Making specific features less discriminative to improve point-based 3D object recognition

Edward Hsiao, Alvaro Collet and Martial Hebert
Robotics Institute, Carnegie Mellon University, USA
{ehsiao,acollet,hebert}@cs.cmu.edu

Abstract

We present a framework that retains ambiguity in feature matching to increase the performance of 3D object recognition systems. Whereas previous systems removed ambiguous correspondences during matching, we show that ambiguity should be resolved during hypothesis testing and not at the matching phase. To preserve ambiguity during matching, we vector quantize and match model features in a hierarchical manner. This matching technique allows our system to be more robust to the distribution of model descriptors in feature space. We also show that we can address recognition under arbitrary viewpoint by using our framework to facilitate matching of additional features extracted from affine transformed model images. The evaluation of our algorithms in 3D object recognition is demonstrated on a difficult dataset of 620 images.

1. Introduction

Recognizing and estimating the 3D pose of an object from a single image has many applications in computer vision, robotics, and augmented reality. Unlike many other computer vision tasks such as object categorization and segmentation, the application of 3D recognition to robotics and manipulation requires a much higher standard of performance. Consumers will not use a robot if there is only a 90 percent chance that their objects will be found and retrieved when they ask for it. Even with these higher standards, many current state-of-the-art 3D recognition systems do not fair well under real-world conditions. In this paper, we propose two algorithms that can be used to improve many existing point-based systems regardless of the type of features or method of matching used. We validate our methods on a family of SIFT-based systems [2, 4, 9].

The general paradigm for 3D object recognition [2, 4, 21] is to first generate correspondences between image features and model features, and then to use the 3D positions associated with the matched model features to estimate the pose of an object by enforcing geometric constraints. Given a set of perfect correspondences between 3D points and 2D projections, the problem of determining the pose of a calibrated camera has been extensively studied. The main problem that remains unsolved in 3D object recognition is the problem of automatically generating enough reliable correspondences. Even though techniques like RANSAC are able to deal with incorrect correspondences, often there are just not enough correspondences to begin with. If enough correspondences are provided, recovering the pose is essentially solved and we show that the various methods for recovering pose have very similar recognition performance.

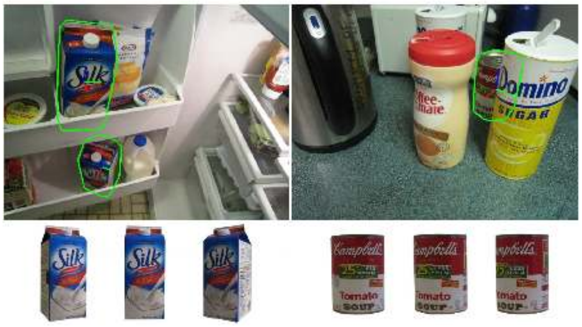

Figure 1. Examples of recognized objects with our improvements. The bottom row of images shows the closest views of the object used to generate the 3D model.

In many of the current point-based 3D object recognition systems [2, 4, 11], specific point-to-point correspondences from 3D model to 2D image are obtained initially by matching discriminative features. Much of the recent research in local image features [12, 14] has been to design descriptors that are as discriminative and robust as possible to obtain point-to-point matches. In this paper, we claim that a small amount of generalization by quantizing the descriptors can significantly improve the robustness of matching and thus the performance of specific object recognition.

One main issue that arises with manmade objects is that there are inevitably locations on the object that have similar local appearances. Features extracted from these locations have similar descriptors, and in the extreme case, the descriptors may be exactly the same. Most current

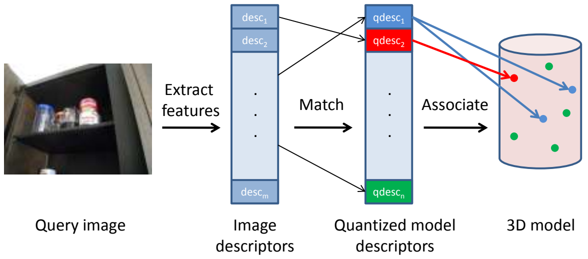

Figure 2. Quantization framework. Features are extracted from a query image and then matched to a set of quantized model descriptors. Each quantized model descriptor is associated with all of its possible locations on the 3D model, which allows similar features to be matched.

algorithms perform matching discriminatively; ambiguous matches are often discarded because they are assumed to arise from background clutter. This is exemplified by the ratio test [12], which compares the distance to the closest neighbor with the distance to the second closest neighbor. Discriminative matching prevents features with similar descriptors from being matched, even though these features contain rich information about the pose of the object. The presence of similar features is an inherent issue in matching and no amount of tuning parameters or design of local features can circumvent this problem.

