A Fast Multi-model Approach for Object Duplicate Extraction

Paolo Piccinini*, Andrea Prati+, Rita Cucchiara*
* D.I.I. Modena - + Di.S.M.I. Reggio Emilia - Univ. of Modena and Reggio Emilia
{paolo.piccinini, andrea.prati, rita.cucchiara}@unimore.it

Abstract

This paper presents an innovative approach for localizing and segmenting duplicate objects for industrial applications. The working conditions are challenging, with complex heavily-occluded objects, arranged at random in the scene. To account for high flexibility and processing speed, this approach exploits SIFT keypoint extraction and mean shift clustering to efficiently partition the correspondences between the object model and the duplicates onto the different object instances. The re-projection (by means of an Euclidean transform) of some delimiting points onto the current image is used to segment the object shapes. This procedure is compared in terms of accuracy with existing homography-based solutions which make use of RANSAC to eliminate outliers in the homography estimation. Moreover, in order to improve the extraction in the case of reflective or transparent objects, multiple object models are used and fused together. Experimental results on different and challenging kinds of objects are reported.

1. Introduction

Object extraction and localization are crucial tasks for several computer vision applications, such as object recognition [2], image/video retrieval [9], automatic robot navigation [13], or pick-and-place for industrial applications [11]. While in image/video retrieval the basic objective is to identify the single best instance of the target/query object, in the case of pick-and-place applications, for example, the aim is not limited to count and classify the first (or best) instance, but to determine the locations, orientations and sizes of all (or most of) the duplicates/instances. Object duplicates can have different sizes, poses and orientations, and they can be seen from different viewpoints and under different illuminations. Thus, their extraction can be very challenging, especially accounting for partially-visible objects.

Most of the picking systems consider the case of well separated objects, well aligned on the belt and with a synchronized grasping of the objects. However, there are several applications in which this approach will be insufficient, since forcing the objects to stay well separated and aligned on the belt will waste space and time of the process. Moreover, there can be objects which need to and are convenient to be kept in bins, for saving time and/or for hygienic reasons, as shown in Fig. 1.

(a) (b)

Figure 1. Examples of complex situations.

This system addresses object duplicate extraction under the following challenging requirements: (i) there is no assumption on the type and complexity of the target objects, including reflective or semi-transparent objects which may cause false detections; (ii) the required working speed is very high; a fast extraction technique should be adopted to work more than a hundreds of objects per minute; (iii) objects can be severely occluded (see Fig. 1); the system should also work in extreme conditions, such as in the case of *bin picking* applications [11].

Sometimes the matching between image and 3D CAD models is provided; this way is often not affordable since 3D models are not always available and their acquisition is expensive and time consuming; moreover, time constraints would make unsuitable complex fitting of 3D models, which also need to be invariant to projective 3D transformations.

The approach described in this paper is meant to tackle all these points by proposing a feature-based segmentation technique capable to segment multiple occluded objects. When objects are complex, reflective, low-textured and heavily occluded, very few distinctive feature points can be extracted from the image. Having few features to be matched with the original sample, segmentation of multiple

instances of the object is not straightforward. This approach uses very simple features, based on single-point SIFT [10] feature detector, which has proven to be robust and sufficiently general. Then, this work proposes the use of a voting scheme to cluster the matched feature points among the different instances. Once clustered, these points vote for *principal points* of the object, i.e. points which characterize and delimit the object shape, such as, for instance, the four corners in the case of a rectangular-like shape. These points can be used as both delimiters of the object for the segmentation and picking points, depending on the end effector of the robot. In addition, to improve the accuracy in the case of low-textured, reflective or semi-transparent objects (in which few "strong" matches are found), multiple models of the target object can be used.

2. Related Works

Object duplicate extraction can be described with three main phases. The first aims at defining and computing a proper similarity measure between the target object (or part of it) and the object duplicates in the image, and two state-of-the-art techniques have been proposed. The first is the *Bag of Words* (BoW) model, which is based on the histogram of local features [12]. However, since the BoW model is based on histograms, its main weakness relies in the lack of spatial information, which makes it unreliable in cluttered scenes, especially when object duplicates are present. Conversely, the *Part Based Model* considers spatial information of the local features (like in the "Star Model" proposed in [1]).

