Appearance-based Object Reacquisition for Mobile Manipulation

Matthew R. Walter
MIT CS & AI Lab (CSAIL)
Cambridge, MA, USA
mwalter@csail.mit.edu

Yuli Friedman, Matthew Antone
BAE Systems
Burlington, MA, USA
yuli.friedman@baesystems.com
matthew.antone@baesystems.com

Seth Teller
MIT CS & AI Lab (CSAIL)
Cambridge, MA, USA
teller@csail.mit.edu

Abstract

This paper describes an algorithm enabling a human supervisor to convey task-level information to a robot by using stylus gestures to circle one or more objects within the field of view of a robot-mounted camera. These gestures serve to segment the unknown objects from the environment. Our method's main novelty lies in its use of appearance-based object "reacquisition" to reconstitute the supervisory gestures (and corresponding segmentation hints), even for robot viewpoints spatially and/or temporally distant from the viewpoint underlying the original gesture. Reacquisition is particularly challenging within relatively dynamic and unstructured environments.

The technical challenge is to realize a reacquisition capability robust enough to appearance variation to be useful in practice. Whenever the supervisor indicates an object, our system builds a feature-based appearance model of the object. When the object is detected from subsequent viewpoints, the system automatically and opportunistically incorporates additional observations, revising the appearance model and reconstituting the rough contours of the original circling gesture around that object. Our aim is to exploit reacquisition in order to both decrease the user burden of task specification and increase the effective autonomy of the robot.

We demonstrate and analyze the approach on a robotic forklift designed to approach, manipulate, transport and place palletized cargo within an outdoor warehouse. We show that the method enables gesture reuse over long timescales and robot excursions (tens of minutes and hundreds of meters).

1. Introduction

This paper presents a vision-based approach to task-oriented object recognition that enables a mobile robot to perform long-time-horizon tasks under the high-level direction of a human supervisor. 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 periods of time and viewpoint changes, are challenging for robots that operate with imprecise knowledge of location within dynamic, uncertain environments.

Poor absolute localization precludes reliance upon the ability to build and maintain a global map of the objects of interest. Instead, our "reacquisition" strategy relies solely upon images taken from cameras onboard the robot. In this paper, we describe 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 indicates an object by circling it in an image acquired from a camera mounted to the robot. The circling gesture provides a manual segmentation of the object. The system combines image-space features in a so-called *view* to build a model of the object's appearance that it later employs for recognition. As the robot moves about the environment and the object's appearance changes (e.g., due to variations in scale and viewing direction), the algorithm opportunistically and automatically incorporates novel object views into its appearance models. We show that, by augmenting object appearance models with new views, the system significantly improves recognition rates over long time intervals and spatial excursions.

We demonstrate our reacquisition strategy on a robotic forklift that operates within outdoor, semi-structured environments. The vehicle autonomously manipulates and transports cargo under the direction of a human supervisor, who uses a combination of stylus gestures and speech to convey tasks to the forklift via a hand-held tablet. We present the results of a preliminary experiment in which the forklift is tasked with recalling a number of different objects placed ambiguously in a scene. We evaluate the

performance of the system's object recognition function, and evaluate the effect of incorporating different viewpoints into object appearance models. We conclude by discussing the method's limitations and proposing 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., [19, 12]). 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 [14] 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 [8] 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. [4] 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.

Considerable effort has been devoted to vision-based recognition to facilitate human interaction with robotic vehicles. Much of this work focuses on detecting people in the robot's surround and recognizing the faces of those who have previously interacted with the robot [1, 3, 20, 11]. Having developed an ability to detect human participants, several groups [18, 3, 13, 20, 16, 11, 2] have described vision algorithms that track body and hand gestures, allowing the participant to convey information to the robot. In addition to detecting the location and pose of human participants, 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). Our work similarly enables human participants to teach objects to the robot, using speech as a means of assigning task information. However, in our case, the user identifies objects by indicating their location on images of the scene.

Haasch et al. [9] detect hand-pointing gestures from which the region of the environment in which the object may lie is inferred. They then compare this spatial information and the user's verbal cues against a model of the scene to determine whether the referenced object is already known. Object recognition relies upon Normalized Cross-Correlation matching. If the object is thought to be new, the algorithm incorporates verbal cues to refine its location (e.g., based upon specified color) and instantiates a new model using local appearance information and information derived from user speech. As the authors note, the demonstrated system supports only a small number of objects. Furthermore, the results are limited to uncluttered indoor scenes in which the objects are in clear view of the robot, rather than outdoor, unstructured environments.

