One-shot Visual Appearance Learning for Mobile Manipulation

Matthew R. Walter[†], Yuli Friedman[*], Matthew Antone[*], and Seth Teller[†]

[†]MIT CS & AI Lab (CSAIL), Cambridge, MA, USA
{mwalter, teller}@csail.mit.edu
[*]BAE Systems, Burlington, MA, USA
{yuli.friedman, matthew.antone}@baesystems.com

Abstract

We describe a vision-based algorithm that enables a robot to robustly detect specific objects in a scene following an initial segmentation hint from a human user. The novelty lies in the ability to "reacquire" objects over extended spatial and temporal excursions within challenging environments based upon a single training example. The primary difficulty lies in achieving an effective reacquisition capability that is robust to the effects of local clutter, lighting variation, and object relocation. We overcome these challenges through an adaptive detection algorithm that automatically generates multiple-view appearance models for each object online. As the robot navigates within the environment and the object is detected from different viewpoints, the one-shot learner opportunistically and automatically incorporates additional observations into each model. In order to overcome the effects of "drift" common to adaptive learners, the algorithm imposes simple requirements on the geometric consistency of candidate observations.

Motivating our reacquisition strategy is our work developing a mobile manipulator that interprets and autonomously performs commands conveyed by a human user. The ability to detect specific objects and reconstitute the user's segmentation hints enables the robot to be situationally aware. This situational awareness enables rich command and control mechanisms and affords natural interaction. We demonstrate one such capability that allows the human to give the robot a "guided tour" of named objects within an outdoor environment and, hours later, to direct the robot to manipulate those objects by name using natural language instructions.

We implemented our appearance-based detection strategy on our robotic manipulator as it operated over multiple days in different outdoor environments. We evaluate the algorithm's performance under challenging conditions that include scene clutter, lighting and viewpoint variation, object ambiguity, and object relocation. The results demonstrate a reacquisition capability that is effective in real-world settings.

1 Introduction

Increasingly, robots are seen, not only as machines used in isolation for factory automation, but as aides that facilitate humans through a range of daily activities. Whether an autonomous wheelchair that augments the facilities of disabled patients, a robotic forklift that manipulates and transports cargo under the direction of a human supervisor, or a humanoid that assists astronauts in dexterous manipulation, robots need to exhibit situational awareness. Robots must be able to formulate spatially extended, temporally persistent models of their surround if they are to serve as effective partners.

This paper describes a vision-based algorithm that enables robots to automatically model the adaptive appearance of *a priori* unknown objects in their surround based upon a single user-provided segmentation cue. The robot can then call upon these models to reacquire the objects after extended excursions in both space and time. As we describe, this situational awareness capability enables more effective interaction between robots and humans, allowing humans to efficiently convey object- and task-level information to the robot. The challenge is to achieve the level of recall that is necessary to reacquire objects under conditions typical of unprepared, dynamic environments. In order to reliably detect objects with minimal effort on the part of the user, the algorithm automatically builds and maintains a model that offers robustness to the effects of clutter, occlusions, and variations in illumination and viewpoint.

More specifically, we present a one-shot learning algorithm that automatically maintains an adaptive visual appearance model for each user-indicated object in the environment, enabling object recognition from a usefully wide range of viewpoints. The user provides a manual segmentation of an object by circling it in an image from a camera mounted on the robot. The primary novelty of our method lies in the automatic generation of multiple-view, feature-based object models that capture variations in appearance that result from changes in scale, viewpoint, and illumination. This automatic and opportunistic modeling of an object's appearance enables the robot to recognize the presence of the object in an image and thereby reconstitute the user's segmentation hints and task information, even for viewpoints that are spatially and temporally distant from those of the original gesture. As we show, this ability allows for more effective, scalable command capabilities. We describe a scenario in which the user gives a mobile manipulator a guided tour of objects in an outdoor environment and, at a later time, directs the robot to reacquire and manipulate these objects by speaking their name.

Our work is motivated by the development of a mobile manipulation platform that is designed to manipulate and transport objects in an unprepared, outdoor environment under minimal direction from a human supervisor [Teller et al., 2010]. As a demonstration of the effectiveness of the algorithm in this scenario, we present an analysis of the performance of the reacquisition algorithm under a variety of conditions typical of outdoor operation including lighting and viewpoint variations, scene clutter, and unobserved object relocation. We describe conditions for which the method is well-suited as well as those for which it fails. In light of these limitations, we conclude with a discussion on directions for future work.

1.1 Related Work

An extensive body of literature on visual object recognition has been developed over the past decade, with much of the recent work focusing on methods that are robust to challenges like illumination and viewpoint variation, scene clutter, and partial occlusion. Generalized algorithms are typically trained to identify abstract object categories and delineate instances in new images using a set of exemplars that span the most common dimensions of variation, including 3D pose, illumination, and background clutter. Training samples are further diversified by variations in the instances themselves, such as shape, size, articulation, and color. The current state-of-the-art involves learning relationships among constituent object parts represented using view-invariant descriptors. The notable work in this area includes that of Savarese and Fei-Fei [2007], Hoiem et al. [2007], Liebelt et al. [2008]. Rather than *recognition* of generic categories, however, the goal of our work is the *reacquisition* of specific previously observed objects. We therefore still require invariance to camera pose and lighting variations, but not to intrinsic within-class variability.

Lowe [2001] introduces the notion of collecting multiple image views to represent a single 3D object, relying on SIFT feature correspondences to recognize new views and to decide when the model should be augmented. Gordon and Lowe [2006] describe a more structured technique for object matching and pose estimation that uses bundle adjustment to explicitly build a 3D model from multiple uncalibrated views. They also utilize SIFT correspondences for recognition but further estimate the relative camera pose via RANSAC and Levenberg-Marquardt optimization. Collet et al. [2009] extend this work by incorporating Mean-Shift clustering to facilitate registration of multiple instances during recognition. They demonstrate high precision and recall with accurate pose in cluttered scenes amid partial occlusions, changes in view distance and rotation, and varying illumination. As with our work, their approach is motivated by the importance of object recognition for robotic manipulation. While similar in this respect, Collet et al.'s work along with the other techniques differ fundamentally from ours in their reliance upon an extensive initial training phase. During this step, these methods build object representations offline in a "brute-force" manner by explicitly acquiring views from a broad range of different aspects. In contrast, we describe a one-shot learning algorithm that opportunistically builds object models online while the robot operates.

With respect to the goal of detecting the presence of specific objects within a series of images, our work has similar objectives to that of visual tracking. In visual tracking, an object is manually designated or automatically detected based on appearance or motion characteristics, and its state is subsequently tracked over time using visual and kinematic cues [Yilmaz et al., 2006]. General tracking approaches assume small temporal separation with limited occlusions or visibility loss, and therefore slow visual variation, between consecutive observations [Comaniciu et al., 2003]. These trackers tend to perform well over short time periods but are prone to failure when an object's appearance changes or it disappears from the camera's field of view. To address these limitations, "tracking-by-detection" algorithms adaptively model variations in appearance online based upon positive detections [Lim et al., 2004; Collins et al., 2005] that differentiate the object from the background. These self-learning methods extend the duration for which an object can be tracked, but they tend to "drift" as they adapt to incorporate the appearance of occluding objects or the background when the object disappears from view. Grabner et al. [2008] alleviate the drifting that stems from using labeled self-predictions for training by proposing a semi-supervised approach that incorporates a trained prior with a model trained online based upon unlabeled samples. Kalal et al. [2010] similarly utilize unlabeled examples as part of the online training, but propose the additional use of structure that exists between these examples to improve the track-by-detect accuracy. Alternatively, Babenko et al. [2009] describe a solution that utilizes multiple-instance learning rather than fixed positively or negatively labeled examples for online training. Like similar approaches, these algorithms improve robustness and thereby allow an object to be tracked over longer periods of time despite partial occlusions and frame cuts. However, they are still limited to relatively short, contiguous video sequences. Although we use video sequences as input, our approach does not rely on a temporal sequence and is therefore not truly an object "tracker"; instead, its goal is to identify designated objects over potentially disparate views.

More closely related to our reacquisition strategy is the recent work by Kalal et al. [2009] that combines an adaptive tracker with an online detector in an effort to improve robustness to appearance variation and frame cuts. Given a single user-provided segmentation of each object, their Tracking-Modeling-Detection algorithm utilizes the output of a short-term tracker to maintain a model of each object that consists of suitable image patch features. Meanwhile, they employ this model to learn an online detector that provides an alternative hypothesis for an object's position, which is used to detect and reinitialize tracking failures. The algorithm maintains each model through a growing phase in which tracked features that are deemed appropriate are added and a pruning phase that removes features based upon the output of the detector. Growing the model allows the algorithm to adapt to appearance variations while the pruning step alleviates the effects of drift. While we do not rely upon a tracker, we take a similar approach of learning an object detector based upon a single supervised example by building an image-space appearance model online. Unlike Kalal et al.'s solution, however, we impose geometric constraints to validate additions to the model, which reduces the need to prune the model of erroneous appearance representations.