The second phase exploits the similarity measure to locate the duplicates. The basic idea is to find matches between the rough model of the object expressed in terms of local visual features (directly extracted by a sample image of it, and not by complicated synthetic models), and the current image. For instance, affine covariant regions provide a set of points invariant to scale, rotation and translation. Local descriptors, such as SIFT (Scale Invariant Feature Transform) [10] are extracted and the Generalized Hough Transform or a probabilistic model [7] can then be applied in order to localize the position of the objects. In fact, as demonstrated in [5], this approach is prone to errors in those applications where images may lack enough distinctive features. To overcome to this problem, Hess and Fren [5] propose to further improve the registration between two images by exploiting, together with the SIFT-based distrinctive features, a refined local set of features in which the SIFT matching criteria are applied on a region centred on the global keypoints. Some proposals exploit local features to locate objects [6], but since they use very specific features (such as round holes), they cannot be easily extended for whichever type of object.

Eventually, the third phase starts from object location (e.g. the object's center) to segment the whole shape as accurate as possible and has been rarely adopted in the literature, since the need for obtaining the extraction of the whole shape exists only in those applications where object's encumbrance must be estimated.

Other works follow an approach similar to our, such as [13], where a PCA-SIFT approach is used to identify multiple instances of the same object in real time and a voting scheme similar to ours is used, but the achieved clustering is used only for localization purposes and not for detecting overlapped objects. Leordeanu and Hebert in [8] present a spectral method for finding consistent correspondences between two sets of features. The nice mathematical framework presented there permits to include also spatial relationships under the form of affinity between two correspondences. This is also our case but the graph-based approach proposed for finding the optimal solution is likely to be too onerous for our time constraints. Moreover, it is not clear whether this method can handle duplicate objects.

3. Single Model Approach

The final objective of the system is to segment as many objects as possible in cluttered scenes such as those reported in Fig. 1. In order to segment multiple objects in a cluttered scene, a first solution is to use a single object model M, which consists simply in an image containing a single object on a plain background. The object model can be captured of whichever size and orientation, and under generic illumination.

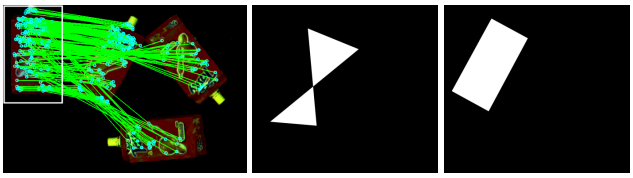
Our proposal goes through the following two phases:

1. *Feature extraction and matching*: significant features are extracted from both the object model and the current image; given a proper similarity measure, features are matched between the model and the current image, and the best correspondences are retained;

2. *Object segmentation*: given the set of correspondences, it is possible to compute a registration transform between the model and the segmented object in the current image.

The point 1 is achieved by using th SIFT feature detector and the 2NN (two nearest neighbors) heuristic proposed in [10]. Similarly to [8], we can define two sets of features P of n_P data features extracted on the current image I, and Q of n_Q model features extracted on the model image M. Let $\mathcal{M} = \{m_1, \ldots, m_N\}$ be the set of N matches (or assignments) found between the set P and the set Q, where each match m_i contains the (x, y) coordinates on the two reference systems: $m_i = \left\langle \left(x_i^M, y_i^M\right), \left(x_i^I, y_i^I\right) \right\rangle$. Given this set, the simplest approach for computing the registration transform between M and I (point 2 above) is to estimate the

planar homography using a least squares approach. Alternatively, direct linear transform (DLT) or singular value decomposition (SVD) can be used to solve efficiently the system of linear equations obtained by the matches [5].



(a) Matched features (b) Homography without RANSAC (c) Homography with RANSAC

Figure 2. Wrong homography due to outliers.

Unfortunately, all these approaches are very sensitive to outliers in the set of correspondences. For instance, Fig. 2(b) shows an example of incorrect homography obtained by least square method using all the matches: it is evident that some of these matches are outliers in the estimation of the homography's parameters. A well known method to deal with the outliers is provided by RANSAC [3] which finds a set of inlier correspondences which can be used for computing the transform as described above.