In this paper, we argue that matching is more robust if we do not commit initially to specific point-to-point correspondences. Instead, if a match is ambiguous, we claim that the image feature should be associated with a set of possible locations on the model, retaining the ambiguity of the correspondence until hypothesis testing. Figure 2 illustrates our proposed framework. Given a candidate pose of an object, the correspondence ambiguity can be resolved as the one which best fits the hypothesized pose. Finally, the candidate pose with the greatest evidence after considering multiple hypotheses is chosen as the pose of the object.

We propose to maintain feature ambiguity by quantizing the features on a model. Each quantized feature is associated with a descriptor and all of its possible model locations. These quantized features are still matched discriminatively, but the quantization allows us to associate a feature on a query image with multiple locations on a model. Because retaining feature ambiguity increases the potential number of outliers, we demonstrate an efficient way to handle these additional correspondences.

Another issue that arises in the real world is that objects in unstructured environments can appear in any orientation and position, often significantly different from the images used to train the model. Accounting for all possible viewpoints is infeasible, yet a 3D recognition system must still recover the object pose given a finite set of training images. In the past, this has been addressed by using affine invariant features [13], affine invariant patches [21] and view clustering [11]. Here we take the approach of simulating novel viewpoints [6, 8, 15, 18] and adding features extracted from affine transformed training images to our model. One problem with this approach is that the number of features on the model increases significantly, with many features having similar descriptors. We show that our quantization framework facilitates matching to these features and that handling viewpoint in this way can significantly increase the performance of 3D object recognition.

2. Preserving ambiguity by quantization

Vector quantization of features has been used widely in the computer vision literature for categorization tasks such as scene recognition [20] and object categorization [28]. Many of the algorithms used for these tasks fall in the realm of the Bag of Words approach, where a dictionary of visual features is learned through clustering and new images are categorized by comparing histograms of quantized visual words. In these cases, quantization is used as a way to generalize and be robust to intra-category variations.

Most related to our work are methods that employ geometric reasoning on visual words [1, 26] for image retrieval and category recognition. However, for the task of specific 3D object recognition, the prevailing view is to use highly discriminative features. As a result, multiple features with similar appearance on a model are rarely matched.

In this paper, we claim that these similar features are essential to obtain reliable 3D object recognition. We introduce ambiguity into the matching process by quantizing the model features and associating each quantized descriptor with potentially multiple locations on the model. When an image feature is matched to a quantized feature, it is associated with all the possible locations of that feature. During hypothesis testing, the most likely correspondence given the current pose can then be determined. Our framework allows us to choose the most likely hypothesis given what we have seen, and combines both ambiguous and unique features in a unified framework.

2.1. Hierarchical mean-shift quantization

In general, it is very difficult to choose the number of feature clusters *a priori* as different models have different number of features and degree of feature similarity. We choose the mean shift algorithm because it clusters features based on the similarity of the descriptors in feature space. The bandwidth parameter of mean shift is a rough indication of the desired intra-cluster variation and is more relevant to set than the number of clusters.

In our implementation, we use a dual-bandwidth approach where features are quantized in a hierarchical manner [17] using two levels of mean shift with bandwidths r_1 and r_2, such that $r_1 < r_2$. Clustering in this way allows matching to be more robust to the distribution of descriptors

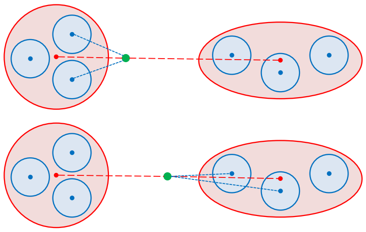

Figure 3. Example of quantized matching where (top) a query feature in green matches at a coarser level, but does not match at a finer level by the ratio test and on (bottom) the vice versa is true.

in feature space. Our quantization scheme results in three levels of quantized features, where the finest level $l = 0$ corresponds to the original features. Each quantized feature q_i^l at level l is then associated with a set of 3D positions on the model corresponding to all the features in that cluster.

2.2. Discriminative hierarchical matching (DHM)

For the task of image retrieval, a common technique is hierarchical matching on vocabulary trees [17] which assigns a visual word to every image feature. However, for the task of object recognition, most image features arise from background clutter. Assigning a visual word to every image feature can increase the number of outlier correspondences by orders of magnitude, making RANSAC intractable. We propose to perform discriminative matching on each level of the hierarchy to limit matches to background clutter.

Candidate correspondences are obtained by independently matching the image features with each of the three levels of features. A feature is matched on a particular level if it satisfies the ratio test within that level. The matched image feature is then associated with all the possible 3D locations of its corresponding quantized model feature. The final set of correspondences is obtained by aggregating all candidate correspondences at all levels, removing any duplicate point correspondences.