It is worth noting research in object category recognition that similarly seeks to learn to detect an object's class based upon as few as one training example. Underlying these approaches is the view that certain knowledge generalizes across object classes and that information useful in describing one object category is also helpful in representing another. Transfer learning provides a formal mechanism for bootstrapping a detector for a new object class from existing object class learners. Fei-Fei et al. [2003] utilize a constellation model [Fergus et al., 2003] to represent object categories as a collection of parts, each described in terms of appearance and spatial position. Having learned the distribution over model parameters for a set of objects, they treat this distribution as a prior over parameters for a new object class. This prior provides a means for transferring knowledge from the set of learned models. They then use variational Bayesian methods to learn the posterior over shape and appearance for this new class based upon a small set of training examples. Alternatively, Stark et al. [2009] propose a parts-based representation similar to a constellation model to describe object classes entirely in terms of shape information, which they argue is well suited to transfer to new models. This representation has the advantage that it can be factored on a per-part basis. Given a new object class, the authors demonstrate that this factorization enables them to more efficiently learn the corresponding model from those of known object categories with few training examples. The factorized representation has the added benefit that it allows partial knowledge to be transferred from a set of known object models to a new object class, which supports sharing shape information between possibly disparate object classes.

Meanwhile, considerable effort has been devoted to utilize vision-based recognition and tracking to facilitate human-robot interaction. While much of this work focuses on person detection, various techniques exist for learning and recognizing inanimate objects in the robot's surround. Of particular relevance are those in which a human partner "teaches" the objects to the robot, typically by pointing to a particular object and using speech to convey object-specific information (e.g., color, name) and tasks [Haasch et al., 2005; Breazeal et al., 2004]. Our work similarly enables human participants to teach objects to the robot, using speech as a means of conveying information. However, in our case, the user identifies objects by indicating their location within images of the scene. Additionally, the aforementioned research is limited, at least in implementation, to uncluttered indoor scenes with a small number of objects, whereas we focus on reacquisition in outdoor, semi-structured environments.

2 Reacquisition Methodology

This section presents an overview of our approach to object reacquisition in the context of autonomous mobile manipulation. We introduce the overall system and describe how the incorporation of object reacquisition improves the effectiveness and competency of the robot.

2.1 A Human-Commandable Robot for Mobile Manipulation

Our object reacquisition strategy is motivated by our ongoing development of a robotic forklift (Figure 1) that autonomously manipulates cargo within an outdoor environment under the high-level direction of a human supervisor [Teller et al., 2010]. The user conveys *task-level* commands to the robot that include picking up, transporting, and placing desired palletized cargo from and to truck beds and ground locations. The system supports the user's ability to command the robot with speech and simple pen-based gestures through a handheld tablet interface [Correa et al., 2010].

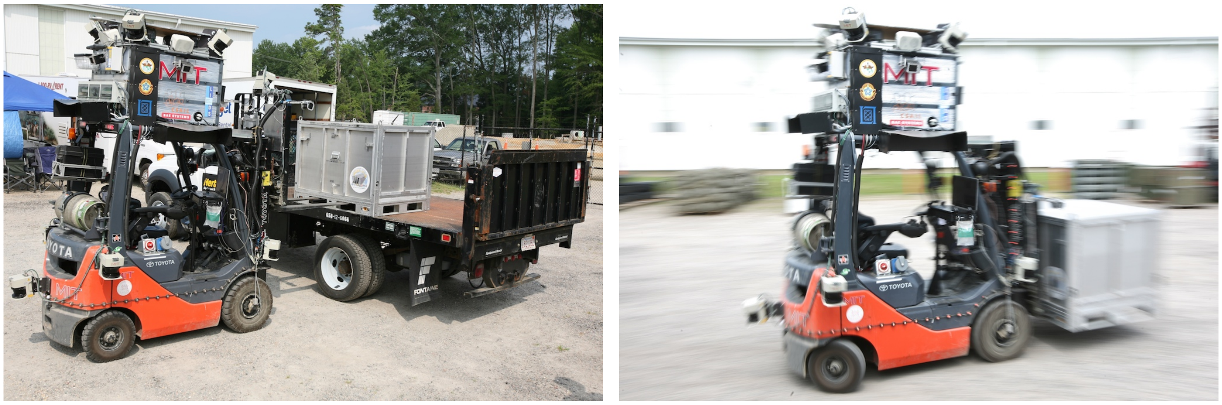
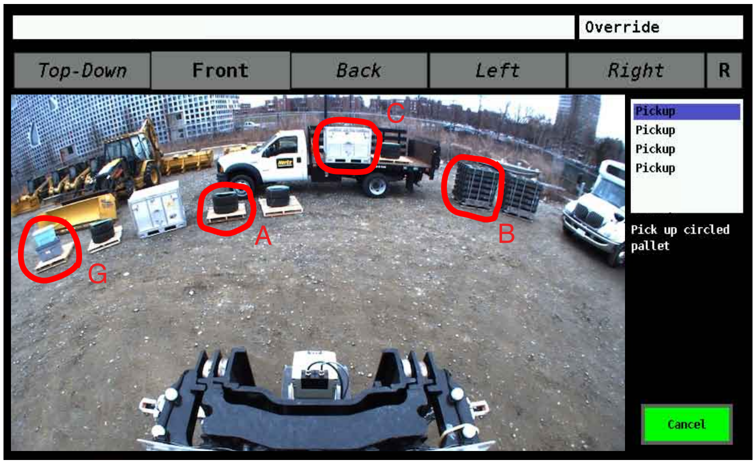


Figure 1: The robotic forklift is capable of autonomously manipulating and transporting cargo based upon image-space segmentation and voice commands.

For example, the user can identify a specific pallet to pick up or a desired destination by circling the appropriate location within a robot-mounted camera image shown on the tablet (Figure 2). The user can also summon the robot to one of several named locations in the environment by speaking through the tablet.



Pick up circled pallet

Figure 2: The tablet interface displaying a view from the robot's forward-facing camera along with user gestures (red). Depending on the accompanying speech, the gestures may indicate objects to manipulate or they may serve as segmentations of named objects as part of a guided-tour.

The platform (Figure 1) is equipped with four cameras that face forward, backward, to the left, and to the right, several LIDARs, encoders mounted to the two front wheels, an IMU, and a GPS. The system uses GPS for coarse localization (3 m–5 m) and otherwise relies upon dead reckoning for navigation, obstacle detection, and pallet detection and manipulation. For more information, the reader is referred to the paper by Teller et al. [2010]. As Walter et al. [2010c] discuss in more detail, the system takes as input a single image-space segmentation of a pallet and is capable of autonomously detecting, tracking, and manipulating cargo. After the user circles the desired object in the image, the system infers the corresponding volume of interest in the robot's surroundings based upon known camera calibration. The robot then actively sweeps this volume using a single planar LIDAR mounted with its scan plane parallel to the plane formed by the two lifting forks.

Scans culled from the volume of interest are fed into a classification and estimation framework that detects pallet structure and tracks its pose as the robot proceeds with manipulation. In our tests [Walter et al., 2010c], this system successfully detected and engaged pallets at a rate of 95%.

2.2 Increased Autonomy Through Object Reacquisition

The system performs the manipulation and navigation tasks autonomously with little effort required on the part of the user, for example, finding and safely engaging the pallet based solely upon a single gesture. Nevertheless, extended tasks such as moving multiple objects previously required that the user specify each object in turn, thus necessitating periodic albeit short intervention throughout. By introducing the ability to reacquire objects of interest in the environment, however, our system allows the user to command the manipulation of multiple objects in rapid succession at the outset, after which the system effectively reconstitutes each gesture and manipulates the corresponding pallet. Thus, the system is able to utilize valuable, yet unobtrusive, segmentation cues from the user to execute extended-duration commands [Walter et al., 2010a].

More generally, a robot's ability to recognize specific objects over extended periods of time enables rich, efficient command mechanisms, in essence by allowing the user to share their world model with the robot. For example, we describe a guided tour scenario in which the user teaches specific objects in the environment to the robot by circling their position in an image from one camera and speaking their label [Walter et al., 2010b]. The system then builds and maintains an appearance model for the named object that is shared across cameras. The system also binds to each object an approximate position (3 m–5 m precision) and orientation suitable for manipulation based upon the estimated absolute pose of the robot. At a later point in time, the user can command the robot to manipulate an object simply by referring to it by name (e.g., "bot, pick up the tire pallet"). Tellex et al. [2011] utilize this shared object model as part of a novel language grounding algorithm that is capable of understanding much richer natural language instructions. The primary technical challenge for object detection is to achieve a reacquisition capability that operates across sensors and that is sufficiently robust to local clutter and appearance variation to be useful in practice. We show that the incorporation of opportunistically captured multiple views provides robustness to viewpoint and lighting variations.

Through our previous work [Teller et al., 2010; Walter et al., 2010c], we have developed the underlying system's ability to interpret high-level commands that dictate navigation and manipulation, and to efficiently use segmentation cues provided by the human operator. The method that we describe in this paper builds upon these capabilities by providing a means of automatically generating these valuable segmentation cues based upon a learned model of a specific object's appearance. The concept is to employ this capability to help close the loop for manipulation by first directing the robot to the object's approximate location and then performing segmentation. As with the original human-in-the-loop system, this segmentation yields the volume of interest on which the pallet detection and tracking algorithm operates.

Our proposed reacquisition system (Figure 3) relies on a synergy between the human operator and the robot, with the human providing initial visual cues (thus easing the task of automated object detection and segmentation) and the robot maintaining persistent detection of the indicated objects upon each revisit, even after sensor coverage gaps (thus alleviating the degree of interaction and attention that the human need provide).