The result of estimating the planar homography from the matches in Fig. 2(a) (the model is in the top-left corner, bounded in white) with RANSAC and least square estimation is shown in Fig. 2(c). Although the result is appreciable, this approach still presents some drawbacks. The first is that, in the case of multiple instances of the same object, the SIFT does not guarantee to find all the correspondences on the same instance. Even though RANSAC can, at a certain level, handle this situation by iteratively estimate the most consistent set of matches (as in the case of Fig. 2(c)), it is not able to cope with a large number of outliers due to the presence of multiple instances, as shown in the example of Fig. 3(b).

A possible solution, described in the next Section, is to cluster the matched features $\mathcal{M}$, and then perform RANSAC for each cluster of features separately. In this manner, multiple homographies can be estimated, one for each detected instance of the object. Nevertheless, even though it leads to better results (as shown in Fig. 3(c)), this approach splits the set $\mathcal{M}$ in several clusters, and this gives fewer points on which the homographies' parameters are estimated, resulting, typically, in less accurate results.

3.1. Segmentation of Multiple Instances

Given the drawbacks of the aforementioned approaches, a 2-steps method based on a voting scheme which allows us to estimate the locations and the orientations of duplicate objects has been developed. The first step (clustering of the matched features) allows the estimation of the object center's position. In the second step, for each cluster

(a) Matched features (b) Hom.-RANSAC all matches (c) Hom.-RANSAC clustered

Figure 3. Wrong homography due to multiple instances.

the positions of a fixed number of *principal points* are computed in order to define the object delimiting shape and/or the grasping points for a bin-picking application.

Instead of applying RANSAC on the entire set $\mathcal{M}$, we can first partition the set on S subsets $\mathcal{M}^i$ possibly containing only the features belonging to a single object/instance O^i. The clustering of features can be performed by considering, similarly to [13], the relative position of each feature with respect to the center of the object. In other words, given the object model and having selected manually its center P_0, the vector distance between each feature of the model and P_0 is computed and stored. Then, given the matched features m_y in $\mathcal{M}$ and assuming a pure roto-translational (i.e., Euclidean) transform, these vectors are reported in the current image by exploiting the main orientation of each keypoint provided by SIFT. Given the approximations introduced by the image noise, the features localization and the Euclidean transform assumption, the center estimation is not accurate. To deal with this, the center's estimates are clustered by using the Mean Shift [4], and only the centers with a minimum number of contributing matches are considered as correct. This minimum number depends on the object nature: increasing the object's complexity, the minimum number is decreased. With the mean shift clustering, the set $\mathcal{M}$ of correspondences is partitioned in subsets $\mathcal{M}^i$ for each of the duplicate object O^i found in the current image.

(a) Matched features (b) Homography-RANSAC clustered (c) Our approach

Figure 4. Wrong homography due to few collinear matches.

By running the RANSAC and least square on each subset $\mathcal{M}^i$, the resulting homographies are generally more accurate, as shown in Fig. 3(c), where only one of the homographies is drawn. This method will be hereinafter called *RANSAC clustered*. However, this approach still presents the problem of being inaccurate if few matches per instance

are available and/or if these matches are fairly collinear (see Fig. 4(b)), where the few matches on the middle object are collinear and the resulting homography is imprecise. Moreover, RANSAC's result is unpredictable due to its random sampling procedure, which might be a problem in industrial applications.

Our proposal is to further relax the problem's conditions by assuming a complete Euclidean transform for all the pixels of the model (not only the center). This means that we only consider in-the-plane rotations and translations, not permitting the model to scale (reasonable condition if we assume that the objects are more or less at the same distance from the camera) or to rotate (much) out of the image plane. It is also worth noting that a precise segmentation is not required since our goals are to find the grasping points and to evaluate occlusions (in order to avoid the picking of covered objects).

