In Springer Tracts for Advanced Robotics, 2011

Vision-based Reacquisition for Task-level Control

Matthew R. Walter, Yuli Friedman, Matthew Antone, and Seth Teller

Abstract We describe a vision-based algorithm that enables a robot to "reacquire" objects previously indicated by a human user through simple image-based stylus gestures. By automatically generating a multiple-view appearance model for each object, the method can reacquire the object and reconstitute the user's segmentation hints even after the robot has moved long distances or significant time has elapsed since the gesture. We demonstrate that this capability enables novel command and control mechanisms: after a human gives the robot a "guided tour" of named objects and their locations in the environment, he can dispatch the robot to fetch any particular object simply by stating its name. We implement the object reacquisition algorithm on an outdoor mobile manipulation platform and evaluate its performance under challenging conditions that include lighting and viewpoint variation, clutter, and object relocation.

1 Introduction

This paper describes a vision-based algorithm that enables a human to efficiently convey object- and task-level information to a robot. The ability to understand and execute long task sequences (e.g., in which individual tasks may include moving an object around in an environment) offers the potential of more natural interaction mechanisms as well as a reduced burden for the human. However, achieving this ability and, in particular, the level of recall necessary to reacquire objects after extended time periods and viewpoint changes, are challenging for robots that operate with imprecise knowledge of absolute location within dynamic, uncertain environments.

Matthew R. Walter, Seth Teller
MIT CS & AI Lab (CSAIL), Cambridge, MA, USA, e-mail: {mwalter, teller}@csail.mit.edu

Yuli Friedman, Matthew Antone
BAE Systems, Burlington, MA, USA, e-mail: {yuli.friedman, matthew.antone}@baesystems.com

We present an algorithm that automatically learns a 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 multi-view, feature-based object models that capture variations in appearance due to scale and viewpoint changes. This automatic and opportunistic modeling of an object's appearance enables the robot to reconstitute the user's gesture (and the corresponding 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; in particular, 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 name.

Progressing from a preliminary demonstration of this approach [15], we analyze 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. 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 and using view-invariant descriptors to represent these parts (e.g., [13, 9]). 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 [10] 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 [6] describe a more structured technique for object matching and pose estimation that explicitly builds a 3D model from multiple uncalibrated views using bundle adjustment, likewise establishing SIFT correspondences for recognition but further estimating the relative camera pose via RANSAC and Levenberg-Marquardt optimization. Collet et al. [3] extend this work by incorporating Mean-Shift clustering to facilitate registration of multiple instances during recognition, demonstrating high precision and

recall with accurate pose in cluttered scenes amid partial occlusions, changes in view distance and rotation, and varying illumination. All of the above techniques build object representations offline through explicit "brute-force" acquisition of views spanning a fairly complete set of aspects, rather than opportunistically as in our work.

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 [16]. Generally, tracking approaches assume a relatively small temporal separation, and therefore slow visual variation, between consecutive observations; they can therefore evolve appearance models gradually and incrementally over time, e.g., through the use of multiple instance learning [1]. 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.

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 [7, 2]. 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

Our object reacquisition strategy is motivated by our ongoing development of a robotic forklift [14] that autonomously manipulates cargo within an outdoor environment under the high-level direction of a human supervisor. 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. Using a handheld tablet interface (Fig. 1), the user identifies a specific pallet or destination by circling it in an image from one of four cameras mounted to the robot. The user can also summon the robot to one of several named locations in the environment by speaking to the tablet.

The system performs these tasks autonomously with little effort required on the part of the user (e.g., 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