Our algorithm maintains visual appearance models of the objects that the user indicates. When the robot returns to the scene, it can still recall, recognize, and act upon the object even when errors and drift in its navigation degrade the precision of its pose estimates. In fact, the algorithm utilizes dead-reckoned pose estimates only to suggest the creation of new appearance models; it

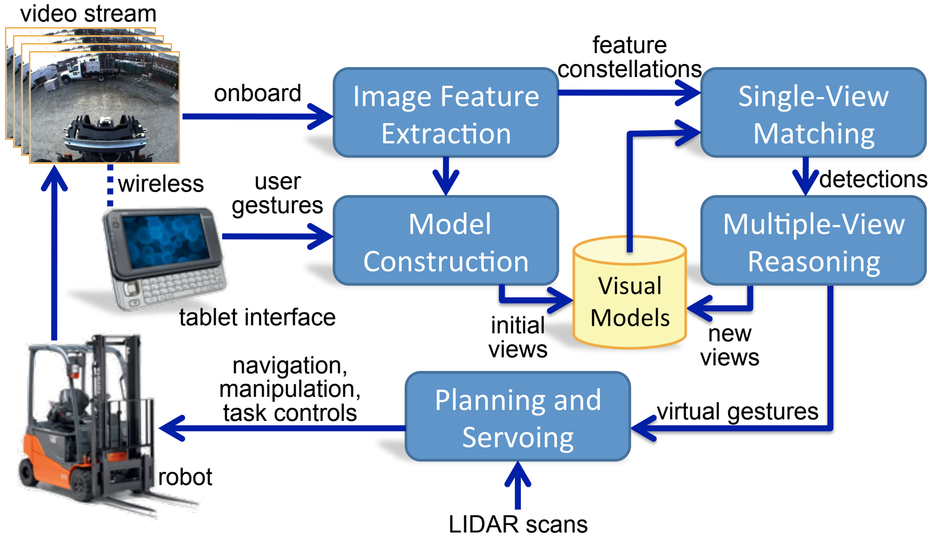


Figure 3: Block diagram of the reacquisition process.

uses neither pose information nor data from non-camera sensors for object recognition. The specific guided tour scenario, however, relies upon the ability to determine the robot's absolute position and heading to within 3 m–5 m and 20 degrees, respectively. These requirements result from the LIDAR-based pallet detection and servoing capabilities, which are most effective when the robot is within this region about the pallet.

3 Visual Appearance for Object Reacquisition

Our algorithm for maintaining persistent identity of user-designated objects in the scene is based on creating and updating appearance models that evolve over time. We define a *model* $\mathcal{M}_i$ as the visual representation of a particular object i, which consists of a collection of views, $\mathcal{M}_i = \{v_{ij}\}$. We define a *view* v_{ij} as the appearance of a given object at a single viewpoint and time instant j (i.e., as observed by a camera with a particular pose at a particular moment).

The method constructs object appearance models and their constituent views from 2D constellations of keypoints, where each keypoint consists of an image pixel position and a descriptor that characterizes the local intensity pattern. Our algorithm searches new camera images for each model and produces a list of visibility hypotheses based on visual similarity and geometric consistency of keypoint constellations. New views are automatically added over time as the robot moves; thus the collection of views opportunistically captures variations in object appearance due to changes in viewpoint and illumination.

3.1 Model Initiation

The algorithm processes each image as it is acquired to detect a set $\mathcal{F}$ of keypoint locations and scale invariant descriptors. We use Lowe's SIFT algorithm for moderate robustness to viewpoint and lighting changes [Lowe, 2004], but any stable image features may be used. In our application, the user initiates the generation of the first appearance model with a gesture that segments its location in a particular image. Our system creates a new model $\mathcal{M}_i$ for each indicated object. The set of SIFT keypoints that fall within the gesture in that particular frame form the new model's first view v_{i1}.

In addition to a feature constellation, each view contains the timestamp of its corresponding image, a unique identifier for the camera that acquired the image, the user's 2D gesture polygon, and the 6-DOF inertial pose estimate of the robot body.

Algorithm 1 Single-View Matching

```
Input: A model view $v_{ij}$ and camera frame $\mathcal{I}_t$
Output: $\mathcal{D}_{ijt} = \left(H^\star_{ij}, c^\star_{ij}\right)$
$\mathcal{F}_t = \{(x_p, f_p)\} \leftarrow \texttt{SIFT}(\mathcal{I}_t)$;
$\mathcal{C}_{ijt} = \{(s_p, s_q)\} \leftarrow \texttt{FeatureMatch}(\mathcal{F}_t, \mathcal{F}_{ij})\ s_p \in \mathcal{F}_t, s_q \in \mathcal{F}_{ij}$;
$\forall\, x_p \in \mathcal{C}_{ijt},\ x_p \leftarrow \texttt{UnDistort}(x_p)$;
$\mathcal{H}^\star_{ijt} = \{H^\star_{ijt}, d^\star_{ijt}, \tilde{\mathcal{C}}^\star_{ijt}\} \leftarrow \{\}$;
for $n = 1$ to $N$ do
    Randomly select $\hat{\mathcal{C}}_{ijt} \in \mathcal{C}_{ijt}$, $|\hat{\mathcal{C}}_{ijt}| = 4$;
    Compute homography $\hat{H}$ from $(x_p, x_q)$ in $\hat{\mathcal{C}}_{ijt}$;
    $\mathcal{P} \leftarrow \{\}, \hat{d} \leftarrow 0$;
    for $(x_p, x_q) \in \mathcal{C}_{ijt}$ do
        $\hat{x}_p \leftarrow \hat{H} x_p$;
        $\hat{x}_p \leftarrow \texttt{Distort}(\hat{x}_p)$;
        if $d_{pq} = |x_q - \hat{x}_p| \le t_d$ then
            $\mathcal{P} \leftarrow \mathcal{P} + (x_p, x_q)$;
        $\hat{d} \leftarrow \hat{d} + d_{pq}$;
    if $\hat{d} < d^\star_{ij}$ then
        $\mathcal{H}^\star_{ijt} \leftarrow \{\hat{H}, \hat{d}, \mathcal{P}\}$;
$c^\star_{ijt} = |\tilde{\mathcal{C}}^\star_{ijt}|/(|v_{ij}| \min(\alpha|\tilde{\mathcal{C}}^\star_{ijt}|, 1)$
if $c^\star_{ijt} \ge t_c$ then
    $\mathcal{D}_{ijt} \leftarrow \left(H^\star_{ijt}, c^\star_{ijt}\right)$
else
    $\mathcal{D}_{ijt} \leftarrow ()$
```

3.2 Single-View Matching

Single-view matching forms the basis of determining which, if any, models are visible in a given image. Outlined in Algorithm 1, the process searches for instances of each view by matching its feature constellation to those in the image. For a particular view v_{ij} from a particular object model $\mathcal{M}_i$, the goal of single-view matching is to produce visibility hypotheses and associated likelihoods of that view's presence and location in a particular image.

As mentioned above, we first extract a set of SIFT features $\mathcal{F}_t$ from the image captured at time index t. For each view v_{ij}, our algorithm matches the view's set of descriptors $\mathcal{F}_{ij}$ with those in the image $\mathcal{F}_t$ to produce a set of point-pair correspondence candidates $\mathcal{C}_{ijt}$. We evaluate the similarity s_{pq} between a pair of features p and q as the normalized inner product between their descriptor vectors f_p and f_q, where $s_{pq} = \sum_k (f_{pk} f_{qk}) / \|d_p\| \|d_q\|$. We exhaustively compute all similarity scores and collect in $\mathcal{C}_{ijt}$ at most one pair per feature in $\mathcal{F}_{ij}$, subject to a minimum threshold. Appendix A lists this and the other specific parameter settings that were used to evaluate the algorithm in Section 4.

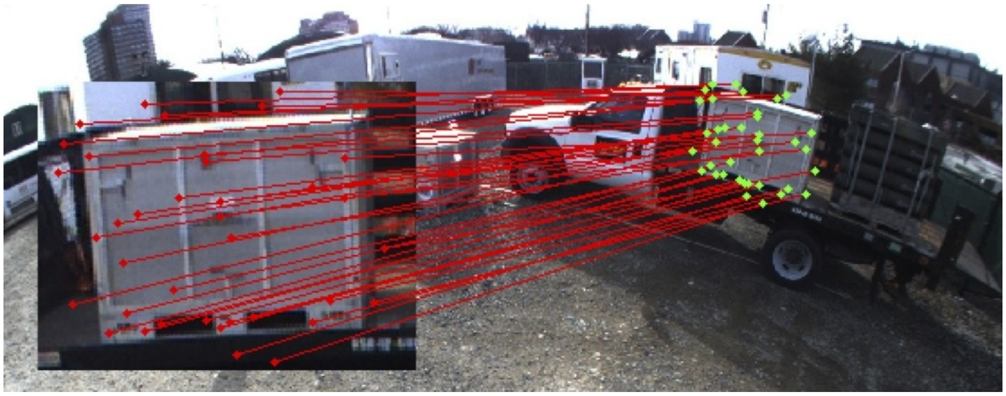


Figure 4: A visualization of an object being matched to an appearance model (inset) derived from the user's stylus gesture. Lines denote the correspondence between SIFT features within the initial view to those on the object in the scene.