Similarly, Ghidary et al. [7] combine single-hand and two-hand gesture detections with spoken information to localize objects using depth-from-focus. Objects are then added to an absolute spatial map that is subsequently used to recall location. In contrast with our reacquisition effort, their work relies upon accurate robot localization and performs little if any vision-based object recognition.

Breazeal et al. [3] use computer vision to facilitate a robot's ability to learn object-level tasks from a human partner via social interaction. As one modality of this interaction, the work uses vision to track people and their pointing gestures, as well as to track objects in the robot's surround. This information is then used to detect participants' object-referential deictic gestures and their focus of attention as the robot learns labels of unknown objects. The authors provide an example in which a human teaches the robot the names of colored buttons, and later asks it to act on these buttons by name. The demonstrated application of the work is a stationary humanoid robot situated in an indoor environment.

2. Reacquisition Methodology

A motivation for our vision-based approach to object reacquisition is our work developing an autonomous forklift that operates in outdoor semi-structured environments typical of disaster relief and military supply chain efforts [21]. The system autonomously performs material handling tasks under the direction of a human supervisor who conveys task-level commands to the robot through a combination of speech and stylus gestures via a wireless handheld tablet

Figure 1. The robotic forklift manipulates and transports cargo under the direction of a human supervisor who uses speech and stylus gestures on a hand-held tablet to convey commands.

(Figure 1) [5]. The requirements of the target application (i.e., the system must operate in existing facilities with minimal preparation, interaction must require little training, and the interface must scale to allow simultaneous control of multiple vehicles) lead to a design goal in which increasing autonomy is entrusted to the robot. The remainder of this section describes a general strategy for vision-based object reacquisition that is consistent with this goal.

2.1. Task-Level Command Interface

The supervisor 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 within the environment. A handheld tablet interface displays live images from one of four robot-mounted cameras. The supervisor indicates a particular pallet to pick up by circling its location in one such image. Similarly, the user designates, by circling in an image, a ground or truckbed location where a pallet should be placed. The supervisor can also summon the robot to one of several named locations in the environment by speaking to the tablet.

Conveying any one of these tasks requires little effort on the part of the human participant. We have developed the robot's capability to autonomously resolve the information necessary to perform these tasks (e.g., by using its LIDARs to find and safely engage the pallet based solely upon a single gesture). Nevertheless, tasks that are more complex than that of moving a single pallet from one location to another necessitate the supervisor's periodic, albeit short, intervention until the robot has finished. Consider, for example, that the supervisor would like the robot to transport the four pallets highlighted in Figure 2 to four different storage locations in the warehouse. The aforementioned command interface requires that, for each pallet, the supervisor must:

1. Circle the desired pallet in the image.
   [*wait for the robot to pick up the pallet*]

2. Summon the robot to the desired destination.
   [*wait for the robot to navigate to the destination*]

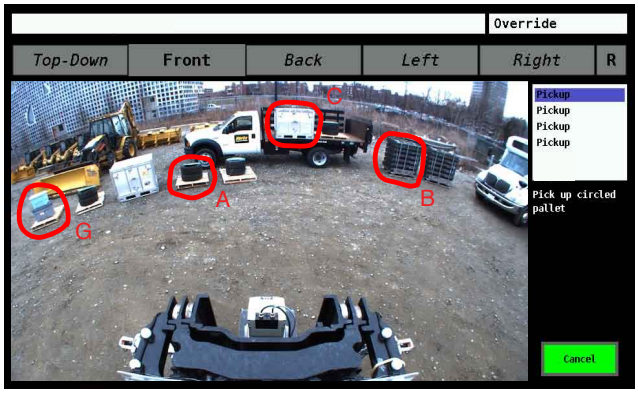


Figure 2. The tablet interface displaying a view from the robot's forward-facing camera along with user gestures (red). The ability to reacquire scene objects allows the supervisor to indicate multiple objects, and specify the intended destination of each, in rapid succession, rather than having to wait for one transport task to finish before commanding the next one.