Figure 3 illustrates the reason for our choice of hierarchical clustering. In Figure 3 (top), the query feature in green is equidistant to the centers of the two fine clusters (blue), but it is significantly closer to the coarse cluster (red) on the left than the coarse cluster on the right. At this stage, it is impossible to disambiguate the correspondences in the two fine clusters, so the quantized matching returns all the candidate locations of the coarse cluster on the left for later processing. Conversely in Figure 3 (bottom), the query feature is equidistant to the centers of the two coarse clusters, but will match at the fine cluster level. If there were only one level of clustering, one of these two situations would result in no correspondence.

2.3. View-constrained RANSAC

Quantized matching drastically increases the number of outliers as all potential locations on the model for a particular quantized feature that do not correspond to the actual location are incorrect. This is a significant issue as the number of iterations of RANSAC needed to guarantee a consistent set of inliers increases dramatically with the number of times a feature is repeated. If each feature is repeated α times, then approximately α^n times more iterations are needed to guarantee the same level of performance from RANSAC, where n is the sample size.

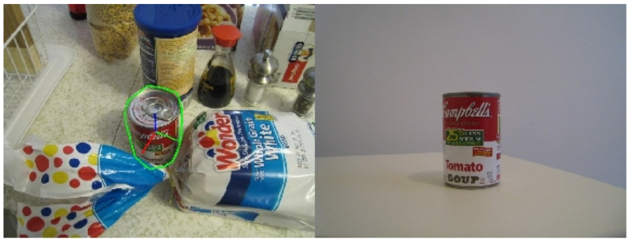

Figure 4. Example of tomato soup can recognized (left) at a viewpoint significantly different from the closest model view (right).

Prior to the advent of highly discriminative locally invariant features, such as SIFT, local features were mostly shape-based and very ambiguous (e.g., corners, high curvature points, curve inflections). Given that one-to-one matching was infeasible, it was not uncommon for the co-visibility [19] of model features to be used as a constraint to reduce the search space. This constraint avoided attempting to estimate an object's pose from a set of features that were not simultaneously visible on the model. In early literature on the topic, methods such as interpretation trees [5], Hough transforms [5], alignment [7] and grouping [10] were used to address feature ambiguity.

In this paper, we introduce a modified version of RANSAC, termed view-constrained RANSAC, to again exploit the co-visibility of model features. In practice, this is implemented by maintaining the set of views for which each point is visible when generating the 3D models. We will refer to the set of cameras for which a point P_i is visible in as its view set V_i. The view-constrained RANSAC algorithm begins by choosing a correspondence $C_{i,j}$ between an image point p_i and a model point P_j at random from the set of candidate matches, C. Only points P_k with a view set V_k that overlaps with the view set of the selected model point P_j are retained. The view-constrained set of correspondences $C_{vc(j)}$ for a model point P_j is defined as

$$C_{vc(j)} = \{C_{i,k} : V_j \cap V_k \neq \emptyset \wedge k \neq j\} . \qquad (1)$$

The remaining $n - 1$ points needed to generate a pose hypothesis are then selected at random without replacement from the view-constrained set of points $C_{vc(j)}$. The process is repeated for a fixed number of iterations and the pose with the greatest consistent evidence is selected.