Since many similar-looking objects may exist in a single image, $\mathcal{C}_{ijt}$ may contain a significant number of outliers and ambiguous matches. We therefore enforce geometric consistency on the constellation by means of random sample consensus (RANSAC) [Fischler and Bolles, 1981] with a plane projective homography H as the underlying geometric model [Hartley and Zisserman, 2004]. Our particular robot employs wide-angle camera lenses that exhibit noticeable radial distortion. Before applying RANSAC, we correct the distortion of the interest points to account for deviations from standard pinhole camera geometry, which enables the application of a direct linear transform to estimate the homography.

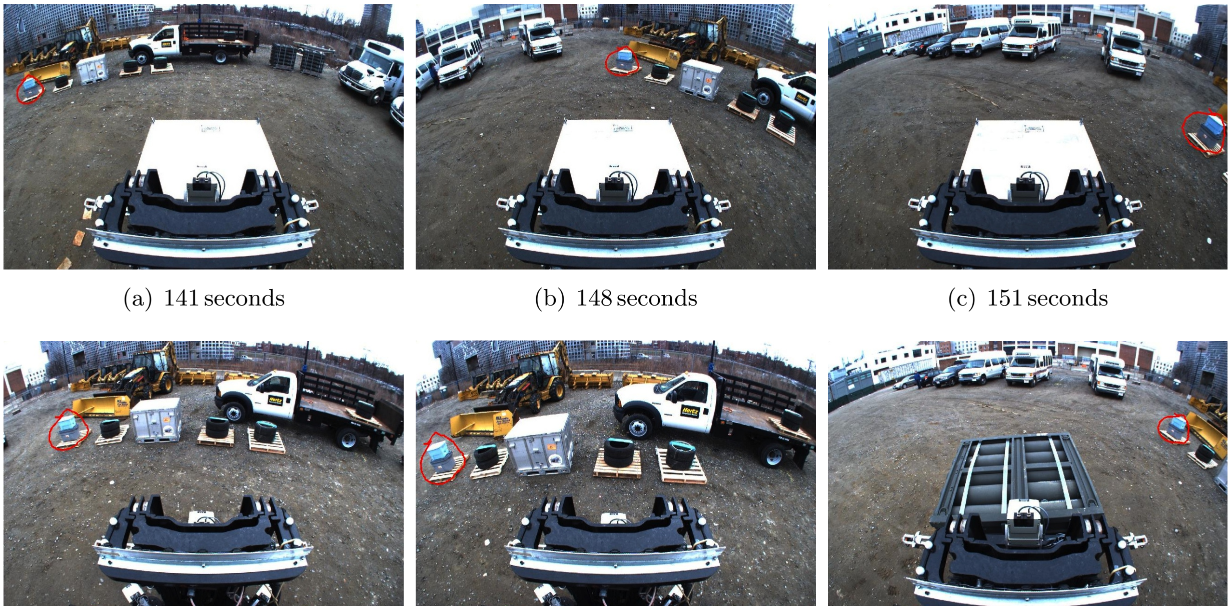
At each RANSAC iteration, we select four distinct (un-distorted) correspondences $\hat{\mathcal{C}}_{ijt} \in \mathcal{C}_{ijt}$ with which we compute the induced homography $\hat{H}$ between the current image and the view v_{ij} (Line 7). We then apply the homography to all matched points within the current image, re-distort the result, and classify each point as an inlier or outlier according to its distance from its image counterpart and a pre-specified threshold t in pixel units (Lines 12 and 13). As the objects are non-planar, we use a loose value for this threshold in practice to accommodate deviations from planarity due to motion parallax.

RANSAC establishes a single best hypothesis for each view v_{ij} that consists of a homography $H^{\star}_{ijt}$ and a set of inlier correspondences $\tilde{\mathcal{C}}^{\star}_{ijt} \in \mathcal{C}_{ijt}$ (Figure 4). We assign a confidence value c_{ijt} to the hypothesis, which represents the proportion of inliers to total points in v_{ij} as well as the absolute number of inliers $c_{ijt} = |\text{inliers}|/(|v_{ij}| \min(\alpha|\text{inliers}|, 1)$ (Line 17). If the confidence is sufficiently high per a user-defined threshold t_c, we output the hypothesis.

3.3 Multiple-View Reasoning

The above single-view matching procedure may produce a number of match hypotheses per image and does not prohibit detecting different instances of the same object. Each object model possesses one or more distinct views, and it is possible for each view to match at most one location in the image. Our algorithm reasons over these matches and their associated confidence scores to resolve potential ambiguities, thereby producing at most one match for each model and reporting its associated image location and confidence.

First, all hypotheses are collected and grouped by object model. To each "active" model (i.e., a model for which a match hypothesis has been generated), we assign a confidence score equal to that of the most confident view candidate. If this confidence is sufficiently high as specified by a



(a) 141 seconds

(b) 148 seconds

(c) 151 seconds

(d) 288 seconds

(e) 292 seconds

(f) 507 seconds

Figure 5: New views of an object annotated with the corresponding reprojected gesture. New views are added to the model when the object's appearance changes, typically as a result of scale and viewpoint changes. Times shown indicate the duration since the user provided the initial gesture. Note that the object was out of view during the periods between (c) and (d), and (e) and (f), but was reacquired when the robot returned to the scene.

threshold t_c^{match}, we consider the model to be visible and report its current location, which is defined as the original 2D gesture region transformed into the current image by the match homography associated with the hypothesis.

Note that while this check ensures that each model matches no more than one location in the image, we do not impose the restriction that a particular image location match at most one model. Indeed, it is possible that running the multiple-view process on different models results in the same image location matching different objects. However, we have not found this to happen in practice, which we believe to be a result of surrounding contextual information captured within the user gestures.

3.4 Model Augmentation

As the mobile manipulator navigates within the environment, an object's appearance changes due to variations in viewpoint and illumination. Furthermore, the robot makes frequent excursions—for example, moving cargo to another location in the warehouse—that result in extended frame cuts. When the robot returns to the scene, it typically observes objects from a different vantage point. Although SIFT features tolerate moderate appearance variability due to some robustness to scale, rotation, and intensity changes, the feature and constellation matches degenerate with more severe scaling and 3D perspective effects.

To combat this phenomenon and retain consistent object identity over longer time intervals and larger displacements, the algorithm periodically augments each object model by adding new views whenever an object's appearance changes significantly. In this manner, the method opportunis-

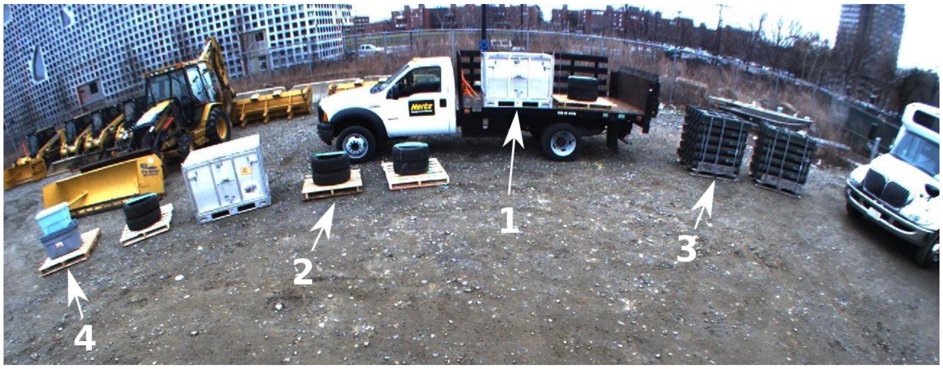


Figure 6: The setup of the first experiment as viewed from the robot's front-facing camera. The scene includes several similar-looking objects to assess the system's behavior in the presence appearance ambiguity.

tically captures the appearance of each object from multiple viewing angles and distances. This increases the likelihood that new observations will match one or more views with high confidence and, in turn, greatly improves the overall robustness of the reacquisition. Figure 5 depicts views of an object that were automatically added to the model based upon appearance variability.

The multiple-view reasoning signals a positive detection when it determines that a particular model $\mathcal{M}$ is visible in a given image. We examine each of the matching views v_j for that model and consider both the robot's motion and the geometric image-to-image change between the v_j and the associated observation hypotheses. In particular, we evaluate the minimum position change $d_{\min} = \min_j \|p_j - p_{\text{cur}}\|$ between the robot's current position p_{cur} and the position p_j associated with the j^{th} view, along with the minimum 2D geometric change $h_{\min} = \min_j \text{scale}(H_j)$ corresponding to the overall 2D scaling implied by match homography H_j. If both $d_{\min}$ and $h_{\min}$ exceed pre-specified thresholds, signifying that no current view adequately captures the object's current image scale and pose, then we create a new view for the model $\mathcal{M}$ using the hypothesis with the highest confidence score.

In practice, the system instantiates a new view by generating a "virtual gesture" that segments the object in the image. SIFT features from the current frame are used to create a new view as described in Section 3.1, and this view is then considered during single-view matching (Section 3.2) and during multiple-view reasoning (Section 3.3).

4 Experimental Results

We conducted two sets of experiments to illustrate and validate our reacquisition algorithm on real data streams collected by our robotic forklift within outdoor environments. The first of these focused on demonstrating the advantages of multiple-view reasoning over single-view matching in temporally local reacquisition; the second was designed to evaluate performance under more challenging conditions that include differences in sensors, illumination, and relative object pose from initial training through reacquisition. Appendix A lists the parameter settings that were used.