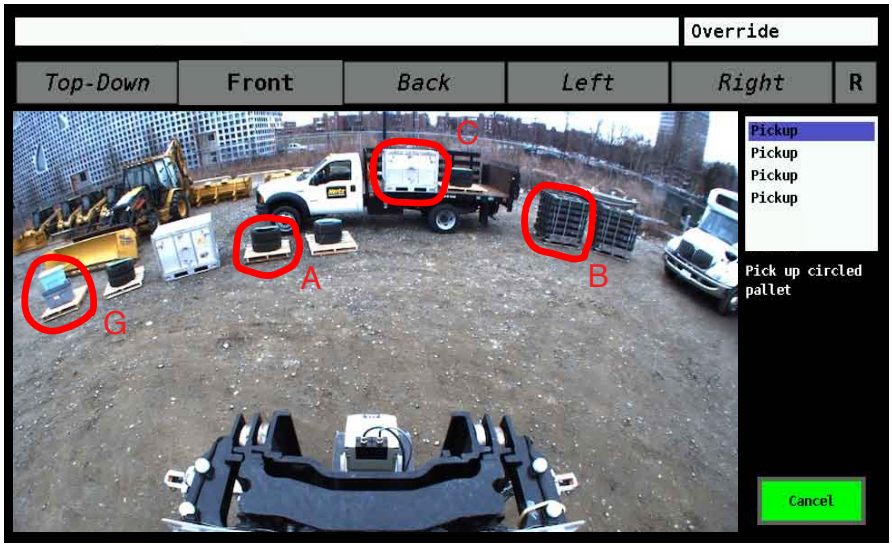


Fig. 1 The tablet interface displaying a view from the robot's forward-facing camera along with user gestures (red).

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 the gesture and manipulates the pallet.

Thus, the system is able to utilize valuable yet unobtrusive segmentation cues from the user to execute extended-duration commands. The ability to recognize particular objects across spatial and temporal excursions allows for other higher-level control mechanisms besides task-specific reacquisition. One example that we support is a guided tour scenario in which the user identifies specific objects in the environment by circling their position in an image from one camera and speaking their label. The system then builds and maintains an appearance model for the named object that is shared across cameras. At a later point in time, the user can command the robot to manipulate an object by referring to it by name (e.g., "bot, pick up the tire pallet"). The primary technical challenge is achieving 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.

Our proposed reacquisition system (Fig. 2) 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).

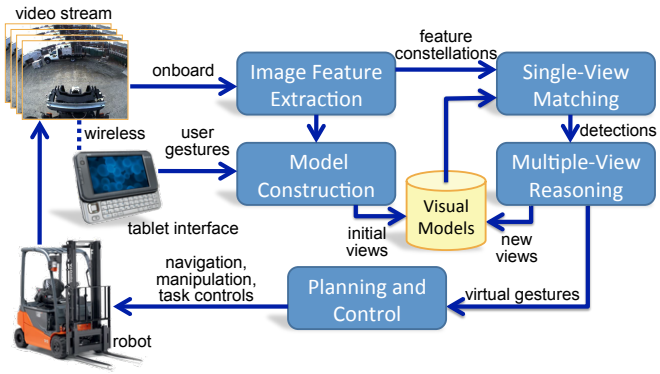


Fig. 2 Block diagram of the reacquisition process.

Our algorithm maintains visual appearance models of the initially indicated objects so that 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 system degrade the precision of its measured position and heading. In fact, the algorithm utilizes dead-reckoned pose estimates only to suggest the creation of new appearance models; it uses neither pose information nor data from non-camera sensors for object recognition. The robot thus handles, without human intervention, a longer string of sequential commands than would otherwise be possible.

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* $\mathscr{M}_i$ as the visual representation of a particular object i, which consists of a collection of views, $\mathscr{M}_i = \{v_{ij}\}$. 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).

Object appearance models and their constituent views are constructed from 2D constellations of keypoints, where each keypoint comprises an image pixel position and an invariant descriptor characterizing the intensity pattern in a local neighborhood. Our algorithm searches each new camera image 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

As each camera image is acquired, it is processed to detect a set $\mathscr{F}$ of keypoint locations and scale invariant descriptors; we use Lowe's SIFT algorithm for moderate robustness to viewpoint and lighting changes [11], 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 $\mathscr{M}_i$ for each indicated object, and any SIFT keypoints and corresponding descriptors that fall within the gesture at that particular frame are accumulated to form the new model's first view v_{i1}.

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

3.2 Single-View Matching

The basic operational unit in determining whether and which models are visible in a given image is feature constellation matching of a single view to that image. For a particular view v_{ij} from a particular object model $\mathscr{M}_i$, 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, a set of SIFT features $\mathscr{F}_t$ is extracted from the image captured at time index t. For each view v_{ij}, our algorithm matches the view's set of descriptors $\mathscr{F}_{ij}$ with $\mathscr{F}_t$ to produce a set of point-pair correspondence candidates $\mathscr{C}_{ijt}$. 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 $\mathscr{C}_{ijt}$ at most one pair per feature in $\mathscr{F}_{ij}$, subject to a minimum threshold.

Since many similar-looking objects may exist in a single image, $\mathscr{C}_{ijt}$ 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) [5] with a plane projective homography H as the underlying geometric model [8]. Our particular robot employs wide-angle camera lenses that exhibit noticeable radial distortion, so before applying RANSAC, we un-distort the points, thereby correcting deviations from standard pinhole camera geometry and allowing the application of a direct linear transform for homography estimation.