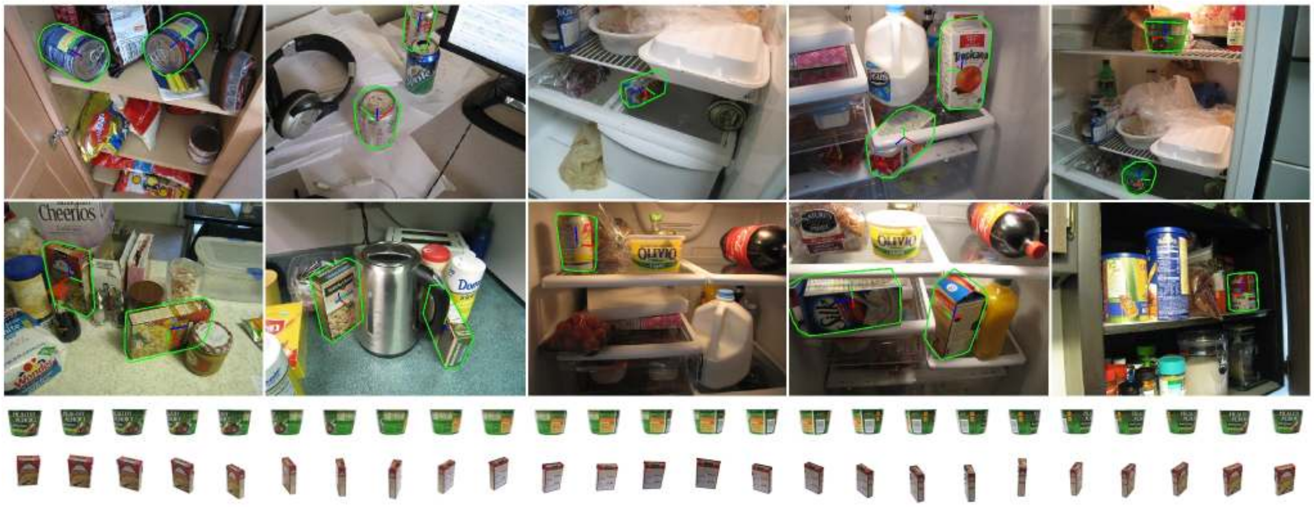

Figure 5. Example detections from our challenging dataset. The images were taken in cluttered environments with different lighting conditions and with the objects under various viewpoints and occlusions. The bottom two rows show the views used to generate the models for two objects.

3. Viewpoint Change

In unstructured environments, objects may appear in an image with a viewpoint significantly different from the images used to generate the 3D models. A naïve solution to this problem is to incorporate images of the object from all possible viewpoints, although densely sampling the view space would require a very large number of images.

A more tractable approach to account for viewpoint change is to simulate novel viewpoints by applying affine or perspective transformations to the model images [6, 8, 18]. Viewpoint simulation has been used to determine a keypoint's repeatability [6] and to model a keypoint's local appearance [8, 18]. Recently, Morel *et al.* [15] demonstrated that directly matching features extracted from these simulated viewpoints significantly outperformed the state-of-the-art affine invariant features [13] under large viewpoint change. Matching is performed by extracting features from a finite set of affine transformations of both model and query images and then comparing all sets of features.

Our approach is inspired by Morel *et al.* and incorporates features extracted from affine warped images onto the 3D models. An affine transformation A can be decomposed as

$$A = \lambda R(\psi) \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} R(\phi), \quad (2)$$

using Singular Value Decomposition (SVD), where $R(\psi)$, $R(\phi)$ are rotation matrices, $\lambda > 0$, and $t \geq 1$. In this decomposition, λ corresponds to the zoom and $R(\psi)$ corresponds to the planar rotation of the camera. For the case of SIFT-based systems, we can ignore these terms as SIFT features are both scale and rotation invariant. However, other types of features may require sampling the whole space of transformations. The remaining terms in the decomposition correspond to the camera viewpoint, where $t = \frac{1}{\cos(\theta)}$ is the tilt of the camera and ϕ is the longitude angle.

We consider tilts of $t = \{1, \sqrt{2}, 2\}$ corresponding to latitude angles of $\theta = \{0, 45, 60\}$ degrees in our implementation. For each t, we follow Morel *et al.* and sample the longitude angles ϕ by an arithmetic series $\phi = \{0, b/t, ..., kb/t\}$ for $b = 72$ degrees and $k = \lfloor 180 \cdot t/b \rfloor$. Each pair $\{t, \phi\}$ specifies an affine transformation $A_{t,\phi}$ which we use to transform a model image I:

$$I_{t,\phi}(x, y) = I(A_{t,\phi}(x, y)). \quad (3)$$

From the affine transformed image $I_{t,\phi}$, we extract SIFT features and compute the locations of each keypoint $p^i = A^{-1}_{t,\phi} p^i_{t,\phi}$ on the original image. We refer to these features as simulated affine (SA) features.

A problem that arises with using SA features is that the total number of features on the model may increase by an order of magnitude or more. A typical model with SA features contains tens of thousands of features, many of which have similar descriptors. Pruning these features is very difficult because there is no clear metric as to when two features are similar enough to remove one of them. Our quantization framework facilitates matching to similar features and results in a seamless integration of the SA features into a recognition system.

4. Experimental Results

In order to validate our method's performance in feature-based object recognition, two sets of experiments were conducted. The first set evaluates our algorithm's ability to recognize objects in images, while the second set evaluates the algorithm's accuracy in recovering the full pose (3D position and orientation) of objects in images. Given that our

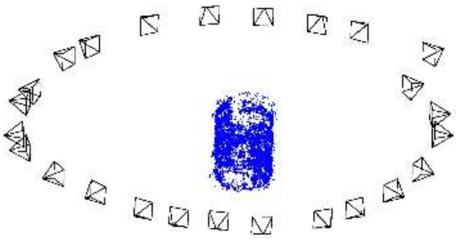

Figure 6. 3D model of the tomato soup can from 25 images.

methods can be easily used to extend any point-based 3D object recognition algorithm, we use three state-of-the-art algorithms (Gordon and Lowe [4], EPnP [9], and Collet *et al.* [2]) as our baseline systems. We incorporate the SA features and quantization separately (SA, Q) and together (SA+Q) to show their performance gains in complex scenes.