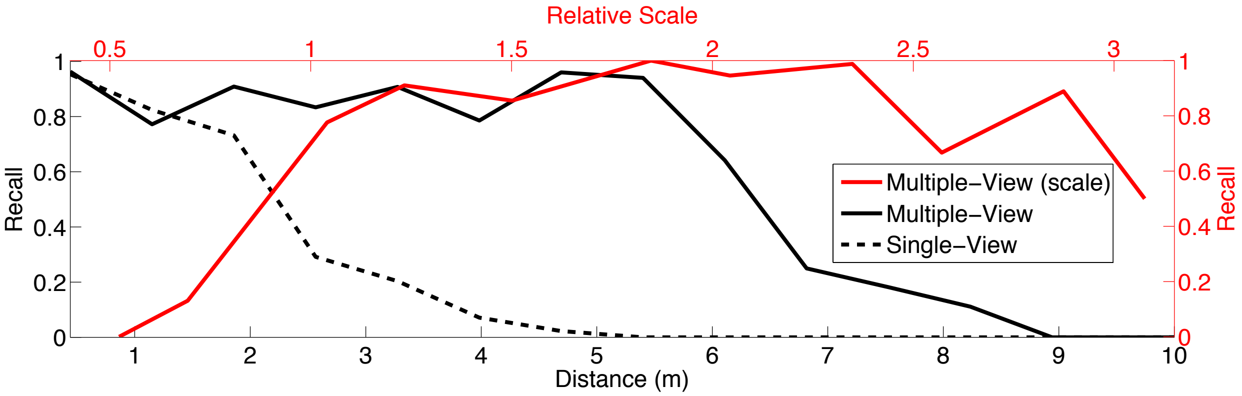


Figure 7: Recall as a function of the robot's distance (bottom label) and scale change (top label) from the original gesture position. A relative scale of one does not match zero distance due to imprecise measurement of scale in the ground truth bounding boxes.

4.1 Single vs. Multiple Views

Mimicking the scenario outlined in Section 2, we arranged the environment as shown in Figure 6, with nine pallets, seven on the ground and two on a truck bed. The objects were chosen such that all but one (containing boxes) had a similar-looking counterpart in the scene. The robot moved each of the four objects to another location in the warehouse approximately 50 m away, according to the order indicated in Figure 6. After transporting each pallet, the forklift returned roughly to its starting position and heading, with pose variations typical of autonomous operation. Full-resolution (1296 × 964) images from the front-facing camera were recorded at 2 Hz. The overall experiment lasted approximately 12 minutes. We manually annotated each image with a bounding box for each viewed object and used these annotations as ground truth to evaluate performance. Here, a detection is deemed positive if the center of the reprojected (virtual) gesture falls within the ground truth bounding box.

Figure 7 indicates the detection rate for all four objects as a function of the robot's distance from the location at which the original gesture was made. Note that single-view matching yields recognition rates above 0.6 when the images of the scene are acquired within 2 m of the single-view appearance model. Farther away, however, the performance drops off precipitously, mainly due to large variations in scale relative to the original view. On the other hand, multiple-view matching yields recognition rates above 0.8 up to distances of 5.5 m from the point of the original gesture and detections up to nearly 9 m away. The plot shows corresponding recall rates as a function of relative scale, which represents the linear size of the ground truth segmentation at the time of detection relative to its initial size.

The improvement in recognition rates at greater distances suggests that augmenting the model with multiple object views facilitates recognition across varying scales and viewpoints. Figure 8 indicates the number of views that comprise each model as a function of time since the original gesture was provided. Object 2, the pallet of tires near the truck's front bumper, was visible at many different scales and viewpoints during the experiment, resulting in a particularly high number of model views. Despite the presence of three similarly-looking objects including an adjacent pallet of tires, there were no false positives. The same is true of Objects 3 and 4.

The experiment helps validate the reacquisition method's tolerance to object ambiguity and large changes in viewpoint and scale. It does not, however, examine the effects of illumination variation. The test was conducted on an overcast day with relatively uniform illumination, conditions

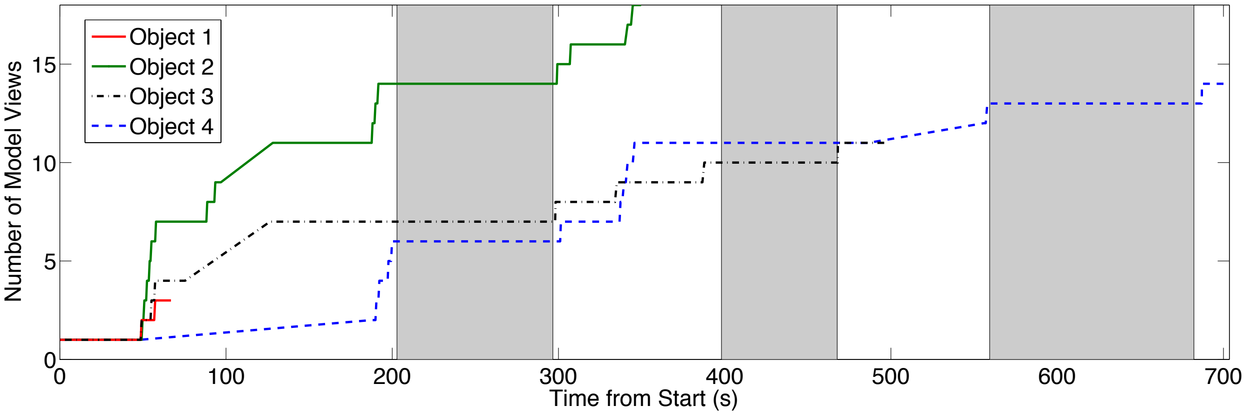


Figure 8: The number of views comprising the appearance model for each object as a function of time since the original user gestures were made. Gray bands roughly indicate the periods when the objects were not in the camera's field of view. The object numbering matches that in Figure 6.

that are near-ideal for image feature matching. Furthermore, the evaluations do not explore the ability to share models across cameras. The next section presents the results of a set of experiments that evaluate the method's performance under these and other challenging conditions.

4.2 Varying View Conditions

We consider another, more extensive set of experiments that evaluate the method's robustness to typical conditions that include variations in illumination along with the effects of scene clutter, object ambiguity, and changes in context. The data was acquired over the course of a week at an active, outdoor military storage facility where the robot was operating (Figure 9). The environment consisted of more than one hundred closely spaced objects that included washing machines, generators, tires, engines, and trucks, among others. In most cases, objects of the same type with nearly identical appearance were placed less than a meter apart (i.e., well below the accuracy of the robot's absolute position estimate). In addition to clutter, the data sets were chosen for the presence of lighting variation that included global brightness changes, specular illumination, and shadow effects, along with viewpoint changes and motion blur. The data sets are representative of many of the challenges typical to operations in unprepared, outdoor settings.

We structured the experiments to emulate the guided tour interaction in which the user names and segments objects within an image as the robot drives by them, and later directs the robot to retrieve one or more of the objects by name. A number of conditions can change between the time that the object is first indicated and the time it is reacquired, including the physical sensor (right-facing vs. front-facing camera), illumination, object positions within the environment, aspect angle, and scale.

Several video clips collected at 2 Hz were paired with one another in five combinations. Each pair consisted of a short "tour" clip acquired from the right-facing camera and a longer "reacquisition" clip acquired from the front-facing camera. Figure 10 shows a few of the images from the five scenarios, including those from the tour and reacquisition phases. Ground truth annotations were manually generated for each image in the reacquisition clips and were used to evaluate performance in terms of precision and recall. We used the metric employed in the PASCAL challenge [Everingham et al., 2010] to deem a detection correct, requiring that the area of the intersection of the detection and ground truth regions exceed a fraction of the area of their union.

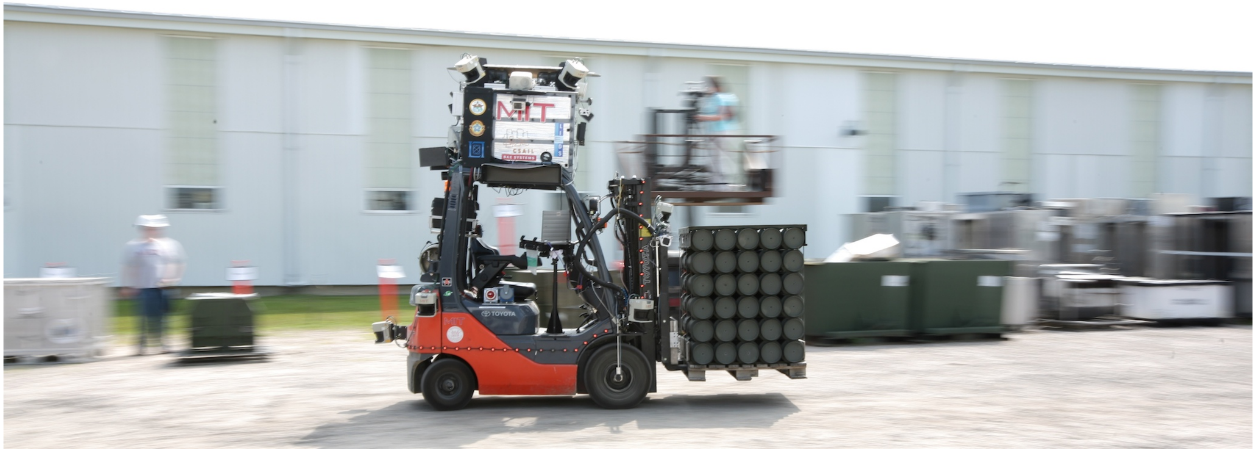


Figure 9: The robot transporting cargo at an active military outdoor storage facility.