At each RANSAC iteration, we select four distinct (un-distorted) correspondences from $\mathscr{C}_{ijt}$ with which we compute the induced homography H between the current image and the view v_{ij}. 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 in pixel units. 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 v_{ij} consisting of a homography and a set of inlier correspondences $\tilde{\mathscr{C}}_{ijt} \in \mathscr{C}_{ijt}$ (Fig. 3). We assign a confidence value c_{ijt} to the hypothesis that represents the proportion of inliers to total points in v_{ij} as well as the absolute number of inliers $c_{ijt} = |\text{inliers}|/(|v_{ij}| \min(\alpha|\text{inliers}|, 1)$. If the confidence is sufficiently high, we output the hypothesis.

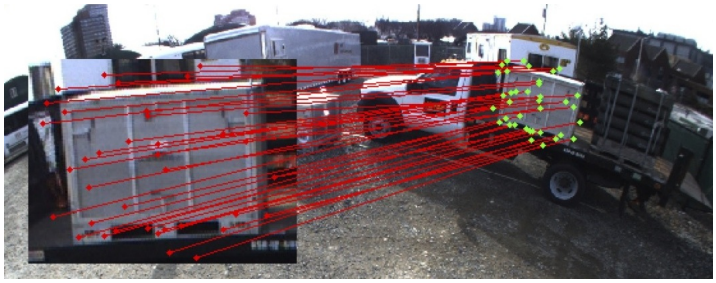


Fig. 3 A visualization of an object being matched to an appearance model (inset) derived from the user's stylus gesture. Red lines denote correspondence between SIFT features within the initial view (red) to those on the object in the scene (green).

3.3 Multiple-View Reasoning

The above single-view matching procedure produces 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 each view can match at most one object in the image with some associated confidence score. Our algorithm reasons over all information at each time step to resolve potential ambiguities, thereby producing at most one match for each model and reporting its associated image location.

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, 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 single-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 robot moves through the environment to execute its tasks, an object's appearance changes due to variations in viewpoint and illumination. Furthermore, when there are gaps in view coverage (e.g., when the robot transports a pallet away from the others and later returns), the new aspect at which an object is observed generally differs from the previous aspect. Although SIFT features are robust to a certain degree of scale, rotation, and intensity changes, thus tolerating moderate appearance variability, the feature and constellation matches degenerate with more severe 3D perspective effects and scaling.

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 any object's appearance has changed sufficiently. This greatly improves the overall robustness of reacquisition, as it opportunistically captures object appearance from multiple aspects and distances and thus increases the likelihood that new observations will match one or more views with high confidence.

When the multi-view reasoning has determined that a particular model $\mathscr{M}$ is visible in a given image, we examine all of that model's matching views v_j 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 $d_{\text{min}} = \min_j \|p_j - p_{\text{cur}}\|$ 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 $h_{\text{min}} = \min_j \text{scale}(H_j)$ 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 a new view is created for $\mathscr{M}$ 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 Sect. 3.1, and this view is then considered during single-view matching (Sect. 3.2) and during multi-view reasoning (Sect. 3.3).

4 Experimental Results

We conducted two sets of experiments to illustrate and validate our reacquisition algorithm on real data streams collected by a robotic platform in outdoor environments. The first of these focused on demonstrating the advantages of multi-view rea-

soning 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.

4.1 Single vs. Multiple Views

Mimicking the scenario outlined in Sect. 2, we arranged the environment as shown in Fig. 1, with nine pallets (seven on the ground and two on a truck bed). The pallet loads were chosen such that all but one pallet (containing boxes) had a similar-looking counterpart in the scene. The robot moved each of the four pallets indicated in Fig. 1 to another location in the warehouse approximately 50 m away. 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 the performance of the algorithms. Here, a detection is deemed positive if the center of the reprojected (virtual) gesture falls within the ground truth bounding box.