4.1. Data set

Many object recognition algorithms work very well in controlled conditions, but fail when faced with real-world scenes with strong illumination and viewpoint changes, occlusions, clutter, and where many instances of the same object might be present. Existing 3D recognition datasets [3, 16] have a large number of objects, but are mainly comprised of objects on monotone backgrounds. For evaluation under a more natural setting, the dataset we collected consists of common household objects in real, cluttered environments under different lighting conditions, occlusions and viewpoints. Household objects and environments are of importance in related fields such as robotics, due to major renewed interest [22, 24] in enabling mobile manipulators to perform useful tasks in unstructured human environments. Regular household objects often contain repeated patterns, logos, text, and are often seen many at a time. These issues cause most object recognition and pose estimation systems to not achieve success rates required for robotic manipulation.

Our dataset contains 10 household objects and 62 images per object, for a grand total of 620 images. For each object, three types of images were taken. 25 images contain one instance of the object, and 25 images contain two instances, both with their ground truth marked as regions and ID within the image. Finally, 12 more images were collected in a calibrated setup and their full 6D poses were ground truthed. The full dataset is available online[1] and a few examples are shown in Figure 5.

4.2. Base systems

The 3D object recognition systems used as baselines in our evaluation are those of Gordon and Lowe [4], EPnP [9], and Collet *et al.* [2]. All of these systems use sparse 3D models of objects with SIFT features for recognition and share a common methodology which we summarize here. The goal of these systems is to estimate a transformation $M = [R, t]$ of a 3D model with respect to the camera frame for each object class instance in the image. This is accomplished by minimizing the sum of reprojection errors between the set of N projected 3D points $\mathbf{P}$ from the model and the set of N 2D points in the image, $\mathbf{p}$. The optimal transformation M^* is defined as:

$$M^* = \arg\min_M \sum_{i=1}^{N} d(\mathbf{p}_i, M\mathbf{P}_i)^2 \qquad (4)$$

The 3D models used in this paper are created with a standard Structure from Motion [25] algorithm from 25 images taken at approximately equally spaced intervals in a circle around each object, as shown in Figure 6. Every 3D point on the model is associated with a corresponding SIFT descriptor. Finally, proper alignment and scale for each model are computed to match the real object dimensions.

When using SA features, we augment the basic 3D model by first extracting SA features from each of the model images. Then, using the estimated camera geometry, we search for correspondences of each SA feature along the epipolar lines in the nearby views. These correspondences are used to triangulate the SA features onto the 3D model.

When incorporating quantization, we use quantized descriptors (Section 2.1) and replace the ratio test with quantized matching (Section 2.2) and RANSAC with view-constrained RANSAC (Section 2.3).

4.2.1 Gordon and Lowe [4]

Gordon and Lowe introduced a fast 3D scene recognition algorithm, which we modify to recognize objects. The algorithm extracts SIFT features from the input image and matches against each object model using the ratio test to obtain a set of candidate 3D-2D correspondences $\mathbf{P} \leftrightarrow \mathbf{p}$. Using RANSAC, a random subset of n points is chosen and used to estimate a pose hypothesis by minimizing the reprojection error with Levenberg-Marquardt. If the number of points consistent with the pose hypothesis is higher than a threshold, a new object instance is created and the pose is refined using all consistent points. This procedure is repeated until the number of unallocated points is lower than a threshold, or the maximum number of iterations has been exceeded.

4.2.2 Enhanced PnP [9]

Enhanced PnP is a non-iterative, $O(n)$ solution to the PnP problem which does not require any initialization and is much faster than standard iterative minimization techniques. The EPnP 3D recognition system we created is similar to that of Gordon and Lowe, but instead of using Levenberg-Marquardt, we use the EPnP algorithm.

[1] http://www.cs.cmu.edu/∼ehsiao/3drecognition/

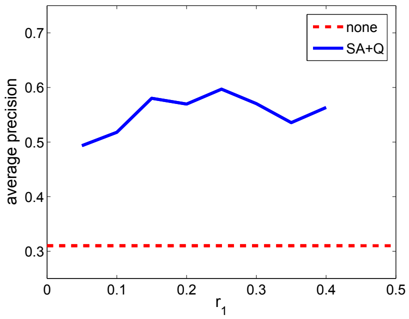

Figure 7. Effect of different bandwidths on the average precision for the orange juice carton using the Collet *et al.* system. We vary the smaller bandwidth, r_1, and choose the larger bandwidth to be $r_2 = 1.5r_1$. There is no significant change in performance for $r_1 \in [0.15, 0.30]$, and even over the entire range, we obtain substantial improvement over the baseline system.