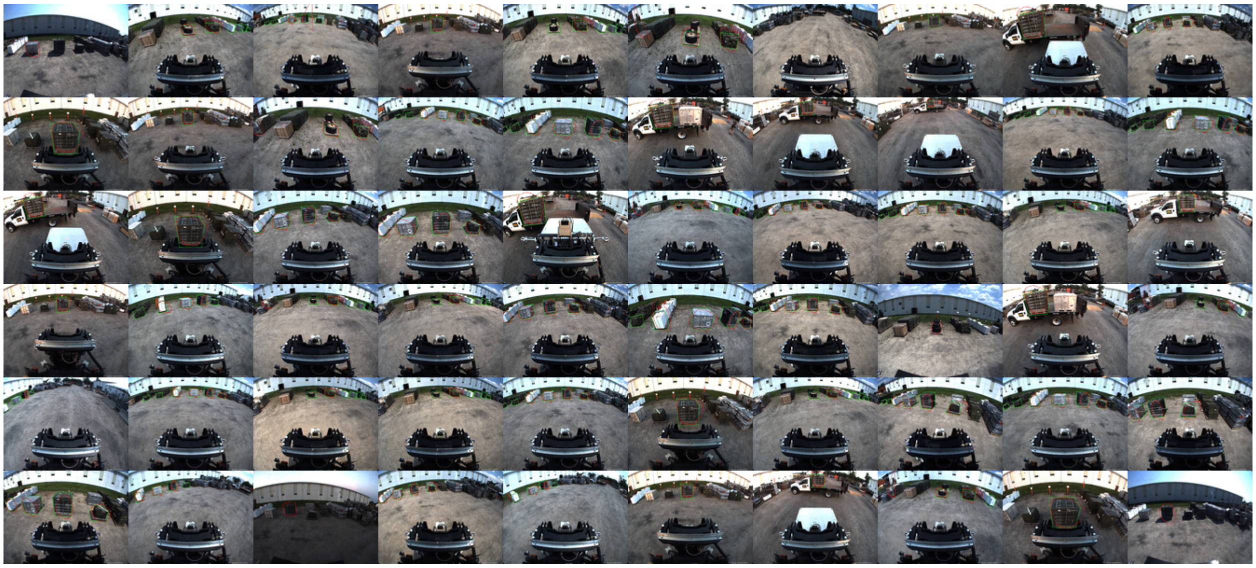


Figure 10: Some of the images acquired from the forward- and right-facing cameras during the five scenarios of the experiment. The right-facing images (those without the forklift structure in the lower portion of the frame) reveal the user gestures that indicate the sole training example for each object. The forward-facing images show model views learned automatically by the algorithm with their reconstituted gestures (red, ellipse-like outlines).

Table 1 lists the scenarios, their characteristics, and the performance achieved by our algorithm. Possible condition changes between tour and reacquisition clips include "sensor" (right vs. front camera), "lighting" (illumination and shadows), "3D pose" (scale, position, aspect angle), "context" (unobserved object relocation with respect to the environment), "confusers" (objects of similar appearance nearby), and "Δt" (intervening hours:minutes). True and false positives are denoted as TP and FP, respectively; "truth" indicates the total number of ground truth instances; "frames" is the total number of images; and "objects" refers to the number of unique object instances that were toured in the scenario. Performance is reported in terms of aggregate precision TP/(TP+FP) and recall TP/truth.

Table 1: Conditions and reacquisition statistics for the different experiment scenarios.

Scenario	Train	Test	Sensor	Lighting	3D pose	Context	Confusers	Δt	Frames	Objects	Truth	TP	FP	Precision	Recall
1	Afternoon	Afternoon	✓		✓		✓	00:05	378	6	1781	964	59	94.23%	54.13%
2	Evening	Evening	✓	✓	✓		✓	00:05	167	1	167	158	0	100.00%	94.61%
3	Morning	Evening	✓	✓	✓		✓	14:00	165	1	165	154	0	100.00%	93.33%
4	Morning	Evening	✓	✓	✓			10:00	260	1	256	242	0	100.00%	94.53%
5	Noon	Evening	✓	✓	✓	✓		07:00	377	1	257	243	0	100.00%	94.55%

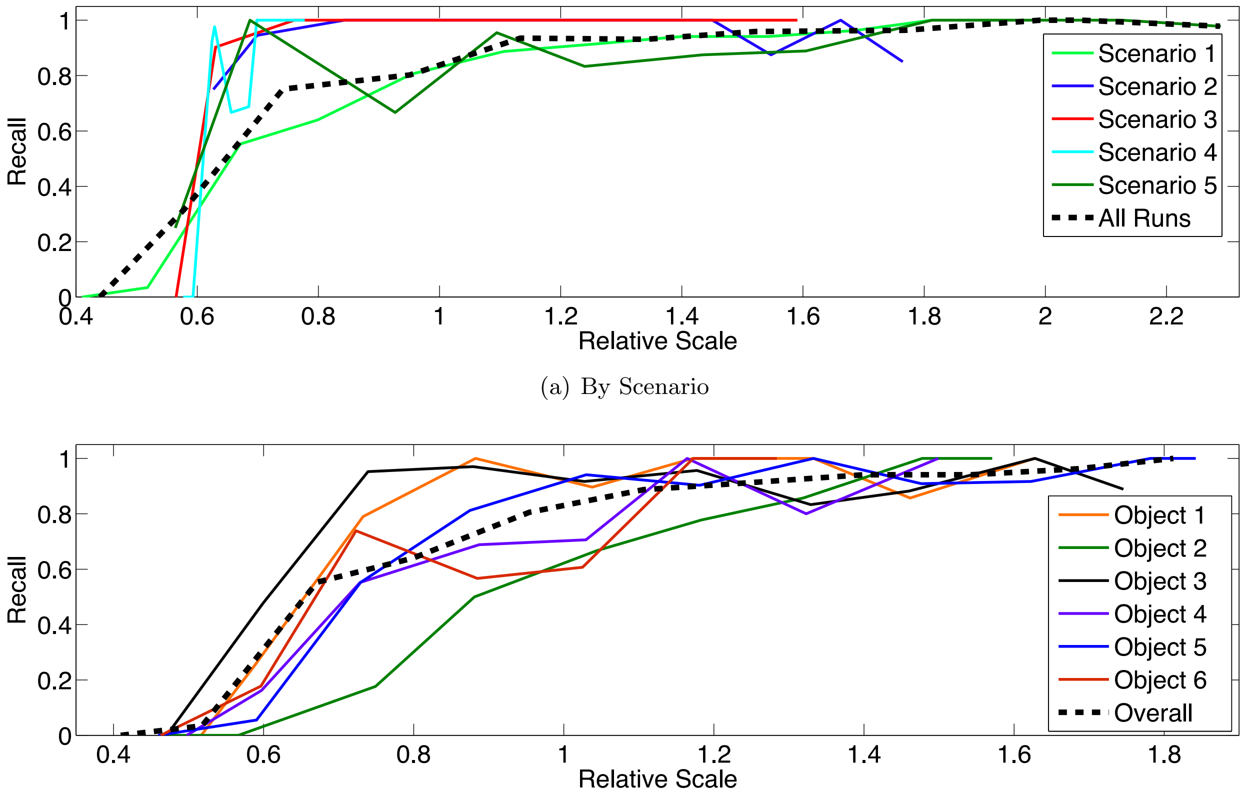
In all five experiments, the method produced few if any false positives as indicated by the high precision figures. This demonstrates that our approach to modeling an object's appearance variation online does not result in the learner drifting, as often occurs with adaptive learners. We attribute this behavior to the geometric constraints that help to prevent occlusions and clutter from corrupting the appearance models. While the algorithm performs well in this respect, it yields a reduced number of overall detections in the experiments that make up the first scenario. This scenario involved significant viewpoint changes between the initial training phase and the subsequent test session. Training images were collected as the robot moved in a direction parallel to the front face of each object and the user provided the initial training example for each object when it was fronto-parallel to the image. As the robot proceeded forward, only views of the object's front face and one side were available for and added to its appearance model. During the testing phase of the experiment, the robot approached the objects from a range of different angles, many of which included views of unmodeled sides of the object. In these cases, the algorithm was unable to identify a match to the learned model. This, together with saturated images of a highly reflective object resulted in an increased false negative rate. While utilizing homography validation as part of the multiple-view matching significantly reduces false matches, it also results in false negatives due to unmodeled 3D effects such as parallax.

We evaluate the relationship between recall rate and the change in scale between an object's initial (scale=1) and subsequent observations. Figure 11(a) plots aggregate performance of all objects for each of the five test scenarios, while Figure 11(b) shows individual performance of each object in Scenario 1. Figure 12 plots the performance of a single object from Scenario 5 in which the context has changed: the object was transported to a different location while nearby objects were moved. Finally, in Figure 13, we report recall rates for this object, which is visible in each of the scenarios.

For the above experiments, we manually injected a gesture for each object during each tour clip—while the robot was stationary—to initiate model learning. We selected a single set of parameters for all scenarios: for single-view matching, the SIFT feature match threshold (dot product) was 0.9 with a maximum of 600 RANSAC iterations and an outlier threshold of 10 pixels; single-view matches with confidence values below 0.1 were discarded. The reasoning module added new views whenever a scale change of at least 1.2 was observed and the robot moved at least 0.5 m. We found that the false positive rate was insensitive to these settings for the reasoning parameters, which were chosen to effectively trade off object view diversity and model complexity (data size).