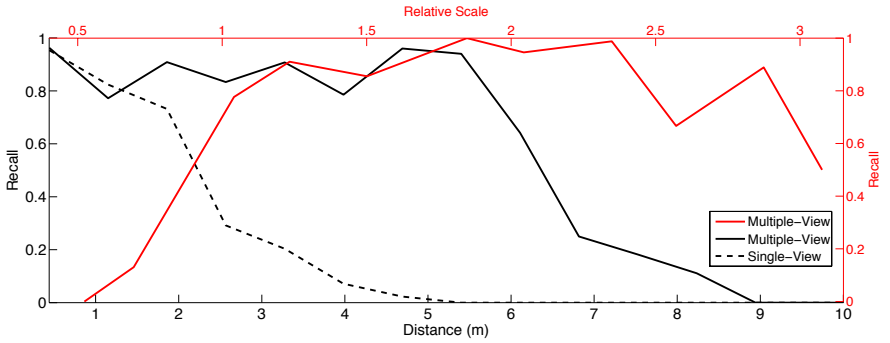


Fig. 4 Recall as a function of the robot's distance (black) and scale change (red) from the original gesture position. A relative scale of one does not match zero distance due to imprecise measurement of scale in the ground truth bounding boxes.

Figure 4 indicates the detection rate, with respect to ground truth, 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.

This set of experiments was conducted under cloudy but bright illumination that is near-ideal for image feature matching. Additionally, all of the images were obtained from one sensor and processed sequentially, thus limiting scale change discontinuities.

4.2 Varying View Conditions

We designed another set of experiments based on data collected in an outdoor lot at various times over multiple days. These experiments were designed to realize the guided tour interaction in which the user indicates the locations and identities of objects as the robot drives by them, and then asks the robot to retrieve one or more of the objects at a later time. 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. 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 [4] 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.

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 environment), "confusers" (identical-looking objects 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.

Plots in the figures depict recall rate as a function of objects' visual scale change between the first observation (scale=1) and every subsequent observation. Figure 5(a) shows aggregate performance of all objects for each of the five test scenarios, while Fig. 5(b) shows individual performance of each object in Scenario 1. Figure 6 shows the performance of a single object from Scenario 5 in which the

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%

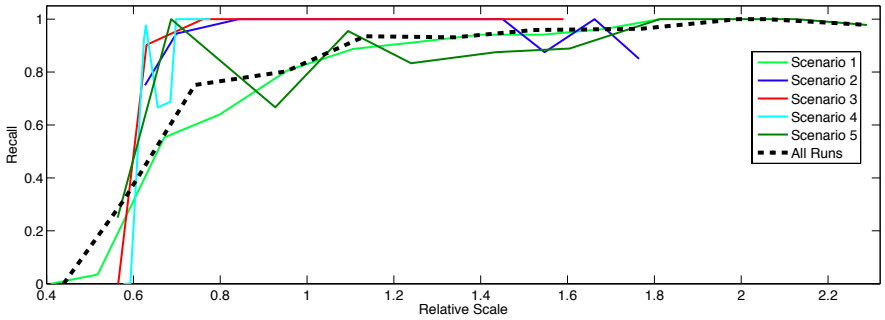
context has changed: the object has been transported to a different location while nearby objects have been moved. Finally, in Fig. 7, we report recall rates for this object, which is visible in each of the scenarios.

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 500 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 when a scale change of at least 1.2 was observed and the robot moved a distance of at least 0.5 m.

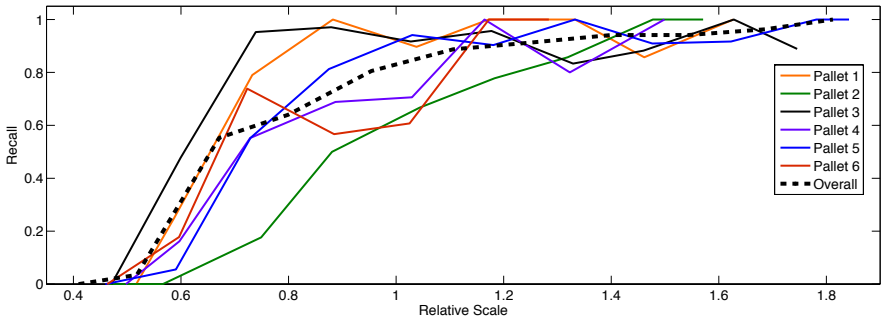
5 Discussion

We described an algorithm for object instance reacquisition that facilitates task-level autonomy with a mobile robot. The system takes as input a coarse user-specified object segmentation in a single image from one of the robot's cameras, then builds an appearance model of that object automatically and online. Multi-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 8 depicts a few examples.

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., 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 9 demonstrates several of these failure conditions.



(a) By Scenario



(b) By Object

Fig. 5 Recall rates as a function of scale change (a) for all objects by scenario, and (b) for each object in Scenario 1.