4.2.3 Collet *et al.* [2]

The algorithm introduced by Collet *et al.* improves on the Gordon and Lowe method by combining RANSAC and mean shift clustering on the set of 3D-2D correspondences. This combination allows for a real-time solution of the correspondence problem, even when there are many instances of the same object present. After extracting 3D-2D correspondences from a new image, the 2D locations $\mathbf{p}$ are clustered using the mean shift algorithm. Each cluster of points p^k is then processed independently by running the Gordon and Lowe pose estimation described in Section 4.2.1. Finally, all detected instances from different clusters with similar estimated pose are merged together, and the instances with the most consistent points survive.

4.2.4 Parameters

The parameters for our experiments were calibrated on images not in the dataset and were kept constant for every system and every object. The mean shift cluster bandwidths used for feature quantization were $r_1 = 0.2$ and $r_2 = 0.3$, although the exact choice has little impact on the overall performance of the system (Figure 7). For matching, we choose a ratio test threshold of 0.8. We also restrict image features to have at most 10 model correspondences in the view-constrained RANSAC to maintain tractability. The evaluation on this dataset was performed only once.

4.3. Object Detection

We first evaluated the performance of each system for object detection. For each detection in an image, we project all the points of the corresponding model onto the image using the recovered pose and calculate the region A inside the convex hull. We use the region overlap criterion [27]

$$\frac{A \cap A_{gt}}{A \cup A_{gt}} > 0.5, \tag{5}$$

between the region A and the ground truth segmentation A_{gt} to determine if an object is correctly detected.

Gordon and Lowe	none	SA	Q	SA+Q
Clam chowder can	0.36	0.56	0.46	**0.79**
Diet coke can	0.09	0.07	0.04	**0.23**
Juice box	0.37	0.44	0.44	**0.71**
Orange juice carton	0.28	0.44	0.33	**0.53**
Pot roast soup	0.32	0.18	0.53	**0.79**
Rice pilaf box	0.63	0.81	0.56	**0.81**
Rice tuscan box	0.50	**0.66**	0.47	0.62
Soy milk can	0.07	0.05	0.14	**0.39**
Soy milk carton	0.44	0.46	0.44	**0.66**
Tomato soup can	0.48	0.48	0.45	**0.72**
Average	0.35	0.41	0.39	**0.62**
EPnP	none	SA	Q	SA+Q
Clam chowder can	0.36	0.52	0.49	**0.79**
Diet coke can	0.08	0.07	0.05	**0.23**
Juice box	0.27	0.35	0.43	**0.73**
Orange juice carton	0.27	0.30	0.27	**0.53**
Pot roast soup	0.32	0.19	0.54	**0.73**
Rice pilaf box	0.60	0.71	0.41	**0.81**
Rice tuscan box	0.45	0.56	0.40	**0.64**
Soy milk can	0.04	0.08	0.17	**0.39**
Soy milk carton	0.28	0.39	0.52	**0.64**
Tomato soup can	0.40	0.55	0.46	**0.75**
Average	0.31	0.37	0.37	**0.62**
Collet *et al.*	none	SA	Q	SA+Q
Clam chowder can	0.37	0.43	0.78	**0.92**
Diet coke can	0.12	0.04	0.28	**0.51**
Juice box	0.33	0.44	0.66	**0.87**
Orange juice carton	0.31	0.48	0.39	**0.61**
Pot roast soup	0.32	0.21	0.67	**0.81**
Rice pilaf box	0.61	0.76	0.71	**0.96**
Rice tuscan box	0.49	0.60	0.51	**0.80**
Soy milk can	0.06	0.03	0.27	**0.57**
Soy milk carton	0.36	0.46	0.63	**0.88**
Tomato soup can	0.45	0.47	0.76	**0.92**
Average	0.34	0.39	0.57	**0.78**

Table 1. Average precision by object for the three base systems: (top) Gordon and Lowe, (middle) EPnP, and (bottom) Collet *et al.* We demonstrate the improvements of simulated affine features (SA), quantization (Q), and the combination of the two (SA+Q).

Figure 8 shows the averaged Precision/Recall plots for the three baseline systems. To summarize the performance of all the objects for each baseline system, we use the Average Precision corresponding to the area underneath the Precision/Recall curve. The results are shown in Table 1.

From the table, the performance of the baseline systems is very similar when none of our algorithms are incorporated. EPnP and the Gordon and Lowe system show similar performance gains when augmented with the proposed methods, suggesting that matching has a larger impact on the performance of 3D recognition than the particular choice of pose estimation algorithm. Collet *et al.*'s system, which combines RANSAC and mean shift clustering, shows further improvement once SA features and quantization are added. The use of mean shift clustering in conjunction with RANSAC reduces the outlier-inlier ratio in each cluster, and makes RANSAC more tractable with the significant increase in correspondences added by our algorithms. Some example detections are shown in Figure 5.