5 Discussion

We described an algorithm for object instance reacquisition that enables mobile manipulation. The system takes as input a coarse, user-specified object segmentation in a single image from



(a) By Scenario

(b) By Object

Figure 11: Recall rates as a function of scale change (a) for all objects by scenario, and (b) for each object in Scenario 1.

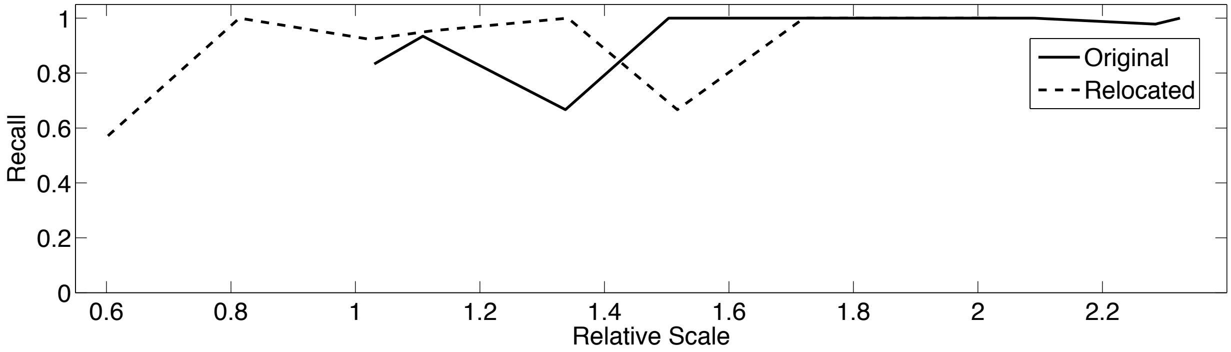


Figure 12: Recall rates as a function of scale change for an object in different positions and at different times. The pallet was on the ground during the tour and reacquired 7 hours later both on the ground and on a truck bed.

one of the robot's cameras, then builds an appearance model of that object automatically and online. Multiple-view models enable robust, long-term matching with very few false positives despite the presence of drastic visual changes resulting from platform motion, differing sensors, object repositioning, and time-varying illumination. Figure 14 displays a few examples. Each row corresponds to a different object model with the left-most image showing the user segmentation

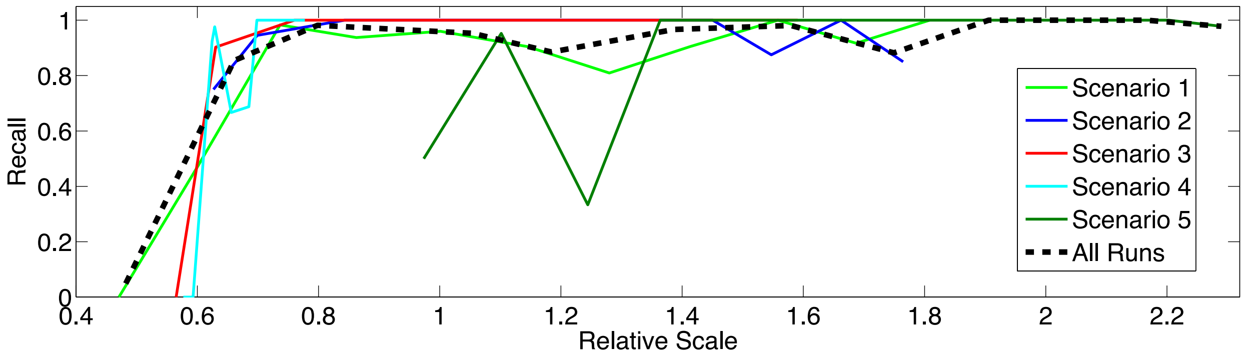


Figure 13: Recall rates as a function of scale change for a single object across all scenarios.

and those to the right the subsequent reacquisitions.

Despite its successes, our approach has several shortcomings. For one, end-to-end performance is limited by the reliability of low-level feature extraction and matching. While SIFT keypoints exhibit good robustness to moderate scaling, global brightness changes, and in-plane rotation, they are confounded by more substantial variations due to parallax, lens distortion, and specular reflections (e.g., as seen with the metallic pallet). Longer exposure times under low-light conditions amplify additive noise and motion blur, further degrading frame-to-frame keypoint consistency; similarly, pixel saturation due to very bright or dark objects (e.g., as observed with the washing machines) reduces contrast and diminishes the number and quality of extracted keypoints. Figure 15 demonstrates several of these failure conditions.

Another limitation of our approach lies in the implicit assumption that observed objects are sufficiently planar that their frame-to-frame motion is best described by a plane projective homography. This is certainly untrue for most real objects, and while maintaining multiple object views improves robustness to non-planarity, our matching algorithm remains sensitive to drastic parallax, particularly when the original segmentation engulfs scenery distant from (e.g., behind) the object. As we saw with the results from Scenario 1, this planarity assumption increases the rate of false negatives when the algorithm is invoked to recognize objects from viewpoints too different from those used for training. One way to address this is to incorporate 3D information from LIDAR scans or structure-from-motion point clouds into the appearance models, which is a subject of ongoing research. Another strategy would be to relax rigid geometric constraints in favor of more qualitative graph matching.

Our multiple-view representation treats each view as an independent collection of image features and, as a result, the matching process scales linearly with the number of views. We suspect that computational performance can be greatly improved through a bag-of-words representation that utilizes a shared vocabulary tree for fast (sub-linear) matching [Nistér and Stewenius, 2006]. Robust statistical optimization via iteratively reweighted least squares could also improve the runtime performance and determinism of the approach over RANSAC-based constellation matching.

6 Conclusion

This paper presented a one-shot reacquisition algorithm that exploits robot motion to automatically model the adaptive appearance of objects based upon a single, user-provided training example. The algorithm enables a robot to formulate spatially extended, temporally persistent representations of their surround that they can then call on to reacquire objects. This capability enables human users

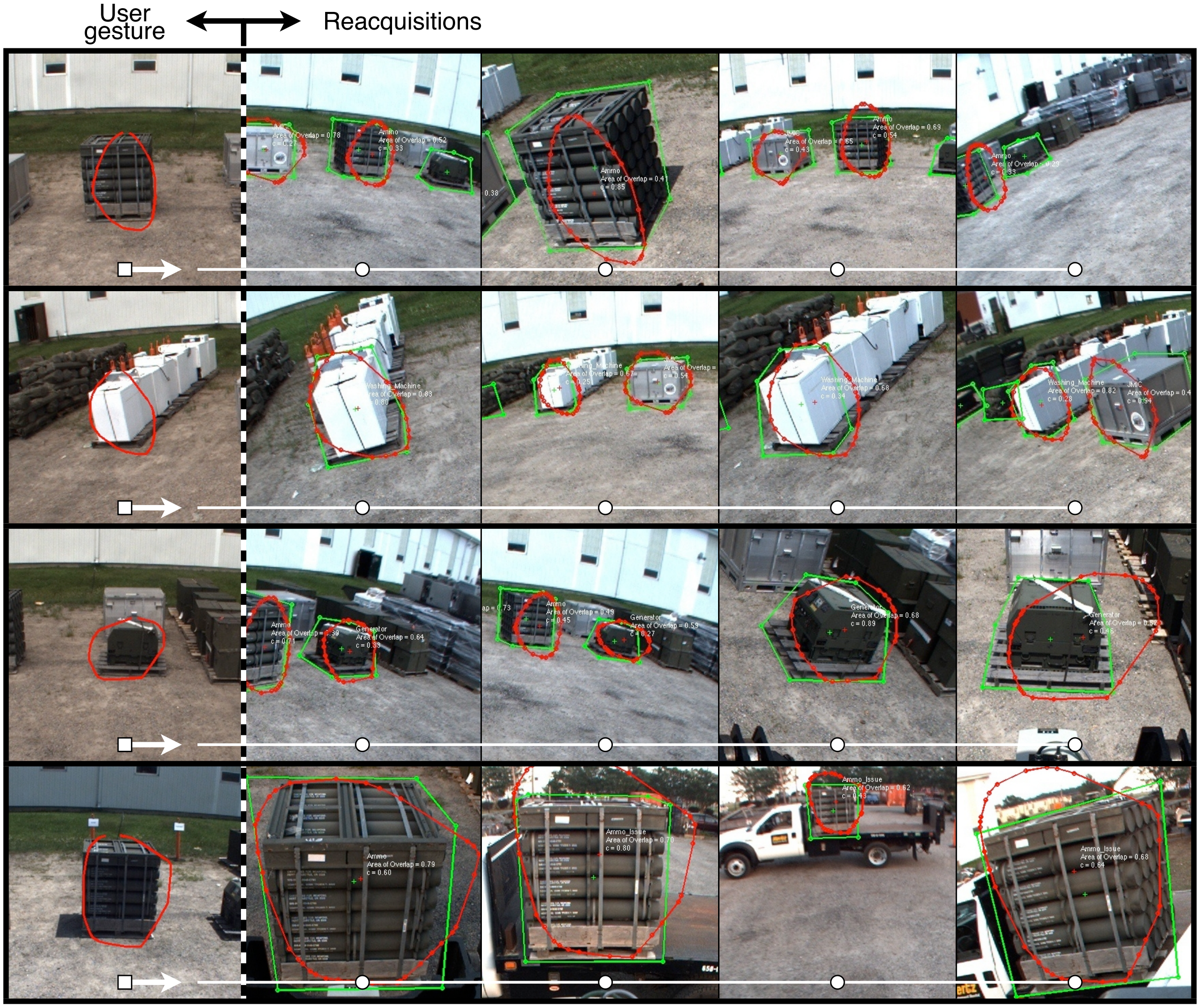


Figure 14: Images that correspond to positive detections over variations in relative pose, lighting, and scale. Each row represents a different object model, with the left-most image displaying the user's initial gesture and the subsequent images exhibiting the reconstituted gesture (red, ellipse-like outlines) and the ground truth segmentation (green simple polygons). Note that all images are shown at the same pixel scale.

to more effectively and efficiently command robots, as we demonstrate on a mobile manipulation platform.