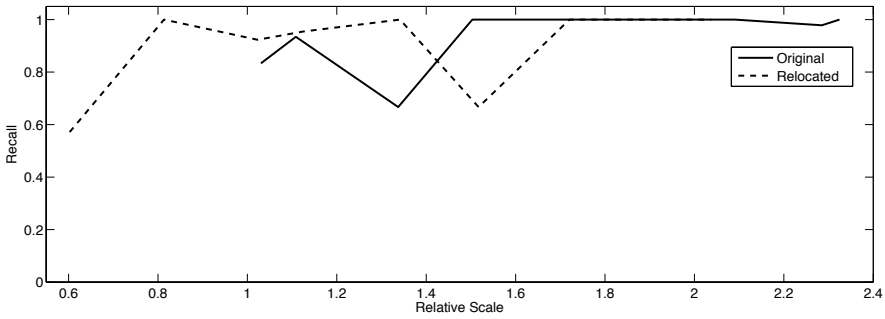


Fig. 6 Recall rates as a function of scale change for an object in different positions and at different times. The pallet was on the ground during the tour and reacquired 7 hours later both on the ground and on a truck bed.

Another limitation of our approach lies in the implicit assumption that observed objects are planar and thus that their frame-to-frame motion is best described by a plane projective homography. This is certainly untrue for most real objects, and

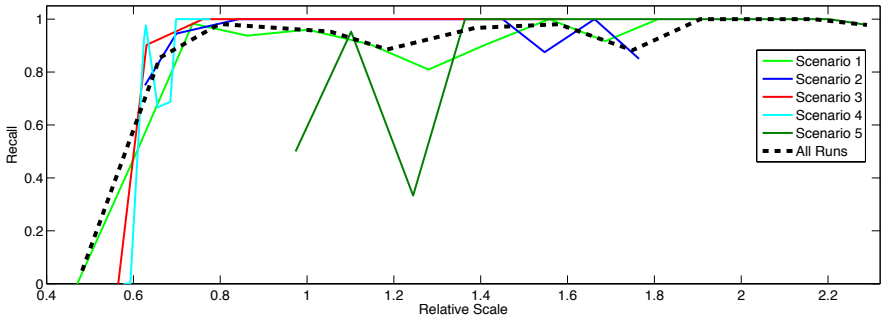


Fig. 7 Recall rates as a function of scale change for a single object across all scenarios.

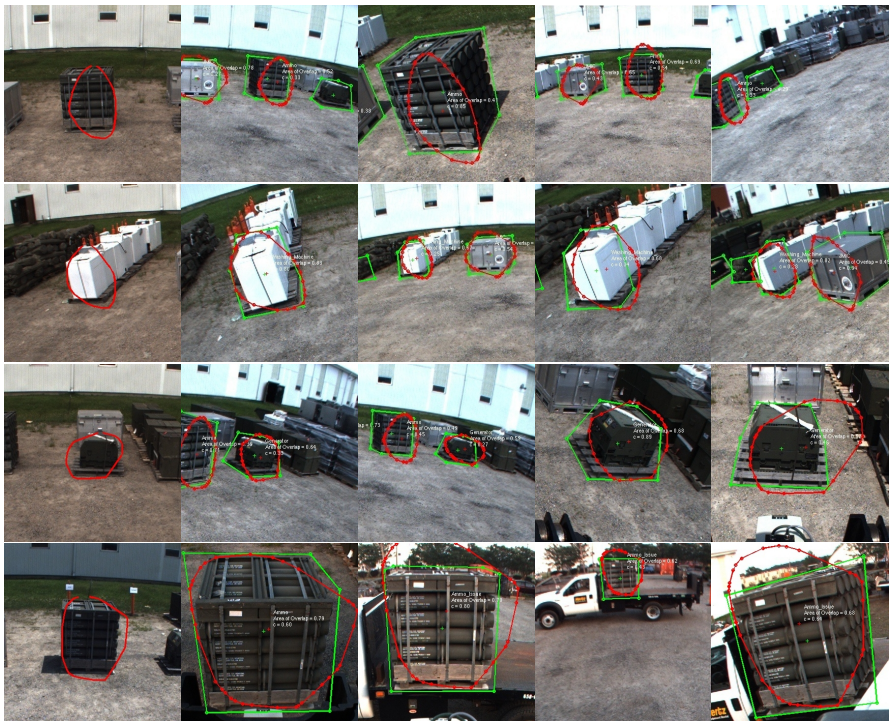


Fig. 8 Images corresponding to positive detections over variations in relative pose, lighting, and scale. Each row corresponds to a different object, with the left-most image displaying the user's initial gesture and the subsequent images displaying the reconstituted gesture (red) and the ground truth (green). Note that all images are shown at the same pixel scale.