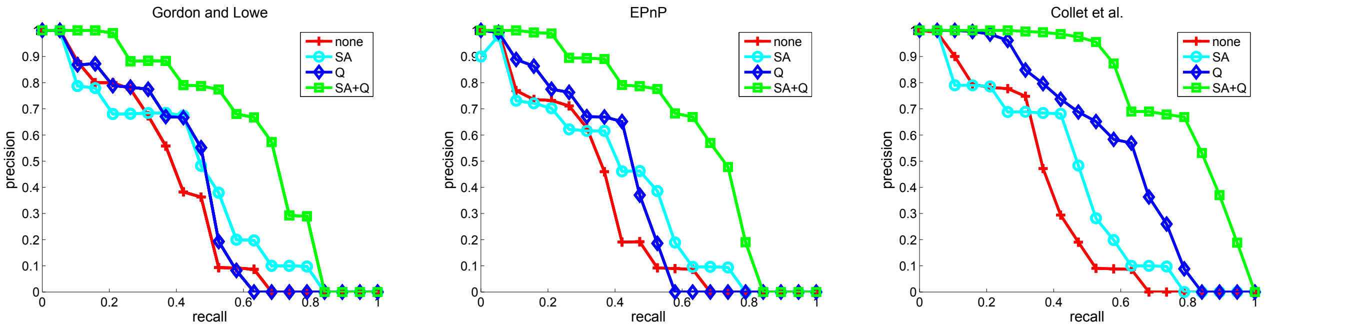

Figure 8. Averaged Precision/Recall plots: (left) Gordon and Lowe, (center) EPnP, and (right) Collet *et al.* For each plot, we show the improvements from simulated affine features (SA), quantization (Q) and the combination of the two (SA+Q).

Correct detections (%)	none	SA	Q	SA+Q
Gordon and Lowe	63	80	69	88
EPnP	61	73	70	88
Collet *et al.*	65	75	74	92

Table 2. Detections (%) within 5 cm and 22.5 degrees of the true pose. SA+Q gives significant improvement over the baseline.

Translation error (cm)	none	SA	Q	SA+Q
Gordon and Lowe	1.17	1.31	1.10	1.12
EPnP	1.19	1.20	1.15	1.13
Collet *et al.*	1.22	1.30	1.12	1.18
Rotation error (degrees)	none	SA	Q	SA+Q
Gordon and Lowe	4.59	5.25	4.77	5.17
EPnP	4.73	5.66	4.47	5.18
Collet *et al.*	4.84	5.04	4.87	5.31

Table 3. Translation error in cm (top) and rotation error in degrees (bottom) for the correct detections. SA+Q approximately maintains the accuracy while improving the recognition rate.

The objects which show the most improvement, as expected, are the objects with repeated patterns (e.g., diet coke can, soy milk can). Recognizing some of these objects is already very difficult, as they have particularly few features. With repeated patterns on them as well, most systems are unable to generate enough correspondences to estimate a reliable pose. Our improvements on the Collet *et al.* system increase the performance of the diet coke can by over four times and that of the soy milk can by over nine times.

The remaining objects which do not have repeated patterns also benefit significantly from the addition of SA features and quantization, doubling the performance for almost all the objects compared to the Collet *et al.* system. Objects such as the juice box and the pot roast soup have large regions where there is tiny text. Given that these regions look similar locally, most systems cannot find enough unique correspondences in these areas. Quantization addresses this issue and uses these features for pose estimation.

Figure 9 shows example failures of the system. The first two images are false positives due to a planar pose ambiguity described and addressed in [23]. In the center image, the system detects the repeated pattern on the object correctly, but chooses the wrong side of the object because it fails to incorporate matches from other sides of the object. Finally, the last two images show examples where the objects were not detected. In these images, the lighting conditions and viewpoints are too different from the images used to generate the model and there are not enough correct matches to estimate a pose.

4.4. Pose accuracy

In this section we evaluate the accuracy of the pose recovered from the recognition systems. To conduct this experiment, we extrinsically calibrated a camera and ground truthed the objects in 12 poses each. For 8 of the object poses, we placed the object at 0.5 m from the camera and rotated it standing upright at intervals of 45 degrees. The remaining 4 poses were with the object lying on the table and were rotated at 90 degree intervals.

We evaluate the pose for both rotation and translation error. We compute the translation error as the Euclidean distance and the rotation error as the quaternion angle $2\arccos(q^T q_{gt})$ from the ground truth pose. For this set of experiments, we measure the translation error on the plane of the table and consider the error of objects which were detected within 5 cm and 22.5 degrees of the true pose.

Table 2 shows the percentage of correct detections for each of the systems. Out of 120 total experiments per system, the baseline systems retrieved less than two-thirds correctly. SA features and quantization boosted recognition rate to close to 90 percent for each of the systems. It is worth mentioning that some of the instances that were not detected correspond to poses where only the repeated pattern is visible; in these cases, it is impossible even for a human to disambiguate.