We offered a detailed analysis of our reacquisition algorithm by evaluating its performance under the challenging conditions that are typical of outdoor, unstructured environments. Our opportunistic image-based approach performed well over a wide range of lighting, scale, and context changes. The results demonstrate that the algorithm offers sufficient robustness to real-world conditions to be usable in practice.

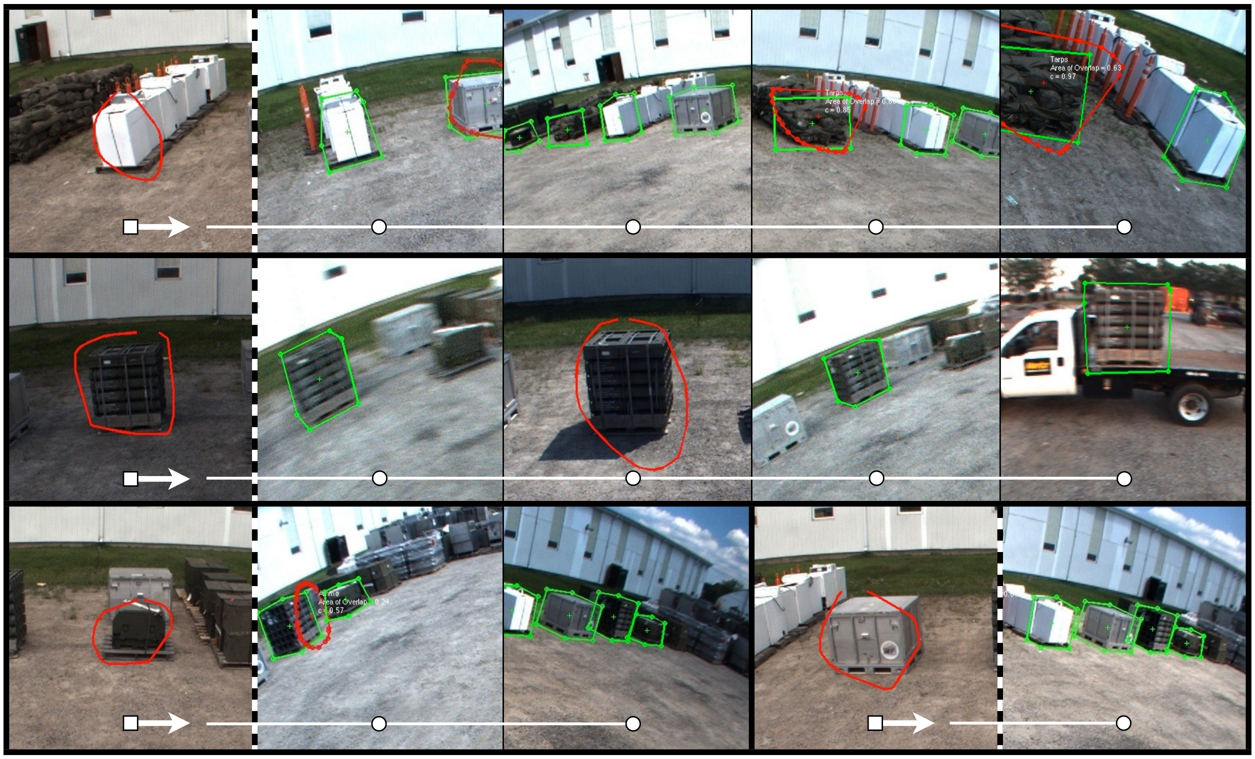


Figure 15: Images that depict failed detections of four toured objects. We have found that our algorithm fails to reacquire objects for which limited contrast due to over-saturation or poor illumination yields few keypoints. This sensitivity is exacerbated when there is a significant variation in illumination between the initial segmentation and the reacquisition phase.

Acknowledgments

We gratefully acknowledge the support of the U.S. Army Logistics Innovation Agency (LIA) and the U.S. Army Combined Arms Support Command (CASCOM).

This work was sponsored by the Department of the Air Force under Air Force Contract FA8721-05-C-0002. Any opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

A Appendix

The following table lists the algorithm parameter settings that were used in the implementations that we evaluate in Section 4.

References

Babenko, B., Yang, M.-H., and Belongie, S. (2009). Visual tracking with online multiple instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 983–990, Miami, FL.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., and

Table 2: Parameter settings.

Parameter	Description	Setting
N	Number of RANSAC iterations for single-view matching (Alg. 1, line 5).	600
s_{pq}^{min}	Minimum dot product allowable distance between SIFT feature matches (Alg. 1, line 2).	0.9
t_d	Maximum distance in pixels of projected interest points for RANSAC inliers (Alg. 1, line 12).	10.0 px
t_c	Minimum confidence threshold for homography validation (Alg. 1, line 18).	0.10
t_c^{match}	Minimum confidence threshold for a visible model match.	0.10
h_{min}	Minimum scale variation between an existing model view and a new view for model augmentation	1.20
d_{min}	Minimum displacement of the robot between new and existing views for model augmentation.	0.50 m

Chilongo, D. (2004). Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(2):315–348.

Collet, A., Berenson, D., Srinivasa, S. S., and Ferguson, D. (2009). Object recognition and full pose registration from a single image for robotic manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 48–55, Kobe, Japan.

Collins, R. T., Liu, Y., and Leordeanu, M. (2005). Online selection of discriminative tracking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1631–1643.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577.

Correa, A., Walter, M. R., Fletcher, L., Glass, J., Teller, S., and Davis, R. (2010). Multimodal interaction with an autonomous forklift. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 253–250, Osaka, Japan.

Everingham, M., Van Gool, L., Williams, C. K., and Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal on Computer Vision*, 88(2):303–338.

Fei-Fei, L., Fergus, R., and Perona, P. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1134–1141, Nice, France.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–271, Madison, WI.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Gordon, I. and Lowe, D. G. (2006). What and where: 3D object recognition with accurate pose. In *Toward Category-Level Object Recognition*, pages 67–82. Springer-Verlag.

Grabner, H., Leistner, C., and Bischof, H. (2008). Semi-supervised on-line boosting for robust tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–247, Marseille, France.

Haasch, A., Hofemann, N., Fritsch, J., and Sagerer, G. (2005). A multi-modal object attention system for a mobile robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2712–2717, Edmonton, Alberta, Canada.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision.* Cambridge University Press, second edition.

Hoiem, D., Rother, C., and Winn, J. (2007). 3D LayoutCRF for multi-view object class recognition and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN.

Kalal, Z., Matas, J., and Mikolajczyk, K. (2009). Online learning of robust object detectors during unstable tracking. In *On-line Learning for Computer Vision Workshop*, Kobe, Japan.

Kalal, Z., Matas, J., and Mikolajczyk, K. (2010). P-N learning: Bootstrapping binary classifiers by structural constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–56, San Francisco, CA.

Liebelt, J., Schmid, C., and Schertler, K. (2008). Viewpoint-independent object class detection using 3D feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK.

Lim, J., Ross, D., Lin, R.-S., and Yang, M.-H. (2004). Incremental learning for visual tracking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 793–800, Vancouver, B.C., Canada.

Lowe, D. G. (2001). Local feature view clustering for 3D object recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 682–688, Kauai, HI.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110.

Nistér, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, New York, NY.

Savarese, S. and Fei-Fei, L. (2007). 3D generic object categorization, localization and pose estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Stark, M., Goesele, M., and Schiele, B. (2009). A shape-based object class model for knowledge transfer. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 373–380, Kyoto, Japan.

Teller, S., Walter, M. R., Antone, M., Correa, A., Davis, R., Fletcher, L., Frazzoli, E., Glass, J., How, J. P., Huang, A., Jeon, J. h., Karaman, S., Luders, B., Roy, N., and Sainath, T. (2010). A voice-commandable robotic forklift working alongside humans in minimally-prepared outdoor environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 526–533, Anchorage, AK.

Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., and Roy, N. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1507–1514, San Francisco, CA.

Walter, M. R., Friedman, Y., Antone, M., and Teller, S. (2010a). Appearance-based object reacquisition for mobile manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, San Francisco, CA.

Walter, M. R., Friedman, Y., Antone, M., and Teller, S. (2010b). Vision-based reacquisition for task-level control. In *Proceedings of the International Symposium on Experimental Robotics (ISER)*, New Dehli, India.

Walter, M. R., Karaman, S., Frazzoli, E., and Teller, S. (2010c). Closed-loop pallet engagement in unstructured environments. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5119–5126, Taipei, Taiwan.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys*, 38(4).