3. Circle the location on the ground where the robot should place the pallet.
   [*wait for the robot to place the pallet*]

4. Summon the robot back to the truck for the next pallet.
   [*wait for the robot to navigate back to the truck*]

Though each one of these operations requires little effort on the part of the supervisor, the supervisor is periodically involved throughout the whole process of transporting the pallets, an operation that can take tens of minutes in a typical military warehouse. A better alternative would be to allow the supervisor to specify all four tasks at the outset, by identifying the objects in the image and speaking their destinations in rapid succession. In order to achieve this means of interaction, the robot must be capable of recognizing the specified objects upon returning to the scene.

2.2. Reacquisition

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 long sensor coverage gaps (thus alleviating the degree of interaction and attention that the human need provide).

We use our application scenario of materiel handling logistics as a motivating example: the human visually indicates pallets of interest and their intended drop-off destinations to the robot through gestures on a touch screen interface. Once the initial segmentation is provided, the robot continues to execute the task sequence, relying on reacquisition for object segmentation as needed.

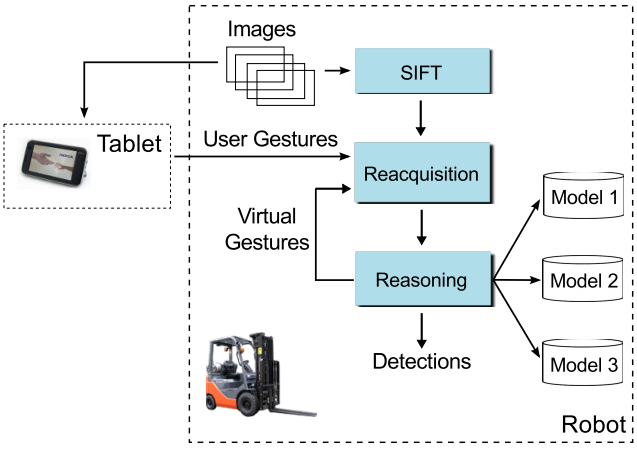


Figure 3. Block diagram of the reacquisition process.

Our algorithm maintains visual appearance models of the initially indicated objects so that when the robot returns from moving a given pallet, it can still recall, recognize, and act upon the next pallet 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* $\mathcal{M}_i$ 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).

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 around that position. The user provides the initial object segmentation by circling its location in a particular image, thereby initiating the generation of the first appearance model. Our algorithm searches each subsequent 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 views together capture 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 dense set $\mathcal{F}$ of keypoint locations and scale invariant descriptors; we use Lowe's SIFT algorithm for moderate robustness to viewpoint and lighting changes [15], but any stable image features may be used. In our logistics application, a handheld tablet computer displays current video views from the robotic forklift's onboard cameras, and the operator circles pallets of interest with a stylus. Our system creates a new model $\mathcal{M}_i$ for each indicated object. 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.

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 $\hat{\mathcal{C}}_{ijt} \in \mathcal{C}_{ijt}$, $|\bar{\mathcal{C}}_{ijt}| = 4$;
  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 $d_{pq} = |x_q - \hat{x}_p| \le t$ **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

The basic operational unit in determining whether and which models are visible in a given image is constellation matching of a single view to that image through a process outlined in Algorithm 1. 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, a set of SIFT features $\mathcal{F}_t$ is extracted from the image captured at time index t. For each view v_{ij}, our algorithm matches the view's set of descriptors $\mathcal{F}_{ij}$ with $\mathcal{F}_t$ to produce a set of point pair correspondence candidates $\mathcal{C}_{ijt}$. The similarity score metric s_{pq} between a given pair of features p and q is 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 possible similarity scores and collect in $\mathcal{C}_{ijt}$ at most one pair per feature in $\mathcal{F}_{ij}$, subject to a minimum threshold.

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) [6] with a plane projective homography H as the underlying geometric model [10]. Our particular robot employs wide-angle camera lenses that exhibit noticeable radial distortion, so before applying RANSAC, we un-distort them, 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 (undistorted) correspondences from $\mathcal{C}_{ijk}$ with which we compute the induced homography H between the current image and the view v_{ij}. We then apply H 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 prespecified 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.