while maintaining multiple object views does improve 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. One way to address this is to incorporate 3D information from LIDAR scans or structure-

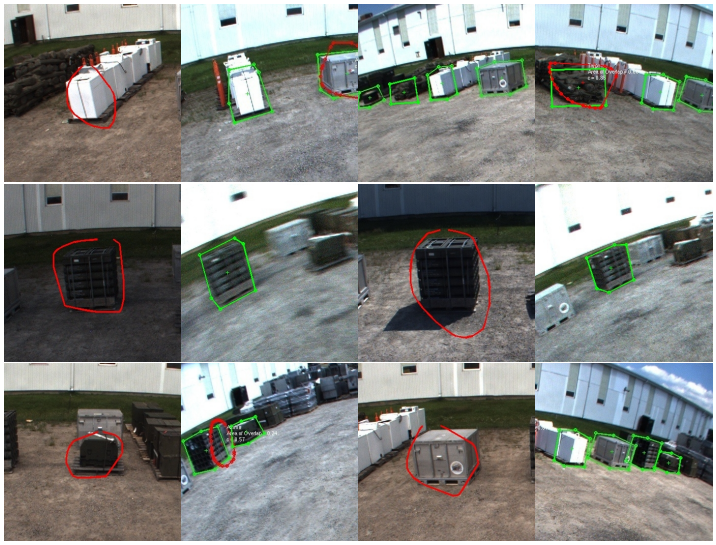


Fig. 9 Images that depict failed detections of the toured object. We have found that our algorithm fails to reacquire objects for which limited contrast due to over-saturation or poor illumination yield few keypoints. This sensitivity is exacerbated when there is a significant variation in illumination between the initial segmentation and the reacquisition phase.

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 currently 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 [12]. Robust statistical optimization via iteratively reweighted least squares could also improve both runtime performance and determinism of the approach over RANSAC-based constellation matching.

6 Conclusion

This paper presented 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, though it has obvious limitations. We are currently developing a reacquisition method that models each object's 3D structure to improve system performance and robustness.

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.

References

1. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2009)
2. Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., Chilongo, D.: Tutelage and collaboration for humanoid robots. Int'l J. of Humanoid Robotics **1**(2), 315–348 (2004)
3. Collet, A., Berenson, D., Srinivasa, S., Ferguson, D.: Object recognition and full pose registration from a single image for robotic manipulation. In: Proc. IEEE Int'l Conf. on Robotics and Automation, pp. 48–55 (2009)
4. Everingham, M., Van Gool, L., Williams, C., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. Int'l J. on Computer Vision **88**(2), 303–338 (2010)
5. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
6. Gordon, I., Lowe, D.: What and where: 3D object recognition with accurate pose. In: Toward Category-Level Object Recognition, pp. 67–82. Springer-Verlag (2006)
7. Haasch, A., Hofemann, N., Fritsch, J., Sagerer, G.: A multi-modal object attention system for a mobile robot. In: Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems, pp. 2712–2717 (2005)
8. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press (2004)
9. Hoiem, D., Rother, C., Winn, J.: 3D LayoutCRF for multi-view object class recognition and segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
10. Lowe, D.: Local feature view clustering for 3D object recognition. In: Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 682–688 (2001)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int'l J. of Computer Vision **60**(2), 91–110 (2004)
12. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
13. Savarese, S., Li, F.: 3D generic object categorization, localization and pose estimation. In: Proc. Int'l. Conf. on Computer Vision, pp. 1–8 (2007)
14. Teller, S., Walter, M.R., Antone, M., Correa, A., Davis, R., Fletcher, L., Frazzoli, E., Glass, J., How, J., Huang, A., Jeon, J., Karaman, S., Luders, B., Roy, N., Sainath, T.: A voice-commandable robotic forklift working alongside humans in minimally-prepared outdoor environments. In: Proc. IEEE Int'l Conf. on Robotics and Automation, pp. 526–533 (2010)
15. Walter, M.R., Friedman, Y., Antone, M., Teller, S.: Appearance-based object reacquisition for mobile manipulation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition Work., pp. 1–8 (2010)
16. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38**(4), 13 (2006)