Table 3 shows the average translation error in cm and average rotation error in degrees. Despite the average rotation error being slightly higher with our proposed methods, this error of less than a degree is well within the uncertainty of the manual ground truth. Importantly, we were able to achieve a higher recognition rate while maintaining essentially equivalent pose accuracy.

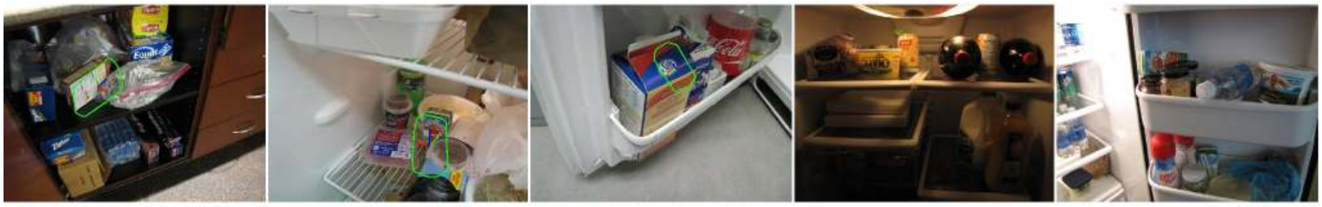

Figure 9. Examples of misdetection with Collet *et al.* and SA+Q. In the first two images, the point matches are on only one side of the object, resulting in a planar pose ambiguity. For the third image, the system finds a repeated pattern on the wrong side of the object. In the last two images, the system does not find the objects due to significant lighting and viewpoint changes from the model training images.

5. Conclusion

The main contribution of this paper is to show that not committing to specific point-to-point correspondences until the hypothesis verification step can significantly improve the performance of recognition. We develop a framework in which features are quantized and matched in a hierarchical manner. To maintain the tractability of RANSAC, we propose a view-constrained RANSAC method to reduce the ratio of potential outliers to inliers. We show that incorporating features from affine transformed images is a way to address viewpoint change and that matching to these features is facilitated by the quantization framework. Our results on a difficult dataset demonstrate that quantization combined with SA features can significantly improve the performance of current state-of-the-art 3D recognition systems.

Acknowledgements

This work was supported in part by the National Science Foundation under ERC Grant No. EEEC-0540865. Edward Hsiao is partially supported by a National Science Foundation Graduate Research Fellowship. Alvaro Collet is partially supported by a Caja Madrid Fellowship.

References

[1] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[2] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *ICRA*, 2009.

[3] J. Geusebroek, G. Burghouts, and A. Smeulders. The Amsterdam library of object images. *IJCV*, 61(1), 2005.

[4] I. Gordon and D. G. Lowe. What and where: 3d object recognition with accurate pose. In *Toward Category-Level Object Recognition*, 2006.

[5] W. Grimson, T. Lozano-Pérez, and D. Huttenlocher. *Object recognition by computer*. MIT Press, 1990.

[6] S. Hinterstoisser, S. Benhimane, and N. Navab. N3m: Natural 3d markers for real-time object detection and pose estimation. In *ICCV*, 2007.

[7] D. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *IJCV*, 5(2), 1990.

[8] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 28(9), 2006.

[9] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *IJCV*, 81(2), 2009.

[10] D. G. Lowe. *Perceptual organization and visual recognition*. PhD thesis, Stanford University, 1984.

[11] D. G. Lowe. Local feature view clustering for 3D object recognition. In *CVPR*, 2001.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.

[13] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1), 2004.

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10), 2005.

[15] J. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2), 2009.

[16] S. Nene, S. Nayar, and H. Murase. Columbia object image library (COIL-100). *Technical Report CUCS-006-96, Columbia University*, 1996.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[18] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *PAMI*, 99(1), 2009.

[19] H. Plantinga and C. Dyer. Visibility, occlusion, and the aspect graph. *IJCV*, 5(2), 1990.

[20] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44(19), 2004.

[21] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *CVPR*, 2003.

[22] A. Saxena, J. Driemeyer, and A. Ng. Robotic Grasping of Novel Objects using Vision. *IJRR*, 27(2), 2008.

[23] G. Schweighofer and A. Pinz. Robust pose estimation from a planar target. *PAMI*, 28(12), 2006.

[24] S. Srinivasa, C. Ferguson, D. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M. VandeWeghe. HERB: A Home Exploring Robotic Butler. *Autonomous Robots*, In Press.

[25] R. Szeliski and S. Kang. Recovering 3D shape and motion from image streams using non-linear least squares. In *CVPR*, 1993.

[26] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *CVPR*, 2006.

[27] G. Willems, T. Tuytelaars, and L. Gool. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In *ECCV*, 2008.

[28] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.