The RANSAC procedure produces a single best hypothesis for v_{ij} consisting of a homography and a set of inlier correspondences $\tilde{\mathcal{C}}_{ijt} \in \mathcal{C}_{ijt}$ (Figure 4). We assign a confidence value c_{ijt} to the hypothesis that incorporates 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.

3.3. View Context

Though our system allows a region in a single image to match multiple similar-looking objects, in practice we observe that multiple hypotheses are rarely necessary, even when the scene contains identical-looking objects. The reason for this is that user gestures are typically liberal, generally containing both the object of interest and some portion of the immediately surrounding environment. While the selected object may not itself be visually distinct from other nearby objects, its context (i.e., the appearance of its support surface and background) typically provides additional discriminating information in the form of feature descriptors and constellation shape.

Figure 4. 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).

3.4. Multiple-View Reasoning

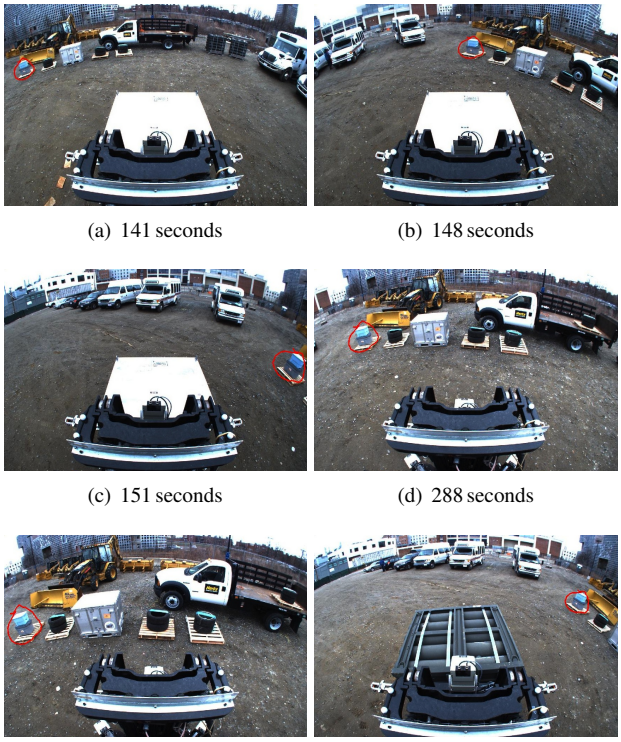
The above single-view matching procedure produces a number of match hypotheses per image and does not prohibit detecting different instances of the same object. Each object model possesses one or more distinct views, and each view can match at most one, though possibly different object in the image with 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. For each "active" model $\mathcal{M}$ (i.e., a model for which a match hypothesis has been generated), we assign the model a confidence score equal to that of the most confident view candidate. If $\tilde{c}$ exceeds a threshold, we consider this 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 hypothesis's associated match homography.

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 matching process on different models results in the same image location being matched with different objects. However, we have not found this to happen in practice, which we believe to be a consequence of the context information captured by the user gestures as discussed in Section 3.3.

3.5. Model Augmentation

As the robot moves through the environment to execute its tasks, each 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.



(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 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. Figure 5 depicts views of an object that were automatically added to the model based upon appearance variability.

When the multi-view reasoning has determined that a particular model $\mathcal{M}$ is visible in a given image, we examine all of that model's matching views v_j and determine both the robot body motion and the geometric image-to-image change between the v_j and the associated observation hypotheses. In particular, we determine the minimum position change $d_{\text{min}} = \min_j \|p_j - p_{\text{cur}}\|$ where p_{cur} is the current position of the robot and p_j is the position of the robot when the j^{th} view was captured, as well as the minimum 2D geometric change $h_{\text{min}} = \min_j \text{scale}(H_j)$ where $\text{scale}(H_j)$ determines 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 $\mathcal{M}$ 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 multi-view reasoning (Section 3.4).

4. Experimental Results

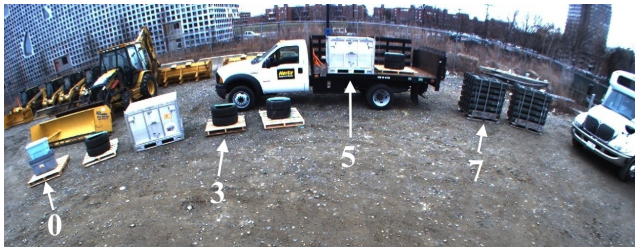


Figure 6. The experimental setup as viewed from the forklift's front-facing camera. The scene includes several similar-looking pallets and loads.