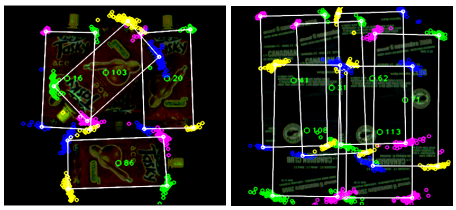
Once the matches have been partitioned, the aforementioned procedure is repeated for L relevant points of the object:

1. during the definition of the object's model, the user can select L *principal points* $\mathcal{P} = \{P_0, \ldots, P_{L-1}\}$, where P_0 is the center of the object and the other points represent both other grasping points and points delimiting the objects, such as extrema points of the oriented bounding box;

2. for each $m_y \in \mathcal{M}^i$, the estimate for each of the L principal points is computed; let us define as $P^i_{j,y}$ the estimate obtained from match m_y of instance O^i for the principal point P_j, with $j = 1, \ldots, L$;

3. $L-1$ mean shift algorithms are issued to find the best estimate $P^i_{j,*}$ for P_j in O^i; in this case, the mean shift is not employed for clustering, so a simpler technique (e.g., to compute the average location) should be enough; however, computational complexity of mean shift with some tens of points, as in most of our cases, is negligible;

4. the $L-1$ estimates $P^i_{j,*}$ are used to obtain the segmentation of O^i.



(a) (b)

Figure 5. Examples of the segmentation results.

Fig. 4(c) shows how we can solve with our approach the problem inherent to homography. Additionally, Fig. 5 shows two examples of the result achieved with this procedure. The large green circles represent the estimates $P^i_{0,*}$ of the object's center and the values close to them are the number of matches assigned to that object through the clustering. The small circles in blue, magenta, yellow and green represent the estimates $P^i_{j,y}$ of the other four principal points (the extrema of the bounding box, in these examples). Note that the estimates are mixed up and fairly distributed, but the mean shift is nonetheless capable to act correctly. The white lines connect the estimates $P^i_{j,*}$ of the principal points and are the boundaries of the final segmentation. It is evident that this approach is able to segment also very occluded objects, as shown in Fig. 5(b).

4. Multi-Model Extension

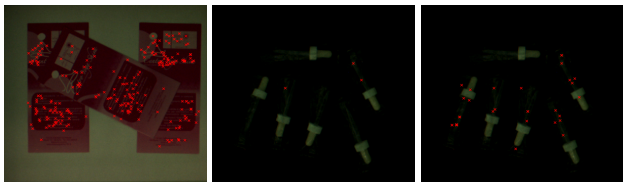
The approach described in the previous section is able to segment duplicate objects even if heavily occluded, but only in the case that both the model and the current image are described by a high number of features (Fig. 6(a)). When the number of features is lower due to almost untextured objects or poor illumination (Fig.6(b)), the extraction algorithm is likely to fail due to the following limitations:

1. the SIFT algorithm relies on the gradient and the textured areas of the object with high contrast;

2. our tests have shown that SIFT is not robust enough to local brightness changes, especially in the case of reflective or semi-transparent objects, resulting in too few and unreliable matches (see Fig. 6(b));

3. our approach becomes less reliable when the model has few matches, due to the few votes in determining the principal points of the object;

4. the SIFT algorithm was designed to give only a single match (i.e., the best one) for each keypoint. When the model has a low number of keypoints and the matches are partitioned onto different duplicates it is not possible to perform a reliable estimation of the principal points.



(a) (b) (c)

Figure 6. Examples of objects with a high and low number of matches.

To increase the number of the available matches for the extraction, the current image is matched with N_M different models of the same object. All the models are taken under free environmental illumination and using different object orientations in order to be robust to the reflexes created by a possible transparent container.

Each considered model generates a certain number of keypoints which are then matched with those of the current image, generating the set of matches $\mathcal{M}^j$ for the model M^j. All the matches obtained by different models can be merged in a new set $\widetilde{\mathcal{M}} = \bigcup_{j=1}^{N_M} \mathcal{M}^j$, whose cardinality is typically larger than that of $\mathcal{M}$ defined before. For instance, Fig. 6(c) shows the matches obtained with 5 models on the same image of Fig. 6(b). The merged set $\widetilde{\mathcal{M}}$ is then used (with the same procedure described in Sect. 3) to obtain a better result, in terms of both the number of segmented objects and the accuracy in the extraction.

5. Experimental Results

In order to validate our approach, experiments over very different kinds of objects are performed. The experiments have been divided in two test campaigns: the first aiming at demonstrating the accuracy of the choice for single model approach and the voting scheme for object duplicate extraction (Section 5.1); the second aiming at demonstrating the improvement achieved by using multiple objects (Section 5.2). Moreover, tests are performed on different objects, both high-textured (juice boxes, packs of nutella and advertisement flyers) and low-textured (shampoos, droppers, and mascaras). These objects have different characteristics, such as: reflectiveness (shampoos and mascaras), semi-transparency (droppers) and non-regular shape (droppers and mascaras).

