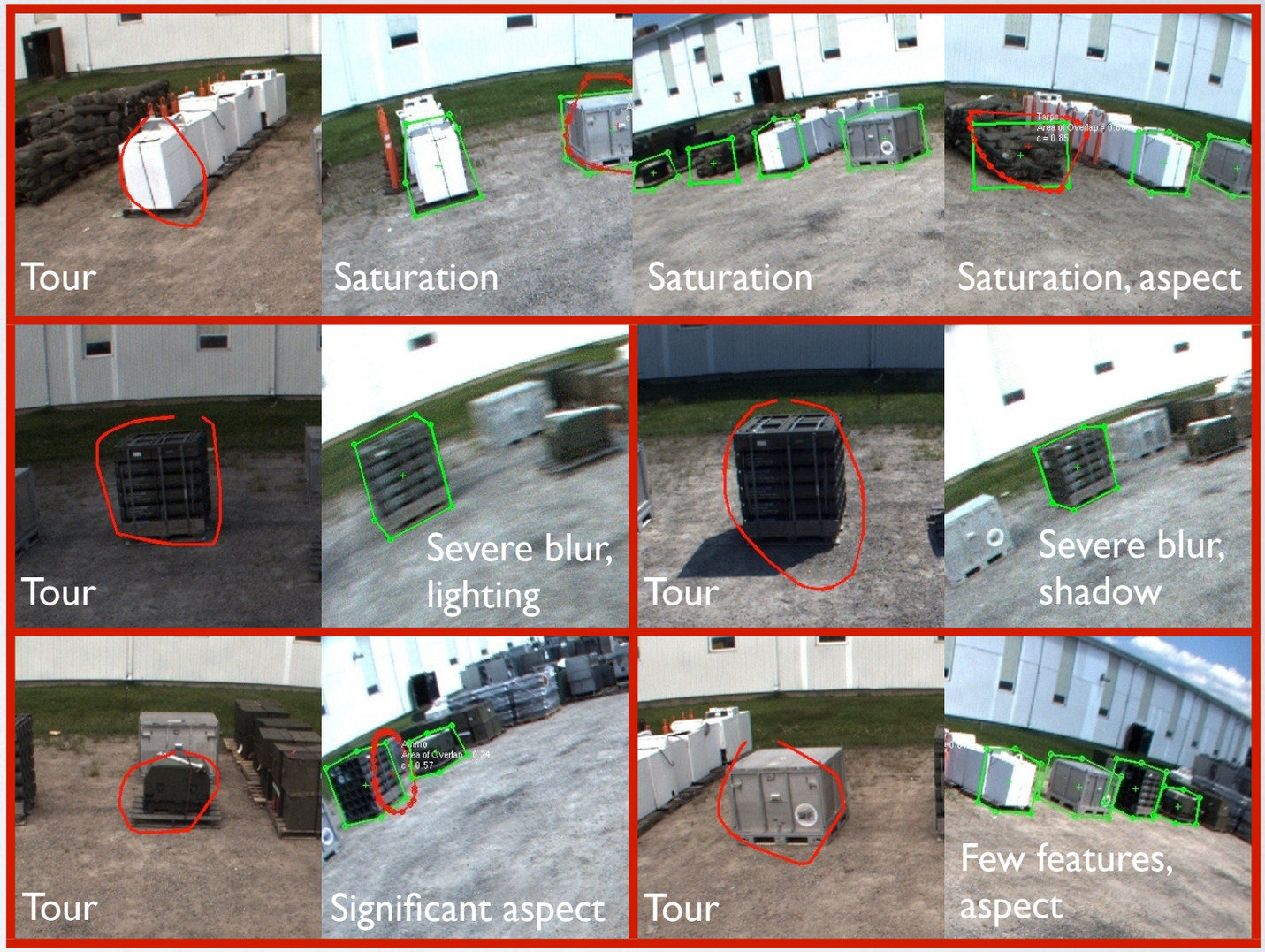


SUCCESS CASES

FAILURE CASES AND LIMITATIONS

Limited by reliability of low-level feature extraction

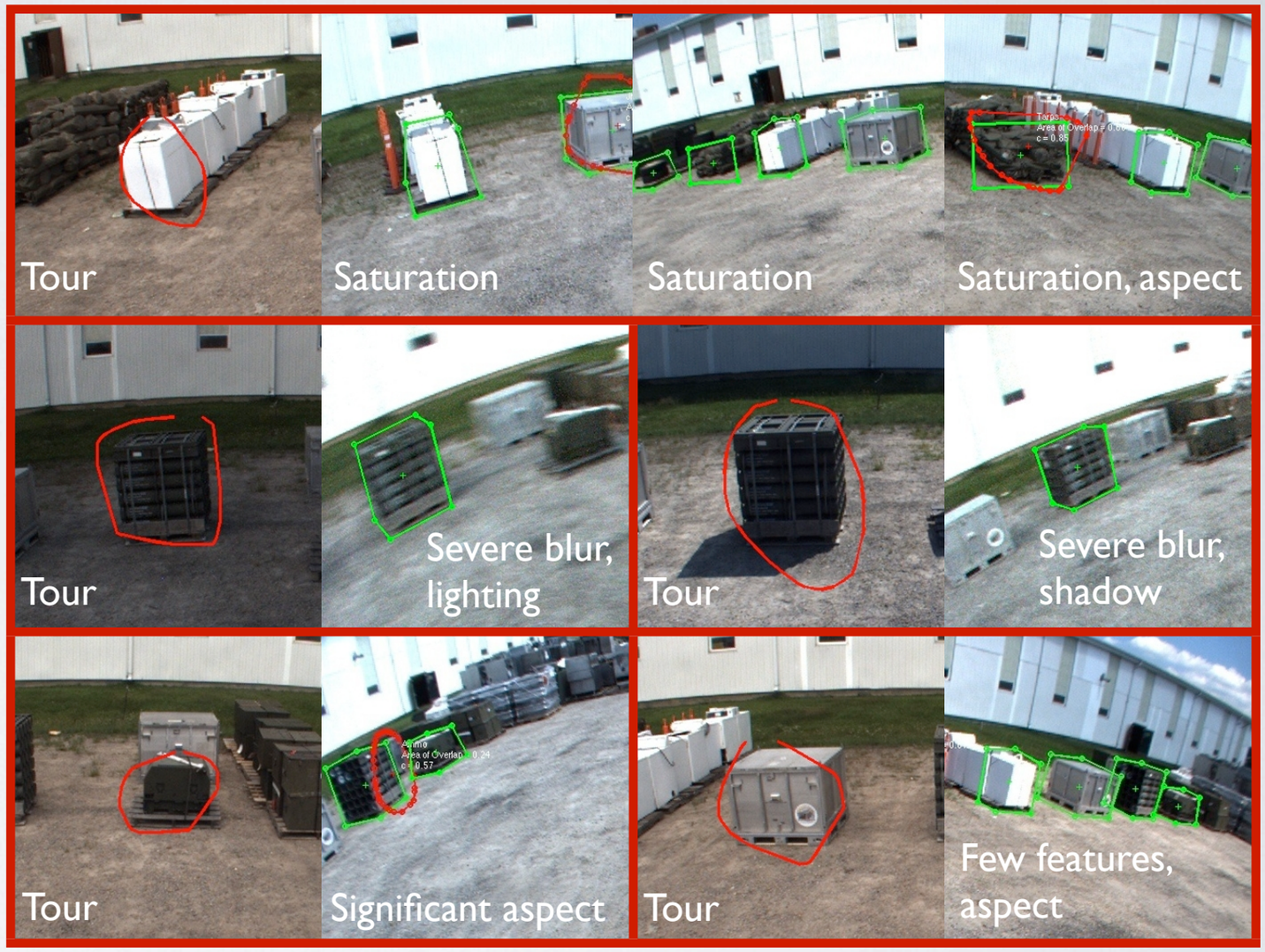
- SIFT is robust to
  - Moderate scaling
  - Global brightness change
  - In-plane rotation
- Matching degrades under
  - Parallax
  - Lens distortion
  - Specular reflection
  - Saturation (fewer features)



FAILURE CASES AND LIMITATIONS

Limitations of appearance-based matching

- Assumption that all interest points lie in a common plane
  - Homography estimation is approx.
  - Sensitive to parallax
  - In-plane rotation
- Limited computational scalability: Views are treated independently



CURRENT & FUTURE WORK

Improvements to vision-based reacquisition

- Shared feature vocabulary across views & models
- Structure-from-motion model estimation
  - Finite number of views addresses scalability
  - Full 3D model improves robustness to parallax
- Utilize LIDAR data
  - Automatic segmentation
  - Augment model with LIDAR-based descriptor

CURRENT & FUTURE WORK

Endowing greater situational awareness

- Sharing models among multiple robots
- Generalize to object category recognition
- Grounding natural spatial language to provide richer dialogue

CONCLUSIONS

- Vision-based object matching method
  - Online, appearance-based object reacquisition
  - Automatically and opportunistically maintains multiple views to increase robustness
- Shared situational awareness yields richer interaction (guided-tour)
- Analyzed performance of real-world conditions

QUESTIONS?

mwalter@csail.mit.edu

http://people.csail.mit.edu/mwalter