We conducted a preliminary analysis of the single-view and multiple-view object reacquisition algorithms on images collected with the forklift in an outdoor warehouse. Mimicking the scenario outlined in Section 2.1, the environment was arranged with nine loaded pallets placed within view of the forklift's front-facing camera, seven pallets on the ground and two on a truck bed (Figure 6). The pallet loads were chosen such that all but one pallet were similar in appearance to another in the scene, the one outlier being a pallet of boxes.

The experiment consisted of going through the process of moving the four pallets indicated in Figure 6 to another location in the warehouse, approximately 50 m away from

the scene. The first pallet, pallet 5, was picked up autonomously from the truck while the remainder were transported manually, in order 3, 7, and then 0. 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.

For ground truth, we manually annotated each image to include a bounding box for each viewed object. We used this ground truth to evaluate the performance of the reacquisition algorithms. A detection is deemed positive if the center of the reprojected (virtual) gesture falls within the ground truth bounding box.

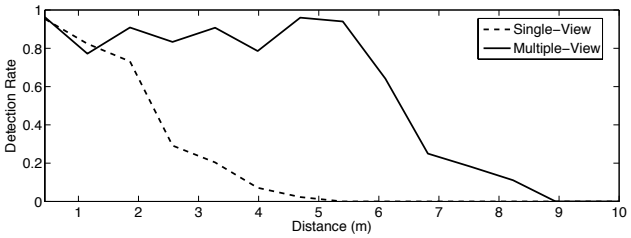


Figure 7. Probability of detection as a function of the robot's distance from the original gesture position.

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. Detection rate is expressed with respect to the ground truth annotations. 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 improvement in the multiple-view recognition rates at greater distances suggests that augmenting the model with different views of the object facilitates recognition across different 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. Pallet 3, the pallet of tires near the truck's front bumper, was visible from many different scales and viewpoints during the experiment, resulting in a particularly high number of model views.

5. Discussion

We described an algorithm for object recognition that maintains an image-space appearance model of environmental objects in order to facilitate a user's ability to command a mobile robot. The system takes as input a single coarse segmentation of an object in one of the robot's cameras, specified by the user in the form of a image-relative stylus gesture. The algorithm builds a multi-view object appearance model automatically and online, enabling object recognition despite changes in appearance resulting from robot motion.

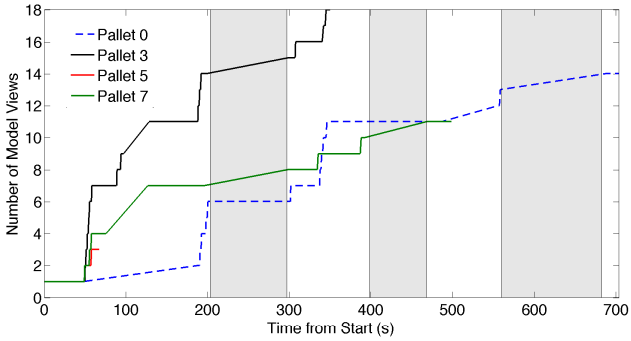


Figure 8. The number of views comprising the appearance model for each pallet, as a function of time since the original user gestures were made. Gray bands roughly indicate the periods when the pallets were not in the camera's field-of-view. Pallets are numbered from left to right in the scene.

As described, the reacquisition algorithm is in its early development and exhibits several limitations that we are currently addressing. For one, we assume that the interest points on the 3D objects culled within the (virtual) gesture are co-planar, which is not the case for most real-world objects. While maintaining different object views improves robustness to non-planarity, our homography-based matching algorithm remains sensitive to parallax, particularly when the gesture captures scenery distant from the object. One way to address these issues would be to estimate a full 3D model of the object's geometry and pose by incorporating LIDAR returns into object appearance models. A 3D model estimate would not only improve object recognition, but would also facilitate subsequent manipulation.

Additionally, our multiple-view model representation currently treats each view as an independent collection of image features and, as a result, the matching process scales poorly with the number of views. We suspect that the computational performance can be greatly improved through a "bag of words" representation that utilizes a shared vocabulary tree for fast matching [17].