The accuracy of our approach can be measured by means of three different metrics: the precision/recall at *object-level*, the precision/recall at *pixel-level* and the accuracy of the *center location*. The first metrics accounts for how many correct objects are segmented (where correct segmentations are evaluated by the operator), while the second considers the pixel-by-pixel segmentation. In this case, the precision and the recall are computed with reference to all the objects to be segmented, and thus the recall tends to be very low if some objects are missed. Finally, the last metrics is more application-oriented, thinking to a pick-and-place application where the accuracy in determining the grasping point (e.g., the center) is crucial. All these measures are computed with respect to a manually-determined ground truth.

5.1. Tests for Single Model Approach

Results for the first test campaign are summarized in Table 1. We compared our proposal with the use of homography-based segmentation by either RANSAC on all the matches (*all RS* in Table 1) or RANSAC on clustered matches (*clus RS*), as described previously. Since the RANSAC applied on all the matches finds a single instance of the object, the precision at object-level is close to 100%, but the recall at object-level is very low. Instead, the RANSAC applied to clustered matches shows a poor accuracy in identifying the center. This is due to the fact that this algorithm finds more objects than the version on all the matches (precision/recall both at object- and pixel-level are higher), but the resulting homographies are less accurate since they are estimated by fewer matches.

Fruit Juices					
	Object-level		Pixel-level		Center
	Precision	Recall	Precision	Recall	Mean
all RS	100.00%	25.00%	22.95%	23.66%	5.41 px
clus RS	91.67%	82.50%	77.43%	79.93%	18.97 px
Ours	**97.37%**	**92.50%**	**88.55%**	**87.64%**	**5,76 px**
Nutella Packs					
all RS	100.00%	15.38%	13.23%	14.46%	6.98 px
clus RS	66.67%	33.85%	38.96%	35.82%	17.24 px
Ours	**97.84%**	**86.78%**	**82.87%**	**83.13%**	**3.86 px**
Paper Flyers					
all RS	90.00%	16.36%	17.46%	15.72%	14.68 px
clus RS	74.00%	64.91%	71.31%	68.46%	22.27 px
Ours	**96.15%**	**90.91%**	**86.35%**	**89.39%**	**2.66 px**

Table 1. Experimental results for single model approach.

Our approach outperforms the other two in all the cases and for every metrics, with average precision and recall at object-level of 97.84% and 86.78%, at pixel-level of 82.87% and 83.13%, and an average center's distance of 3.86 pixels. Additionally, our approach is much faster than the others since it avoids both the iterative procedure of RANSAC and the least square estimation. On average, our system takes about 0.8 seconds for each image to segment a number of objects between 3 and 10, while the RANSAC-based approaches take 4.41 sec. and 4.27 sec., in the case of clustered and non-clustered matches, respectively. These time performances have been obtained on a standard Windows XP PC with Core Duo at 2.4 Ghz processor and 2 GB of memory.

5.2. Multiple Model Approach

Table 2 shows the values of the precision/recall for both pixel-level and object-level as a function of the number of models. In most of the cases the precision at object-level tends to be very close to 100% and generally the recall increases with the number of models. For shampoos and droppers, the table shows a very particular trend for precision/recall for both pixel-level and object-level. Regarding the shampoos, the precision at object-level and at pixel-level have the same behavior, that is, increasing the number of models the precision tends to decrease. This is due to the fact that, by increasing the matches for low-textured