The experiments described here provide only initial insights into the performance of the reacquisition algorithm. While not exhaustive, the results suggest that the contribution of an automatic multiple-view appearance model significantly improves object recognition rates by allowing the system to tolerate variations in viewing direction and scale. Finally, we plan additional experiments to better understand

the robustness of the reacquisition system to conditions typical of the unstructured outdoor environment in which the robot operates, and to evaluate the effects of such factors as lighting variation and scene clutter, particularly involving objects with nearly identical appearance.

6. Conclusion

This paper proposed an interface technique in which valuable user input can be reused by capturing the visual appearance of each user-indicated object, searching for the object in subsequent observations, and reassociating the newfound object with the existing gesture (and its semantic implications). We presented preliminary results that employ the reacquisition algorithm for a robotic forklift tasked with autonomously manipulating cargo in an outdoor warehouse. In light of these results, we discussed the reacquisition method's limitations and proposed possible solutions.

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] L. Aryananda. Recognizing and remembering individuals: Online and unsupervised face recognition for humanoid robot. In *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*, volume 2, pages 1202–1207, Lausanne, Oct. 2002.

[2] A. Bauer, K. Klasing, G. Lidoris, Q. Mühlbauer, F. Rohrmüller, S. Sosnowski, T. Xu, K. Kühnlenz, D. Wollherr, and M. Buss. The Autonomous City Explorer: Towards natural human-robot interaction in urban environments. *Int'l J. of Social Robotics*, 1(2):127–140, Apr. 2009.

[3] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo. Tutelage and collaboration for humanoid robots. *Int'l J. of Humanoid Robotics*, 1(2):315–348, 2004.

[4] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, pages 48–55, May 2009.

[5] A. Correa, M. R. Walter, L. Fletcher, J. Glass, S. Teller, and R. Davis. Multimodal interaction with an autonomous forklift. In *Proc. ACM/IEEE Int'l Conf. on Human-Robot Interaction (HRI)*, pages 253–250, Osaka, Japan, Mar. 2010.

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

[7] S. Ghidary, Y. Nakata, H. Saito, M. Hattori, and T. Takamori. Multi-modal interaction of human and home robot in the context of room map generation. *Autonomous Robots*, 13(2):169–184, Sep. 2002.

[8] I. Gordon and D. Lowe. What and where: 3D object recognition with accurate pose. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, pages 67–82. Springer-Verlag, 2006.

[9] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A multi-modal object attention system for a mobile robot. In *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*, pages 2712–2717, Edmonton, Alberta, Aug. 2005.

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

[11] M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, H. Gotoda, Y. Shirai, and H. Ueno. Adaptive visual gesture recognition for human-robot interaction using a knowledge-based software platform. *Robotics and Autonomous Systems*, 55(8):643–657, Aug. 2007.

[12] D. Hoiem, C. Rother, and J. Winn. 3D LayoutCRF for multi-view object class recognition and segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[13] J. Kofman, X. Wu, T. Luu, and S. Verma. Teleoperation of a robot manipulator using a vision-based human-robot interace. *IEEE Trans. on Industrial Electronics*, 52(5):1206–1219, Oct. 2005.

[14] D. Lowe. Local feature view clustering for 3D object recognition. In *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 682–688, 2001.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. of Computer Vision*, 60(2):91–110, Nov. 2004.

[16] K. Nickel and R. Stiefelhagen. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing*, 25(12):1875–1884, Dec. 2007.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, Washington, DC, 2006.

[18] D. Perzanowski, A. Schultz, W. Adams, E. Marsh, and M. Bugajska. Building a multimodal human–robot interface. *Intelligent Systems*, 16(1):16–21, Jan.-Feb. 2001.

[19] S. Savarese and F. Li. 3d generic object categorization, localization and pose estimation. In *Proc. Int'l. Conf. on Computer Vision (ICCV)*, pages 1–8, 2007.

[20] R. Stiefelhagen, H. Ekenel, C. Fügen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel. Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot. *IEEE Trans. on Robotics*, 23(5):840–851, Oct. 2007.

[21] S. Teller et al. A voice-commandable robotic forklift working alongside humans in minimally-prepared outdoor environments. In *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, May 2010.