Fruit Juices					
	Object-level		Pixel-level		Center
# models	Precision	Recall	Precision	Recall	Mean
1	97.37%	92.50%	88.55%	87.64%	5.76 px
2	100.00%	92.86%	97.22%	92.83%	3.53 px
3	100.00%	100.00%	96.45%	93.01%	4.21 px
Nutella Packs					
1	97.84%	86.78%	82.87%	83.13%	3.86 px
2	100.00%	86.78%	96.61%	73.67%	1.69 px
3	100.00%	86.78%	96.34%	83.29%	1.39 px
4	100.00%	86.78%	96.00%	97.36%	1.56 px
Paper Flyers					
1	96.15%	90.91%	86.35%	89.39%	2.66 px
2	100.00%	90.91%	96.03%	98.62%	1.96 px
3	100.00%	90.91%	96.20%	99.37%	1.47 px
4	100.00%	90.91%	96.47%	99.36%	1.42 px
5	100.00%	90.91%	96.54%	99.46%	1.35 px
Mascaras					
1	100.00%	22.22%	56.44%	73.12%	5.05 px
2	100.00%	44.44%	73.26%	97.78%	4.32 px
3	100.00%	55.56%	71.79%	96.60%	5.55 px
4	100.00%	77.78%	68.52%	95.97%	6.48 px
5	100.00%	88.89%	67.64%	97.44%	3.39 px
Shampoos					
1	100.00%	5.56%	99.33%	96.34%	1.24 px
2	100.00%	5.56%	98.77%	97.47%	1.08 px
3	100.00%	16.67%	89.78%	92.53%	5.58 px
4	100.00%	27.78%	88.02%	91.31%	6.68 px
5	100.00%	27.78%	87.44%	90.32%	7.21 px
6	100.00%	38.89%	86.78%	90.63%	7.81 px
7	100.00%	38.89%	87.34%	91.25%	7.60 px
8	88.89%	44.44%	77.56%	81.28%	6.28 px
9	90.91%	55.56%	79.29%	85.57%	5.14 px
10	90.91%	55.56%	80.13%	86.66%	4.07 px
11	91.67%	61.11%	80.49%	87.48%	4.35 px
Droppers					
1	-	0.00%	-	-	-
2	80.00%	22.22%	71.58%	57.35%	6.32 px
3	77.78%	38.89%	68.06%	54.17%	6.40 px
4	81.82%	50.00%	72.62%	60.05%	6.45 px
5	85.71%	66.67%	76.59%	66.29%	6.15 px

Table 2. Experimental results for multiple model approach.

objects, the false matches increase as well. However, the number of objects correctly segmented (i.e., the recall at object-level) increases accordingly, and this is an important cue for our application. It is worth noting that, due to their transparency, the system is not capable to segment droppers using a single model, resulting in a 0% in all the measures. The mean accuracy of center location is generally very good (in the worst case the mean distance is 6.33 pixels). However, the standard deviation of the distance is limited, which means that the influence of the number of models on this measure is negligible.

Summarizing, these results demonstrate that the more models are used, the higher will be the recall at object-level (number of segmented objects) and the precision at pixel-level (better extraction). These results have been obtained with a standard dual-core PC equipped with a GP-GPU (General Purpose Graphical Processing Unit) on which a specialized version of the SIFT algorithm which allows us to process one frame in 0.417 sec on average, which allows us to reach the requested processing speed. It is worth noting that the processing speed is not significantly affected by the number of models since their keypoints can be precomputed and stored.

References

[1] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[2] V. Ferrari, T. Tuytelaars, and L. Gool. Simultaneous object recognition and segmentation from single or multiple model views. *Int. J. Comput. Vision*, 67(2):159–188, 2006.

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

[4] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, Jan 1975.

[5] R. Hess and A. Fern. Improved video registration using non-distinctive local image features. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR '07*, pages 1–8, June 2007.

[6] T. Knoll and R. Jain. Recognizing partially visible objects using feature indexed hypotheses. *IEEE Journal of Robotics and Automation*, 2(1):3–13, Mar 1986.

[7] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int'l Journal of Computer Vision*, 77(1-3):259–289, May 2008.

[8] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Proc. of IEEE Intl Conference on Computer Vision*, pages 1482–1489, 2005.

[9] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262 – 282, 2007.

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

[11] K. Rahardja and A. Kosaka. Vision-based bin picking: recognition and localization of multiple complex objects using simple visual cues. In *in 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1448–57. IEEE Press, 1996.

[12] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int'l Journal of Computer Vision*, 73(2):213–238, June 2007.

[13] S. Zickler and M. Veloso. Detection and localization of multiple objects. In *Proc. of 6th IEEE-RAS International Conference on Humanoid Robots*, pages 20–25, Dec. 2006.